General Relativity and Gravitation 1992

Proceedings of the Thirteenth International Conference on General Relativity and Gravitation held at Cordoba, Argentina, 28 June–4 July 1992

Edited by

R J Gleiser

C N Kozameh

O M Moreschi

Facultad de Matemática, Astronomía y Física
Universidad Nacional de Córdoba
Córdoba, Argentina

Institute of Physics Publishing
Bristol and Philadelphia

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library.

ISBN: 0 7503 0261 5

Library of Congress Cataloging-in-Publication Data are available

Published by Institute of Physics Publishing, wholly owned by
The Institute of Physics, London.

Institute of Physics Publishing, Techno House, Redcliffe Way, Bristol BS1 6NX, UK

US Editorial Office: Institute of Physics Publishing, The Public Ledger Building, Suite 1035, Independence Square, Philadelphia, PA 19106, USA

Printed in the UK by Galliard (Printers) Ltd, Great Yarmouth, Norfolk

CONTENTS

PREFACE

The scientific program of **GR13** maintained the traditional format of the Conferences of the International Society on General Relativity and Gravitation. The program consisted of plenary sessions with invited talks, workshops, and concurrent poster sessions. A printed book of abstracts, contributed for presentation at the Conference, was distributed among Conference participants.

The Proceedings volume contains the **GR13** plenary lectures along with summaries of the highlights of the workshops provided by their Chairmen. We are very much aware of the effort required of the plenary speakers and of the workshop Chairmen, and are grateful to all of them for their contributions to the Proceedings.

The major advances since the previous GR Conference were reviewed at the plenary talks, while the more technical aspects as well as ongoing research were presented at the workshop and poster sessions. There was, for the first time, a special 'discussion' session. We think it was worthwhile and encourage similar sessions in future Conferences.

This was the first GR Conference held in Latin America and the participants from this region, particularly the graduate students, had the unique opportunity to meet and exchange ideas with the leading researchers in this field. We are grateful to the International Society of General Relativity and Gravitation for having selected Cordoba as the site of the **GR13** Conference.

We would also like to acknowledge the sponsorship of a large number of institutions which are listed below in alphabetical order. It would not have been possible to organize **GR13** in Argentina without their support. Equally important was the special dedication offered by the Local Organizing and International Scientific Committees and by so many of our colleagues and students, as well as by the administrators directly involved in the Conference. We regret that we cannot mention everybody, although we are convinced that their help was essential to **GR13**. However, we consider it necessary to express our special appreciation to Robert Wald, who chaired the Scientific Committee, to Alan Held, our link with the GRG Society, and to the **GR13** secretaries, Ruth Bestgen, Marta Garcia, Alberto Gattoni and Victoria Paganini.

Throughout the organization of **GR13** we received strong support and encouragement from the Rector of the University of Cordoba, Professor Francisco Delich. The executive and legislative branches of the government of the Province of Cordoba provided much needed financial and organizational support. Special mention goes to Professor Ronald Rivas O'Neall, the Secretary of the President of the State Senate, for his invaluable assistance.

R J Gleiser
C N Kozameh
O M Moreschi

GR13 was held under the auspices of the International Society of General Relativity and Gravitation, Centro Latinoamericano de Física, Secretaría de Ciencia y Tecnología de Argentina, Ministerio de Educación de Argentina, Universidad Nacional de Córdoba, Facultad de Matemática, Astronomía y Física, Gobierno de la Provincia de Córdoba and Municipalidad de Córdoba.

The Conference was co-sponsored by

- Academia Nacional de Ciencias de Argentina.
- Cámara de Diputados de la Provincia de Córdoba.
- Cámara de Senadores de la Provincia de Córdoba.
- Consejo de Investigaciones Científicas y Tecnológicas de la Provincia de Córdoba (CONICOR).
- Consejo Nacional de Investigaciones Científicas y Técnicas de Argentina (CONICET).
- Facultad de Matemática, Astronomía y Física(FaMAF).
- Fundación Antorchas.
- Gobierno de la Provincia de Córdoba.
- International Centre for Theoretical Physics (ICTP).
- Instituto de Intercambio Cultural Argentino Norteamericano (IICANA).
- Instituto Nacional de Tecnología Industrial (INTI).
- Italian Embassy in Argentina.
- Municipalidad de la Ciudad de Córdoba.
- Organization of American States (OAS).
- A Rosemblum's Family.
- Secretaría de Ciencia y Tecnología de Córdoba.
- The International Union of Pure and Applied Physics (IUPAP).
- The Gravity Research Foundation.
- Universidad Nacional de Córdoba (UNC).

PART 1
PLENARY LECTURES

Paper presented at the 13th Int. Conf. on General Relativity and Gravitation
Cordoba, Argentina, 1992: Part 1, Plenary Lectures

Laser interferometric gravitational wave detectors

K Danzmann
Max-Planck-Institut für Quantenoptik, D-8046 Garching, Germany

1. Introduction

More than 70 years ago, Gravitational Waves have been predicted as one of the consequences of Einstein's Theory of General Relativity [1]. They are clearly one of the fundamental building blocks of our theoretical picture of the universe and there is some circumstantial evidence pointing to their existence [2]. But in spite of numerous attempts over the last 30 years, their direct detection remains as one of the great unsolved problems of experimental physics.

1.1. The birth of gravitational astronomy

Today, the technology seems to be in hand to finally tackle this problem and this article is aiming at outlining the current efforts to bring non-linear gravity into confrontation with experiment through the detection of gravitational waves. But the final aim is not a mere proof of the existence of gravity waves, rather to make them useful for observational astronomy through the creation of a world-wide network of detectors. We have to realize that the information carried by gravitational waves is complementary to the information carried by electro-magnetic radiation. Whereas electro-magnetic radiation is an incoherent superposition of radiation mostly emitted by thermally excited atoms and high-energy electrons, it is the coherent, bulk motion of huge amounts of mass that produces significant levels of gravity waves. Electro-magnetic radiation is easily scattered and absorbed, but gravitational radiation is transmitted almost undisturbed through all forms and amounts of intervening matter [3]. The introduction of Gravitational Astronomy would thus literally open a new window to the universe [4].

2. The detection of gravitational waves

Gravitational waves change the metric of space-time. They can be detected through the strain in space created by their passage. Consider a gravity wave impinging perpendicular to the plane of a circle, see Fig. 1. During the first half-cycle of the wave, the circle will be deformed into a standing ellipse and during the second half-cycle into a horizontal ellipse. It is the fractional change in diameter that is commonly quoted as a measure for the amplitude, $h = 2dL/L$, of a gravitational wave. The principle behind the detection of gravity waves is thus a simple length measurement. The problem is that the length change is so small. As an example, consider a supernova in a not too distant galaxy. This might produce a relative length change here on earth

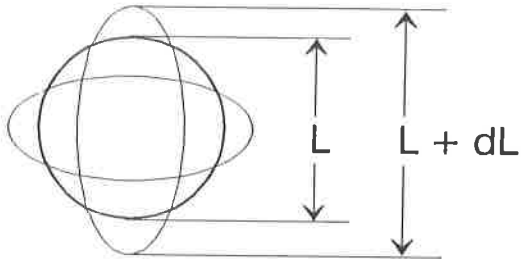

Figure 1. Gravitational waves change distances by squeezing space.

of 1 part in 10^{21}. Such a relative length change corresponds to the diameter of one hydrogen atom on the distance from here to the sun.

2.1. Bar antennas

The history of attempts to detect gravity waves began in the 1960s with the famous bar experiments of Joseph Weber [5]. Even though these experiments have not yet detected gravitational waves, they had the undisputed effect of alerting the scientific community to the possibility of experimentally detecting gravity waves. Bar detectors have in the meantime been developed and refined in several places all over the world. Being supercooled to mK temperatures and equipped with very sophisticated length transducers [6], they have now reached a sensitivity ($h \approx 10^{-18}$ for millisecond pulses) where they could expect to see the next supernova in our own galaxy, and they are likely to remain an important ingredient of the world-wide gravity wave-watch. But being resonant devices, they are in practice sensitive only in a relatively narrow band around their central frequency. They are also limited in their sensitivity through the quantum-mechanical uncertainty of their mechanical state, although this limitation may in principle be overcome by QND techniques, and so their usefulness will in all likelihood remain limited for the forseeable future.

2.2. Laser interferometers

Although the seeds of the idea can already be found in early papers by Pirani [7] and Gertsenshtein and Pustovoit [8], it was really in the early 1970s when the idea emerged that laser interferometers might have a better chance of detecting gravity waves, mainly promoted by Weiss [9] and Forward [10]. Large interferometers would offer the additional advantage of having broad-band sensitivity and they would not be limited by the uncertainty principle until well below a sensitivity of 10^{-23}. A Michelson interferometer measures the phase difference between two light fields having propagated up and down two perpendicular directions, i.e. essentially the length difference between the two arms. This is exactly the quantity that would be changed by the passage of a properly oriented gravitational wave, see Fig. 2.

Immediately obvious at this point is the need for long interferometer arms. The quantity measured is the absolute phase difference between the fields. But the gravity wave induces a fractional length change. So the phase difference measured can be increased by increasing the arm length or, equivalently, the interaction time of the light

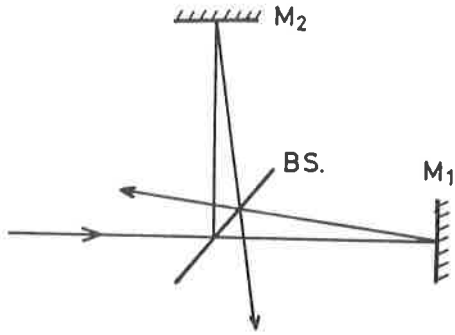

Figure 2. Michelson interferometer.

with the gravity wave. This works up to an optimum for an interaction time equal to half a gravity wave period. For a gravity wave frequency of 1 kHz this corresponds to half a millisecond or an arm length of 75 kilometers.

2.3. Long light-path

While it is clearly impractical to build such a large interferometer, there are ways to increase the interaction time without increasing the physical arm length beyond reasonable limits. Historically, two approaches have emerged: storing the light in the arms in resonant optical Fabry-Perot cavities [11] and literally folding the light back and forth in optical delay lines [12]. Nowadays this distinction is beginning to disappear, because with the development of Dual Recycling [13], a new optical technique, we now have a hybrid arrangement in our hands that combines the advantages of both approaches and more.

3. Prototypes

Several small prototypes of laser interferometric gravitational wave detectors have been developed in the world, a delay-line based interferometer with 30 m arm length at the Max-Planck-Institut für Quantenoptik in Garching, a Fabry-Perot based instrument with 10 m arm length at the University of Glasgow, a delay-line based instrument with 10 m arm length at the Institute for Space and Astronautical Science in Tokyo, and a Fabry-Perot based instrument with 40 m arm length at the California Institute of Technology. With arm lengths on the order of a few tens of meters these prototypes are clearly too small to permit observations of real gravity waves. Their sensitivities have continually been improved over the years, and the larger ones have all reached sensitivities for millisecond pulses roughly equivalent to the best bar detectors, but in addition they are broad-band devices.

While the absolute sensitivity reached by the prototype detectors is certainly encouraging, it is much more important that they are well-understood devices. That is, the various physical processes creating noise sources at the various frequencies have to

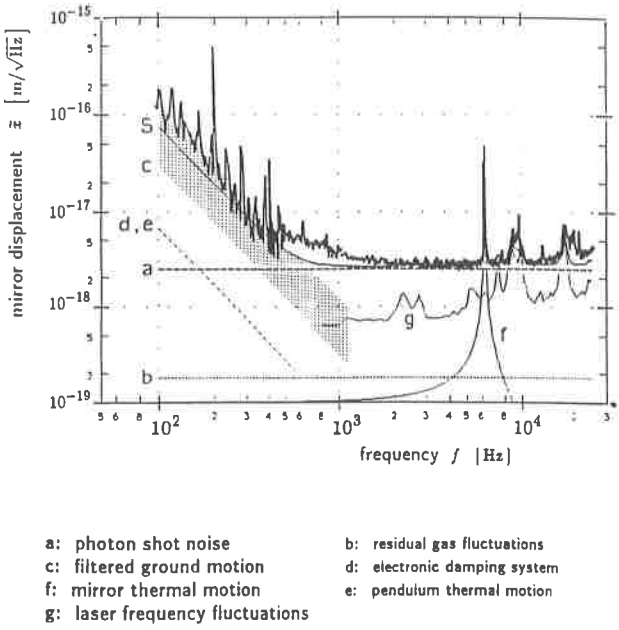

Figure 3. Noise analysis of the Garching 30-m prototype.

be identified in order to find ways to improve on those. In Fig. 3 we see such a noise analysis for the Garching 30-m prototype [14].

Shown is the spectral density of the apparent mirror displacement as expected from the most important noise sources. For comparison, the measured spectral density of displacement noise is also given. Very good agreement between the measured and the expected noise is found. The additional sharp peaks at frequencies of a few hundred Hertz are due to violin string resonances of the suspension wires holding the mirrors. The sensitivity of the prototype is presently limited by residual ground motion at frequencies below 1 kHz (labeled c in Fig. 3), by the photon shot noise corresponding to the available laser power at frequencies between 1 kHz and 6 kHz (labeled a), by the thermallly excited internal mechanical vibration of the mirrors in a narrow peak at 6 kHz (labeled f), and by residual frequency fluctuations of the laser at higher frequencies (labeled g). At the present level of sensitivity, the refractive index fluctuations due to the Brownian motion of the residual gas in the vacuum pipe (labeled b) are unimportant. Also unimportant at this level are the above-resonant wing of the thermally excited mirror suspension pendulum resonance (labeled e), the sub-resonant wing of the mirror internal mechanical resonance (labeled a), and the noise introduced by the electronic damping system for the mirror suspension (labeled d). While the particular noise sources are different, the level of understanding reached for the other major prototypes is comparable to the one shown for the Garching prototype.

4. Large interferometer projects

After almost two decades of research on small prototypes, the time had come to proceed towards the construction of full-scale interferometers with arm lengths of several kilometers, and several such proposals were submitted at the end of the 1980s. Out of ten research groups in Germany and Britain, the GEO collaboration was formed [15], aiming at the construction of a laser interferometer with 3 km arm length near Hannover in the German state of Niedersachsen. The French-Italian VIRGO collaboration, comprising 9 research groups from both countries, is planning a detector with 3 km arm length near Pisa [16]. The American LIGO project [17], with scientists at MIT and Caltech, wants to build two detectors, with 4 km arm length each. Just recently, the Australian AIGO collaboration has proposed a 3-km detector near Perth [18].

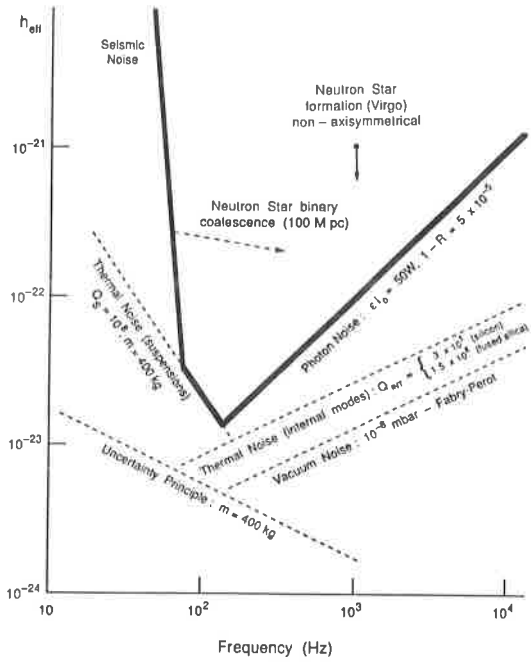

Figure 4. Noise sources relevant for laser interferometric gravitational wave detectors.

As an example, Fig. 4 shows the significance of the various noise sources for the final design sensitivity of a fully optimized GEO-type interferometer to broad-band gravitational wave bursts. At low frequencies, the sensitivity will be limited by the thermal noise of the mirror suspension, at intermediate frequencies by thermal noise due to internal mechanical mirror vibrations, and at high frequencies by photon shot noise. In the following, the main problems encountered in the design of such an interferometer and the envisioned solutions for the major noise sources are highlighted, sometimes using the GEO project as an example. But it should be emphasized that the problems

are common to all projects and that there actually is a very strong coordination and sharing of tasks between the various collaborations. An overview of the current status (summer of 1992) of the world-wide efforts is given at the end.

5. Vibration isolation

We are trying to measure very small length changes due to the action of gravitational waves and so it is extremely important to ensure that no other effect moves the test masses. By far the largest disturbance is the random ground motion because of the natural seismic activity. The spectral density of displacement due to seismic ground motion typically falls of as

$$\tilde{x}(f) \approx 10^{-7}\mathrm{m}/\sqrt{\mathrm{Hz}}\left[\frac{1\mathrm{Hz}}{f}\right]^2. \qquad (1)$$

At a frequency of 1 kHz, this is about 10^7 times larger than the effect we are trying to measure and we require an effective way of vibration isolating the test masses.

Passive vibration isolation is, in principle, straightforward [19]. The object to be isolated gets suspended by a pendulum. If a disturbance shakes the suspension point of the pendulum with a frequency below its resonant frequency, then the pendulum will transmit the disturbance unattenuated. But a disturbance above the resonant frequency f_o will get attenuated with the ratio $(f_o/f)^2$ up to a frequency Qf_o, where the frequency dependence changes to a linear slope. The resonant frequency f_o should thus be as low as possible. Due to practical limitations on the pendulum length to around one meter, horizontal resonant frequencies are usually limited to around 1 Hz. Vertical resonant frequencies are normally a bit higher because the spring has to be stiff enough to support the full load. Several of these stages can be cascaded to achieve a very steep fall-off above the highest normal mode of the coupled oscillator system.

5.1. Stacks

A very simple, yet very effective way of achieving vibration isolation at moderately high frequencies (above 100 Hz) is the use of vibration isolation stacks. A detailed analysis of stack systems has recently been carried out by Cantley *et al* [20]. Stacks are basically alternating layers of a high-density material (like lead) and rubber. Each layer acts as a 3-dimensional pendulum, although with a fairly low Q. But since it is easy to use several such layers, a rather steep fall-off towards higher frequencies can be achieved. As an example, Fig. 5 shows a comparison between calculated and measured transmissibility of a small 4-layer lead-rubber stack intended for the Garching prototype.

Because of their low mechanical Q-factor, such stacks show a rather high thermal noise. The last two stages of the suspension should accordingly be very high-Q wire pendulums of low resonant frequency. These will also give an additional $1/f^4$ filtering in the horizontal direction where it is most critical (this is the direction of the laser beam; in principle, vibration isolation in the vertical direction is only required because of the unavoidable cross-coupling between the two degrees of freedom).

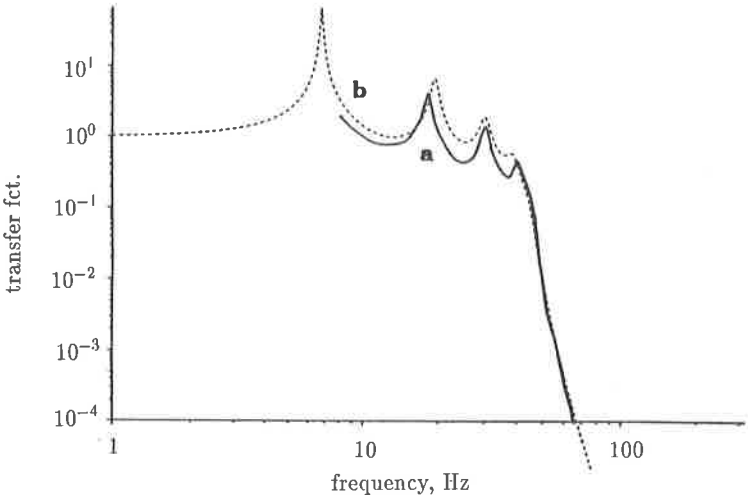

Figure 5. Transmissibility of a 4-layer stack; (a) measured, (b) calculated.

5.2. Low-frequency isolation

Extending vibration isolation down to lower frequencies becomes increasingly difficult. The ground noise spectrum increases towards lower frequencies like $1/f^2$ and the isolation decreases because one is moving closer to the highest normal mode of the pendulum chain. Some very promising work on low frequency passive isolation has been done in Pisa [21]. The Superattenuator uses a cascade of an inverted pendulum, 7 gas springs, and a wire pendulum. The use of gas springs has the advantage of offering isolation in the vertical direction comparable to that in the horizontal. This is further improved through the addition of magnetic anti-springs acting in the vertical direction. This design should achieve efficient vibration isolation all the way down to about 10 Hz. The control problems associated with this approach seem tractable and it could be an alternative to the use of stacks.

Another way of achieving isolation at very low frequencies would be the use of active or combined active/passive vibration isolation systems. Some very promising work, aiming at achieving isolation all the way down to 1 Hz, is going on at the Joint Institute for Laboratory Astrophysics in Boulder, Colorado [22].

6. Position control and feed-back

An interferometer with as many components and degrees of freedom as a gravitational wave detector requires sophisticated control systems to keep all parameters at their optimal operating points. The guiding principle behind the position control of the optical components is easily stated: At low frequencies the optical components must be rigidly held relative to each other to keep them from drifting and to prevent the interferometer and the recycling cavities from losing lock. On the other hand, at higher frequencies, where gravitational wave signals could be detected (above 100 Hz, or, possibly, above 10 Hz), the test masses must be totally free and the system used

to prevent them from moving at low frequencies must not exert any residual forces at high frequencies. Three main classes of control systems are used:

6.1. Local controls

The suspension of the optical components via high-Q pendulums is an efficient way to isolate them from high-frequency vibrations and pendulum thermal noise. But at the resonant frequencies of the undamped suspension, large vibration amplitudes can occur that will greatly exceed the dynamic range of the detector (a pendulum with a Q of 10^7 can, at its resonance, lead to an amplification of up to 10^7). The resonance must thus be damped. But damping it in the usual dissipative way will degrade the Q and introduce thermal noise. The damping is instead done in an active and frequency-selective way. This technique is routinely used on the prototypes. The position of the masses is sensed with a low-noise local sensor. This signal is then electronically filtered to a narrow frequency range around the resonance and then fed back to a force transducer acting on the mass selectively in this band around the resonant frequency only. On the prototypes the position sensing is usually done via shadow sensors consisting of a LED-photodetector pair with the light path partly interrupted by a movable vane mounted on the test mass. The force is applied through a coil acting on a magnet on the mass. Such systems typically show a sensing noise of around $10^{-11}\mathrm{m}/\sqrt{\mathrm{Hz}}$. In the full scale detector we are trying to measure displacements smaller than $10^{-20}\mathrm{m}/\sqrt{\mathrm{Hz}}$. So the servo gain would have to roll off by 9 orders of magnitude in the small frequency range from a few Hz to 100 Hz, which is clearly a formidable task. The problem can be solved by sensing the motion of, and applying feed-back to, a higher stage in the suspension system and using the passive $1/f^2$ filtering of each stage. Such systems are currently being tested in the prototypes.

6.2. Global controls

In order to optimally align an optical interferometer, and keep it aligned regardless of drift or stability of the optical components, some kind of automatic alignment system is required. The guiding principle is that the alignment signals for such a system should be derived directly from the existing interfering beams without introducing additional components into the high-sensitivity/high-intensity part of the interferometer. Suitable feed-back should then be applied to all the relevant optical components.

For example, consider the case of two interfering beams, where a differential high frequency phase modulation is applied and the overall phase difference is determined by coherently demodulating the intensity of the interfered output. Relative angular misalignment introduces a differential phase gradient between the two beams which can be sensed using a split photodiode and coherently demodulating as before. Lateral misalignment may be detected using another split photodiode and a suitable lens arrangement to cause laterally offset beams to converge.

An illustration of this techniques for a Fabry-Perot cavity is given in Fig. 6. All 4 degrees of freedom necessary for alignment can be extracted. An extension of this technique can be used to automatically align all degrees of freedom for all components of the interferometer. A system similar to the automatic alignment system intended for the large interferometer is being installed and tested in the Glasgow prototype [23].

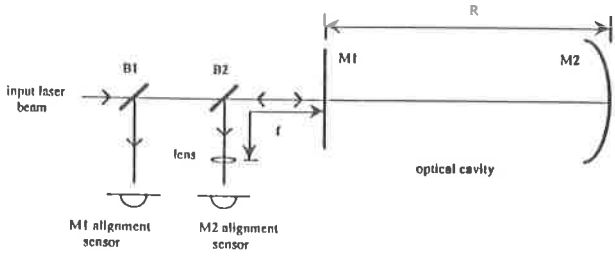

Figure 6. Automatic alignment of a Fabry-Perot cavity.

6.3. Fast feed-back

While high-frequency ground motions can be efficiently suppressed by the suspension system, there are very-low frequency seismic motions (1 Hz and less) that are very difficult to isolate against. This very-low frequency seismic motion may lead to rms differential arm length changes of several microns. Unsuppressed, these would lead to a severe limitation of the sensitivity of the detector, because an interferometer operating away from its null fringe by an amount δx becomes sensitive to the intensity fluctuations δI of the laser according to

$$h_{\text{noise}} = \frac{\delta x}{\ell} \frac{\delta I}{I} \tag{2}$$

The sensing in this case is no problem, because the main interferometer output itself provides the signal. But these deviations from the null have to be suppressed by at least 10^6 which requires a fast servo with a bandwidth of order a kHz. This feed-back is best applied relative to a reaction mass suspended from the same isolation system to avoid coupling ground motion back in through the actuator. An alternative would be feed-back relative to ground using constant-gradient coils. These coils exert a force essentially independent of the position of the coil. Ground vibrations are thus decoupled from the test mass if the constant-gradient volume is large enough to accomodate the residual motion.

7. Thermal noise

7.1. Mirror internal noise

The mirrors are macroscopic objects and, correspondingly, have internal mechanical vibration modes. Even for perfectly well isolated mirrors, these resonances will still get thermally excited. At the resonant frequencies the mechanical vibration amplitudes are too large and will overwhelm any mirror motion due to a possible gravitational wave. The only solution is to shift all mechanical resonances out of the frequency range of interest by a suitable choice of mirror shape and material. The best compromise is obtained for cylindrical mirrors with a length about half the diameter. For synthetic quartz this yields resonant frequencies of a few kHz, - above the interesting frequency window.

But even though the resonances are outside the observation window, the sub-resonant wings of those resonances do cause a stochastic motion of the mirror surface. The strain spectral density due to these motions is given by

$$\tilde{h} \approx \sqrt{\frac{16kT}{\pi^3 \rho {v_s}^3 Q_{\mathrm{INT}} \ell^2}}, \tag{3}$$

where ρ is the density, ℓ the arm length, v_s the sound velocity, and Q_{INT} the mechanical quality factor of the mirror material. Note that this expression is independent of mirror size. It is mandatory to use a material with very low internal damping (very high Q). Single crystal Silicon suggests itself for the non-transmitting components with a Q of up to 10^8, but even synthetic quartz gives a Q of a few hundred thousand, whereas low-expansion Zerodur only has a Q of about 1000. Special attention has to be paid to the way of suspending the mirrors, because any way of dissipating energy, like friction of a rubbing suspension wire, will immediately destroy an internal Q as high as this. The problem of material Q as a function of experimental parameters is currently being investigated [24].

It should be noted that the noise density due to thermal mirror noise will be strictly constant in frequency only if the internal damping mechanism has viscosity-like behaviour. Sub-resonant thermal noise like this has never been measured directly, but it is highly likely that the mechanical mirror resonance will behave much more like a harmonic oscillator with a complex spring constant [25]. In this case the thermal noise would actually increase from the quoted level towards lower frequencies like the inverse of the square-root of the frequency.

7.2. Suspension noise

Thermal noise is also present in the last stage of the vibration isolation system. This will be a simple wire pendulum made from a sling supporting the mirror. In this case the resonant frequency is about 1 Hz, and the frequency window of observation is on the above-resonant wing. The spectral density of apparent strain noise due to this effect is given by

$$\tilde{h} \approx \sqrt{\frac{16kT\omega_o}{mQ_S\omega^4\ell^2}}, \tag{4}$$

where m is the mirror mass, ω_o the resonant frequency and Q_S the mechanical quality factor of the suspension pendulum. This Q can be much higher than the internal Q of the wire material because most of the energy of the pendulum is in the form of potential and kinetic energy of the swinging bob and not in the elastic energy of a bent wire. But it is extremely important that the wire support points be properly designed to avoid friction. Pendulum Qs as high as 10^7 have been experimentally observed [26], (meaning that a 1 Hz pendulum will oscillate for more than 4 months before its amplitude has decreased to one-third).

8. The laser source

The sensitivity of a simple Michelson interferometer with optimized arm length to gravitational wave bursts is limited by the photon shot noise to

$$h_{DL} \approx 2.4 \times 10^{-21} \left[\frac{\epsilon I_o}{50\,\mathrm{W}}\right]^{-1/2} \left[\frac{f}{1\,\mathrm{kHz}}\right]^{3/2}, \tag{5}$$

where ϵ is the quantum efficiency of the detector, I_o is the laser output power, and f is the center frequency of the burst. Green light and a bandwidth of half the center frequency have been assumed. The first problem to be solved is thus the construction of a laser with sufficient output power in a stable single transverse and longitudinal mode. Moreover, frequency as well as amplitude of the laser have to be stabilized to unprecedented values.

8.1. Output power

Currently all operating prototypes use Argon ion lasers as light sources. The most powerful commercially available lasers of this type offer a single-mode output power of about 5 W. The coherent addition of several such lasers phase-locked to a master oscillator to reach higher output power has been demonstrated experimentally [16]. But this approach does not seem promising because of the poor energy efficiency of these lasers (only about 10^{-4}), their complexity and high cost of operation, and their poor free-running noise that requires elaborate means for stabilization. The laser sources under development for all large projects are all-solid state YAG lasers pumped by laser diodes. YAG lasers have traditionally been pumped by discharge lamps, but the dramatic advances in the development of laser diodes in the last few years have now made it possible to replace the noisy and inefficient lamps with diode lasers [28].

The laser developed in the GEO project is based on a diode-pumped miniature monolithic ring laser oscillator with an output power of a few hundred milliwatts. The oscillator incorporates an electro-optic phase modulator to permit fast tuning and easy frequency stabilization. The output of this oscillator is then amplified in a diode-pumped YAG rod enclosed in a ring resonator. The oscillator is operational and by the end of 1992 the final laser system is expected to deliver a single-mode output power of more than 50 W at a wavelength of 1064 nm. The fundamental wavelength of this laser could be used in an interferometer, offering advantages concerning ease of operation, fundamental absorption in optical components, and required figure accuracy of the optics. Nevertheless, the GEO and LIGO projects are investigating the option of doubling the frequency to obtain light in the green at 532 nm. Because of the shorter wavelength in the green only half the power is required to reach the same shot-noise limited sensitivity. Also the optical components can be smaller because the diffraction-limited beam-diameter is smaller in the green by the square-root of two. Doubling efficiencies around 50 percent have been achieved for powers around 10 W and much more seems possible [29].

8.2. Frequency noise

A perfect interferometer is entirely insensitive to frequency fluctuations of the laser. But in a real interferometer, noise signals can be created if at the output of the interferometer there is interference of lightbeams that have a different history. Such a situation can arise if the storage times in the arms are not identical, or if stray light can reach the output through a path different from that of the main beam. If the relative amplitude of stray light capable of interfering with the main beam is σ, then the arm length change $\frac{\delta\ell}{\ell}$ simulated by a frequency change $\frac{\delta\nu}{\nu}$ of the laser is

$$\frac{\delta\ell}{\ell} = \sigma\frac{\delta\nu}{\nu}, \tag{6}$$

if stray-light with a path-difference of one full round-trip dominates.

The prototypes use relatively noisy Argon lasers that require sophisticated frequency stabilization techniques. Figure 7 shows as an example the unstabilized frequency noise of the laser previously used on the Garching prototype, curve (a). Curve (b) shows the noise after prestabilization onto a rigid Fabry-Perot reference resonator, and curve (c) shows the noise after final stabilization onto the average arm length of the 30-m interferometer. A lowest value of about $5 \times 10^{-3}\mathrm{Hz}/\sqrt{\mathrm{Hz}}$ is reached in the relevant frequency range. The Argon laser in the Glasgow prototype, being stabilized onto the 10 m long Fabry-Perot cavity in one of the interferometer arms using similar techniques, reaches a frequency noise of a few times $10^{-5}\mathrm{Hz}/\sqrt{\mathrm{Hz}}$.

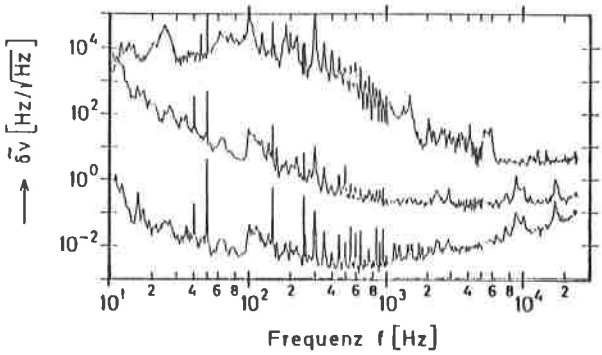

Figure 7. Frequency noise of the Garching Innova 90-5 laser:
(a) upper curve: unstabilized laser,
(b) middle curve: laser stabilized onto a 25-cm rigid Fabry-Perot,
(c) lower curve: laser stabilized onto the interferometer arm length.

For a full-scale detector, the frequency-noise out of the laser must be smaller than $10^{-6}\mathrm{Hz}/\sqrt{\mathrm{Hz}}$. But for the diode-pumped YAG laser the unstabilized frequency noise is orders of magnitude smaller than for an Argon laser. So achieving the same gain in the feed-back loop as for the Argon laser is already enough to reach the desired stability goal.

9. Recycling

Clearly the sensitivity of a simple Michelson interferometer is not sufficient, even if very strong lasers are used. Two techniques have been developed that improve the interferometer sensitivity to a level that would allow the detection of gravitational wave signals with high confidence. These techniques are known as Power Recycling and Signal Recycling, and the combination of both as Dual Recycling [13].

Power Recycling makes use of the fact that the interferometer output is held on a dark fringe by a feed-back loop and almost all the light goes back towards the input, i.e. the locked interferometer behaves as a mirror. By placing a mirror in the input of the interferometer (see Fig. 8), a resonant optical cavity can be formed that uses the whole locked interferometer as an end mirror. So the circulating light power inside the

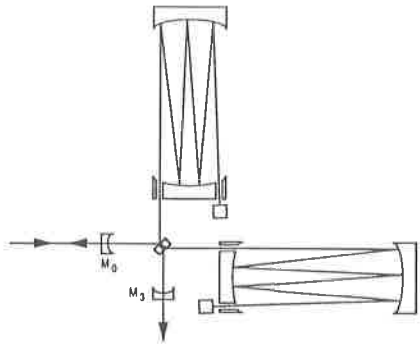

Figure 8. Dual recycled interferometer.

interferometer will be higher than the laser power by the inverse of the losses in the interferometer.

Signal Recycling works similarly, except that it leads to a resonant enhancement of the signal instead of the light. A gravity wave shaking the mirrors will phase modulate the reflected laser light, or in other words create side-bands of the laser frequency. These side-bands exit through the output port of the interferometer. By placing another mirror there (see Fig. 8), a resonant cavity for the signal-containing side-bands is formed. Depending on the reflectivity of this mirror, the detector can be made to operate narrow-band or broad-band, and by changing the position of this mirror the interferometer can be tuned to specific gravitational wave frequencies. As an additional advantage, this configuration greatly reduces the power losses due to bad interference, because the light can no longer escape the interferometer through the output port, which is now closed by the signal recycling mirror. Just as the signal recycling cavity enhances light at its resonant mode and frequency, it suppresses light which, through aberrations, got diffracted into non-resonant modes of the light field.

The shot noise-limited sensitivity of a Dual Recycled interferometer to gravitational wave bursts is given by

$$h_{\rm DL} \approx 10^{-22} \left[\frac{f}{1\,{\rm kHz}}\right] \left[\frac{\epsilon I_o}{50\,{\rm W}}\right]^{-1/2} \left[\frac{1-R}{5\times 10^{-5}}\right]^{1/2} \left[\frac{\ell}{3\,{\rm km}}\right]^{-1/2}, \qquad (7)$$

where f is the center frequency of the burst, ϵ is the quantum efficiency of the detector, I_o is the laser output power, ℓ is the arm length, and R is the mirror reflectivity. Green light and a bandwidth of half the center frequency have been assumed.

9.1. Mirror losses

In order to make these recycling techniques work, we need mirrors with extremely small losses. This requires substrates with a microroughness on the order of an Angstrom and reflective coatings with very small scatter and absorption. Fortunately, the last few years have brought us tremendous advances in the art of making superpolishes and supercoatings. Mirrors with reflection losses of much less than 50 parts per million are now available from several sources.

But it is not just the linear reflection losses that are responsible for the total losses in the Power Recycling cavity. One of the "mirrors" of this cavity is actually a very complicated object, - an interferometer locked to a dark fringe. So any effect that degrades the interfering wavefronts will let light leak out the wrong port of the interferometer. Though some of this light can be recovered by the Signal Recycling mirror, this is a serious loss process for the Power Recycling cavity.

There are two main reasons for a degradation of the interference in the interferometer: wavefront deformation because of imperfections of optical elements and because of thermal effects in the optical elements. Both processes can be addressed through optical modeling and numerical wavefront propagation calculations.

9.2. Thermal Distortions

Thermal effects can arise because of absorption in the optical coatings or in the bulk of elements used in transmission. This will cause a deformation of the substrates and/or a lensing effect through thermally induced refractive index changes. Figure 9 is the result of a model calculation including thermal effects for the main mirrors [30]. It shows the interference minimum as a function of the thermally induced mirror deformation. For increasing light power, the interference quality deteriorates in a very non-linear fashion, even approaching chaotic looking behaviour above a certain threshold.

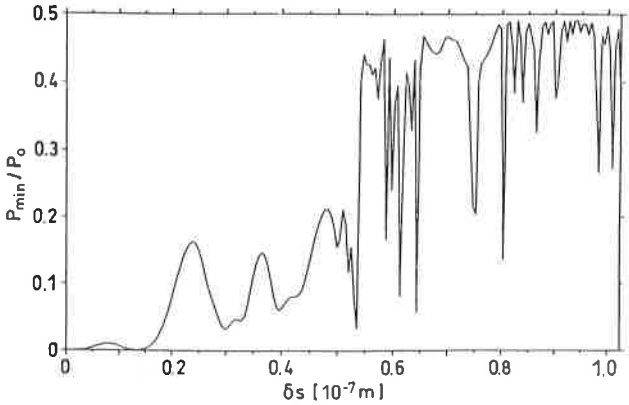

Figure 9. Power loss due to thermal effects:
interference minimum as a function of local mirror deformation.

To this end, we have studied the absorption in coatings produced in ion beam sputtering chambers. Fortunately, the absorption in modern coatings is only on the order of very few parts per million and thermal effects seem very tractable. Absorption in the bulk of the beam-splitter will probably be the limiting process at circulating powers of many kilowatts.

9.3. Mirror quality

Existing supermirrors with almost negligible losses have been developed for applications using spot sizes of a millimeter or so. For large laser interferometers the beam size will

be on the order of several centimeters and the high demands on the surface quality now extend to much larger lateral dimensions. Surface deformations on length scales of several centimeters will behave just as the microroughness does for the smaller beams. But it is not the substrate alone that determines the surface quality. The reflective dielectric multilayer stack coated onto the mirror can show variations of its effective thickness that may overwhelm the surface variations of the substrate.

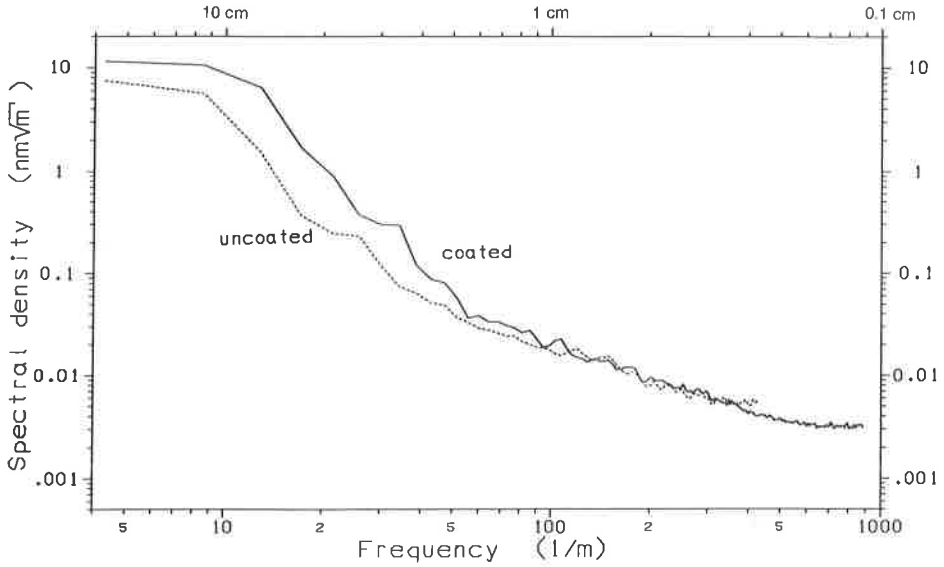

Figure 10. Surface errors of a Garching prototype mirror.

9.4. Optical modeling

Optical modeling is an important activity in all large interferometer projects. Using a numerical wavefront-propagation code running on the Garching Cray Y-MP, we are studying the effects of surface deformations on the light fields in the interferometer. Currently this code can propagate optical wavefronts on a 4000 x 4000 grid. So it is possible to even include scattering to angles large enough for the light to miss the end mirror entirely and to hit the vacuum tube after a few hundred meters. The code takes as its input assumed or measured surface profiles of the relevant mirrors. Surface deviations from the ideal shape can be measured with today's state-of-the-art to an accuracy of a few Angstroms over a field of a quarter of a meter with a lateral resolution of half a millimeter [31]. As an example, Fig. 10 shows the spectral density of surface errors as a function of spatial frequency for one of the 25-cm mirrors in the Garching prototype detector. A strong increase of the errors for large spatial wavelengths is obvious. It can also be seen that the process of coating deposition used in this case has increased the long-wavelength errors considerably. The code permits us to simulate all kinds of mirror distortions and to predict the performance of a specific

mirror in practice. Through an R+D contract with Zeiss we are addressing the mirror manufacturing and coating geometry problem and we are confident that in 1992 the first prototype mirrors with the required quality will be finished.

In addition to these static simulations, we are interested in optimizing the whole optical lay-out of the interferometer. Through a combination of numerical and analytical modeling, we are able to take into account all relevant physics, including thermal effects and the important power recovery through signal recycling. We are especially interested in finding the best compromise between storage time (or number of bounces) in the arms and storage time in the signal recycling cavity.

10. Status of efforts in the world

10.1. Proposals

The American LIGO proposal [17] calls for the construction of two detectors with 4 km arm length. It was approved in the fall of 1991 and funds have been appropriated by Congress. Two sites were selected in the spring of 1992, one in Hanford, Washington and the other in Livingston, Louisiana. Construction is expected to begin in 1993 and commissioning may start as early as 1997.

An Australian proposal (AIGO) for a 3-km detector near Perth [18] has been submitted, but so far has not been able to obtain approval.

In Japan, a proposal for an intermediate 100-m interferometer (TENKO-100) [32] has been funded. Construction may be finished in 1994. In parallel, plans are being developed for a 3-km interferometer.

The French-Italian VIRGO collaboration [16] has proposed a 3-km interferometer to be built near Pisa and the German-British GEO collaboration [15] has proposed a 3-km interferometer to be built near Hannover. Both interferometers have been coordinated under the EUROGRAV framework. A consolidation and tighter integration of European efforts is currently being prepared. The French government has recently approved funding and a decision from the remaining relevant funding bodies in Europe is expected in 1992.

10.2. A world-wide network

All these projects are not in competition with each other. On the contrary, each of the projects is crucially dependent on the others. To sort out gravitational wave events from the ever-present noise background requires observation in coincidence of several detectors. So two gravitational wave detectors are the absolute minimum to even prove the existence of gravitational waves. But to fully unravel the information contained in the signals with respect to the source direction, time structure and polarization requires a world-wide network of four detectors [33].

If all goes well, this network can be in place by the end of this decade, and at the beginning of the next millenium we may be able to mark the beginning of the age of Gravitational Astronomy.

Acknowledgments

I am indebted to the whole GEO collaboration for help in preparing this article, especially to my colleagues in Glasgow and Garching.

References

1. Misner C W, Thorne K S and Wheeler J A 1973 *Gravitation* (San Francisco: Freeman)
2. Taylor J H and Weisberg J M 1989 *Astrophys. J.* **345** 434
3. Thorne K S 1983 *Gravitational Radiation* ed N Deruelle and T Piran (Dordrecht: North Holland)
4. Thorne K S 1992 *Recent Advances in General Relativity* ed A Janis and J Porter (Boston: Birkhauser) pp 196-229
5. Weber J 1960 *Phys. Rev.* **117** 306
6. Michelson P F, Price J C and Taber R C 1987 *Science* **237** 150
7. Pirani F A E 1956 *Acta Physica Polonica* **15** 389
8. Gertsenshtein M E and Pustovoit V I 1963 *JETP* **16** 433
9. Weiss R 1972 *Quart. Progr. Rep. of RLE, MIT* **105**, 54
10. Moss G E, Miller L R and Forward R L 1971 *Appl. Opt.* **10** 2495
11. Drever R W P 1990 *The Detection of Gravitational Waves* ed D Blair (Cambridge: Cambridge University Press) pp 306-28
12. Winkler W 1990 *The Detection of Gravitational Waves* ed D Blair (Cambridge: Cambridge University Press)
13. Meers B J 1988 *Phys. Rev. D* **38** 2317; Strain K A and Meers B J 1991 *Phys. Rev. Lett.* **66**, 1391
14. Shoemaker D, Schilling R, Schnupp L, Winkler W, Maischberger K and Rüdiger A 1988 *Phys. Rev. D* **38** 423; Niebauer T M, Schilling R, Danzmann K, Rüdiger A and Winkler W 1991 *Phys. Rev. A* **43** 5022
15. Hough *et al* 1989 *Poposal for a Joint German-British Interferometric Gravitational Wave Detector* Report No. MPQ 147 (Munich: Max-Planck-Institut für Quantenoptik)
16. Bradascia C *et al* 1989 *Proposal for the Construction of a Large Interferometric Detector of Gravitational Waves* proposal to CNRS and INFN; Bradascia C *et al* 1991 *Gravitational Astronomy* ed D E McClelland and H Bachor (Singapore: World Scientific)
17. Vogt R E, Drever R W P, Raab F J, Thorne K S and Weiss R 1989 *Laser Interferometer Gravitational-Wave Observatory* proposal to the National Science Foundation (California Institute of Technology); Abramovici A *et al* 1992 *Science* in press
18. Sandeman R J, Blair D G and Collett J 1991 *Australian International Gravitational Research Centre* proposal to the CRC (Australian National University); McClelland D E *et al* 1991 *Gravitational Astronomy* ed D E McClelland and H A Bachor (Singapore: World Scientific)
19. Robertson N A 1990 *The Detection of Gravitational Waves* ed D Blair (Cambridge: Cambridge University Press)
20. Cantley C *et al* 1992 *Rev. Sci. Instr.* **63** 2210
21. DelFabbro R *et al* 1988 *Rev. Sci. Instr.* **59** 292; Bradascia C *et al* 1989 *Phys. Lett. A* **137** 329
22. Nelson P G 1991 *Rev. Sci. Instr.* **62** 2069
23. Morrison E *et al* 1992 to be published
24. Logan J E, Robertson N A, Hough J and Veitch P J 1991 *Phys. Lett. A* **161** 101
25. Saulson P R 1991 *Phys. Rev. D* **42** 2437
26. Martin W 1978 (Glasgow: Ph.D. Thesis)
27. Kerr G A and Hough J 1989 *Appl.Phys. B* **49** 491
28. Schütz I, Welling H and Wallenstein R 1990 *Advanced Solid State Lasers* (Salt Lake City)
29. Byer R L 1992 private communication.
30. Winkler W, Danzmann K, Rüdiger A and Schilling R 1991 *Phys. Rev. A* **44** 7022
31. Freischlad K, Küchel M, Wiedmann W, Kaiser W and Mayer M 1990 *Proc. SPIE* **1332** 8
32. Kawashima N 1991 *Gravitational Astronomy* ed D E McClelland and H A Bachor (Singapore: World Scientific); Fujimoto M, Ohashi M, Mio N and Tsubono K 1991 *Gravitational Astronomy* ed D E McClelland and H A Bachor (Singapore: World Scientific)
33. Gürsel Y and Tinto M 1990 *Phys. Rev. D* **40** 3884

Paper presented at the 13th Int. Conf. on General Relativity and Gravitation
Cordoba, Argentina, 1992: Part 1, Plenary Lectures

Gravitational lensing

Jürgen Ehlers and Peter Schneider
Max–Planck–Institut für Astrophysik, Karl–Schwarzschild–Str. 1, W–8046 Garching, Germany

1. Introduction

1.1. Historical Remarks

In his Opticks, published in 1704, Sir Isaac Newton already asked: "Do not Bodies act upon Light at a distance, and by their action bend its Rays?" This question was taken up only in 1784 by Henry Cavendish and, independently in 1801, by Johann von Soldner. They both computed the bending of the orbits of "lights corpuscles" by a spherical body of mass M and found the "Newtonian" deflection angle

$$\hat{\alpha} = \frac{2GM}{c^2 r} = \frac{R_s}{r} \quad , \tag{1}$$

provided $r \gg R_s$ is the impact parameter and the photon's speed at infinity is c; R_s denotes what we now call the Schwarzschild radius of the deflecting body. At the time this prediction could not be tested and was apparently soon forgotten. The question of gravitational light deflection was raised anew by Albert Einstein. In 1907 he rediscovered the law (1) by combining heuristically Maxwell's equations with his principle of equivalence, which led to an effective index of refraction

$$n = 1 - \frac{U}{c^2} \tag{2}$$

of a vacuum gravitational field depending on the gravitational potential U. Finally, in 1915, in possession of his field equation, he noted almost in passing that space curvature doubles the bending, so that (1) and (2) have to be changed into

$$\hat{\alpha} = \frac{2R_s}{r} \quad , \quad n = 1 - \frac{2U}{c^2} \quad , \tag{3}$$

results now verified by VLBI to within an uncertainty of about 10^{-3}.

Sir Arthur Eddington (1920) and O. Chwolson (1924) realised that light deflection may cause several images ("fictitious binaries", e.g.) of a source, and Einstein (1936) computed the change of apparent brightness due to differential deflection by a point mass. The observability of such effects was considered very unlikely, though, due to the small probability for sufficient alignment of sources and deflectors, taken to be stars in our galaxy. However, in 1937 in two remarkably prescient papers Fritz Zwicky considered the possible astronomical importance of gravitational light bending by external galaxies and concluded that "the probability that nebulae which act as gravitational

lenses will be found becomes practically a certainty". Possible astrophysical and cosmological applications of gravitational lensing have been elaborated theoretically since the early sixties. At GR4, held in London in 1965, Sjur Refsdal showed, among other things, how masses of galaxies and the Hubble constant might be measured by lensing.

The first verification of Zwicky's prediction occured in 1979 when D. Walsh, R.F. Carswell and R.J. Weymann tentatively interpreted a "double quasar" as a pair of images of one quasar and within one year A. Stockton and P. Young *et al* identified the lensing galaxy. Since then at least 7 firm cases of multiply imaged quasars and about 20 gravitationally lensed images (arcs, rings) of galaxies or radio lobes of quasars have been found. Theoretical activity has, of course, increased rapidly in connection with these discoveries.

1.2. The astrophysical significance of lensing

Gravitational lensing may serve to

- determine masses of galaxies and clusters and the distribution of matter in them, including dark matter;
- explain luminous arcs and rings;
- observe distant sources, magnified by gravitational lenses which act like (cheap) telescopes;
- estimate sizes of Ly–α clouds.

By microlensing, one may

- identify dark deflection bodies;
- measure the size and structure of different emission regions particularly of QSOs;
- study the graininess (stellar population) of deflecting systems.

Successful models of gravitational lens cases also serve to confirm our picture of the universe, especially the cosmological nature of QSO redshifts. Moreover, taking into account statistically the light deflection caused by inhomogeneously distributed matter provides corrections to the ideal observable relations of Friedmann universe models. Finally, such successful applications provide evidence, if only indirect one, for the validity of the law of gravity on galactic and supergalactic scales.

This brief report outlines the theoretical framework and describes some of the applications of the, by now quite extensive, field of gravitational lensing, with emphasis on the role of GR. For more details and references, we refer to a forthcoming book (Schneider, Ehlers and Falco 1992, hereafter SEF), the conference reports Moran, Hewitt & Lo (1992), Mellier *et al* (1990), Kayser & Schramm (1992), the contributions by Blandford and Fort to GR12, and the excellent review article by Blandford and Narayan (1992). The notation here follows that of SEF.

2. Theoretical framework

2.1. From Maxwell's equation to geometrical optics; Fermat's principle in GR

A locally approximately plane electromagnetic wave in an arbitrary spacetime obeys, in leading WKB order, Maxwell's vacuum field equations if and only if

$$F_{\alpha\beta} = 2R\left\{e^{iS} k_{[\alpha} A_{\beta]}\right\} \quad ,$$

where

$$k_\alpha = S_{,\alpha} \quad , \quad g^{\alpha\beta} S_{,\alpha} S_{,\beta} = 0 \quad , \tag{4a}$$

$$k^\alpha A_\alpha = 0 \quad , \quad k^\beta A_{\alpha;\beta} = -\frac{1}{2} A_\alpha k^\beta{}_{;\beta} \quad . \tag{4b}$$

According to the eikonal equation (4a) for the (real) phase S, the light rays, defined as integral curves of $\dot{x}^\alpha = k^\alpha$, are null geodesics; (4b) says that the (complex) amplitude A_α is transverse, $k^\alpha A_\alpha = 0$, and propagates along the rays such that the polarization vector is parallel on the rays and $-A^*_\alpha A^\alpha$ changes inversely to the area of the cross section of an "infinitesimal" ray bundle. The energy tensor of such a wave, averaged "over a wavelength", behaves like that of a stream of "photons" with 4–momentum $\hbar k^\alpha = P^\alpha$ and conserved 4–current number density $N^\alpha = -(2\hbar^2)^{-1} P^\alpha$; $T^{\alpha\beta} = N^\alpha P^\beta$. A radiation field may be represented classically either as a photon gas or as an incoherent ensemble of waves. It then further follows that I_ω/ω^3, where I_ω is the specific intensity and ω the circular frequency, is an observer independent invariant which is constant on each ray if there is no interaction with matter.

For gravitational lens theory, the most useful way to characterise light rays is in terms of the following version of Fermat's principle, due essentially to H. Weyl (1917!):

In a conformally static spacetime with metric ($c = 1$)

$$ds^2 = \Omega^2 \left\{ e^{2U} d\eta^2 - e^{-2U} dl^2 \right\} \quad , \tag{5}$$

where the scalar U and the Riemannian 3–metric dl^2 depend on the spatial coordinates only — the conformal factor $\Omega > 0$ may depend on all coordinates — a spacetime null curve $x^\alpha(u)$ is a geodesic if and only if its spatial projection $x^a(u)$ obeys the variational principle

$$\delta \int e^{-2U} dl = 0 \quad ; \tag{6}$$

the variation is to be done with fixed end points. Since $\eta_a = \int e^{-2U} dl$, according to (5), is the arrival time of a photon emitted at $\eta = 0$, (6) expresses the stationarity of the arrival time η_a. (For a generalisation to arbitrary spacetimes, see Perlick 1990 and SEF.)

The foregoing statements about light rays, polarization vectors and intensities, combined with approximations and statistics, form the basis of gravitational lens theory.

2.2. Time delay and Fermat potential

Let some domain of spacetime contain a point source S, a point observer O and, between S and O, a matter distribution D, small compared to the distances of D from S and O. Because of light deflection by D, O may possibly see several images of S, and if the luminosity of S changes, this change might be noticed by O in the various images at different times. The observer may be able to measure the angles between the images and between the images and D, the fluxes of the various images, the redshifts of S and D, the spectra and polarisations of the light, and the time delays between images. For an extended source, O may also observe the shapes of the images of S. One basic task of lens theory is to establish relations between (i) the observables just listed, (ii) a model of D, and (iii) a model of the universe as far as it affects those relations.

To do this requires obviously more or less drastic approximations. The least problematic one is the use of geometrical optics (reasons for this are given, e.g., in SEF). To proceed, we assume the metric to be approximated, in the relevant part of spacetime containing the light rays from S to O, by (Futamase 1989, Jacobs *et al* 1991)

$$ds^2 = R^2(\mu)\left\{(1+2U)\,d\eta^2 - (1-2U)\,dl_k^2\right\} \quad , \tag{7}$$

where dl_k^2 is a metric of constant curvature, $k = 1, -1, 0$, R the expansion function of a "background" Friedmann cosmological model with conformal time η, and U the (Newtonian) gravitational potential of the deflecting, localized matter D. Except for some neighbourhood of D, $U \ll 1$, and (7) is nearly a Robertson–Walker metric. During the time light passes D, R and U are nearly constant, and in the region where most of the deflection takes place, dl_k^2 is nearly Euclidean, so there (7) reduces to the linearized metric of a weak source. It is, then, reasonable (and usual) to assume that the relevant light rays proceed from the source to the neighbourhood of D as in the background, get deflected and retarded near D, and then proceed undisturbed to the observer.

The metric (7) is conformally static, thus Fermat's principle applies. To use it, we consider null curves consisting of two smooth, geodesic pieces, one from the source to an event D near the deflector, and one from there to the observer, with a corner at D. Their projections into the "comoving" 3–space with metric dl_k^2 are spatial geodesics with lengths l_d, l_s, l_{ds}, see Fig. 1.

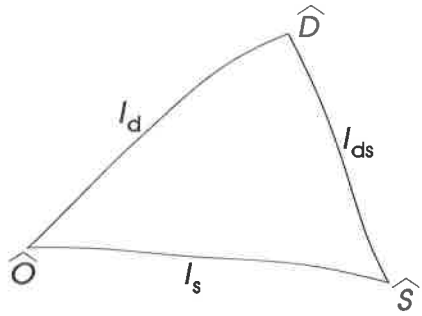

Figure 1.

$\hat{O}$, $\hat{S}$, $\hat{D}$ are the points in the comoving space representing O, S, D. The unperturbed path proceeds smoothly from $\hat{S}$ to $\hat{O}$, of course.

On each ray, since $ds^2 = 0$, we have by (7) $\int d\eta = \int dl - 2\int U dl$. Measured in proper time at the observer when $R = R_0$, the arrival time delay Δt of the broken ray relative to the unperturbed ray is therefore

$$\Delta t = R_0 \Delta\eta = R_0(l_d + l_{ds} - l_s - 2\int U dl) \quad .$$

With a little trigonometry, the Newtonian expression for U and the assumption that the deflector is transparent and geometrically thin (SEF), one arrives at a formula for the arrival time delay:

$$\Delta t = \frac{(\vec{\theta} - \vec{\beta}^2)^2}{2H_0\,(\chi_d - \chi_s)} - 2R_s\,(1+z_d)\,\Psi\,(\vec{\theta}) + \text{const.} \tag{8}$$

$\vec{\theta}$ denotes the (vectorial) angular distance of an image from the center of D, $\vec{\beta}$ the (unobservable) angular distance of the unperturbed source from the deflector, H_0 the Hubble constant, R_s the Schwarzschild radius of the deflector and

$$\Psi(\vec{\theta}) = \int \frac{dm'}{M} \ln\left|\vec{\theta} - \vec{\theta}'\right| \tag{9}$$

is a deflection potential, here defined in terms of the fraction $\frac{dm'}{M}$ of mass of D contained in a solid angle $d^2\theta'$.

The χ–terms in (8) arise as follows. The auxiliary distances l in Fig. 1 can be related to angular diameter distances D_d, D_s, D_{ds} as defined in the background universe of eq. (7). They are also unobservable, theoretical quantities. To relate them to the observable redshifts of source and deflector, z_s and z_d, one has to specify how the metric of the inhomogeneous universe is to be split into a background part and a perturbation. One way of doing this is to assume that, on average, the expansion rate of the actual universe (on the scale, say, of galaxies) equals that of the background Friedmann universe, $\langle\theta_{\text{inh.}}\rangle = \theta_F$, and $\langle\dot{u}^\alpha_{\text{inh.}}\rangle = 0$ ($\dot{u}^\alpha$ = 4–acceleration), and that the averaged area $\int dA_{\text{inh.}}$ of a cross section z = const. of a typical past light cone equals that of the background universe. Let ϱ_b be the density of that background, and let the fraction $\tilde{\alpha}$ of all mass be distributed smoothly, so that the factor $1 - \tilde{\alpha}$ is bound in clumps. The angular diameter distance will then refer to "empty cones", i.e. to ray bundles between clumps where the density is $\tilde{\alpha}\varrho_b$; these D's are called Dyer–Roeder distances. Then one gets the useful relation (SEF)

$$(1 + z_d)\,\frac{D_d D_s}{D_{ds}} = \frac{c}{H_0}\,\left[\chi_d - \chi_s\right]^{-1} \quad , \tag{10}$$

where $\chi(z; \Omega, \tilde{\alpha})$ is a function of $\tilde{\alpha}$, the present density parameter Ω, and the redshift and $\chi_d \equiv \chi(z_d; \tilde{\alpha}, \Omega)$ etc. This equation has been used to obtain (8).

The function $\Delta t(\vec{\theta}, \vec{\beta})$ clearly exhibits the influence of the lens (Ψ, R_s) and of the cosmological model (H_0, χ) on the time delay and thus of the lens mapping (see below); it deserves the name Fermat potential.

2.3. The lens mapping and related equations

Eq. (8) is valid for all kinematically conceivable, broken light rays. The actual, physical light rays follow via Fermat's principle, (6), which here gives $\frac{\partial}{\partial\vec{\theta}}\,\Delta t = 0$, hence

$$\vec{\beta} = \vec{\theta} - 2R_s H_0\,(1 + z_d)\,(\chi_d - \chi_s)\,\frac{\partial\Psi}{\partial\vec{\theta}} \quad . \tag{11}$$

It can be used to express angular distances between several images of a source in the form

$$\vec{\theta}_{ij} = \vec{\theta}_i - \vec{\theta}_j = g(z_d, z_s, \Omega, \tilde{\alpha})\,R_s H_0\,(\vec{\alpha}(\vec{\theta}_i) - \vec{\alpha}(\theta_j))$$

where $\vec{\alpha} = 2\,\frac{R_s}{D_a}\,\frac{\partial\Psi}{\partial\vec{\theta}}$ is the deflection angle. Moreover, using (11) in (8), one gets

$$\Delta t_{ij} = 2R_s\,(1 + z_d)\,h\,(\vec{\theta}_i, \vec{\theta}_j) \quad .$$

In the last two equations, g depends on the cosmological model, and h and $\vec{\alpha}$ depend on the deflector model. These equations show how, in principle, the deflector mass or R_s, H_0, Ω, $\tilde{\alpha}$ can be obtained from lens observations, provided one has a reliable model for the relative mass distribution in the lens. In practice, many difficulties have to be overcome (see the references cited).

2.4. Two general theorems, magnification

The lens mapping

$$f : \mathbb{R}^2 \to \mathbb{R}^2 , \quad \vec{\theta} \to \vec{\beta} \tag{12}$$

given by eq. (11) is, for localized lenses, invertible for large $\vec{\theta}$. In general, however, $f^{-1}(\vec{\beta}) = \left\{\vec{\theta}_1, \ldots, \vec{\theta}_{2N+1}\right\}$; the number of images is odd (Burke 1981). The inverse μ_{ij} of the Jacobian matrix

$$\beta_{ij} = \frac{\partial \beta_i}{\partial \theta_j}$$

describes the distortion of an image compared to that of the undistorted source which, of course, is not observable. However, the μ_{ij} at several images can be composed to give the relative distortion of the images.

Since lensing does not change the frequencies, constancy of I_ω implies that the flux magnification (relative to the unlensed case) is given by the absolute value of

$$\mu = \det \ \mu_{ij}$$

while sgn μ gives the parity of an image. In the case of transparent lenses, at least one image (with positive parity) is not demagnified, $\mu \geq 1$; generically $\mu > 1$ for at least one image (Schneider 1984).

2.5. Critical curves and caustics

Rescaling the variables $\vec{\beta}$, $\vec{\theta}$, one may rewrite the lens mapping (11) as

$$f : \vec{x} \to \vec{y} = \nabla \left(\frac{1}{2} \vec{x}^2 - \Psi (\vec{x}, \vec{p}) \right) \quad . \tag{13}$$

Here, $\vec{p}$ denotes parameters on which a lensing configuration may depend.

Alternatively, the source position $\vec{y}$ corresponding to an image $\vec{x}$ can be characterized by

$$\nabla \phi = 0 , \quad \phi = \frac{1}{2} (\vec{x} - \vec{y})^2 - \Psi (\vec{x}; \vec{p}) \quad . \tag{14}$$

(As before, the gradient operator refers to $\vec{x}$.)

To avoid confusion, we shall use the terminology appropriate to the *physical meaning* of the symbols; thus $\vec{y}$ is called the source position, $\vec{x}$ the image position (although according to mathematical terminology $\vec{x}$ is the pre–image, $\vec{y}$ the image variable for the map f defined by (13)).

One basic question of lens theory is: How many images exist for a given source, i.e. what is the set $f^{-1}(\vec{y})$, and how does it depend on $\vec{y}$? For localized sources the deflection potential Ψ increases like ln r, the deflection angle decreases like r^{-1}. Therefore,

as one would expect, the map f is bijective for large $|\vec{x}|$. Also, f is surjective, i.e. for any source position $\vec{y}$, there is at least one image $\vec{x}$ such that $\vec{y} = f(\vec{x})$. Thirdly, if the Jacobian matrix $\left(\frac{\partial \vec{y}}{\partial \vec{x}}\right)$ is invertible at $\vec{x}$, the map f is locally invertible at $\vec{x}$. Therefore, if $\vec{y}$ moves inwards from infinity along a curve, there will for a while be a unique image $\vec{x}$ until possibly a point is reached where $\det\left(\frac{D\vec{y}}{D\vec{x}}\right) = 0$. The set of all points $\vec{x}$ where this equation holds is called the *critical set*, its image the *caustic set* of f. In the theory of singularities of maps one studies the behaviour of f and f^{-1} near critical and caustic points, respectively, in general. Lens theory is concerned with the special case of *gradient maps*, see eq. (14).

One may study such maps from several points of view, two of which we mention.

Firstly, one may consider f as defining a 2–surface in $\mathbb{R}^4 = \{(\vec{x}, \vec{y})\}$. On this 2–surface, the sympletic form $\Omega = dy_i \wedge dx_i$ vanishes; thus the 2–surface is a Lagrangean submanifold of $(\mathbb{R}^4, \Omega)$. The lens map may then be viewed as a *projection* of this 2–surface onto the $\vec{y}$–plane. This is useful since it gives a "quasi–intuitive" insight into the way of how $f^{-1}(\vec{y})$ changes with $\vec{y}$, and since such *Lagrangean maps* have been studied extensively by mathematicians.

In the second view, based on (14), one thinks of $\vec{x} \to \phi(\vec{x}, \vec{y}; \vec{p})$, for each fixed value of the *control parameters* $(\vec{y}, \vec{p})$, as a surface in $(\vec{x}, \phi)$–space. For large $\vec{x}$, this arrival time surface approaches the paraboloid $\phi = \frac{1}{2}(\vec{x} - \vec{y})^2$. The images, in this context also called *states*, then correspond to those points where the tangent plane to the surface is parallel to the plane $\nabla\phi = 0$. (This second view corresponds to the so–called "static model" of catastrophe theory, which is popular also in discussions of phase transitions.)

By eqs. (13) and (14) the Jacobian matrix of the lens mapping has components

$$\frac{\partial y_i}{\partial x_k} = \phi_{ik} = \delta_{ik} - \Psi_{ik} \quad .$$

Since for $x \to \infty$, $\Psi_{ik} \to 0$, the critical set is bounded and, of course, closed, hence compact. The caustic set is compact and in addition of measure zero (Sard's theorem).

At a regular, i.e. non–critical image $\vec{x}^{(0)}$, the magnification μ is finite, and near $\vec{x}^{(0)}$, f is locally diffeomorphic. At a critical image $\vec{x}^{(0)}$, the magnication is formally infinite; however, if instead of a point source an extended source is considered or if wave optics is taken into account, the "physical" magnification becomes finite, but in general large.

It follows from the above that if $\vec{y}$ moves, the number of its images can change only if $\vec{y}$ enters or crosses the caustic set. Thus knowing the critical and caustic sets provides a qualitative overview of a lens mapping.

Generically, the critical set of a lens mapping consists of finitely many closed, smooth, compact curves without end points. The images of critical curves, the caustics, may have finitely many *cusps* (spikes).

Whenever a source point crosses a caustic where the latter is smooth, two new images appear on opposite sides of the critical curve; the corresponding segments of critical curves which consist of "double images", are called *folds*. Near and inside a cusp, a source point has three images which merge when the source reaches the cusp, and only one image survives if the source has passed through the cusp. Both folds and cusps are stable, i.e., they are preserved under all small deformations of the deflection potential, and smooth maps $\mathbb{R}^2 \to \mathbb{R}^2$ have no other kinds of stable singularities than folds and cusps (Whitney 1955). Near a fold, the amplification diverges like the inverse square root of the distance from the caustic; near a cusp, it diverges even stronger.

These facts can be established by approximating the Fermat potential at a critical point by a polynomial of suitably high order and then studying the resulting representative mapping.

If one considers not a single lens mapping, but a family of those depending on some parameters $\vec{p}$, the critical curves and caustics depend on the value of $\vec{p}$. Qualitative changes of the pattern of critical curves and caustics, called *metamorphoses*, can occur for particular values $\vec{p}^{(0)}$ and at special points $(\vec{x}^{(0)}, \vec{y}^{(0)})$. These higher–order singularities are also useful to survey lensing models; they are needed to study caustics and self–intersections of null cones of spacetimes, but we shall not pursue this topic here, but refer the reader to Chap. 6 of SEF (see also Blandford & Narayan 1986).

2.6. Remarks on mutliple lens plane theory

The theory outlined so far effectively deals with cases where, because of the thin–lens assumption and small deflection angles, the lensing matter may be thought of as being distributed on a single lens plane between source and observer. Instead, one may consider the more general case where several such lens planes occur (Blandford *et al* 1986, Kovner 1987, SEF), an idealisation which better approximates the real situation where clumps at all redshifts influence light propagation.

The resulting mapping may again be obtained via Fermat's principle. It is, of course, more complicated than the single plane case. In particular, in contrast to (14) the mapping is no longer a gradient map. (Question: is it still Lagrangean?) The odd number theorem remains valid, and recently it has been shown that this also holds for the magnification theorem stated in 2.4 (S. Seitz *et al* 1992).

3. Remarks about the observational situation

Gravitational lens phenomena, predicted long ago, have been observed in a variety of appearances. First, there are cases of strong lensing; by that we mean that a deflector is sufficiently massive and compact to split images of a background source, or at least to distort them in such a way that the lens phenomenon is readily identified or at least suspected. Examples of strong lensing are multiply imaged QSOs, the luminous arcs in clusters of galaxies, and the radio rings.

In contrast to strong lensing, weak lensing cannot be identified in individual sources, but only by considering a sample of sources. For example, background galaxies are distorted by a foreground cluster in a characteristic way (images are preferentially elongated in a direction tangent to the cluster center). However, since galaxies are not intrinsically round, the distortion cannot be identified from any individual galaxy image, but only the statistical analysis of the images of background sources around a cluster yields evidence for the coherent image distortion. Weak lensing particularily means that the magnifications of the images are not very much larger than unity.

A third class of lensing phenomena is microlensing: light bundles from distant compact sources 'feel' the graininess of the matter distribution in intervening galaxies. The stars in such galaxies cause only small deflections, but their differential deflection can be sufficient to yield considerable magnification.

This is not the place to provide an – even only partially – complete review of observed lensing phenomena and specific objects; the reader is referred to the reviews mentioned

in the introduction, and to the report by Fort (1990) from the last GR conference and references therein. Instead, we will pick out a few aspects which may be of particular interest to relativists.

3.1. Multiple QSOs

Up to now, we know about seven firm cases of multiply imaged QSOs, and about as many good candidates for this effect. The criteria by which a candidate lensing system becomes a secure case of lensing are not easily explicitly stated, but having two QSO images with the same redshift and similar spectra is not sufficient to enter the class of lens system. In fact, the lensing community has learned its lesson from objects like 1145−071, where two closely spaced QSO images have the same redshift, similar spectra, but only one of them is a strong radio source, with very strict limits on the radio flux ratio. Although one could construct complicated (and artificial) lens models which account for the observations, this system most likely is a binary QSO, rather than a lensed source. In order for a multiple QSO to be accepted as a lens system, one or more of the following criteria should be satisfied in addition to the similarity of spectra: (a) in the case of two images, (a1) they both should be radio QSOs, and the radio characteristics should agree (spectral slope, morphology in resolved sources), (a2) a potential lens should be seen between or near the images, (a3) both QSOs should show other relatively rare properties, such as broad absorption lines, (correlated) rapid variability, high polarization etc. (b) If more than two images are seen, the lensing hypothesis becomes more probable a priori, in particular if the geometrical arrangement of the images is according to the expectations from 'generic elliptical' lens models.

The two doubles confirmed as lens systems both have a visible galaxy between the images; in addition, for 0957+561, both components are radio sources, and VLBI has revealed the similar milliarcsecond radio structure (also showing the different parities of the two images), which yields particularily striking evidence for the lensing nature of this system. For the one confirmed triple, 2016+112, there are several objects close to the images, two of which have been identified as galaxies (at different redshifts); in this case, light from the QSO is probably deflected more than once. The four known quadruples, the best-known of which is the so-called Einstein cross 2237+0305, have their images arranged in a form which is predicted by all canonical models, i.e., either rather symmetrically, in which case the images have comparable brightness, or two of the images are close together (as in the case of the so-called 'triple-QSO' 1115+080) and are much brighter than the remaining two, indicating that the close pair of images lies close to a critical curve of the lens. For 2237+0305 the lens is easily observable (this lens system will be discussed in more detail below), and the lens in 1115+080 has also been found. The image separation of these multiply imaged QSOs range from slightly more than one arcsecond to about six arcseconds.

Except for 2016+112, a case which is theoretically not well understood, all multiply imaged QSOs have an even number of images, in contrast to the theoretical expectation mentioned in Sect.2.4. However, the odd number theorem made no prediction about the relative brightnesses of the images. It is currently believed that the 'missing image' is very close to the center of the lensing galaxy where the surface mass density is so high as to yield a strong demagnification (overfocusing) of this central image. In fact, none of the multiple QSO lens systems yields a hint of a finite core radius of the lens.

There are several candidates for multiply imaged QSO, mostly doubles (since for them verification is most difficult). Perhaps the most interesting case is 2345+007, a

double QSO with 7.3 arcsecond separation, the largest known. Spectral similarities are striking; the report (Nieto et al 1988) that one of the images is indeed a close double has not been confirmed by other groups. No sign of any lens has been obtained in deep imaging of the field; however, a strong absorption line system at redshift 1.49 indicates matter in the lightbeam in one of the two images (that which is suspected as double). As will be discussed below, clarification of the lensing nature of 2345+007 is of vital importance for an application to the determination of intervening Lyα clouds. The status of the double system 1635+267 ($\Delta\theta = 3.8$ arcseconds) is similar to that of 2345+007: the spectra of both images are remarkably similar, and this system is an excellent candidate for lensing. Recently, a very close pair ($\Delta\theta \approx 0.45$ arcseconds) has been identified, both through ground-based observations (Magain et al. 1992) and from the Space Telescope gravitational lens snapshot survey (Maoz *et al* 1992); the QSO has a redshift of 3.8, making it the lens candidate with the largest redshift. The small image separation observed in that system is close to the most probable separation as expected from theory, if standard assumptions about the distribution of galaxies in the universe (i.e., the lens population) are made.

3.2. Rings

From symmetry, if a source is situated behind a (sufficiently compact) spherical deflector, its image will be a ring. However, exact symmetry is not needed for the formation of ring-shaped images, if extended sources are considered. Five ring images have been found to date (see contributions in Kayser and Schramm 1992), all by radio observations. In the first case (1131+0456), optical imaging suggests a similar ring-like structure for the optical source (Hammer & LeFevre 1991), but the best argument in favour of lensing is detailed modeling: since the image of an extended source contains much more information than a few point images, typical lens models for rings are highly overdetermined. The reconstruction of the image from a simple lens model, applied to data in two wavebands and to the polarized flux, provides a powerful tool not only to constrain lens parameters (such as mass and ellipticity), but also to confirm the lensing nature (Kochanek *et al* 1988); recently, this inversion technique has been greatly improved to account for finite resolution maps (Kochanek & Narayan 1992). For a second ring, MG1654+1346, the lensing nature is indisputable: one of the two radio lobes of a QSO at redshift 1.7 is mapped into a ring image by a galaxy at redshift 0.25, centered on the ring. A third ring source, PKS1830−211, one of the strongest radio sources in the sky, contains two compact components which are rapidly variable; monitoring of this source and detailed modeling could make this an ideal target for applying the lens method to determine the Hubble constant (see below).

3.3. Arcs and arclets

In about 10 clusters of galaxies, long and narrow images (arcs) have been observed. Whereas the nature of these images has been controversial right after their discovery, the redshift measurement of one of these arcs have ruled out interpretations which put the arc at the same redshift as the cluster. Several such redshift measurements are available today, all with higher values than the cluster in which the arc is seen.

Presently, the nature of the arcs is interpreted as galaxies at high redshift being gravitationally lensed by foreground clusters. The highly elongated shape of the images

is due to the distortion by the lens mapping, with the source being situated close to a cusp singularity of the lens. (There are also "straight arcs", which are most conveniently interpreted as sources lying close to so-called beak-to-beak singularities, one of the types of metamorphoses mentioned in Sect. 2.5.) Note that the length of the arc in A370 is about 21 arcseconds, and its width is roughly 2 arcseconds, so that the deformation by lensing is indeed huge; correspondingly, the total flux of the image is much larger than that of the unlensed source. It is only this high magnification which allows spectroscopy of such intrinsically faint extended sources – clusters of galaxies forming arcs provide us with (cheap) 'natural telescopes'.

If a cluster is sufficiently compact to form such spectacular long arcs, it will also deform other background sources lying close to the direction to the center of the cluster. One example of this effect can be seen in A370, where various elongated blue images are observed. These images share the property that they are elongated in the tangential direction with respect to the cluster center. Such tangential elongation is expected from light deflection. Indeed, the redshift of one of these arclets in A370 has been measured to be $z_s = 1.305$ which thus stems from a background source. Since the density of faint background galaxies is very large (Tyson 1988), these sources can in principle be used to 'map' the mass distribution of compact clusters. This is not a trivial task, both from observation and theory. Very deep images have to be obtained in order for the source density to be sufficiently high. Sources (galaxies) are intrinsically not round, and therefore intrinsic ellipticity has to be disentangled from lens-induced deformation. This clearly is an ambitious statistical problem, treated in considerable depth in the literature (e.g., Kochanek 1990, Miralda-Escudé 1991), but has been demonstrated to work in practice (Tyson, Valdes & Wenk 1990), thus constituting a nice example of weak lensing. In principle, the redshift distribution of this faint background galaxy population can be obtained from such studies of arclets around a sample of clusters of galaxies.

3.4. Microlensing

The graininess of the matter in galaxies affects the light bundles of sufficiently small sources traversing such a galaxy. The relevant scale for the source is about $3 \times 10^{16} \sqrt{M/M_\odot}$ cm, so that the optical (and X-ray) continuum emission from AGNs can be affected by this effect, whereas the broad line region is probably too large to be affected as a whole (differential effects, however, affecting the shape of the broad emission lines, can occur). Since the relative alignment of source and lensing galaxy changes in time, so does the magnification: microlensing leads to a lens-induced variability of sources seen through a galaxy. This effect, in general, is extremely difficult to distinguish from intrinsic variability of sources, but for multiply imaged QSOs, such microlensing could be observed: an intrinsic variation of the source will be seen in all images, with their respective time delay, whereas variability due to microlensing is uncorrelated between the individual images.

The best lens system for observing microlensing is 2237+0305, owing to the small redshift ($z_d \approx 0.04$) of the lens (so that velocities in the lens plane correspond to large effective velocities in the source plane) and to the fact that the time delay of the images is of the order of one day, so that intrinsic flux variations of the source will be seen nearly simultaneously in all four images. In fact, uncorrelated flux variations in at least three images have been observed (Corrigan et al. 1991), ranging up to half a magnitude. This is a clear signature of microlensing.

It is relatively difficult to obtain quantitative conclusions from such microlensing observations. One can compare observed lightcurves with numerical simulations (see Wambsganss 1990 and references therein), but due to the statistical nature of the effect definite conclusions will only be possible if a sample of microlensing events is observed, which means that a sufficiently long data track of the source must be available. Nevertheless, the observations of microlensing in 2237+0305 have led to the following conclusions: the observed flux variations are compatible with the picture in which microlensing is produced by a normal stellar mass spectrum (Wambsganss, Paczyński & Schneider 1991, Witt, Kayser & Refsdal 1992). The size of the emission region of the optical continuum source of 2237+0305 is just compatible with a simple accretion disc model (Rauch & Blandford 1991, Jaroszyński, Wambsganss & Paczyński 1992); the observed variations lead to an upper limit of the source size.

In addition, microlensing can be used to obtain information about the brightness structure of sources. For example, the lightcurve of an extended source crossing a caustic is a convolution of its one-dimensional brightness profile and a universal function, the point-source magnification function near folds. Hence, from a well-observed high-magnification event in an observed microlensing light curve, the one-dimensional brightness profile of a source could in principle be reconstructed (Grieger 1990). The broad-line region in QSOs, probably too extended to be magnified as a whole by microlensing, has intrinsic structure. Therefore, differential magnification across the broad line region can affect the line profiles of the broad emission lines (Schneider & Wambsganss 1990). In fact, line profile differences in 2237+0305 have been observed by Fillipenko (1989). For further applications of microlensing, see Sect. 12.4 of SEF.

4. Applications of gravitational lensing effects

As mentioned before, gravitational light deflection does not only lead to spectacular effects as multiple images and rings, but has become a useful tool for astrophysics. Here, we want to mention only some of the potential applications which have been suggested in the literature, concentrating on those which might interest the cosmologists most.

4.1. Determination of the Hubble constant

As suggested as early as 1964 by S. Refsdal, the difference in light travel time between any two images of a multiply imaged QSO can be used to determine, at least in principle, the Hubble constant, i.e., the overall scale of a gravitational lens arrangement (from dimensional arguments, $H_0 \propto 1/\Delta t$). The only lens system which is sufficiently intrinsically variable and sufficiently monitored by now to detect correlated flux variations in both images (with a respective time delay) is 0957+561, but even in this system it has been extremely difficult to measure Δt. Only recently, by using a sophisticated statistical method applied to optical and radio data (Press, Rybicki & Hewitt 1992a,b) a reliable estimate has been obtained, $\Delta t = 540 \pm 12$ days. The correlation of flux of both images is not perfect, indicating that at least one of the images is affected by microlensing (Schild & Smith 1991).

The determination of the Hubble constant from the measurement of the time delay, as discussed above, requires knowledge about the mass distribution of the lens. However, in this respect the system 0957+561 is a fairly complicated one, since the cluster

in which the main lensing galaxy is embedded makes a significant contribution to the deflection. Unfortunately, basically nothing is known about the mass distribution of the cluster. The construction of lens models proceeds by assuming that the deflection caused by the cluster varies slowly over the region of the size of the image separation; the contribution by the cluster is then described by the lowest order terms of a Taylor expansion of the deflection angle around the center of the main galaxy (the validity of this approach can be questioned, see Kochanek 1991a). A 'plausible' ansatz for the shape of the mass distribution in the lens galaxy is chosen, compatible with what is known about the matter distribution in elliptical galaxies, and the parameters of the model are varied to obtain a match of all observables with the model. Needless to say that such an approach yields models which are far from being unique. Even worse, there are invariance transformations of lens model parameters which have no impact on the observables (Gorenstein, Falco & Shapiro 1988; e.g., adding a uniform mass sheet to the lens acts like a Gaussian thin lens and is equivalent, in terms of the lens model, to decreasing the separation between lens and source). This degeneracy can be broken by obtaining additional observables; in the case of 0957+561, the velocity dispersion of the lens galaxy (a measure for the lens mass) has been observed, thus allowing the degeneracy to be broken. It thus seems that a point estimate of the Hubble constant is possible from this lens system. Taking the lens model from Falco, Gorenstein & Shapiro (1991), assuming that the velocity dispersion measures the total mass of the lens galaxy, and taking a cosmological model with $\Omega_0 = 1$, one obtains a value for H_0 which is less than $50\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ (the 'best' value being closer to 40). The uncertainties, however, are still considerable. First, as already mentioned, the lens model is not unique, and one can construct equally good models which would yield different values for H_0. Second, it is not clear whether the velocity dispersion of the stars probes the total matter of the lens galaxy; if the stellar mass is more centrally concentrated than the dark matter in the galaxy, the effective dispersion can be larger than the measured value by up to a factor $\sqrt{1.5}$. Third, additional inhomogeneities around the line-of-sight to the QSO can perturb the propagation of light and thus affect the lens mapping; in this respect it is interesting to know that spectroscopy of the galaxies around 0957+561 indicates that they do not all belong to a cluster at $z_d = 0.36$, but that there seems to be an additional concentration at a higher redshift of about 0.54.

With all the difficulties and uncertainties mentioned, it might appear that the determination of the Hubble constant from lensing will not yield more accurate values than the classical method of 'climbing up the distance ladder'. However, it should be stressed that the latter method measures the Hubble constant from nearby objects, whereas lensing permits to obtain measurements of H_0 on truly cosmic scales. It is by no means evident that both methods should yield the same results, in fact: since the probability that a galaxy (on which an observer is situated to measure H_0) is placed in an overdense region of the universe is high, the local Hubble expansion is likely to be slower than the mean Hubble expansion, yielding a systematically higher value of H_0 from local measurements. This was demonstrated quantitatively by Turner, Cen & Ostriker (1992). Hence, measuring H_0 from lensing might provide one of the few ways to obtain that value of H_0 which enters the Friedmann equations.

4.2. Constraints on the cosmological constant

Gravitational lensing provides a method to constrain the magnitude of a possible cosmological constant. Three different approaches have been used to obtain such con-

straints: as pointed out by Gott, Park & Lee (1989), the lensing behaviour of galaxies changes drastically if the cosmological constant is so large as to have an antipode at a redshift smaller than that of the source in a lens system. The confirmed lens system with the highest redshift ($z_s = 3.27$) is 2016+112, which implies that $q_0 > -2.3$; this estimate can be improved once lens systems with sources at higher redshifts are confirmed.

A positive cosmological constant λ increases the volume elements corresponding to a fixed redshift interval and solid angle. Assuming no strong evolution of the number and mass distribution of galaxies, the probability for any given source to be multiply imaged can be calculated as a function of source redshift and λ. Lensing probabilities become very large once λ approaches unity (for flat universes, i.e., $\Omega_0 + \lambda = 1$; see, e.g., Fukugita & Turner 1991, Fukugita *et al* 1992). Whereas existing lens surveys are burdened with selection biases (e.g., Kochanek 1991b), it seems unlikely that the results obtained so far are compatible with large λ cosmologies.

The effect of λ on the volume per redshift element depends on z. If a multiply imaged QSO is selected without regards to a lens (thus excluding systems like 2237+0305 for which the discovering observation was targeted to the lens galaxy), one can calculate the probability distribution for the lens redshift, which depends on λ; in particular, the probability is shifted towards larger redshifts of the lens for increasing λ. From available data, a large λ model can nearly be ruled out, and a few more lens systems adequate for this test can yield strong constraints on λ (Kochanek 1992). Note that these statistical effects are much more sensitive to λ than to Ω_0. However the uncertainties in lensing statistics are considerable (e.g., Mao 1991).

4.3. Mass distribution in lenses

Modeling an observed lens system yields values of the parameters of a parametrized lens model. Needless to say that there are many possible choices for the lens parameters, and so values obtained for physical parameters are not unique. The question is whether there are some physical parameters which are robust against variations of the parametrization of the lens model.

For double QSOs, there is not much we can learn about the matter distribution of the lens without a priori assumptions about its characteristics. An example of this can be found in Borgeest (1986). What is fairly well determined is that part of the mass of the lens which lies in a cylinder with axis along the line-of-sight and with radius being the mean of the angular separations of the two images from the center of the lensing galaxy (this mass estimate includes the mass from, e.g., a cluster in which the main lens galaxy is embedded).

The situation improves if quadruple QSOs are considered. In that case, the mass of the lens lying within a cylinder with radius being the mean of the separation of the images from the lens center is well determined. The best example of this is again 2237+0305: Observations with Space Telescope have revealed the image positions to very high accuracy. On the other hand, the closeness of the lens allows a detailed map of its brightness distribution. Assuming a constant mass-to-light ratio (M/L), surface brightness is directly translated into surface mass density. As has been shown recently (Rix, Schneider & Bahcall 1992), the assumption of constant M/L, i.e., a model with just three parameters (the other two being the unknown source position), is able to reproduce the positions of the images (8 observational constraints!) to within very

high accuracy. From that we can conclude that a constant M/L is a good assumption for this lens galaxy, and its value agrees with that obtained from stellar dynamics for 2237+0305 (Foltz *et al* 1992) and for other galaxies. The mass within the inner kpc of this galaxy is found to be $(2.16 \pm 0.04) \times 10^{10} M_{\odot}$ for a Hubble constant of 50, i.e., accurate to within 2%. This model reproduces the observed brightness ratios of the images only to within a factor of 1.5, which however is not surprising, since the flux of the images has been observed to be affected by microlensing.

For ring images, detailed reconstruction can yield fairly good estimates of the mass within the ring and the orientation and ellipticity of the mass distribution. The observational predictions of such lens models are very insensitive to the distribution of the matter inside the ring and nearly unaffected by the matter outside the ring.

Detailed modeling of arcs in several clusters of galaxies has provided us with some crucial information about the mass distribution in clusters: First, the amount of dark matter, inferred previously from studies of the dynamics of cluster galaxies, agrees with that obtained from the lens models (as is the case for galaxy-mass lenses, lens models for arcs are far from unique, with the mass inside a circle of radius given by the distance of the arc from the center of the cluster being the most robust parameter). Second, the dark matter is not tied to individual galaxies, but spread more evenly throughout the clusters. Third, the mass inside clusters of galaxies must be distributed more compactly than previously thought; otherwise, clusters would not be able to form caustics and thus multiple images. Once a complete sample of clusters (selected, say, by their X-ray flux; it seems that the X-ray luminosity of a cluster is a good indicator of its lensing power) is scrutinized for the occurrence of arcs, statistical information about the mass spectrum of clusters, their core radii etc. will be available (see Bergmann 1992, Wu and Hammer 1992).

4.4. Dark matter in our Galaxy

The rotation curve of our Galaxy and that of other spirals indicates the presence of dark matter. It is unclear what the nature of this dark matter is; candidates are: weakly interacting elementary particles, brown dwarfs and 'Jupiters', or black holes. If the dark matter is composed of such compact objects, these can lead to lensing effects of background sources: for stellar mass lenses, the corresponding angular separation of split images would be much too small to be detectable, but the magnification could be observed. The problem with this idea is that the probability of finding a single lens sufficiently close (within its Einstein radius) to the line of sight of an extragalactic object is about 10^{-6}. Hence, for observing this effect a huge number of sources must be photometrically monitored.

Paczyński (1986) suggested monitoring of the stars of a nearby galaxy, the Large Magellanic Cloud or M31. In fact, two groups are currently carrying out such an observational program (see the corresponding contributions in Kayser & Schramm 1992). The difficulties for such a program are enormous; we want to mention only a few of them. First, many stars are intrinsically variable, and it will be a major task to separate intrinsic variability from lens-induced one. A magnification event will be a 'once in a lifetime' event, that is, events will not recur. The program will work by comparing the brightnesses of sources from consecutive observations. The required large number of stars implies that the amount of data produced in the course of the program will be huge.

On the other hand, such an experiment is probably the only way to clarify the nature of the dark matter in the halo of our Galaxy. It will be sensitive to compact objects of masses between $10^{-6}M_\odot$, set by the time resolution, and about $1M_\odot$, determined by the duration of the experiment (note that this is the relevant mass range for 'brown dwarfs': objects with mass smaller than $\sim 10^{-7}M_\odot$ will have evaporated during the lifetime of our galaxy, whereas objects with mass $\gtrsim 0.1M_\odot$ burn hydrogen and shine, see De Rujula, Jetzer & Masso 1992). Paczyński (1991) suggested a calibration experiment, to observe the galactic bulge stars. There, we know the minimum density of lenses along the line of sight (the normal disk population of stars), and the probability of lensing turns out to be quite similar to that of the halo. Hence, this calibration experiment can be used to test the sensitivity of the observational program, including the software for data analysis, and may detect a population of brown dwarfs in the disk. Whatever the outcome of the program will be, it certainly will produce the most useful database for stellar variability and is thus worth the effort.

4.5. Massive black holes in the universe

If some fraction of the matter in the universe is contained in compact objects (defined such that their mass lies inside their Einstein radius; note that even red giants are compact objects with this definition), lensing can in principle reveal such a cosmic population (Press & Gunn 1973). Compact objects with mass $\gtrsim 10^{11}M_\odot$ lead to image splitting with angular separation $\gtrsim 1$ arcsecond, and their density can be constrained by the fraction of multiple QSOs. This fraction is small (at least for normal QSOs; the apparantly most luminous ones can have a larger fraction of lens systems due to the amplification bias, see Sect. 12.5 of SEF) and can be accounted for by the normal galaxy population (within all the uncertainties mentioned). Certainly, a mass fraction $\Omega \gtrsim 0.01$ in compact objects with mass $\gtrsim 10^{12}M_\odot$ can be excluded from existing data. The Space Telescope snapshot survey will be most valuable in obtaining tighter constraints (Bahcall *et al* 1992).

Lower mass objects can be constrained by high resolution observations. Using a sample of VLBI observations of compact radio sources, Kassiola, Kovner & Blandford (1991) have excluded a mass fraction $\Omega \gtrsim 0.4$ of compact objects in the mass range $10^7M_\odot \lesssim M \lesssim 10^9M_\odot$. In principle, using a homogeneous sample of compact radio sources, the mass range can be lowered to about $10^5M_\odot$, and a sample of about 1000 objects can limit Ω in that mass range to better than 0.01.

If the dark matter in galaxies is made of high mass black holes with $M \sim 10^6M_\odot$, their presence should show up in high-resolution VLBI observations of multiply imaged QSOs (Wambsganss & Paczyński 1992), as they would leave an imprint on the VLBI maps which can be detected by comparing the milliarcsecond structure of the images. Compact objects of about a solar mass can in principle be detected by their microlensing power, leading to lens-induced variability. But, as discussed in Sect. 1.4, it is extremely difficult to distinguish this effect from intrinsic variability. If gamma-ray bursts originate from sources at cosmological distances, for which the evidence is now accumulating (e.g., Mao & Paczyński 1992), lensing of a burst source by a compact object will lead to two images, which cannot be spatially resolved by gamma-ray telescopes, but they can be resolved in time: the time delay is about $50(M/M_\odot)$ seconds. Hence, a cosmologically significant density of compact lenses in the appropriate mass range can be detected from recurrent bursts (see Mao 1992, Narayan & Wellington 1992 for details).

4.6. Dark matter on large scales

The universe is inhomogeneous on scales larger than galaxy clusters; however, the mass distribution on such large scales is no longer sufficiently compact to yield strong lensing events. Still, it may be possible to detect the impact of such large scale inhomogeneities on the propagation of light rays. At least two such possibilities have been discussed recently in the literature.

A light bundle propagating through the universe is affected by the shear (Weyl focusing) produced by the inhomogeneities. Thus, an intrinsically round source will attain a slight ellipticity. So far, the situation is similar to the formation of arclets in clusters, except that here no 'center of gravity' can be identified in general. But, as demonstrated in Blandford *et al* (1991, see also Miralda-Escudé 1991b, Kaiser 1992), the shear of large-scale inhomogeneities will lead to a nonzero two-point correlation function of the ellipticities of images from high-redshift sources. The sky is fairly densely covered with faint galaxies, which are supposed to have redshifts $\gtrsim 1$ (Tyson 1988); this source population is thus the natural candidate for such an investigation, which, due to the faintness of the galaxies and their intrinsic ellipticities will be difficult to carry out. The shape of the two-point correlation function reflects the spectrum of the inhomogeneities of large-scale matter in the universe.

Another method for detecting large-scale matter inhomogeneities has been recently pointed out by Bartelmann & Schneider (1992). Motivated by the claim of Fugmann (1990) that there is a large-scale density excess of galaxies from the Lick catalogue around high-redshift flat-spectrum radio sources, we studied the lensing effect of the large-scale matter distribution in the universe on high-redshift objects. Although the corresponding magnifications are very small, they can be detected in a statistically significant way if the source population has a steep intrinsic luminosity function (or if multiple waveband amplification bias is taken into account, see Borgeest, von Linde & Refsdal 1991). In this case, the sources will be preferentially observed behind overdense structures. If one now assumes (as is generally done) that galaxies find themselves in overdense regions, they trace the overdensities and a large-scale correlation of high-redshift sources and galaxies is expected. These effects are, however, tedious and require robust statistical methods to extract them from observations.

4.7. The size of Lyα clouds

Each QSO at sufficiently high redshift, such that the Lyα emission line can be observed from Earth, shows a large number of narrow absorption lines shortward of the Lyα emission line. This so-called 'forest' is interpreted as being due to intervening material in the line-of-sight to the QSO, and the absorption lines being due to the Lyα transition. It is less clear what the nature of the absorbers is; one can measure the redshift, equivalent width (yielding the column density of neutral hydrogen) and line width. Gravitational lens systems provide us with an invaluable tool to determine the size of this absorbing material.

Consider a lens system in which a high-redshift QSO has two observable images. The transverse separation between the corresponding light bundles varies as a function of redshift: it is zero at the source, and largest at the redshift of the lens. Suppose an absorption line is seen in the spectra of both images, at the same redshift; then the size of the absorbing material must be at least as large as the separation between the two light bundles at the respective redshift. Hence, redshift coincidences of absorption

lines leads to a lower bound on the size of the absorbing material. If, in addition, the equivalent widths of the corresponding lines in both spectra are strongly correlated, it could be concluded that the two light bundles actually have crossed the same cloud (as opposed to crossing two different clouds at the same redshift).

Two lens systems have turned out to be useful for such studies. As mentioned above, the lensing nature of one of them, 2345+007, has not yet been totally clarified, but it is an excellent candidate system. The large angular separation of 7.3 arcseconds makes this a very useful target for these spectroscopic investigations (Foltz *et al* 1984), although the faintness of the QSO images renders such an investigation difficult. If this is indeed a lens system, and if the lens redshift is 1.49 as suggested by the strong iron absorption at this redshift, the size of Lyα clouds can be estimated to be $10\,\mathrm{kpc} \lesssim R \lesssim 20\,\mathrm{kpc}$ (Bajtlik, personal communication). A second system, UM673 with an image separation of 2.2 arcseconds, is brighter and thus higher-quality spectroscopy can be obtained. Results of such a study are reported in Smette *et al* (1992); due to the smaller image separation, only sharp lower limits can be obtained, which are completely compatible with the value quoted above.

One can turn this argument around: suppose the size of the Lyα clouds were known; then, a correlation analysis of the absorption line spectra of multiple QSOs could be used to decide whether the multiple system is due to lensing. Note that for this kind of argument, the size of the clouds need to be known only approximately, since the transverse separation of light rays corresponding to multiple QSO images differs greatly for physical pairs of QSOs and for lensed sources. Thus, this Lyα diagnostics may turn out to be the most powerful tool to decide about the lensing nature of multiple QSO systems.

5. Concluding remarks

The investigation of gravitational lensing has provided us with a powerful tool in extragalactic astronomy. Currently, the interplay of observations and theory is most fruitful for both: lensing has triggered renewed interest in old topics like simple geometrical optics or light propagation in curved spacetime, whereas on the other hand, theoretical predictions can be tested observationally only if highest quality data are obtained. It is possible that gravitational lensing effects (microlensing) will shed light on the structure of QSOs and can yield interesting cosmological information, most noticibly about the Hubble constant. The history of gravitational lensing has provided us with quite a few surprises – arcs were not predicted, for example – and some visionary ideas have come true finally, such as the detection of microlensing, first predicted by Chang and Refsdal in 1979. On the other hand, we see a great potential of applying statistical lensing to well-defined samples of sources once these become available. Whatever the main direction of future research in lensing, exciting results can be anticipated.

References

Bahcall, J.N. *et al* : 1992, ApJ 387, 56.
Bartelmann, M. & Schneider, P.: 1992, AA submitted.
Bergmann, A.: 1992, PhD Thesis, Stanford University.
Blandford, R.D. & Narayan, R.: 1986, ApJ 310, 568.

Blandford, R.D. & Narayan, R.: 1992, Ann. Rev. Astr. Ap. (in press)(BN).
Blandford, R.D., Saust, A.B., Brainerd, T.G. & Villumsen, J.V.: 1991, MNRAS 251, 600.
Borgeest, U.: 1986, ApJ 309, 467.
Borgeest, U., von Linde, J. & Refsdal, S.: 1991, AA 251, L35.
Burke, W.L.: 1981, ApJ 244, L1.
Chang, K. & Refsdal, S.: 1979, Nature 282, 561.
Corrigan, R.T. *et al* : 1991, Astron. J. 102, 34.
De Rujula, A., Jetzer, Ph. & Masso, E.: 1992, AA 254, 99.
Dyer, C.C. & Roeder, R.C.: 1973, ApJ 180, L31.
Ehlers, J. & Schneider, P.: 1986, AA 168, 57.
Falco, E.E., Gorenstein, M.V. & Shapiro, I.I.: 1991, ApJ 372, 364.
Filippenko, A.B.: 1989, ApJ 338, L49.
Foltz, C.B., Weymann, R.J., Röser, H.-J. & Chaffee, F.H.: 1984, ApJ 281, L1
Foltz, C.B., Hewett, P.C., Webster, R.L. & Lewis, G.F.: 1992, ApJ 386, L43.
Fort, B.: 1990, in: *General Relativity and Gravitation*, Ashby, N., Bartlett, D.F. & Wyss, W. (eds.), Cambridge University Press.
Fugmann, W.: 1990, AA 240, 11.
Fukugita, M., Futamase, T., Kasai, M. & Turner, E.L.: 1992, ApJ 393, 3.
Fukugita, M. & Turner, E.L.: 1991, MNRAS 253, 99.
Futamase, T. & Sasaki, M: 1989, Phys. Rev. D 40, 2502.
Futamase, T.: 1989, MNRAS 237, 187
Gorenstein, M.V. *et al* : 1988, ApJ 334, 42.
Gorenstein, M.V., Falco, E.E. & Shapiro, I.I.: 1988, ApJ 327, 693.
Gott, J.R., Park, M.-G. & Lee, H.M.: 1989, ApJ 338, 1.
Grieger, B.: 1990, in: Mellier *et al* 1990, p.198.
Hammer, F. & Le Fevre, O.: 1990, ApJ 357, 38.
Hammer, F. & Le Fevre, O.: 1991, ESO Messenger 63, 59.
Hewitt, J.N. *et al* : 1988, Nature 333, 537.
Huchra, J. *et al* : 1985, Astron. J. 90, 691.
Jacobs, M.W., Linder, E.V. & Wagoner, R.V.: preprint Stanford, ITP–905
Jaroszyński, M., Wambsganss, J. & Paczyński, B.: 1992, preprint.
Kaiser, N.: 1992, ApJ 388, 272.
Kassiola, A., Kovner, I. & Blandford, R.D.: 1991, ApJ 381, 6.
Kayser, R. & Schramm, T.(ed.): 1992, *Proceeding of the Hamburg conference on gravitational lensing*, Springer-Verlag (in press).
Kent, S.M. & Falco, E.E.: 1988, Astron. J. 96, 1570.
Kochanek, C.S.: 1990, MNRAS 247, 135.
Kochanek, C.S.: 1991a, ApJ 382, 58.
Kochanek, C.S.: 1991b, ApJ 379, 517.
Kochanek, C.S.: 1992, ApJ 384, 1.
Kochanek, C.S., Blandford, R.D., Lawrence, C.R. & Narayan, R.: 1988, MNRAS 238, 43.
Kochanek, C.S. & Narayan, R.: 1992, preprint.
Langston, G.I. *et al* : 1990, Nature 344, 43.
Magain, P., Surdej, J., Vanderriest, C., Pirenne, B. & Hutsemekers, D.: 1992, AA 253, L13.
Mao, S.: 1991, ApJ 380, 9.
Mao, S.: 1992, ApJ 389, L41.
Mao, S. & Paczyński, B.: 1992, ApJ, submitted.
Maoz, D. *et al* : 1992, ApJ 386, L1.
McKenzie, R.H.: 1985, J. Math. Phys. 4, 1194.
Mellier, Y., Fort, B. & Soucail, G.(ed.): 1990, *Gravitational Lensing*, Lecture Notes in Physics 360, Springer-Verlag (Berlin).
Miralda-Escudé, J.: 1991a, ApJ 370, 1.
Miralda-Escudé, J.: 1991b, ApJ 380, 1.
Moran, J.M., Hewitt, J.N. & Lo, K.Y.(ed.): 1989, *Gravitational Lenses*, Lecture Notes in Physics 330, Springer-Verlag (Berlin).
Narayan, R.: 1992, in: Kayser & Schramm 1992.
Narayan, R. & Schneider, P.: 1990, MNRAS 243, 192.

Narayan, R. & Wellington, S¿: 1992, ApJ, in press.
Nieto, J.-L. *et al* : 1988, ApJ 325, 644.
Ostriker, J.P. & Vietri, M.: 1985, Nature 318, 446.
Paczyński, B.: 1986, ApJ 304, 1.
Paczyński, B.: 1991, ApJ 371, L63.
Perlick, V.: 1990, Class. Quantum Grav. 7, 1319.
Press, W.H. & Gunn, J.E.: 1973, ApJ 185, 397.
Press, W.H., Rybicki, G.B. & Hewitt, J.N.: 1992a, ApJ 385, 404.
Press, W.H., Rybicki, G.B. & Hewitt, J.N.: 1992b, ApJ 385, 416.
Rauch, K. & Blandford, R.D.: 1991, ApJ 381, L39.
Refsdal, S.: 1964, MNRAS 128, 307.
Rix, H.-W., Schneider, D.P. & Bahcall, J.N.: 1992, ApJ, in press.
Roberts, D.H. *et al* : 1985, ApJ 293, 356.
Schild, R. & Smith, R.C.: 1991, Astron. J. 101, 813.
Schneider, P.: 1984, AA 140, 119.
Schneider, P., Ehlers, J. & Falco, E.E.: 1992, *Gravitational Lenses*, Springer-Verlag (New York)(SEF).
Schneider, P. & Wambsganss, J.: 1990, AA 237, 42.
Seitz, S. & Schneider, P.: 1992, AA (submitted).
Sommerfeld, A.: 1959, *Vorlesungen über Theoretische Physik, Bd.IV. Optik*, Geest & Portig (Leipzig.
Smette, A. *et al* : 1992, ApJ 389, 39.
Stickel, M., Fried, J. & Kühr, H.: 1988a, AA 198, L13.
Stickel, M., Fried, J. & Kühr, H.: 1988b, AA 206, L30.
Stickel, M., Fried, J. & Kühr, H.: 1989, AA 224, L27.
Stockton, A.: 1980, ApJ 242, L141.
Turner, E.L., Cen, R. & Ostriker, J.P.: 1992, AJ 103, 1427.
Tyson, J.A.: 1988, Astron. J. 96, 1.
Tyson, J.A., Valdes, F. & Wenk, R.A.: 1990, ApJ 349, L1.
Walsh, D.: 1989, in: Moran *et al* 1989, p.11.
Wambsganss, J.: 1990, *Gravitational Microlensing*, Ph.D. Thesis, University of Munich, available as MPA report 550.
Wambsganss, J. & Paczyński, B.: 1992, preprint.
Wambsganss, J., Paczyński, B. & Schneider, P.: 1991, ApJ 358, L33.
Whitney, H.: 1955, Ann. Math. 62, 374.
Witt, H.J., Kayser, R. & Refsdal, S.: 1992, AA, submitted.
Wu, X.P. & Hammer, F.: 1992, preprint.
Yee, H.K.C.: 1988, Astron. J. 95, 1331.
Young, P. *et al* : 1980, ApJ 241, 507.

Paper presented at the 15th Int. Conf. on General Relativity and Gravitation
Cordoba, Argentina, 1992: Part 1, Plenary Lectures

Applications of numerical relativity: critical behavior and black-hole containing spacetimes

Charles R. Evans

Department of Physics and Astronomy
University of N. Carolina, Chapel Hill, NC 27599

Abstract. Several recent applications of numerical relativity techniques are described. The first is the remarkable evidence of critical phenomena in general relativity associated with gravitational collapse of massless fields. We discuss the two currently known examples of such critical behavior: gravitational collapse of massless scalar field and of axisymmetric gravitational waves. A second application is discussed, in which numerical relativity techniques are used to probe the interaction between strong gravitational waves and black holes. Finally, a numerical approach to computing initial data for spacetimes that contain two black holes is described. Such initial data may ultimately serve as a starting point for the computation of the orbital decay and coalescence of a black hole binary.

1. Introduction

Very near or shortly after the turn of the century, the Laser Interferometer Gravitational-wave Observatory (LIGO) is expected to detect signals from compact binary star systems in their final stages of orbital decay and coalescence [1, 2]. The theoretical foundation for making precise predictions of the phase of the orbit during orbital decay, and using such a prediction to fit observed signals to determine stellar and orbital parameters, is the subject of intense present-day scrutiny [3, 4], as various speakers at the conference have attested. In contrast, theoretical predictions of the signals from the coalescence itself, at least for the coalescence of two black holes, have yet to be made. It seems likely that to make these predictions the full machinery of numerical relativity will be required, and then only after further advances have been achieved in supercomputer technology and additional experience is gained in multidimensional simulations.

Since the development of the cosmic censorship hypothesis [5] over two decades ago, no definitive theorems have been advanced to prove that singularities will always be clothed by event horizons. Nonetheless, support for the conjecture is provided by linear stability analysis of black holes and, with few exceptions, by numerical relativity calculations of gravitational collapse. The latter refers to the fact that in nearly every

gravitational collapse calculation in numerical relativity to date, in which the impending development of a singularity was indicated, an event horizon (and a black hole) has been shown to form. A notable exception is the computation carried out by Shapiro and Teukolsky [6] of collapse of a prolate cluster of collisionless matter. This topic is discussed in some detail by Teukolsky in this volume and will not be addressed further here except to note that cosmic censorship and modeling sources of gravitational radiation are precisely the types of issues that numerical relativity, in principle, can serve to address.

Numerical relativity includes, broadly speaking, any large-scale computational approach to a problem in classical general relativity. It usually refers to calculations of the complete Einstein equations, although a variety of distinct schemes (e.g., spacelike 3+1 [7], 2+2, and null-cone [8] decompositions) have been developed to provide numerical solutions. Occasionally, the term is used to refer to multidimensional Newtonian and post-Newtonian calculations as well, such as Finn's [9] models of gravitational radiation from rotating stellar core collapse and Nakamura and Oohara's and Rasio and Shapiro's [10] simulations of binary neutron star merger. It encompasses applications of Regge calculus also [11]. In general, numerical relativity is used to denote numerical approaches to any problem involving at least two nontrivial dimensions and the solution of partial differential equations. Numerical relativity has as its goal to provide solutions of the field equations in circumstances where an analytic approach is not feasible, which typically means spacetimes that lack a high degree of symmetry or that involve strong, dynamic, and nonlinear fields. Some common applications include *i)* modeling sources of gravitational radiation, *ii)* nonspherical gravitational collapse and explorations of cosmic censorship [6], *iii)* inhomogeneous cosmologies, *iv)* black-hole collisions [12], and *v)* strong gravitational wave–black hole interactions.

While this presentation is meant to be partly in the nature of a review, with so many areas of active research it proved necessary to be somewhat selective in the topics to be treated. I have opted to narrow the focus considerably and review *i)* the very recent discovery of critical phenomena in general relativity, *ii)* numerical models of the interaction between strong gravitational waves and black holes, and *iii)* the numerical construction of initial data for spacetimes containing two black holes (a possible starting point for the coalescence problem). With the focus of the article thus restricted, this discussion unavoidably manages to slight the work of many other researchers and we must instead merely refer the reader to several recent, more-complete, book-length reviews [13, 14].

2. Critical behavior in gravitational collapse

Choptuik [15, 16] recently discovered the existence of critical phenomena in general relativity. The critical behavior is associated with spacetimes that come very close to forming a black hole and those that just manage to do so. Choptuik discovered these effects by computing the gravitational collapse of spherically-symmetric wavepackets of massless scalar field $\phi(r,t)$ using a sensitive finite-difference method. The finite difference code is based upon an adaptive-mesh-refinement algorithm developed by Berger and Oliger [17] that is particularly well suited to follow the development of very fine spatial and temporal features. A second example of critical phenomena has been found by

Abrahams and Evans [18] who considered numerical models of collapse of axisymmetric gravitational waves. These spacetimes were also computed by using a finite difference method, but in this case with a modestly-adjustable moving-mesh algorithm and not with a full adaptive-mesh-refinement scheme.

Some of the characteristics of the critical phenomena seen in these simulations can be summarized in general terms. Critical phenomena become evident in the simulations as variations in the properties of spacetimes across a parameter space of spacetimes. While there are many ways of parameterizing spacetimes, we will restrict attention to parameter spaces that are each described by a single parameter. We can consider a set $\mathcal{G}$ of many such distinct one-dimensional parameter spaces $\mathcal{G}_k$ of spacetimes $\mathcal{S}_k[p_k]$. For each (appropriately defined) parameter space $\mathcal{G}_k$, a critical value $p_k^\star$ of the parameter p_k separates the parameter space into a half space $\mathcal{G}_k^+$ of spacetimes that contain a black hole and a half space $\mathcal{G}_k^-$ of spacetimes that do not. In this way, the parameters p_k are associated with variations in the strength of the ensuing gravitational self-interaction. Critical behavior occurs in a neighborhood of the critical parameter value, when $|p_k - p_k^\star| < \delta_k$, and hence in spacetimes that are just on either side of the "edge" of forming a black hole. Certain features of individual spacetimes display variations across a parameter space that can be considered *critical*. For example, in $\mathcal{G}_k^+$, black-hole mass is found to be a critically-behaving quantity with a power-law dependence $C|p_k - p_k^\star|^\beta$ on the separation of p_k from $p_k^\star$ and with a critical exponent β. The critical behavior of black-hole mass is reminiscent of spontaneous magnetization of ferromagnets in statistical physics, and suggests [18] that black-hole mass plays the rôle of order parameter in general relativity. Near $p_k^\star$, whether a black hole forms or not, the individual spacetimes develop a strong-field region $\mathcal{R}$ that exists during the height of the gravitational self-interaction and within a small enough neighborhood of the center of the implosion. In this strong-field region the gravitational field (and any coupled field) develops an oscillatory character. Close examination of these oscillations has revealed the existence of scaling relations that make successive oscillations echoes of each other on progressively smaller spatial and temporal scales.

Choptuik [16] has been able to demonstrate the *universality* of these critical phenomena in scalar field collapse. (It is likely, but not yet established, that universality is a feature of axisymmetric gravitational wave collapse as well.) To discuss universality the more general space $\mathcal{G}$, of many distinct one-parameter spaces $\mathcal{G}_k$, comes into play. Each $\mathcal{G}_k$ contains a critical spacetime $\mathcal{S}_k[p_k^\star]$ associated with a critical parameter value $p_k^\star$. Universality refers to the fact that, in scalar-field collapse at least, the shape of the fields in the critical solution $S_k[p_k^\star]$ deep in the strong-field region (and hence the scaling relation) and the value of the critical exponent for black-hole mass both do not depend on which parameter space, $\mathcal{G}_k$, of $\mathcal{G}$ is examined. In other words the critical phenomena are generic features independent of the details of the initial data. The critical exponent and scaling relation may depend, however, on the type of field (e.g., minimally-coupled and non-minimally-coupled scalar fields, electromagnetic field, gravitational field, etc.) that induces the collapse and, in a related fashion, on the degree of symmetry (or nontrivial dimensionality) of the spacetimes, but in ways that are not yet understood. The two known examples of critical phenomena in gravitational collapse illustrate these issues and the present rudimentary state of our knowledge.

2.1. Critical phenomena in massless scalar field collapse

The initial discovery [16] of critical behavior was found by modeling collapse of spherically-symmetric wavepackets of scalar field. The equation of motion of the scalar field is

$$\phi_{;\mu}{}^{;\mu} = \zeta R\phi, \tag{1}$$

and Choptuik considers [15] both minimally-coupled ($\zeta = 0$) and non-minimally-coupled fields. In spherical symmetry, the line element is taken to have the form

$$ds^2 = -\alpha^2(r,t)dt^2 + a^2(r,t)dr^2 + r^2 d\Omega^2, \tag{2}$$

with α the lapse function and a the radial metric function. This gauge is a generalization of Schwarzschild coordinates for dynamical spacetimes. The lapse is fixed by adopting the polar time slicing condition [19] and the spatial coordinate trajectories are fixed to be normal to the time slices by adopting a vanishing shift vector $\beta^r = 0$. Defining auxiliary variables

$$\Phi = \phi', \qquad \Pi = \frac{a}{\alpha}\dot{\phi}, \tag{3}$$

the equations Choptuik solves (in the minimally-coupled case) are

$$\dot{\Phi} = \left(\frac{\alpha}{a}\Pi\right)', \tag{4}$$

$$\dot{\Pi} = \frac{1}{r^2}\left(r^2\frac{\alpha}{a}\Phi\right)', \tag{5}$$

$$\frac{\alpha'}{\alpha} - \frac{a'}{a} + \frac{1-a^2}{r} = 0, \tag{6}$$

$$\frac{a'}{a} + \frac{a^2-1}{2r} - 2\pi r(\Phi^2 + \Pi^2) = 0, \tag{7}$$

where a dot and a prime denote $\partial/\partial t$ and $\partial/\partial r$, respectively. As mentioned before, finite difference equations are obtained from these partial differential equations and solved by using an adaptive-mesh-refinement algorithm. The adaptive-mesh-refinement algorithm dynamically monitors the local truncation errors and maintains a limit on the size of these errors by producing, as needed, local, nested refinements in the mesh in both space and time. Hence, any sharp spatial and temporal features that might develop can be followed with this scheme, whereas they would otherwise become underresolved and lost in a simulation using a fixed mesh.

A typical one-parameter space of solutions is generated from initial (Cauchy) data that takes the scalar field ϕ to have an initial profile

$$\phi(r,0) = \phi_0 r^3 e^{-[(r-r_0)/\Delta]^q}, \tag{8}$$

and demands, as a condition on Π, that the scalar radiation be purely ingoing initially. There are several parameters in these data, but if r_0, Δ and q are considered fixed, then ϕ_0 serves as a single parameter p characterizing the sequence (i.e., a particular subspace

$\mathcal{G}_k$). The initial data are completed by solving the slicing condition (6) for α and the Hamiltonian constraint (7) for a.

The parameter ϕ_0 is monotonically related to the strength of the self-interaction. For small ϕ_0 the wavepacket implodes, passes through itself and then disperses to infinity, while for sufficiently large ϕ_0 the imploding wavepacket forms a black hole. There exists along the sequence a critical value, $\phi_0 = \phi_0^\star$, at which a black hole first appears and which separates supercritical ($\phi_0 > \phi_0^\star$) solutions from subcritical ($\phi_0 < \phi_0^\star$) ones. The solutions of greatest interest are those with parameter values ϕ_0 close to the critical value $\phi_0^\star$. In regions of parameter space near $\phi_0^\star$, Choptuik has found that the natural variable for characterizing variations in parameter space is

$$\pi = \ln|\phi_0 - \phi_0^\star|. \tag{9}$$

Stated another way, critical features in solutions that lie near the critical point tend to depend linearly on π, and therefore *exponentially* on the initial conditions. Additionally, structures with increasingly finer spatial and temporal scales develop as $\phi_0 \to \phi_0^\star$.

The echoing behavior and scaling relation can best be described in terms of two alternative variables, $X = \sqrt{2\pi}r\Phi/a$ and $Y = \sqrt{2\pi}r\Pi/a$, which are useful because they are invariant with respect to rescalings of the length and time coordinates ($r \to \kappa r$ and $t \to \kappa t$), and hence to rescaling of the mass of the spacetime. In near-critical solutions and in the strong-field region, a number of oscillations of the scalar field appear, with their number being proportional to $|\pi|$. It is therefore conjectured that every critical solution (one from each $\mathcal{G}_k$) will contain an infinite number of echoes. To describe why these are echoes and to express the scaling relation, we need logarithmic spatial and temporal coordinates ρ and τ defined by

$$\rho = \ln r, \tag{10}$$

$$\tau = \ln(T^\star - T), \tag{11}$$

where r is the proper (areal) radius and T is the proper time of the central observer at $r = 0$. The constant $T^\star$ is the finite accumulation time of the echoes in the precisely critical solution and is a value that can be determined in near-critical solutions by fitting. Choptuik finds [16] that an approximate scaling relation holds for the oscillations of the scalar field:

$$X(\rho - \Delta, \tau - \Delta) = X(\rho, \tau), \tag{12}$$

$$Y(\rho - \Delta, \tau - \Delta) = Y(\rho, \tau), \tag{13}$$

for a particular logarithmic scaling constant Δ. This makes the oscillations appear as echoes of one another but on scales progressively finer by a factor $e^{-\Delta}$. Stated another way, if we observe the radial profiles of X and Y at some time T_1, which gives a small interval $\delta T_1 = T^\star - T_1$ before $T^\star$ but is otherwise arbitrary, and again at a second time T_2 with the still smaller interval $\delta T_2 = e^{-\Delta}\delta T_1$ before $T^\star$, then a new feature will have appeared in the later profiles on a finer scale but the new profiles are in fact identical to the earlier ones upon rescaling radially by a factor e^{Δ}.

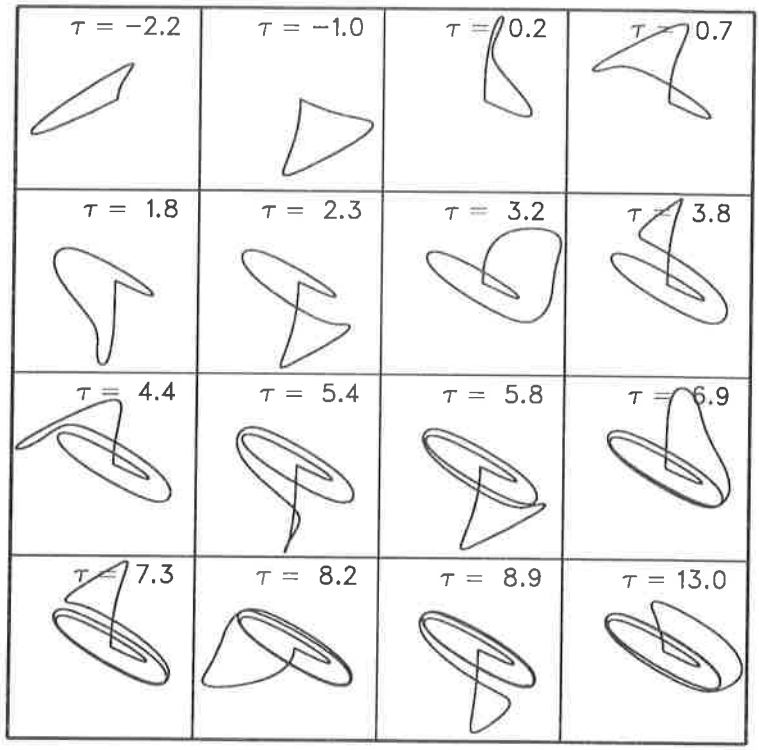

Figure 1. Evolution of the phase portrait of a near-critical solution of the minimally-coupled scalar field–gravity equations. Each portrait is the figure formed by plotting $Y(\rho,\tau)$ versus $X(\rho,\tau)$ at fixed time τ with ρ serving as curve parameter. Phase portraits at successive times are labeled by their values of $|\tau|$. Echoes in the strong-field region appear as self-similar wraps about the nearly elliptical "attractor," whose shape is found to be universal.

The scaling relation is approximate because even in the precisely critical solution the initial oscillations near the outer edge of the strong-field region contain information about the initial data. This information is washed out with each echo, making adherence to the scaling relation progressively tighter. Furthermore, any solution that is near critical, but not precisely critical, will produce only a finite number of echoes as $T^\star$ is approached before it "decides" whether to form a black hole or not.

These echoing properties are illustrated in Figure 1 by plotting, for a particular near-critical solution, the parametric relationship that can be formed between $Y(\rho,\tau)$ and $X(\rho,\tau)$ at fixed time τ with the radial coordinate ρ serving as curve parameter. This produces a phase portrait of the solution at fixed times. The phase portraits at successive times are shown labeled by their values of $|\tau|$. As $T \to T^\star$ a series of oscillations appear, each evident as a loop about $X = Y = 0$. The fact that the loops are self-similar, nearly identically overlapping each other and producing something analogous to an attractor, illustrates the echoing property and the existence of the scaling relation. The value of

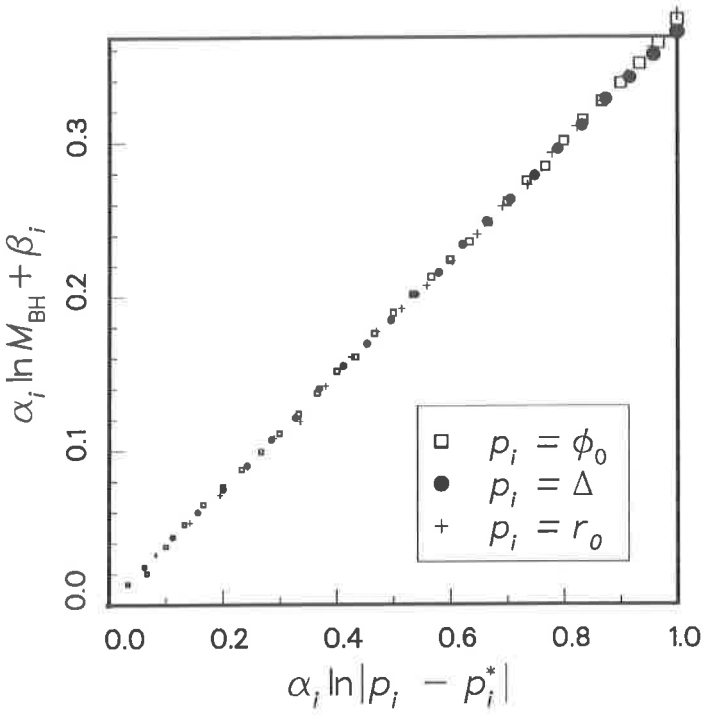

Figure 2. Illustration of the power-law dependence of black-hole mass versus critical separation in parameter space for scalar field collapse. The figure plots $\ln M_{\rm BH}$ versus $\ln(\phi_0 - \phi_0^\star)$. Data from three separate one-parameter spaces are displayed, providing evidence of the universality of the critical behavior of black-hole mass. The slope is the value of the critical exponent, in this case estimated to be $\simeq 0.37$.

Δ in the scaling relation is found to be $\Delta \simeq 3.4$. Both Δ and the profiles, $X(\rho,\tau)$ and $Y(\rho,\tau)$, (and hence the shape of the attractor) are found to be universal, or independent of the family of initial data. Weak dependence on the coupling constant ζ may exist [15].

Equally intriguing behavior is found in one-parameter sequences of solutions from the half-spaces $\mathcal{G}_k^+$ that contain black holes. For the initial data discussed previously, these correspond to $\phi_0 \to \phi_0^\star$ from above. The masses of these black holes have been shown to fit a power law

$$M_{\rm BH} \simeq C|\phi_0 - \phi_0^\star|^\beta, \tag{14}$$

with a critical exponent, $\beta \simeq 0.37$ (see Figure 2). The power-law behavior of $M_{\rm BH}$ is also found to be universal, or independent of the family of initial data, and once again only weak dependence exists, if any, on the coupling constant ζ. The immediate conjecture is that a black hole first appears along any sequence at $p = p^\star$ with infinitesimal mass.

Because of the drastic change in scale with each successive echo ($e^{-\Delta} \simeq 1/30$), the scaling relation probably would not have been discovered without the resolving power afforded by Choptuik's implementation of the Berger-Oliger algorithm. In some of the models Choptuik has computed, the adaptive-mesh-refinement scheme has provided local resolution equivalent to a single uniform grid of 10^9 zones. This has allowed Choptuik to probe points in parameter space with critical separations as small as $|(p-p^*)/p^*| \lesssim 10^{-13}$.

2.2. Critical phenomena in axisymmetric gravitational-wave collapse

Abrahams and Evans [18] have recently found a second example of critical phenomena in general relativity that occurs as axisymmetric *gravitational* wavepackets collapse. These spacetimes are quite distinct physically from Choptuik's spherically-symmetric models of scalar field collapse in two ways: they are source-free, $T^{\mu\nu} = 0$, and have less symmetry (one instead of two Killing vectors). Of course, they also necessarily involve a dynamical degree of freedom of the gravitational field. Several numerical models from a one-parameter family of initial data were discussed at the conference and described elsewhere [20]. While the sequence had a critical parameter value, these particular models each had parameter values far from critical and served merely to show that collapse of gravitational wavepackets would either form a black hole or lead to dispersal of the packet following implosion. In this paper it is now possible to describe the properties of near-critical solutions and to show [18] the existence of critical phenomena similar to that seen in scalar field collapse.

Abrahams and Evans [18] compute axisymmetric, asymptotically-flat vacuum spacetimes using the 3+1 formalism [7]. The coordinates are fixed by adopting the maximal time-slicing condition and the quasi-isotropic spatial gauge. Allowing only one dynamical degree of freedom, the line element takes the form

$$ds^2 = -\alpha^2 dt^2 + \phi^4[e^{2\eta/3}(dr + \beta^r dt)^2$$

$$+r^2 e^{2\eta/3}(d\theta + \beta^\theta dt)^2 + e^{-4\eta/3} r^2 \sin^2\theta d\varphi^2], \tag{15}$$

where α is the lapse function, β^r and β^θ are shift vector components, ϕ is the conformal factor, and η is the even-parity "dynamical" metric function. Maximal slicing results from the condition $K^i{}_i = 0$ on the extrinsic curvature $K^i{}_j$. Abrahams and Evans compute numerical solutions of the following equations:

$$\partial_t \hat{\lambda} = \mathcal{D}_\beta[\hat{\lambda}] - \phi^6(D^r D_r \alpha + 2D^\varphi D_\varphi \alpha)$$

$$+\alpha\phi^6(R^r{}_r + 2R^\varphi{}_\varphi) + \frac{\hat{K}^r{}_\theta}{r}\left[r\partial_r\beta^\theta - \partial_\theta\left(\frac{\beta^r}{r}\right)\right], \tag{16}$$

$$\partial_t \hat{K}^\varphi{}_\varphi = \mathcal{D}_\beta[\hat{K}^\varphi{}_\varphi] - \phi^6 D^\varphi D_\varphi \alpha + \alpha\phi^6 R^\varphi{}_\varphi, \tag{17}$$

$$\partial_t\left(\frac{\hat{K}^r{}_\theta}{r}\right) = \mathcal{D}_\beta\left[\frac{\hat{K}^r{}_\theta}{r}\right] - \frac{1}{r}\phi^6 D^r D_\theta \alpha + \frac{1}{r}\alpha\phi^6 R^r{}_\theta$$

$$+(2\hat{\lambda} - 3\hat{K}^\varphi{}_\varphi)\left[\partial_\theta\left(\frac{\beta^r}{r}\right) - \alpha\frac{K^r{}_\theta}{r}\right], \tag{18}$$

$$\partial_t \eta = \beta^r \partial_r \eta + \beta^\theta \partial_\theta \eta + \partial_\theta \beta^\theta - \beta^\theta \cot\theta + \alpha\lambda, \tag{19}$$

$$\Delta_f^{(3)} \psi = -\frac{1}{4}\psi \left(\Delta_f^{(2)} \eta + \frac{1}{2}\psi^{-8} e^{-2\eta} \hat{K}^i{}_j \hat{K}^j{}_i \right), \tag{20}$$

$$\Delta_f^{(3)} (\alpha\psi) = -\frac{1}{4}\alpha\psi \left(\Delta_f^{(2)} \eta - \frac{7}{2} A^2 K^i{}_j K^j{}_i \right), \tag{21}$$

$$r\partial_r \left(\frac{\beta^r}{r} \right) - \partial_\theta \beta^\theta = \alpha(2\lambda - 3K^\varphi{}_\varphi), \tag{22}$$

$$r\partial_r \beta^\theta + \partial_\theta \left(\frac{\beta^r}{r} \right) = 2\alpha \frac{K^r{}_\theta}{r}, \tag{23}$$

where

$$K^i{}_j K^j{}_i = 2\lambda^2 - 6\lambda K^\varphi{}_\varphi + 6(K^\varphi{}_\varphi)^2 + 2\left(\frac{K^r{}_\theta}{r} \right)^2, \tag{24}$$

and where the transport operator $\mathcal{D}_\beta$ is defined by

$$\mathcal{D}_\beta [u] = \frac{1}{r^2}\partial_r [r^2 \beta^r u] + \frac{1}{\sin\theta}\partial_\theta [\sin\theta \beta^\theta u]. \tag{25}$$

In these equations, $\lambda = K^r{}_r + 2K^\varphi{}_\varphi$, $\hat{K}^i{}_j = \phi^6 K^i{}_j$, D_k is the spatial covariant derivative, $R^i{}_j$ is the spatial Ricci tensor, $A = \phi^2 e^{\eta/3}$, $B = \phi^2 e^{-2\eta/3}$, $\psi = B^{1/2}$ and $\Delta_f^{(3)}$ and $\Delta_f^{(2)}$ are the three- and two-dimensional, flat-space Laplacians, respectively. The analytic properties of this gauge have been discussed previously [21, 22, 19, 23] and the reader is referred to these sources for more details on the coordinates, boundary conditions and asymptotic properties of the gauge. Similarly, portions of the finite difference method have been described elsewhere [21, 22] and will not be elaborated here.

To find Cauchy data for the gravitational field, the freely-specifiable fields, η and $K^r{}_\theta$, are taken to have the form of a *linear* ingoing gravitational wavepacket possessing quadrupolar ($\ell = 2$) angular dependence. The general linear $\ell = 2$ solution is described by a quadrupole moment $I(v)$ of arbitrary profile in advanced time v (or retarded time u). The linear solution involves the quadrupole moment $I(v)$, its first two derivatives, $I^{(1)}(v) \equiv dI(v)/dv$ and $I^{(2)}(v)$, and its integrals, $I^{(-1)}(v) \equiv \int^v dv' I(v')$ and $I^{(-2)}(v)$. Expressions for η and $K^r{}_\theta$ that are consistent with this solution have been found:

$$\eta = \left(\frac{I^{(2)}}{r} - 2\frac{I^{(1)}}{r^2} \right) \sin^2\theta, \tag{26}$$

$$\frac{K^r{}_\theta}{r} = \left(\frac{I^{(2)}}{r^2} - 3\frac{I^{(1)}}{r^3} + 6\frac{I}{r^4} - 6\frac{I^{(-1)}}{r^5} \right) \sin 2\theta. \tag{27}$$

However, since a wavepacket of finite amplitude confined within a finite radius will generate a finite mass, these Cauchy data will be at least slightly nonlinear, depending on the initial amplitude and radius. So to find proper data, the exact Hamiltonian and momentum constraints are solved for ϕ, λ, and $K^\varphi{}_\varphi$, subject to the choice above of η and $K^r{}_\theta$. Not surprisingly, nearly-linear Cauchy data are found still to generate ingoing

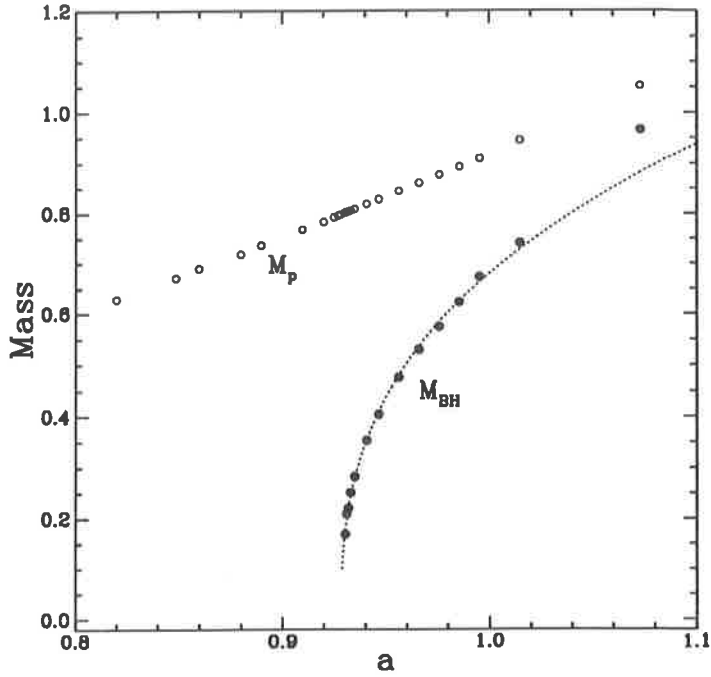

Figure 3. Power-law dependence of black-hole mass versus critical separation in parameter space for axisymmetric gravitational-wave collapse. Remarkably, the value of the critical exponent in this case, $\simeq 0.37$, is indistinguishable from that of scalar field collapse.

solutions. To complete the specification of data, the form of $I(v)$ must be given. A wavepacket with polynomial radial dependence of the form $I^{(-2)}(v) = a\kappa_p L^5[1-(v/L)^2]^6$, for $|v| = |r - r_0| < L$ at $t = 0$, is chosen. Here, κ_p is a constant but a is an amplitude parameter, L is a width parameter, and r_0 is a centering parameter. Each of these might serve as useful parameters of spaces $\mathcal{G}_k$. Initially, L and r_0 have been fixed while a has been chosen to parametrize the Cauchy data and therefore the solutions. Since in the limit $a \to 0$ the mass of the wavepacket is $M_p^{\rm linear} = a^2 L/(2\pi)$, a useful alternative strength parameter is $\Theta(a) = 2\pi M_p/L \simeq a^2$. A wavepacket with $\Theta \ll 1$ only weakly self-interacts [20], escaping to infinity virtually unaffected, while a black hole forms in an evolution where $\Theta \gtrsim 1$, with $M_{\rm BH} \to M_p$ as $\Theta \to \infty$. The critical value along the sequence is found [18] to be $\Theta^\star \simeq 0.80$ ($a^\star \simeq 0.93$).

Like in the scalar field case, supercritical collapse of gravitational wavepackets is found to generate black-hole masses, $M_{\rm BH}$, that are well described by a power law

$$M_{\rm BH} \simeq C(a - a^\star)^\beta. \tag{28}$$

Quite remarkably, the critical exponent value obtained for gravitational wavepacket collapse is also $\beta \simeq 0.37$ and is presently indistinguishable from that seen in scalar field collapse (the estimated numerical uncertainties place the value between 0.35 and 0.39).

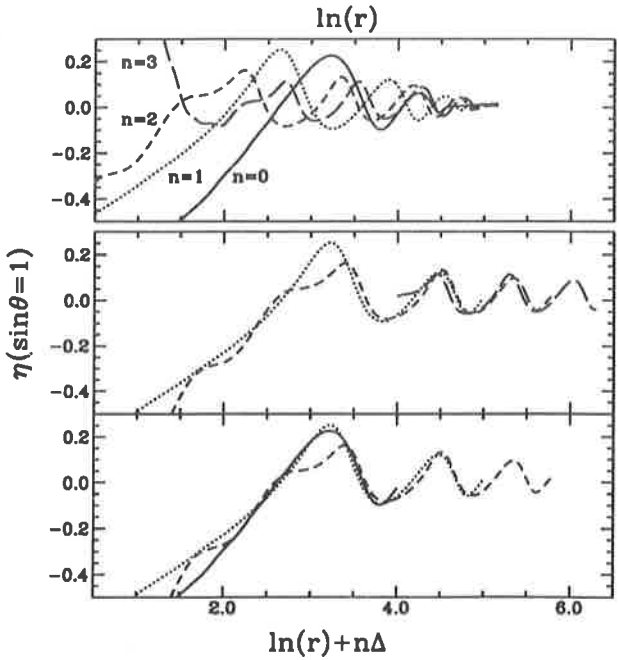

Figure 4. Scaling property of a near-critical solution of axisymmetric gravitational-wave collapse. Radial profiles of the metric function η (along $\theta = \pi/2$) plotted at four times corresponding to alternate maxima of the central value of the lapse function, α_c. The upper panel depicts all four profiles (labeled sequentially $n = 0-3$) plotted versus $\rho = \ln r$. The two lower panels illustrate scaling by overlapping profiles that are shifted by $\rho \to \rho' = \rho + n\Delta$ with $\Delta \simeq 0.6$. Profiles $n = 0, 1, 2$ are plotted in the bottom panel and $n = 1, 2, 3$ in the middle panel.

Figure 3 shows the power-law behavior of black-hole mass found for gravitational-wave collapse. The mass $M_{\rm BH}$ refers to the mass $\sqrt{\mathcal{A}_{ah}/16\pi}$ associated with the area of the apparent horizon $\mathcal{A}_{ah}$, and this is determined at a time $\Delta t = 2\pi/\omega_1^{\ell=2}$ (where $\omega_1^{\ell=2}$ is the real part of the lowest-order $\ell = 2$ quasinormal mode frequency) after the apparent horizon first appears.

Tentative evidence is also observed for a scaling relation on the gravitational field in the strong-field region $\mathcal{R}$. The gravitational field is observed to oscillate on progressively finer spatial and temporal scales. This behavior is evident in examining radial profiles of η (along the equatorial plane $\theta = \pi/2$) as displayed in Figure 4. As can be seen, η exhibits an echoing in $\rho = \ln r$ of the form

$$\eta(\rho - \Delta, t_n) \simeq \eta(\rho, t_{n+1}). \tag{29}$$

The times t_n are found, in this case, by using the central value of the lapse function $\alpha(t, r = 0)$ as a diagnostic to determine the completion of successive oscillations. Once

again a single value of Δ is found to describe the scaling of the echoes. However, in this case, the scaling constant is $\Delta \simeq 0.6$ and it therefore implies a radial scale ratio $e^{\Delta} \simeq 1.8$ which differs considerably from the corresponding value of $e^{\Delta} \simeq 30$ ($\Delta \simeq 3.4$) in scalar-field collapse. This result appears to be robust, having been obtained in simulations with several different resolutions. It is fortunate the scale ratio is much less extreme in this case, since it was obtained using a 2+1 finite difference code without the use of adaptive mesh refinement. Existing 2+1 and proposed 3+1 computations cannot come close to the resolution afforded by adaptive mesh refinement in 1+1, so extension of the adaptive-mesh-refinement scheme to 2D and 3D will be well worthwhile.

2.3. Tentative conclusions

The results obtained so far are suggestive of critical phenomena, but it is fair to ask how close the association really is to standard critical phenomena. Answering this question will likely require the construction of analytic models and more simulations of additional physically-distinct spacetimes with different symmetries and sources. We have seen so far two remarkably similar values of the critical exponent β and two quite different values for the logarithmic scaling factor Δ. The differences may be attributable to the different dimensionalities (or number of Killing vectors) of these two physical models. Do these two physical models then represent different universality classes, despite the nearly identical values of β? The appearance of black holes in only those solutions with $p > p^*$, and the reasonable conjecture that a hole of infinitesimal mass appears at $p = p^*$, suggests that $M_{\rm BH}$ plays the rôle of *order parameter* for these critical phenomena, like spontaneous magnetization M does for ferromagnets below the Curie temperature and like $|\rho - \rho_c|$ does for liquid-gas transitions in the co-existence region. To the extent that $M_{\rm BH}$ can be regarded as the order parameter, it is interesting to note that the critical exponent ($\simeq 0.37$) lies in a range typically observed for β in other critical systems [24]. Choptuik [16] has shown that details inherent in the original data are "washed out" within $\mathcal{R}$ in near-critical evolutions. Information may be steadily lost with each echo as $r \to 0$ and $T \to T^*$ and the rate of loss per echo may depend on the value of Δ. It seems likely that an analogue of the correlation length ξ in statistical systems is the ratio of the radii of the outer edge of the scaling region, $r_{\rm max}$, and the inner edge, r_n, of the innermost echo: i.e., $\xi \sim r_{\rm max}/r_n \sim e^{n\Delta}$. This brings in the scaling variable Δ, and as $p \to p^*$ an ever-larger region (in terms of the scale r_n) becomes "correlated" with self-similar echoes and $\xi \to \infty$.

3. Interaction of black hole spacetimes and gravitational waves

A group of researchers, who are based at or have ties to the US NSF National Center for Supercomputing Applications at the University of Illinois (hereafter called the Illinois group), have recently described [25] numerical models of black-hole spacetimes that interact with finite-amplitude gravitational waves. These are axisymmetric models computed with the Illinois group's new two-dimensional numerical relativity code. This finite difference code also evolves vacuum spacetimes on spacelike time slices by using the 3+1 form of the Einstein equations [7]. The equations for the gravitational field are similar, but not identical to those given in section 2.2.

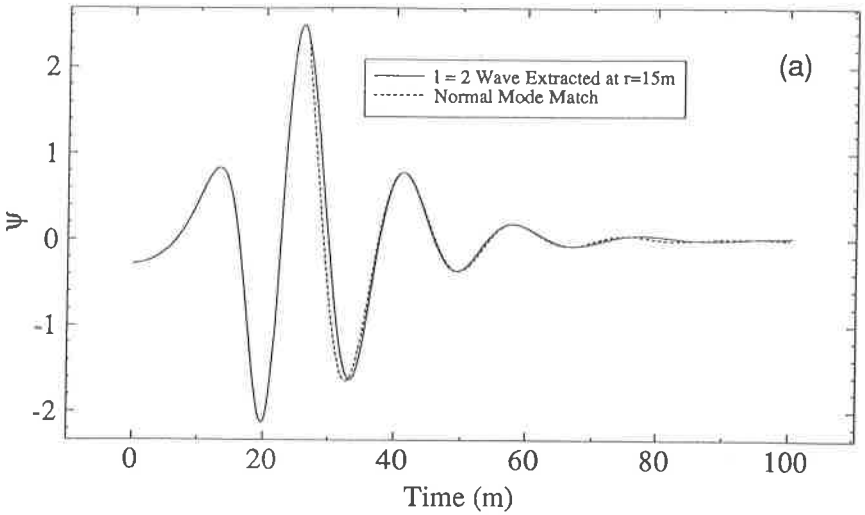

Figure 5. Extracted waveform (solid curve) that is emitted following interaction of a gravitational wave and a black hole. This wave represents a mild but finite amplitude disturbance of the black hole. The normal mode fit (dashed curve) of the waveform is shown to compare well at late times.

The initial data for these models consist of a single black hole superimposed with a time-symmetric gravitational wave (e.g., a Brill wave [26]). The spatial part of the line element takes the form

$$ds^2 = \Psi^4 \left[e^{2q}(d\eta^2 + d\theta^2) + \sin^2\theta \, d\phi^2 \right], \tag{30}$$

where η is a logarithmic radial coordinate. The function $q(\eta, \theta)$ is arbitrary up to satisfying appropriate boundary conditions and $\Psi(\eta, \theta)$ is determined by the Hamiltonian constraint. For $q = 0$ and with an appropriate boundary condition, Ψ becomes the conformal factor for the Schwarzschild solution in isotropic coordinates. When $q \neq 0$ the solution of the Hamiltonian constraint for $\Psi(\eta, \theta)$, along with $q(\eta, \theta)$, corresponds to the superposition of a gravitational wave and a single black hole. Specification of the initial data is completed by assuming time symmetry so that $K_{ij} = 0$. The Brill wave is taken to have the form

$$q = Af(\theta) \left(e^{-[(\eta+\eta_0)/\sigma]^2} + e^{-[(\eta-\eta_0)/\sigma]^2} \right), \tag{31}$$

with parameters A, η_0 and σ specifying the amplitude, range and radial width, respectively. Brill waves with angular dependence $f(\theta) = \sin^n\theta$, for several (even) values of n, have been examined.

The Illinois group describe several simulations with varying initial Brill wave amplitudes and with $n = 4$. Low amplitude Brill waves are shown to excite the fundamental $l = 2$ and $l = 4$ quasi-normal modes of the black hole. The demonstration of an accurate fit by the quasi-normal modes to the late-time behavior of the waveform in these

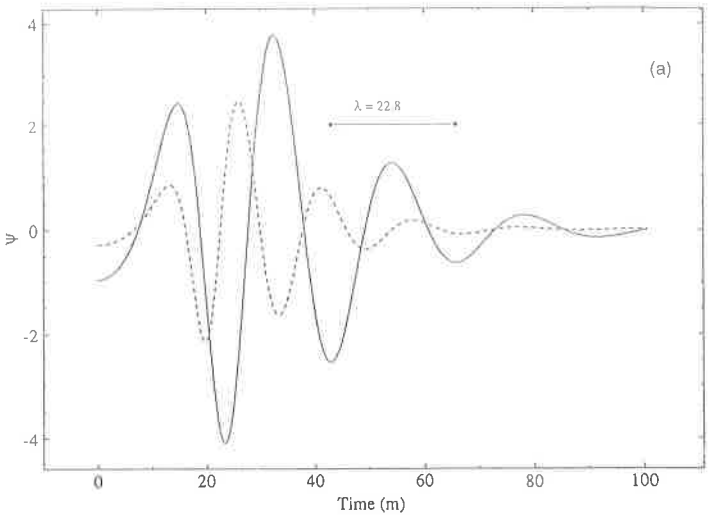

Figure 6. Extracted waveform (solid curve) that is emitted following the interaction between a strong gravitational wave and a black hole. Here the black hole is significantly disturbed by the wave and grows in mass as part of the wave crosses the horizon. The waveform of Figure 5 is overlapped for comparison.

calculations provides an important verification of the code (see Figure 5). A spacetime initially containing a high amplitude Brill wave has also been computed. In this model, the mass associated with the apparent horizon is only 0.59 of the total mass on the initial slice and the initial apparent horizon is observed to be highly distorted. Nevertheless, the gravitational waves shown emerging from the vicinity of the hole quickly develop damped-oscillatory behavior similar to that of quasi-normal modes (Figure 6). The waveform is no longer described well by the quasi-normal modes because the black hole mass grows significantly during the time the waves are being emitted as a significant fraction of the Brill wave is swallowed by the hole. It is as if the initial wave excites oscillations in a bell (the black hole) whose resonance properties are rapidly changing on a time scale of the fundamental oscillation.

In these simulations, the gravitational radiation signal is extracted from data that are available at a finite radius on the spacelike slices. Typically, the fields at distances of $15 - 30M$ from the hole are used to construct a gauge-invariant variable that determines an estimate of the asymptotic waveform. As a check, these researchers also use this boundary data, and data interpolated onto a future-directed null cone, for an integration of the Zerilli equation. The numerically computed Zerilli function, evaluated at larger radii, provides a useful check on the gauge-invariant determination of the waveform.

The intention is to next use this code to compute the head-on collision of two black-holes. This problem was originally studied by Smarr and Eppley [12] over a decade ago, but with less sophisticated techniques and computers. These limitations made it difficult to gauge the accuracy of those calculations. Apart from confirming the previous work,

accurate computation of the axisymmetric collision problem is viewed as a necessary prelude to any attempt to calculate in three dimensions the orbital decay and coalescence of a black-hole binary.

4. Three-dimensional initial data for two black hole collisions

With construction begun on the US Laser Interferometer Gravitational-wave Observatory (LIGO), increasing attention must be devoted to the theoretical task of modeling the orbital decay, coalescence, and gravitational radiation signal of black hole binaries. Before a dynamical evolution can be attempted, initial data must be constructed for the two black holes that satisfies the constraint equations. One means of generating such initial data sets has now been successfully demonstrated [27, 28] that uses the conformal and transverse-traceless decomposition of the constraint equations [7]. This is a conformal imaging procedure that extends Misner's [29] original calculation, of initial data on two asymptotically-flat sheets for a time-symmetric configuration of multiple black holes, to include non-time-symmetric configurations of holes with linear and angular momenta [30, 31, 32].

The procedure involves solving the vacuum Hamiltonian and momentum constraints of 3+1 gravity using the conformal transformations of York [7] to bootstrap into a simultaneous solution. For initial data we take the spatial metric γ_{ij} to be conformally flat, $\gamma_{ij} = \Psi^4\tilde{\gamma}_{ij} = \Psi^4 f_{ij}$ where f_{ij} is the flat three-metric, and the time slice to be maximal $K = K^i{}_i = 0$. The remaining part of the physical extrinsic curvature K_{ij} is then conformally related to the trace-free conformal background extrinsic curvature $\tilde{A}_{ij}$ by $K_{ij} = \Psi^{-2}\tilde{A}_{ij}$. The Hamiltonian and momentum constraints then become

$$\tilde{\nabla}^2\Psi = -\frac{1}{8}\Psi^{-7}\tilde{A}_{ij}\tilde{A}^{ij}, \tag{32}$$

$$\tilde{D}_j\tilde{A}^{ij} = 0, \tag{33}$$

where $\tilde{D}_j$ and $\tilde{\nabla}^2$ are flat-space differential operators. In order for the solutions of these equations to represent black holes, certain boundary conditions have to be assumed. First, boundary conditions are imposed that are consistent with the spacelike slices being asymptotically flat. Second, in order for the source-free equations to describe black holes, the hypersurface must be topologically nontrivial, with each black hole connected to another asymptotically-flat sheet through a throat or Einstein-Rosen bridge. This is the effect of the inner boundary condition used in the work described in the preceding section for one black hole. Here, each black hole on the top sheet may connect to a separate bottom sheet or all of the holes may be connected to the same bottom sheet. The latter approach, with a two-sheeted manifold, has been adopted since it provides an isometry between physical fields on the top and bottom sheets. This isometry implies unique boundary conditions that can be imposed on each of the throats and eliminates the need to include the bottom sheet in the computational domain.

The procedure starts with solutions of the momentum constraint [31] for a single black hole, which have specified (physical) linear and angular momenta:

$$\tilde{A}^p_{ij} = \frac{3}{2r^2}[P_i n_j + P_j n_i - (f_{ij} - n_i n_j)P^k n_k]$$

$$+\frac{3a^2}{2r^4}[P_i n_j + P_j n_i + (f_{ij} - 5n_i n_j)P^k n_k], \tag{34}$$

$$\tilde{A}^s_{ij} = \frac{3}{r^3}(\epsilon_{kim} S^m n^k n_j + \epsilon_{kjm} S^m n^k n_i). \tag{35}$$

The former expression adopts one of two possible isometry conditions (here the negative isometry), which is seemingly most useful physically. The physical linear and angular momenta are given by P^k and S_k, which are not affected by conformal mapping and therefore are not dependent on the solution of the Hamiltonian constraint. These data for $\tilde{A}_{ij}$ are isometric for one hole only. A superposition of two or more such solutions is a solution of the momentum constraint, but does not satisfy the isometry condition. Kulkarni, Shepley and York [32] have shown how higher-order corrections to (34) and (35) can be added, through a method of images for tensors, to construct a solution of the momentum constraint that is inversion symmetric through each hole. This process is rapidly convergent, but analytically nearly intractable. Fortunately, Cook [27] has found a recursive procedure that can be used numerically to generate a discrete solution (i.e., on a computational mesh) to arbitrarily high precision.

Once an inversion-symmetric solution of the momentum constraint is known, the nonlinear Hamiltonian constraint can be solved for Ψ by employing an asymptotically-flat boundary condition $\Psi \to 1$ as $r \to \infty$ and isometry boundary conditions on each of the throats. The conformal factor is then used to "dress" $\tilde{A}_{ij}$ and obtain the physical extrinsic curvature.

Cook [27] has solved these equations for two black hole initial data, subject to the restriction of axisymmetry, for a variety of different black hole linear momenta (aligned along the symmetry axis) and angular momenta (spin axes aligned along the symmetry axis). These results were obtained with two separate finite difference codes, one using Čadež coordinates [33] and one using bispherical coordinates, in order to confirm his results and gauge their precision. In new work, Cook, Choptuik, Dubal, Klasky, Matzner and Oliveira [28] have generalized these results to three dimensions and obtained initial data for two black holes with truly arbitrary linear and angular momenta and masses. Cook and Abrahams [34] have examined these data sets to determine some of the physical properties that are evident on the initial slice, such as asymptotic mass, linear momentum and angular momentum, and areas of apparent horizons.

5. Conclusions

The examples cited here should make evident that numerical relativity has reached a level of sophistication where it is possible both to discover previously unknown physical effects and to state certain results with numerical precision. Even in the interaction of strong gravitational waves with black holes, where we have not yet been surprised with a violation of our physical intuition, the results probe a regime that is far from what we might calculate analytically. Each such solution, where the black hole is strongly perturbed yet fails to produce a naked singularity, lends weight to the notion of relatively generic adherence to the cosmic censorship hypothesis. Finally, the first forays have begun toward designing algorithms and writing codes for three-dimensional numerical relativity calculations. This comes at an opportune time as there are increasing hopes that we will enter the era of gravitational-wave astronomy at the end of the 1990s.

Acknowledgments

The author would like to acknowledge numerous helpful conversations with A. Abrahams, M. Choptuik, G. Cook, E. Seidel and J. York. He is grateful for research support from NSF Grants PHY90-01645 and PHY90-57865 and from the Alfred P. Sloan Foundation.

References

[1] Abramovici, A, Althouse, W E, Drever, R W P, Gürsel, Y, Kawamura, S, Raab, F J, Shoemaker, D, Sievers, L, Spero, R E, Thorne, K S, Vogt, R E, Weiss, R, Whitcomb, S E and Zucker, M E 1992 *Science* **256** 325–33

[2] Abramovici, A 1993, this volume

[3] Cutler, C, Apostolatos, T A, Bildsten, L, Finn, L S, Flanagan, E E, Kennefick, D, Markovic, D M, Ori, A, Poisson, E, Sussman, G J and Thorne, K S 1992 submitted to *Phys. Rev. Lett.*

[4] Cutler, C, Finn, L S, Poisson, E and Sussman, G J 1993 *Phys. Rev.* **D**, in press

[5] Penrose, R, 1969 *Riv. Nuovo Cimento* **1** 252–76

[6] Shapiro, S L and Teukolsky, S A 1991, *Phys. Rev. Lett.* **66** 994–997

[7] York, J W 1979 in *Sources of Gravitational Radiation* (Cambridge: Cambridge University Press) 83–126

[8] Gomez, R, Winicour, J and Isaacson, R A 1992 *J. Comp. Phys.* **98** 11

[9] Finn, L S 1991 in *Nonlinear Problems in Relativity and Cosmology* eds. J Buchler, S Detweiler and J Ipser (New York: New York Acad. Sci.) 156–72

[10] Oohara, K and Nakamura, T 1989 *Prog. Theor. Phys.* **82** 535–54
Nakamura, T and Oohara, K 1989 *Prog. Theor. Phys.* **82** 1066–83
Oohara, K and Nakamura, T 1990 *Prog. Theor. Phys.* **83** 906–40
Nakamura, T and Oohara, K 1991 *Prog. Theor. Phys.* **86** 73–88
Oohara, K and Nakamura, T 1992 *Prog. Theor. Phys.* **88** 307–15
Rasio, F A and Shapiro, S L 1992 *Astrophys. J.* **401** 226–45

[11] Miller, W L 1986, in *Dynamical Spacetimes and Numerical Relativity* edited by J Centrella (Cambridge: Cambridge University Press) 256–303

[12] Smarr, L L 1979, in *Sources of Gravitational Radiation* ed. L L Smarr (Cambridge: Cambridge University Press) 245–74

[13] Evans, C R, Finn, L S and Hobill, D W 1989 *Frontiers in Numerical Relativity* (Cambridge: Cambridge University Press)

[14] d'Inverno, R 1992 *Approaches to Numerical Relativity* (Cambridge: Cambridge University Press)

[15] Choptuik, M W 1992 in *Approaches to Numerical Relativity* edited by R d'Inverno (Cambridge: Cambridge University Press)

[16] Choptuik, M W 1993, *Phys. Rev. Lett.* **70** 9–12

[17] Berger, M J and Oliger, J 1984 *J. Comp. Phys.* **53** 484–512

[18] Abrahams, A M and Evans, C R 1993 submitted to *Phys. Rev. Lett.*

[19] Bardeen, J M and Piran, T 1983 *Phys. Rep.* **96** 205–50

[20] Abrahams, A M and Evans, C R 1992 *Phys. Rev.* **46** R4117–21

[21] Evans, C R 1984 Ph.D. thesis, University of Texas at Austin, unpublished

[22] Evans, C R 1986 in *Dynamical Spacetimes and Numerical Relativity* edited by J Centrella (Cambridge: Cambridge University Press) 3–39

[23] Abrahams, A M and Evans, C R 1988 *Phys. Rev.* **D37** 318–32

[24] Ma, S 1976 *Modern Theory of Critical Phenomena* (Redwood City: Addison-Wesley)

[25] Abrahams, A M, Berstein, D, Hobill, D W, Seidel, E and Smarr, L 1992 *Phys. Rev.* **D45** 3544–58

[26] Brill, D 1959 *Ann. Phys., NY* **7** 466–83

[27] Cook, G B 1991 *Phys. Rev.* **D44** 2983–3000

[28] Cook, G B, Choptuik, M W, Dubal, M R, Klasky, S, Matzner, R A and Oliveira, S R 1993 *Phys. Rev.* **D**, in press

[29] Misner, C W 1963 *Ann. Phys., NY* **24** 102–17

[30] Bowen, J 1979 *Gen. Rel. Grav.* **11** 227–31

[31] Bowen, J and York, J W 1980 *Phys. Rev.* **D21** 2047–56

[32] Kulkarni, A, Shepley, L and York, J W 1983 *Phys. Lett.* **96A** 228–30

[33] Čadež, A 1971 Ph.D. thesis, University of North Carolina, unpublished

[34] Cook, G B and Abrahams, A M 1992 *Phys. Rev.* **D46** 702–13

Paper presented at the 13th Int. Conf. on General Relativity and Gravitation
Cordoba, Argentina, 1992: Part 1, Plenary Lectures

Relativistic dissipative fluids

Robert Geroch

The Enrico Fermi Institute
5640 S. Ellis Ave.,
The University of Chicago
Chicago, Ill 60637

We observe in Nature fluids that manifest dissipation, e.g., the effects of heat conductivity and viscosity. We believe that all physical phenomena are to be described within the framework of General Relativity. What, then, is the appropriate description of a relativistic dissipative fluid? This is not only a question of principle, but also one of practical interest. There exist systems, such as certain neutron stars, in which relativity and dissipation are at the same time significant.

The nonrelativistic theory of a dissipative fluid - the Navier-Stokes theory - has been remarkably successful. The fields of this theory consist of a velocity field of the fluid, two state fields of the fluid (e.g., density and pressure), a heat-flow vector field, and a stress tensor field. The equations of the theory are conservation of mass, energy and momentum, together with two additional equations: One equates the heat-flow to a multiple of the temperature gradient; the other the stress to a multiple of the velocity gradient. There is a natural relativistic generalization of this Navier- Stokes theory, called the Eckart theory[1]. The fields of this theory are a 4-velocity of the fluid, two state fields, a spatial heat-flow field, and a spatial stress field. The equations are essentially those of Navier-Stokes, written in the rest frame of the fluid. Thus, there are three conservation equations, together with equations for the heat-flow and stress. But this Eckart theory turns out to be unsatisfactory. For example, the equations of the theory have no initial-value formulation.

It is not difficult to see what is the problem with the Eckart theory. Consider, for example, the heat equation in one space and one time dimension

$$\frac{\partial T}{\partial t} = \kappa \frac{\partial^2 T}{\partial x^2} \tag{1}$$

where κ is the thermal conductivity. This equation sends signals at arbitrarily large speed. This behavior is not a problem in a nonrelativistic theory, for we have in that case a Newtonian time, which regulates the forward march of the solution. But in a

relativistic theory (such as the Eckart) this behavior means that signals could, at each point of space- time, be sent along all directions in the 3-flat orthogonal to the local fluid 4-velocity at that point. In general, these local 3-flats would not integrate to 3-surfaces, and so the equation could send signals into the past. This behavior, in turn, manifests itself as lack of an initial-value formulation.

It is easy to modify the heat equation to correct its infinite-speed behavior. Introduce a characteristic speed v, and replace Eqn. (1) by

$$\frac{\partial T}{\partial t} = \kappa\left(\frac{\partial^2 T}{\partial x^2} - \frac{1}{v^2}\frac{\partial^2 T}{\partial t^2}\right). \tag{2}$$

Then, provided $v < c$, this equation sends no signal faster than light. But, unfortunately, this new equation appears to be unsuitable for the description of thermal phenomena. For example, it requires as initial data the spatial distribution of both the temperature and its first time derivative. How, physically, does one "set" an initial time-derivative of the temperature?

This situation for the heat equation is rescued by a theorem to the following effect: Consider a solution of Eqn. (1) with initial data $T_0(x)$, and of Eqn. (2) with initial data $T_0(x)$ and $\dot{T}_0(x)$. Then these two solutions are close to each other, except on distance scales of the order of κ/v, and time scales of the order of κ/v^2 [2]. Choosing v to exceed the typical thermal speeds of the particles making up the fluid, then these scales will be less than, respectively, the mean free path and mean free time of the fluid particles. But on these scales, "temperature" does not even make physical sense, and so neither do the solutions of our equations. Thus, we are able to modify the heat equation to achieve causal behavior, the only cost being to change the solutions on scales on which those solutions do not make physical sense anyway.

Similar modifications of the Eckart theory have been proposed[3]. In these new theories, the fields are the same, and the conservation laws remain intact. But the equations relating the heat-flow and stress to the other fluid variables are modified by introducing new terms involving, among other things, the time-derivatives of these two fields. The result, provided the coefficients in these new terms are chosen with sufficient care and to be sufficiently large, is a hyperbolic system. Furthermore, it is believed - although the proof is not complete - that the solutions of this new system are physically realistic. How could such a result be obtained, given that now we cannot expect to have solutions of the old theory - the Eckart theory - for comparison? What is done is the following. Consider any solution of the new system of equations. Write down the expressions that vanish in the case of the Eckart theory - heat-flow minus a multiple of temperature gradient and stress minus a multiple of velocity gradient. Now try to show that these expressions evolve, by virtue of the full equations of the theory, to be "small".

So, it now appears likely that we will have physically acceptable theories for relativistic dissipative fluids. But they will exist only in a rather curious context. For a fixed physical fluid, there will be many different systems of equations to describe that fluid - corresponding to the many different ways that one can add the extra terms to the Eckart equations to obtain a hyperbolic system. But all these systems will give precisely

the same physical results, i.e., the same description of our fluid on scales on which a "fluid description" is valid at all. This is a most surprising resolution of the puzzle of the status of dissipative fluids in general relativity.

References

[1] C. Eckart C. 1940, *Phys. Rev.* **58**, 919.

[2] Reula O., Ortiz O., and Nagy G, *private communication.*

[3] Liu I.S., Muller I., and Ruggeri T. 1986, *Ann. Phys.* **169**, 191;
Geroch R.and Lindblom L. 1990, *Phys. Rev.* D **41**,1885.
Geroch R.and Lindblom L. 1991, *Ann. Phys.* **207**, 394.

Paper presented at the 13th Int. Conf. on General Relativity and Gravitation
Cordoba, Argentina, 1992: Part 1, Plenary Lectures

The interpretation of quantum cosmological models

Jonathan J. Halliwell

Center for Theoretical Physics, Laboratory for Nuclear Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.†

Abstract. We consider the problem of extracting physical predictions from the wave function of the universe in quantum cosmological models. We state the features of quantum cosmology an interpretational scheme should confront. We discuss the Everett interpretation, and extensions of it, and their application to quantum cosmology. We review the steps that are normally taken in the process of extracting predictions from solutions to the Wheeler-DeWitt equation for quantum cosmological models. Some difficulties and their possible resolution are discussed. We conclude that the usual wave function-based approach admits at best a rather heuristic interpretation, although it may in the future be justified by appeal to the decoherent histories approach.

1. Introduction

Quantum mechanics was originally developed to account for a number of otherwise unexplained phenomena in the microscopic domain. The scope of the theory was not thought to extend beyond the microscopic. Indeed, the existence of an external, macroscopic, classical domain was felt to be necessary for the theory's interpretation. This view of quantum mechanics has persisted for a very long time, with not one shred of experimental evidence against it.

Today, however, more ambitious views of quantum mechanics are entertained. The experimental possibilities afforded by SQUIDS suggest that quantum coherence may act on macroscopic scales (Leggett, 1980). Furthermore, the separation of classical and quantum domains is often felt unnatural from a foundational point of view. The domain of applicability of quantum mechanics may therefore need to be extended. But in extrapolating quantum mechanics to the macroscopic scale, where do we stop? At the scale of large molecules? At the laboratory scale? At the planetary scale? There is only one natural place – at the scale of the entire universe. One rapidly arrives, therefore, at the subject of quantum cosmology, in which quantum mechanics is applied to the entire universe. It might only be through this subject that the form of quantum mechanics as we know it may be fully understood.

Yet quantum cosmology was not originally developed with the foundational problems of quantum mechanics in mind. Rather, it was perhaps viewed as a natural area to investigate as part of a more general drive to understand quantum gravity (DeWitt, 1967; Misner, 1969; Wheeler, 1963, 1968). More recently, the success of models like inflation (Guth, 1981) in early universe cosmology has led to a greater desire to understand the initial conditions with which the universe began. On very general grounds, the universe appears to have emerged from an era in which quantum gravitational effects were important. Quantum cosmology, in which both matter and gravitational fields are taken to be quantized, is therefore the natural framework in which to address

† Address after September 1, 1992: Blackett Laboratory, Imperial College of Science and Technology, South Kensington, London, SW7 2BZ, UK.

questions of initial conditions. Indeed, there has been a considerable amount of activity in this subject in recent years.†

Much of this recent attention on quantum cosmology has been focused on obtaining a crude idea of the physical predictions made by certain theories of initial conditions. However, little attention has been devoted to understanding or even clarifying the principles involved in extracting these predictions from the mathematical formalism. This contribution represents a small step towards filling this gap. That is, it is concerned with the *interpretation* of quantum cosmology. In particular, I shall be addressing the question, "How are predictions extracted from a given wave function of the universe?". I will state immediately that there is no wholly satisfactory answer to this question. The aim of this contribution, therefore, is to describe the difficulties involved and the attempts to overcome them, the procedures that are actually used in practice and the motivation behind them.

The interpretation of quantum cosmology raises two particular sets of issues which may be discussed separately. The first set concerns the fact that, unlike ordinary quantum mechanics, the system under scrutiny, the universe, is a closed and isolated system. The special features of (non-relativistic) quantum mechanics applied to such systems are discussed in the accompanying paper by Hartle (Hartle, 1992a). For completeness, some of these features are also covered here, but from a somewhat different angle (namely from the wave function point of view, rather than that of the decoherent histories approach). The second set of issues concerns the fact that the wave function for the system is described not by a time-dependent Schrödinger equation, but by the Wheeler-DeWitt equation. Associated with this is the so-called "problem of time", and the absence of the usual machinery of projection operators, unitary evolution, Hilbert space *etc.* It is this second set of issues that will form the primary focus of this paper.

To focus the discussion, consider the sort of features of the universe one might hope to predict in quantum cosmology. Some of the most important are:

- **Spacetime is Classical.** One of the crudest and most obvious observations about the universe we can make, is that the structure of spacetime is described very accurately by classical laws. The very first requirement of a quantum theory of the universe, therefore, is that it predict this. The question of how classical behaviour emerges in quantum systems is a difficult one, that has not been completely solved even in non-relativistic quantum mechanics. Still, it ought be possible, at least in some crude sense, to see whether a given wave function for the universe is consistent with a prediction of classical spacetime.
- **Initial Inflationary Phase.** Many current observational features of the universe, such as the horizon, flatness and monopole problems, are explained by postulating an inflationary phase at very early times. In models of the universe which admit inflationary solutions, the occurrence and amount of inflation are generally dependent on initial conditions. A correct quantum theory of initial conditions should supply the initial conditions for an inflationary phase to take place.
- **Spectra of Density Fluctuations and Gravitational Waves.** In inflationary models, it is possible to generate gravitational waves and density fluctuations consistent with the observed isotropy of the microwave background, yet sufficiently large

† For general reviews of quantum cosmology, see Fang and Ruffini (1987), Fang and Wu (1986), Halliwell (1990, 1991), Hartle (1985, 1986, 1990), Hawking (1984b), Hu (1989), Laflamme (1991), Page (1986, 1991).

for the formation of large scale structure. These results are obtained by considering the quantum field theory of the fluctuations during the inflationary phase. Again they are initial conditions dependent, and one might hope that quantum cosmology will supply the requisite initial conditions.

- **Low Entropy Initial State.** One of the most striking features of the universe is its time asymmetry. It is very far from equilibrium, indicating that it started out in a very special, low entropy initial state. One would therefore hope to predict this very smooth beginning. This is closely related to the spectra of fluctuations.

In what follows, we will not be concerned with detailed predictions of particular theories of initial conditions. Rather, we will assume that we are given a wave function for the universe (perhaps determined by *some* theory of initial conditions), and discuss the extraction of predictions from it. In particular, we will focus on the emergence of a *semiclassical domain* – an approximately classical spacetime background with quantum fields in it†. This is appropriate for two reasons. First, all observational features of the universe are semiclassical in nature. Second, even if the theory could make observable quantum gravitational predictions, as one would certainly hope of a quantum theory of cosmology, observation of them would be through their correlation with the semiclassical domain. Quite generally, therefore, the emergence of the semiclassical domain is the appropriate thing to look for.

2. Canonical Quantization

In the interests of conciseness, I shall assume that the formalism of quantum cosmology is known, and give only the briefest account here (see for example Halliwell (1990), and the general references cited earlier). For definiteness, we consider the canonical quantization of Einstein gravity coupled to matter for closed universes. The quantum state of the system, the universe, is represented by a wave functional, $\Psi[h_{ij},\phi]$, a functional on superspace, the space of three-metrics h_{ij} and matter fields ϕ on a three-surface. The wave function has no explicit dependence on time. There is therefore no Schrödinger equation, only the constraints of the Dirac quantization procedure,

$$\hat{\mathcal{H}}\Psi[h_{ij},\phi]=0, \qquad \hat{\mathcal{H}}_i\Psi[h_{ij},\phi]=0 \tag{2.1}$$

where $\hat{\mathcal{H}}$ and $\hat{\mathcal{H}}_i$ are the operator versions of the classical constraints,

$$\mathcal{H}=h^{-\frac{1}{2}}\left(\pi^{ij}\pi_{ij}-\frac{1}{2}\pi_i^i\pi_j^j\right)-h^{\frac{1}{2}}\left({}^3R-2\Lambda\right)+\mathcal{H}^{matter}=0 \tag{2.2}$$

$$\mathcal{H}_i=-2\pi_{ij}{}^{|j}+\mathcal{H}_i^{matter}=0 \tag{2.3}$$

(in the usual notation). Eqs.(2.1) are of course the Wheeler-DeWitt equation and the momentum constraints.

† This is, loosely speaking, the same thing as the *quasi-classical* domain defined by Gell-Mann and Hartle (1990) in the context of the decoherent histories approach, but I give it a slightly different name since the wave function approach considered here does not permit it to be defined in quite the same way.

Wave functions satisfying the constraints (2.1) may also be generated using a (complex) path integral representation (Halliwell and Hartle, 1990, 1991). This is used, for example, in the specification of the no-boundary wave function (Hawking, 1982, 1984a; Hartle and Hawking, 1983). The sum-over-histories approach is more general in that it may be used to construct more complicated amplitudes (*e.g.*, ones depending on the three-metric and matter fields on many three-surfaces) and in fact the latter type of amplitude is likely to be the most useful for interpreting the theory. Here we are concerned with describing what has actually been done, and this means discussing single surface amplitudes.

There are other very different ways of quantizing the classical system described by the constraints, (2.2), (2.3). In particular, one could contemplate solving the constraints classically, prior to quantization, and then quantizing an unconstrained theory. The Dirac quantization scheme outlined here is the one most commonly used in practical quantum cosmology, and it is this that we shall use in what follows.

Of course, all known approaches to quantum gravity are fraught with severe technical difficulties, and the formal framework outlined above is no exception. Moreover, a proper theory of quantum gravity, should it exist, might involve substantial departures from this general framework (*e.g.*, string theory). However, there are a number of reasons why it might nevertheless make sense to persevere with this approach to quantum cosmology. Perhaps the most important is that in quantum cosmology one is frequently concerned with *issues*. Many issues, and in particular the interpretational issues considered here, are not very sensitive to the resolution of technical difficulties. Furthermore, they are likely to be equally present in any approach to quantum gravity. No generality is lost, therefore, in employing the formal scheme envisaged in Eq.(2.1).

3. Interpretation

The question of interpretation is that of extracting physical statements about the universe from the wave function, $\Psi[h_{ij}, \phi]$. In non-relativistic quantum mechanics there is a well-defined procedure for achieving this. It is generally known as the Copenhagen interpretation. Although it is frequently regarded as problematic from a foundational point of view, it has been very successful in its physical predictions.† The Copenhagen interpretation involves a number of features and assumptions that should be highlighted for the purposes of this account:

C1. It assumes the existence of an *a priori* split of the world into two parts: a (usually microscopic) quantum system, and an external (usually macroscopic) classical agency.

C2. It concerns systems that are not genuinely closed, because they are occasionally subject to intervention by the external agency.

C3. The process of prediction places heavy emphasis on the notion of *measurement* by the external agency.

C4. Predictions are probabilistic in nature, and generally only have meaning when measurements are performed on either a large ensemble of identical systems, or on the same system many times, each time prepared in the same state.

† Many of the key papers on the Copenhagen interpretation may be found in Wheeler and Zurek (1983).

C5. Time plays a very distinguished and central role.

Quantum cosmology, by contrast, involves a number of corresponding features and assumptions that render the Copenhagen interpretation woefully inadequate:

QC1. It is assumed that quantum mechanics is universal, applying to microscopic and macroscopic systems alike, up to and including the entire universe. There can therefore be no *a priori* split of the universe into quantum and classical parts.

QC2. The system under scrutiny is the entire universe. It is a genuinely closed and isolated system without exterior.

QC3. Measurements cannot play a fundamental role, because there can be no external measuring apparatus. Even internal measuring apparatus should not play a role, because the conditions in the early universe were so extreme that they could not exist.

QC4. The universe is a unique entity. It does not belong to an ensemble of identical systems; nor is it possible to make repeated measurements on it prepared in the same state.

QC5. The problem of time: general relativity does not obviously supply the time parameter so central to the formulation and interpretation of quantum theory.

Of these features, only (QC5) is specific to quantum gravity. The rest would be true of any closed and isolated system described by non-relativistic quantum mechanics, and may be discussed in this context.

The problem of interpretation is that of finding a scheme which confronts features (QC1)–(QC5), yet which may be shown to be consistent with or to reduce to (C1)–(C5) under suitable approximations and restrictions.

Probably the best currently available approach to these difficulties is the decoherent histories approach, which employs not the wave function, but the decoherence functional as its central tool.† It has a number of features which strongly recommend it for quantum cosmology: it specifically applies to closed systems; it assumes no *a priori* separation of classical and quantum domains; it does not rely on the notion of measurement or observation; and its focus on histories rather than events at a single moment of time might sidestep (or at least alleviate) the problem of time in quantum gravity. Unfortunately, a detailed application of this approach to quantum cosmology has not yet been carried out (although efforts in this direction are currently being made (Hartle, 1992b)). Furthermore, my task here is to describe what has actually been done, and for that reason I will describe the somewhat cruder approaches to interpretation based on the wave function. However, I find it useful to remain close to the spirit of the decoherent histories approach, and to think of the wave function approach as an approximation to it (although in a sense yet to be explained). In particular, it is convenient to have as one's aim the assignment of probabilities to histories of the universe.

In order to deal with the issues noted above, conventional wave function-based approaches invoke the Everett (or "Many Worlds") interpretation.†† Above all, the

† The consistent histories approach to quantum mechanics was introduced by Griffiths (1984), and later developed by Omnès (reviewed in Omnés, 1990, 1992). Much of it was developed independently under the name decoherent histories, by Gell-Mann and Hartle (1990). See also Hartle (1990, 1992a). For applications and generalizations, see Blencowe (1991), Albrecht (1990) and Dowker and Halliwell (1992).

†† Everett (1957). A useful collection of papers on the Everett interpretation may be found in DeWitt

Everett interpretation is a scheme specifically designed for quantum mechanical systems that are closed and isolated. Everett asserted that quantum mechanics should be applicable to the entire universe, and there should be no separation into quantum and classical domains. These features of the Everett interpretation are therefore consistent with features (QC1) and (QC2) of quantum cosmology. Furthermore, the state of the entire system should evolve solely according to a wave equation, such as the Schrödinger equation, or in quantum cosmology the Wheeler-DeWitt equation, and there should be no discontinuous changes (collapse of the wave function).

Everett went on to model the measurement process by considering a world divided into a large number of subsystems, and showed how the conventional Copenhagen view of quantum mechanics can emerge. It is this part of the Everett interpretation that leads to its "Many Worlds" feature – the idea that the universe splits into many copies of itself whenever a measurement is performed. It is at this stage, however, that the original version of the Everett interpretation departs in its usefulness from practical quantum cosmology. For the sort of models one is often interested in interpreting in quantum cosmology are minisuperspace models, which are typically very simple, and do not contain a large number of subsystems. Furthermore, the many worlds aspect of the interpretation has, I believe, been rather over-emphasized, perhaps at the expense of undermining the credibility of the overall set of ideas. The Everett interpretation, especially as it applies to practical quantum cosmology, is not so much about many worlds, but rather, about how one might make sense of quantum mechanics applied to genuinely closed systems.

It is, therefore, convenient to pass to a restatement of the Everett interpretation due to Geroch (1984). Geroch translated Everett's modeling of the measurement process using a large number of subsystems into statements about the wave function of the entire system. He argued that predictions for closed quantum systems essentially boil down to statements about "precluded" regions – regions of configuration space in which the wave function for the entire system is very small. From here, it is a small step to a slightly more comprehensive statement given by Hartle, specifically for quantum cosmology (Hartle, 1986). It is the following:

> *If the wave function for the closed system is strongly peaked about a particular region of configuration space, then we predict the correlations associated with that region; if it is very small, we predict the lack of the corresponding correlations; if it is neither strongly peaked, nor very small, we make no prediction.*

This basic idea appears to have been adopted, perhaps provisionally, in most, if not all, attempts to interpret the wave function in quantum cosmology (see also Wada (1988)).

The statement is admittedly rather vague and a number of qualifying remarks are in order. Firstly, to say that a wave function is "strongly peaked" necessarily involves some notion of a measure, and this has to be specified. For example, do we use $|\Psi|^2$? The Klein-Gordon current constructed from Ψ? We will return to this below. Second, the interesting correlations in the wave function are often not evident in the configuration space form. It is therefore appropriate to interpret the above statement rather broadly, as refering to not just the wave function, but any distribution constructed from the wave function. For example, phase space distributions such as the Wigner function, and related functions, have been the focus of attention in a number of pa-

---

and Graham (1973). See also Bell (1981), Deutsch (1985), Kent (1990), Smolin (1984) and Tipler (1986).

pers, and these have sometimes proved quite useful in identifying classical correlations (Anderson, 1990; Calzetta and Hu, 1989; Habib, 1990; Habib and Laflamme, 1990; Halliwell, 1987, 1992; Kodama, 1988; Singh and Padmanabhan, 1989).

For practical quantum cosmology, the implication of the above interpretational scheme is that rather than find the probability distributions for quantities felt to be of interest, as in ordinary quantum mechanics, it is necessary to determine those quantities for which the theory gives probabilities close to zero or one, and hence, for which it makes predictions. On the face of it, this may seem to suggest that the predictive power of quantum cosmology is very limited in comparison to ordinary quantum mechanics. However, by studying isolated systems consisting of a large number of identical subsystems, it may be shown that this interpretation implies the usual statistical interpretation of ordinary quantum mechanics, in which subsystem probabilities are given by $|\psi|^2$ where ψ is the subsystem wave function (Farhi, Goldstone and Gutmann, 1989; Finkelstein, 1963; Graham, 1973; Hartle, 1968). It is in this sense that feature (QC4) of quantum cosmology is reconciled with (C4) of the Copenhagen interpretation.

Given a measure on a set of possible histories for the universe, it is often not peaked about a particular history or family of histories. To obtain probabilities close to one or zero, it is often necessary to restrict attention to a certain subset of the possible histories of the universe, and make predictions within that subset. That is, one looks at *conditional probabilities*. The motivation behind this is anthropic reasoning – we as observers do not look out into a generic universe, but to one in which the conditions for our own existence have necessarily been realized (see, for example, Barrow and Tipler, 1986). Obviously the detailed conditions for our existence could be very complicated and difficult to work out. However, it is possible to get away with exploiting only the weakest of anthropic assumptions in order to make useful and interesting predictions about the universe. It is, for example, extremely plausible that the existence of life requires the existence of stars like our sun. It therefore makes sense to restrict attention only to those histories of the universe which exist long enough for stars to form before recollapsing.

This completes our brief survey of the special features of closed quantum systems. We have discussed features (QC1), (QC2) and (QC4) of quantum cosmology. The status of measurements, (QC3), in the above discussion is admittedly somewhat vague, although it is clear that they do not play a significant role. Their status is perhaps clarified more fully in the decoherent histories approach. Finally, although it is an important issue for the interpretation of quantum cosmology, I have not discussed the problem of time, (QC5). This will be briefly mentioned below, but it would take a lot of space to do it justice. The interested reader is referred to the reviews by Kuchař (1989, 1992), which cover the problem of time and its connections with the interpretation of quantum cosmology. See also Unruh and Wald (1989).

4. Interpretation of Solutions to the Wheeler-DeWitt Equation

We now come to the central part of this contribution, which is to describe how in practice predictions are actually extracted from solutions to the Wheeler-DeWitt equation. As mentioned earlier, the primary aim is to determine the location and features of the semiclassical domain. Different papers in the literature treat the process of prediction in different ways, but it seems to boil down to four distinct steps, which I now describe in turn.

A. Restriction to Perturbative Minisuperspace.

The full theory described by Eq.(2.1) is not only difficult to handle technically, it is not even properly defined. This problem is normally avoided by artificially restricting the fields to lie in the region of superspace in the neighbourhood of homogeneity (and often isotropy). That is, one restricts attention to the finite dimensional subspace of superspace called "minisuperspace", and considers small but completely general inhomogeneous perturbations about it.†

In more detail, the sort of restrictions entailed are as follows. One restricts the three-metric and matter fields to be of the form,

$$h_{ij}(\mathbf{x},t) = h_{ij}^{(0)}(t) + \delta h_{ij}(\mathbf{x},t), \quad \Phi(\mathbf{x},t) = \phi(t) + \delta\phi(\mathbf{x},t) \tag{4.1}$$

Here, the minisuperspace background is described by the homogeneous fields $h_{ij}^{(0)}$ and $\phi(t)$. For example, the three-metric could be restricted to be homogeneous and isotropic, described by a single scale factor a. We will denote the minisuperspace background coordinates by the finite set of functions $q^\alpha(t)$, where $\alpha = 1, \cdots n$. δh_{ij} and $\delta\phi$ are small inhomogeneous perturbations about the minisuperspace background, describing gravitational waves and scalar field density perturbations. They are retained only up to second order in the action and Hamiltonian (and therefore to first order in the field equations). For convenience, we will denote the perturbation modes simply by $\delta\phi$.

With these restrictions on the class of fields considered, the Wheeler-DeWitt equation, after integration over the three-surface, takes the form

$$\left[-\frac{1}{2m_p^2}\nabla^2 + m_p^2 U(q) + H_2(q,\delta\phi)\right]\Psi(q,\delta\phi) = 0 \tag{4.2}$$

Here, ∇^2 is the Laplacian operator in the minisuperspace modes q and we have explicitly included the Planck mass m_p, since it is to be used as a large parameter in a perturbative expansion. H_2 is the Hamiltonian of the perturbation modes, $\delta\phi$, and is quadratic in them. There are more constraint equations associated with the remaining parts of the Wheeler-DeWitt equation, and with the momentum constraints. These are all linear in the perturbations and can be solved, after gauge-fixing. When this is done, only Eq.(4.2) remains, in which $\delta\phi$ may be thought of as a gauge-invariant perturbation variable.

The restriction to perturbative minisuperspace is of course very difficult to justify from the point of view of the full theory. A number of attempts have been made to understand the sense in which minisuperspace models might be part of a systematic approximation to the full theory, but the answer seems to be that their connection is at best tenuous (Kuchař and Ryan, 1986, 1989). What one can say, however, is that solutions to the minisuperspace field equations, including perturbations about them, will (with a little care) be solutions to the full field equations, and thus the lowest order semiclassical approximation to perturbative minisuperspace quantum cosmological models will agree with the lowest order semiclassical approximation to the full theory.

† This general approach has been the topic of many papers, including Banks (1985), Banks, Fischler and Susskind (1985), Fischler, Ratra and Susskind (1986), Halliwell and Hawking (1985), Lapchinsky and Rubakov (1979), Shirai and Wada (1988), Vachaspati and Vilenkin (1988), Wada (1986).

These models may therefore be thought of as useful models in which a number of issues can be profitably investigated, but which may also give some crude predictions about the physical universe.

The essential ideas of the practical interpretational scheme described here will very probably be applicable to situations more general than perturbative minisuperspace, but very little work on such situations has been carried out. We will not go into that here.

B. Identification of the Semiclassical Regime.

The next step involves inspecting the wave function $\Psi(q^\alpha, \delta\phi)$, asking how it behaves as a function of the minisuperspace variables q, and in particular, identifying the regions in which the wave function is exponentially growing or decaying in q, or oscillatory in q.

The regions in which the wave function is rapidly oscillating in q are regarded as the semiclassical domain, in which the modes q, are approximately classical, whilst the perturbation modes $\delta\phi$ need not be. This interpretation comes partly from analogy with ordinary quantum mechanics. But also one can often argue that certain distributions constructed from a rapidly oscillating wave function are peaked about classical configurations. For example, the Wigner function, mentioned earlier, is often used to support this interpretation.

The other regions, in which the wave function tends to be predominantly exponential in behaviour are regarded as non-classical, like the under-the-barrier wave function in tunneling situations. Were this ordinary quantum mechanics, then the wave function would typically be exponentially small in the tunneling regions. One could then invoke Geroch's version of the Everett interpretation, and just that the system will not be found in this region, because it is "precluded". However, a peculiar feature of gravity (readily traced back to the indefiniteness of the action) is that the wave function may be either exponentially small or exponentially large in the regions where it is of exponential form. Nevertheless, one still says that the system is not approximately classical in this regime, even when the wave function is exponentially large. To support this claim, one can argue that, in contrast to the oscillatory regions, a predominantly exponential wave function, either growing or decaying, is *not peaked* about classical configurations.

C. WKB Solution in the Oscillatory Regime.

The next stage of the scheme involves solving in more detail in the region in which the wave function is rapidly oscillating in the minisuperspace variables. This involves focusing on a particular type of state, namely the WKB state,

$$\Psi(q, \delta\phi) = C(q)\ \exp\left(i m_p^2 S_0(q)\right)\ \psi(q, \delta\phi) + O(m_p^{-2}) \tag{4.3}$$

$S_0(q)$ is real, but ψ may be complex. Many models also involve a slowly varying exponential prefactor contributing at order m_p^2, but for the purposes of the present discussion it can be assumed that this is absorbed into C. Also, a possible phase at order m_p^0 depending only on q may be absorbed into ψ, so C may be taken to be real.

Eq.(4.3), it must be stressed, is an *ansatz* for the solution, and many of the predictions subsequently derived depend on assuming this particular form. We will return later to the question of the validity and usefulness of this particular ansatz.

The Wheeler-DeWitt equation may now be solved perturbatively, by inserting the ansatz (4.3), and using the Planck mass m_p as a large parameter. Since this parameter

is not dimensionless, the expansion is meaningful only on length scales much greater than the Planck length. Note also that a double expansion of the full Wheeler-DeWitt equation is involved: a WKB expansion in the Planck mass, and a perturbation expansion in small inhomogeneities about minisuperspace.

Equating powers of the Planck mass, one obtains the following. At lowest order, one gets the Hamilton-Jacobi equation for S_0,

$$\frac{1}{2}(\nabla S_0)^2 + U(q) = 0. \tag{4.4}$$

To next order one obtains a conservation equation for C

$$2\nabla S_0 \cdot \nabla C - C\nabla^2 S_0 = 0 \tag{4.5}$$

and a Schrödinger equation for ψ,

$$i\nabla S_0 \cdot \nabla \psi = H_2 \psi \tag{4.6}$$

Consider now (4.4). As indicated above, it may be argued that a wave function predominantly of the form $e^{im_p^2 S_0}$ indicates a strong correlation between coordinates and momenta of the form

$$p_\alpha = m_p^2 \frac{\partial S_0}{\partial q^\alpha} \tag{4.7}$$

Since $\dot{q}^\alpha = p^\alpha$, (4.7) is a set of n first order differential equations (where, recall, $\alpha = 1, \cdots n$). Furthermore, since by (4.4), S_0 is a solution to the Hamilton Jacobi equation, it may be shown that these equations define a first integral to the classical field equations. The first integral (4.7) may be solved to yield an n-parameter set of classical solutions. It is for this reason that one says that the wave function (4.3) to leading order, corresponds to an ensemble of classical solutions to the field equations.

In ordinary quantum mechanics, this interpretation may be substantiated by subjecting an initial wave function of the form (4.3) to a sequence of approximate position samplings, and showing that the resulting probability distribution for histories is peaked about the set of classical solutions satisfying the first integral (4.7). See Halliwell and Dowker (1992), for example, for efforts in this direction.

The tangent vector to this congruence of classical paths is

$$\nabla S_0 \cdot \nabla \equiv \frac{\partial}{\partial t} \tag{4.8}$$

(4.6) is therefore the functional Schrödinger equation for the perturbation modes along the family of minisuperspace trajectories. This indicates that the perturbation modes are described by quantum field theory (in the functional Schrödinger picture) along the family of classical backgrounds described by (4.7).

D. The Measure on the Set of Classical Trajectories

The final stage is to put a measure on the congruence of classical paths, and to find an inner product for solutions to the Schrödinger equation (4.8). Suppose one chooses some $(n-1)$-dimensional surface in minisuperspace as the beginning of classical evolution. Through (4.7), the wave function then effectively fixes the initial velocities on that surface, in terms of the coordinates. However, the wave function should also provide a probability measure on the set of classical trajectories about which the wave function is peaked. How is this measure constructed from the wave function? The construction of a satisfactory non-negative measure remains an outstanding problem of quantum cosmology, but perhaps the most successful attempts so far involve the Klein-Gordon current,

$$J = \frac{i}{2}(\Psi \nabla \Psi^* - \Psi^* \nabla \Psi) \tag{4.9}$$

It is conserved by virtue of the Wheeler-DeWitt equation (4.2),

$$\nabla \cdot J = 0 \tag{4.10}$$

Choose a family of surfaces $\{\Sigma_\lambda\}$, parametrized by λ, cutting across the flow of J. Then (4.10) suggests that for each λ, a probability measure on the congruence of trajectories is the flux of J across the surface:

$$dP = J \cdot d\Sigma \tag{4.11}$$

This measure is conserved along the flow of J, as is readily shown from (4.10).

The problem with (4.11), however, is that it is not always positive. For example, if the surfaces Σ are taken to be surfaces of constant scale factor, the flow of J typically cuts these surfaces more than once, because of the possibility of expanding and collapsing universes, leading to negative values for (4.11). Furthermore, J vanishes when Ψ is real. Still, some sense may be made out of (4.11) by restricting to the WKB wave functions, (4.3). For these wave functions, the current is

$$J \approx m_p^2 \, |C|^2 \, |\psi|^2 \, \nabla S_0 + O(m_p^0) \tag{4.12}$$

For reasonably large regions of minisuperspace (but not globally), it is usually possible to choose a set of surfaces $\{\Sigma_\lambda\}$ for which $\nabla S_0 \cdot d\Sigma \geq 0$, and thus the probability measure will be positive. Furthermore, this measure implies the standard $|\psi|^2$ measure for the perturbation wave functions, completing the demonstration that quantum field theory for the perturbations emerges in the semiclassical limit.

This approach was described a long time ago by Misner (1972), and developed more recently by Vilenkin (1989). It is problematic for a number of reasons: one is the global problem of choosing the surfaces $\{\Sigma_\lambda\}$; another is that is very tied to the WKB wave functions (4.3). More will be said about this in the next section. Still, it allows predictions to be made in a number of situations of interest.

The "naive" Schrödinger measure is also sometimes proposed in place of (4.11) (*e.g.*, Hawking and Page, 1986, 1988). This is the assertion that the probability of finding the system in a region of superspace of volume dV is $|\Psi|^2 dV$, where Ψ is the wave function of the whole system. This does at least have the advantage that is is everywhere positive. Furthermore, it can be argued to reduce to (4.11) for WKB wave

functions, in the limit in which the volume of superspace dV is taken to be hypersurface of codimension one slightly thickened along the direction of the flow of J. As argued by Kuchař, however, this measure is problematic for other reasons (Kuchař, 1992).

At this stage, it is appropriate to comment on the problem of time. In the Wheeler-DeWitt equation (4.2) (or (2.1)), there is no distinguished variable that plays the role of time. This is the problem of time. In the scheme described above, however, a time parameter has emerged. It is the parameter λ labeling the family of surfaces $\{\Sigma_\lambda\}$, and may be chosen to be the same as the parameter t defined in Eq.(4.8), the affine parameter along the integral curves of ∇S_0. The point to be made is that this parameter has emerged only in the region where the wave function is oscillatory, and in particular, as a consequence of the assumed semiclassical form of the wave function, (4.3). This is therefore an explicit illustration of the point of view, not uncommonly expressed, that time, and indeed spacetime, need not exist at the most fundamental level but may emerge as approximate features under some suitable set of conditions. It is in this sense that feature (QC5) of quantum cosmology may be reconciled with (C5) of ordinary quantum mechanics.

Modulo the above difficulties, Eq.(4.11) is the desired probability measure on possible histories of the universe. It is commonly not normalizable over the entire surface Σ; but this need not matter, because it is *conditional* probabilities that one is typically interested in. Suppose, for example, one is given that the history of the universe passed through a subset s_1 of a surface Σ, and one wants the probability that the universe passed through a subset s_0 of s_1. The relevant conditional probability is,

$$p(s_0|s_1) = \frac{\int_{s_0} J \cdot d\Sigma}{\int_{s_1} J \cdot d\Sigma} \tag{4.13}$$

The set of all histories intersecting Σ could include universes which recollapse after a very short time. A reasonable choice for s_1, therefore, might be universes that exist long enough for stars to form before recollapsing, as discussed earlier. s_0 could then be taken to be the subset of such universes which possess certain features resembling our universe. If the resulting conditional probability turned out to be close to one or zero, this would then constitute a definite prediction.

Steps (A)–(D) above constitute the general interpretational scheme implicit or explicit in most attempts to extract predictions from the wave function of the universe. It is not by any means a consistent interpretational scheme, but is almost a list of rules of thumb inspired by the Everett interpretation, and built on analogies with ordinary quantum mechanics. It has many difficulties, some of which we now discuss.

5. Problems and Objections

We have argued that the WKB wave functions (4.3) correspond to an ensemble of classical paths defined by the first integral (4.7). Strictly speaking, what this means is that the WKB wave function is really some kind of *superposition* of wave functions, each of which corresponds to an individual classical history (like a superposition of coherent states).† A closely related point is the question of why one should be allowed

† For the explicit construction of wave packets in quantum cosmology, see Kazama and Nakayama (1985) and Kiefer (1988).

to study an interpretation based on the WKB form, (4.3): one would expect a more general wave function to be expressed as a *sum* of such terms. In each of these cases, we are acting as if we had a classical statistical ensemble, when what we really have is a superposition of interfering states. Why should it be permissible to treat each term in the sum separately, when strictly they are interfering?

This point concerns the general question of why or when it is permissable to ignore the interference terms in a superposition, and treat each term as if it were the member of a statistical ensemble. Technically, the destruction of interference is generally referred to as decoherence.

Two notions of decoherence have been employed. The most precise is that of the decoherent histories approach, where it enters at a very fundamental level: Interference is most generally and properly thought of as the failure of the probability sum rules for quantum-mechanical histories. Decoherence, as destruction of interference, is therefore best regarded as the recovery of these rules (Gell-Mann and Hartle, 1990).

By contrast, in the wave function approach to quantum cosmology, decoherence appears to have been added as an afterthought, using a different and somewhat vaguer definition: decoherence is held to be related to the tendancy of the density matrix towards diagonality (Joos and Zeh, 1985). It is also associated with the establishment of correlations of a given system with its environment, and with the stability of certain states under evolution in the presence of an environment (Zurek, 1981, 1982; Unruh and Zurek, 1989). These definitions are problematic for a number of reasons. One is that (in ordinary quantum mechanics) the density matrix refers only to a single moment of time, yet the proper definition of interference – the effect one is trying to destroy – is in terms of histories. Another is the question of the basis in which the density matrix should be diagonal.†

Despite the difficulties, the density matrix approach has been the topic of a number of papers on decoherence in quantum cosmology, perhaps because it is technically much simpler (Fukuyama and Morikawa, 1989; Habib and Laflamme, 1990; Halliwell, 1989; Kiefer, 1987; Laflamme and Louko, 1991; Mellor, 1989; Morikawa, 1989; Padmanabhan, 1989; Paz, 1991; Paz and Sinha, 1992). Models are considered in which the variables of interest are coupled to a wider environment, and a coarse-graining is carried out, in which the states of the environment are traced over. In the case of the whole universe, which strictly has no environment, this is achieved quite simply by postulating a sufficiently complex universe with a suitably large number of subsystems, and ignoring some of the subsystems. The typical result of such models is that the interference terms are suppressed very effectively. Furthermore, one and the same mechanism of coarse-graining also causes decoherence in the histories-based definition of the process. It is therefore plausible that a more sophisticated analysis using the decoherent histories approach will lead to the same conclusions. One way or another, one finds some amount of justification for treating the terms in a superposition separately, and treating the set of paths to which a WKB wave function correspond as a statistical ensemble. These arguments have not gained universal acceptance, however (see for example, Kuchař (1992)).

After the presentation of this contribution, J.Ehlers asked about the observational

† The basis issue is discussed, for example, in Barvinsky and Kamenshchik (1990), Deutsch (1985), and Markov and Mukhanov (1988). Also, see Zurek (1992) for a possible reconciliation of the above two differing views of decoherence.

status of quantum cosmology. Since quantum cosmology aspires to make observable predictions, this is obviously a very important question. Interest in quantum cosmology arose partly as a result of the realization that conventional classical cosmological scenarios relied on certain (possibly tacit) assumptions about initial conditions. As is well known, the inflationary universe scenario alleviates the hot big bang model of extreme dependence on initial conditions; but it does not release it from *all* dependence. The amount of inflation, the details of the density fluctuations generated, and indeed, the very occurrence of inflation are initial conditions-dependent. One of the main successes of quantum cosmology (modulo the objections described in this paper) has been to demonstrate that the desired initial conditions for inflation are implied by certain simple boundary condition proposals for the wave function of the universe. This might therefore be regarded as an observational prediction, although it is admittedly a very indirect one.

More direct tests will clearly be very difficult. The universe has gone through many stages of evolution, each of which is modeled separately. In observing the universe today, it is difficult to distinguish between effects largely due to initial conditions and those largely due to dynamical evolution or to the modeling of a particular stage. What is needed is an effect produced very early in the history of the universe, but that is largely insensitive to subsequent evolution. Grishchuk (1987) has argued that gravitational waves might be the sought-after probe of the very early universe. Boundary condition proposals for the wave function of the universe typically make definite predictions about the initial state of the graviton field (*e.g.*, Halliwell and Hawking, 1985). Memory of this initial state could well be preserved throughout the subsequent evolution of the universe, because gravitational waves interact so weakly. Parts of their spectra observable today might therefore contain signatures of the initial state, leading to the exciting possibility of distinguishing observationally between different boundary condition proposals. These ideas are of course rather speculative, and gravitational wave astronomy is still in a very primitive state. Still, quantum cosmology suffers from an acute lack of connections observational cosmology, and any potentially observable effect deserves further study.

L.Smolin commented that the scheme described here is very semiclassical in nature, and suggested that it is perhaps not much more than a classical statistical theory. It is certainly true that it is very semiclassical in nature, that its predictive output has the form of classical statistical theory, and perhaps causes one to wonder how much of it is really quantum-mechanical in nature. However, there *are* some genuinely quantum-mechanical aspects to this predictive scheme. Perhaps the principle one is the prediction of regions in which classical laws are not valid (the regions of superspace in which the wave function is exponential). Determination of the existence and location of these regions requires the quantum theory – their existence and location cannot be anticipated by inspection of the classical theory alone. Furthermore, the existence of these regions underscores the necessity of discussing the issue of initial or boundary conditions from within the context of the quantum theory. For in classical theories of initial conditions (including classical statistical ones), one might be attempting to impose classical initial conditions in a region in which classical laws are quite simply not valid.

Finally, a critical appraisal of the quantum cosmology program (which inspired some of the remarks made here) may be found in Isham (1991).

6. Conclusions

In this talk I have described the heuristic set of rules that have been used so far to make crude but plausible predictions in quantum cosmology. These rules are, however, rather heuristic and semiclassical in nature. The interpretation of the wave function seems to proceed on a case by case basis, and no satisfactory scheme for a completely general wave function is available. I am therefore forced to conclude that quantum cosmology does not yet possess an entirely satisfactory scheme for the extraction of predictions from the wave function.

At the present time, the decoherent histories approach appears to offer the most promising hope of improving the situation. A reasonable expectation is that the heuristic interpretational techniques described here will emerge from this more sophisticated approach. Detailed demonstration of this assertion is very much a matter of current research.

Acknowledgments

I thank Jim Hartle and Raymond Laflamme for very useful discussions on the material of this talk. I am very grateful to the organizers of GR13, and in particular to Carlos Kozameh and Bob Wald, for inviting me to take part in this stimulating meeting. I would also like to thank Jürgen Ehlers for initiating the session on Interpretational Issues in Quantum Cosmology, and for his thought-provoking questions. Finally, I am very grateful to Mario Castagnino, and his colleagues in Buenos Aires and Rosario, for their warm hospitality during my stay in Argentina.

References

Albrecht, A. (1992), to appear in, *Physical Origins of Time Asymmetry*, eds. J.J.Halliwell, J.Perez-Mercader and W.H.Zurek (Cambridge University Press, Cambridge). Two perspective on a decohering spin.

Anderson, A. (1990), *Phys.Rev.* **D42**, 585. On predicting correlations from Wigner functions.

Banks, T. (1985), *Nucl.Phys.* **B249**, 332. TCP, quantum gravity, the cosmological constant and all that...

Banks, T., Fischler, W. and Susskind, L. (1985), *Nucl.Phys.* **B262**, 159. Quantum cosmology in 2+1 and 3+1 dimensions.

Barrow, J.D. and Tipler, F. (1986), *The Anthropic Cosmological Principle* (Oxford University Press, Oxford).

Barvinsky, A.O. and Kamenshchik, A.Yu. (1990), *Class.Quantum Grav.* **7**, 2285. Preferred basis in the many-worlds interpretation of quantum mechanics and quantum cosmology.

Bell, J.S. (1981), in *Quantum Gravity 2: A Second Oxford Symposium*, eds. C.J.Isham, R.Penrose and D.W.Sciama (Clarendon Press, Oxford). Quantum mechanics for cosmologists.

Blencowe, M. (1991), *Ann.Phys.(N.Y.)* **211**, 87. The consistent histories interpretation of quantum fields in curved spacetime.

Calzetta, E. and Hu, B.L. (1989), *Phys.Rev.* **D40**, 380. Wigner distribution and phase space formulation of quantum cosmology.

Deutsch, D. (1985), *Int.J.Theor.Phys.* **24**, 1. Quantum theory as a universal physical theory.

DeWitt, B.S. (1967), *Phys.Rev.* **160**, 1113. Quantum theory of gravity. I. The canonical theory.

DeWitt, B.S. and Graham, N. (eds.) (1973), *The Many Worlds Interpretation of Quantum Mechanics* (Princeton University Press, Princeton).

Dowker, H.F. and Halliwell, J.J. (1992), CTP preprint #2071, Fermilab-pub-92-44A, to appear in *Phys.Rev.D.* The quantum mechanics of history: the decoherence functional in quantum mechanics.

Everett, H. (1957), *Rev.Mod.Phys.* **29**, 454. Relative state formulation of quantum mechanics.

Fang, L.Z. and Ruffini, R.(eds.) (1987), *Quantum Cosmology*, Advanced Series in Astrophysics and Cosmology No.3 (World Scientific, Singapore).

Fang, L.Z. and Wu, Z.C. (1986), *Intern.J.Mod.Phys.* **A1**, 887. An overview of quantum cosmology.

Farhi, E., Goldstone, J. and Gutmann, S. (1989), *Ann.Phys.(N.Y.)* **192**, 368. How probability arises in quantum mechanics.

Finkelstein, D. (1963), *Trans.N.Y.Acad.Sci.* **25**, 621. The logic of quantum physics.

Fischler, W., Ratra, B. and Susskind, L. (1986), *Nucl.Phys.* **B259**, 730. (*errata Nucl.Phys.* **B268** 747 (1986)). Quantum mechanics of inflation.

Fukuyama, T. and Morikawa, M. (1989), *Phys.Rev.* **39**, 462. Two-dimensional quantum cosmology: directions of dynamical and thermodynamical arrows of time.

Gell-Mann, M. and Hartle, J.B. (1990) in, *Complexity, Entropy and the Physics of Information*, Santa Fe Institute Studies in the Sciences of Complexity, vol IX, edited by W.H.Zurek (Addison Wesley); also in, *Proceedings of the Third International Symposium on Foundations of Quantum Mechanics in the Light of New Technology*, edited by S.Kobayashi (Japan Physical Society). Quantum cosmology and quantum mechanics.

Geroch, R. (1984), *Noûs* **18**, 617. The Everett interpretation. (This issue of *Noûs* also contains a number of other papers in which the Everett interpretation is discussed, largely from a philosophical point of view.)

Graham, N. (1973), in *The Many Worlds Interpretation of Quantum Mechanics*, B.S.DeWitt and N.Graham (eds.) (Princeton University Press, Princeton). The measurement of relative frequency.

Griffiths, R.B. (1984), *J.Stat.Phys.* **36**, 219. Consistent histories and the interpretation of quantum mechanics.

Grishchuk, L.P. (1987), *Mod.Phys.Lett.* **A2**, 631. Quantum creation of the universe can be observationally verified.

Guth, A.H. (1981), *Phys.Rev.* **D28**, 347. The inflationary universe: a possible solution to the horizon, flatness and monopole problems.

Habib, S. (1990), *Phys.Rev.* **42**, 2566. The classical limit in quantum cosmology. I. Quantum mechanics and the Wigner function.

Habib, S. and Laflamme, R. (1990), *Phys.Rev.* **42**, 4056. Wigner function and decoherence in quantum cosmology.

Halliwell, J.J. (1987), *Phys.Rev.* **D36**, 3626. Correlations in the wave function of the universe.

Halliwell, J.J. (1989), *Phys.Rev.* **D39**, 2912. Decoherence in quantum cosmology.

Halliwell, J.J. (1990), in *Proceedings of the Seventh Jerusalem Winter School for Theoretical Physics: Quantum Cosmology and Baby Universes*, eds. S.Coleman, J.B.Hartle, T.Piran and S.Weinberg (World Scientific, Singapore). Introductory lectures on quantum cosmology.

Halliwell, J.J. (1991), *Scientific American* **265**, 76. Quantum cosmology and the creation of the universe.

Halliwell, J.J. (1992), CTP preprint #2070, to appear in *Phys.Rev.D.* Smeared Wigner functions and quantum-mechanical histories.

Halliwell, J.J. and Hartle, J.B. (1990), *Phys.Rev.* **D41**, 1815. Integration contours for the no-boundary wave function of the universe.

Halliwell, J.J. and Hartle, J.B. (1991), *Phys.Rev.* **D43**, 1170. Wave functions constructed from an invariant sum-over-histories satisfy constraints.

Halliwell, J.J. and Hawking, S.W. (1985), *Phys.Rev.* **D31**, 1777. Origin of structure in the universe.

Hartle, J.B. (1968), *Am.J.Phys.* **36**, 704. Quantum mechanics of individual systems.

Hartle, J.B. (1985), in *High Energy Physics 1985: Proceedings of the Yale Summer School*, eds. M.J.Bowick and F.Gursey (World Scientific, Singapore). quantum cosmology.

Hartle, J.B. (1986), in *Gravitation in Astrophysics, Cargese, 1986*, eds. B.Carter and J.Hartle (Plenum, New York). Prediction and observation in quantum cosmology.

Hartle, J.B. (1990), in *Proceedings of the Seventh Jerusalem Winter School for Theoretical Physics: Quantum Cosmology and Baby Universes*, eds. S.Coleman, J.B.Hartle, T.Piran and S.Weinberg (World Scientific, Singapore). The quantum mechanics of cosmology.

Hartle, J.B. (1992a), this volume.

Hartle, J.B. (1992b), to appear in *Proceedings of the Les Houches Summer School, 1992.*

Hartle, J.B. and Hawking, S.W. (1983), *Phys.Rev.* **D28**, 2960. Wave function of the universe.

Hawking, S.W. (1982), in *Astrophysical Cosmology*, eds. H.A.Brück, G.V.Coyne and M.S.Longair (Pontifica Academia Scientarium, Vatican City) (*Pont.Acad.Sci.Varia* **48**, 563). The boundary conditions of the universe.

Hawking, S.W. (1984a), *Nucl.Phys.* **B239**, 257. The quantum state of the universe.

Hawking, S.W. (1984b), in *Relativity, Groups and Topology II, Les Houches Session XL*, eds. B.DeWitt and R.Stora (North Holland, Amsterdam). Lectures on quantum cosmology.

Hawking, S.W. and Page, D.N. (1986), *Nucl.Phys.* **B264**, 185. Operator ordering and the flatness of the universe.

Hawking, S.W. and Page, D.N. (1988), *Nucl.Phys.* **B298**, 789. How probable is inflation?

Hu, B.L. (1989), Cornell preprint, CLNS 89/941. Quantum and statistical effects in superspace cosmology.

Isham, C.J. (1991), Imperial College preprint TP/90-91/14, to appear in *Proceedings of the Schladming Winter School, 1991*. Conceptual and geometrical problems in quantum gravity.

Joos, E. and Zeh, H.D. (1985), *Zeit.Phys.* **B59**, 223. The emergence of classical properties through interaction with the environment.

Kazama, Y. and Nakayama, R. (1985), *Phys.Rev.* **D32**, 2500. Wave packet in quantum cosmology.

Kent, A. (1990), *Intern.J.Mod.Phys.* **A5**, 1745. Against many-worlds interpretations.

Kiefer, C. (1987), *Class.Quantum Grav.* **4**, 1369. Continuous measurement of minisuperspace variables by higher multipoles.

Kiefer, C. (1988), *Phys.Rev.* **D38**, 1761. Wave packets in minisuperspace.

Kodama, H. (1988), Kyoto University preprint KUCP-0014. Quantum cosmology in terms of the Wigner function.

Kuchař, K.V. (1989), in *Proceedings of the Osgood Hill Meeting on Conceptual Problems in Quantum Gravity*, eds. A.Ashtekar and J.Stachel (Birkhauser, Boston). The problem of time in canonical quantization of relativistic systems.

Kuchař, K.V. (1992), in *Proceedings of the 4th Canadian Conference on General Relativity and Relativistic Astrophysics*, eds. G.Kunstatter, D.Vincent and J.Williams (World Scientific, Singapore, 1992). Time and interpretations of quantum gravity.

Kuchař, K.V. and Ryan, M.P. (1986), in Yamada Conference XIV, eds. H.Sato and T.Nakamura (World Scientific). Can minisuperspace quantization be justified?

Kuchař, K.V. and Ryan, M.P. (1989), *Phys.Rev.* **D40**, 3982. Is minisuperspace quantization valid? Taub in Mixmaster.

Laflamme, R. (1991), in *Proceedings of the XXII GIFT International Seminar on Theoretical Physics*, St. Feliu de Guixols, Spain, June 3–8, 1991. Introduction and applications of quantum cosmology.

Laflamme, R. and Louko, J. (1991), *Phys.Rev.* **D43**, 3317. Reduced density matrices and decoherence in quantum cosmology.

Lapchinsky, V. and Rubakov, V.A. (1979), *Acta.Phys.Polonica* **B10**, 1041. Canonical quantization of gravity and quantum field theory in curved space-time.

Leggett, A.J. (1980), *Suppl.Prog.Theor.Phys.* **69**, 80. Macroscopic quantum systems and the quantum theory of measurement.

Markov, M.A., and Mukhanov, V.F. (1988), *Phys.Lett.* **127A**, 251. Classical preferable basis in quantum mechanics.

Mellor, F. (1989), *Nucl.Phys.* **B353**, 291. Decoherence in quantum Kaluza-Klein theories.

Misner, C.W. (1969), *Phys.Rev.* **186**, 1319. Quantum cosmology I.

Misner, C.W. (1972), in *Magic Without Magic: John Archibald Wheeler, a Collection of Essays in Honor of his 60th Birthday*, ed. J.Klauder (Freeman, San Francisco). Minisuperspace.

Morikawa, M. (1989), *Phys.Rev.* **D40**, 4023. Evolution of the cosmic density matrix.

Omnès, R. (1990), *Ann.Phys.(N.Y.)* **201**, 354. From Hilbert space to common sense: a synthesis of recent progress in the interpretation of quantum mechanics.

Omnès, R. (1992), *Rev.Mod.Phys.* **64**, 339. Consistent interpretations of quantum mechanics.

Padmanabhan, T. (1989), *Phys.Rev.* **D39**, 2924. Decoherence in the density matrix describing quantum three-geometries and the emergence of classical spacetime.

Page, D.N. (1986), in *Quantum Concepts in Space and Time*, eds. R.Penrose and C.J.Isham (Clarendon Press, Oxford). Hawking's wave function for the universe.

Page, D.N. (1991), in *Proceedings of the Banff Summer Institute on Gravitation*, eds. R.B.Mann and P.Wesson (World Scientific, Singapore). Lectures on quantum cosmology.

Paz, J.P. (1991), *Phys.Rev.* **D44**, 1038. Decoherence and backreaction in quantum cosmology: The origin of the semiclassical Einstein equations.

Paz, J.P. and Sinha, S. (1991), *Phys.Rev.* **D45**, 2823. Decoherence and backreaction in quantum cosmology: multidimensional minisuperspace examples.

Shirai, I. and Wada, S. (1988), *Nucl.Phys.* **B303**, 728. Cosmological perturbations and quantum fields in curved space-time.

Singh, T.P. and Padmanabhan, T. (1989), *Ann.Phys.(N.Y.)* **196**, 296. Notes on semiclassical gravity.

Smolin, L. (1984), in *Quantum Theory of Gravity: Essays in Honour of the 60th Birthday of Bryce DeWitt* (Hilger, Bristol). On quantum gravity and the many-worlds interpretation of quantum mechanics.

Tipler, F. (1986), *Phys.Rep.* **137**, 231. Interpreting the wave function of the universe.

Unruh, W.G. and Wald, R. (1989), *Phys.Rev.* **D40**, 2598. Time and the interpretation of quantum gravity.

Unruh, W.G. and Zurek, W.H. (1989), *Phys.Rev.* **D40**, 1071. Reduction of a wavepacket in quantum Brownian motion.

Vachaspati, T. and Vilenkin, A. (1988), *Phys.Rev.* **D37**, 898. Uniqueness of the tunneling wave function of the universe.

Vilenkin, A. (1989), *Phys.Rev.* **D39**, 1116. The interpretation of the wave function of the universe.

Wada, S. (1986), *Nucl.Phys.* **B276**, 729. Quantum cosmological perturbations in pure gravity.

Wada, S. (1988), *Mod.Phys.Lett* **A3**, 645. Interpretation and predictability of quantum mechanics and quantum cosmology.

Wheeler, J.A. (1963), in *Relativity, Groups and Topology*, eds. C.DeWitt and B.DeWitt (Gordon and Breach, New York). Geometrodynamics and the issue of the final state.

Wheeler, J.A. (1968), in *Batelles Rencontres*, eds. C.DeWitt and J.A.Wheeler (Benjamin, New York). Superspace and the nature of quantum geometrodynamics.

Wheeler, J.A. and Zurek, W.H. (1983), *Quantum Theory and Measurement* (Princeton University Press, Princeton).

Zurek, W.H. (1981), *Phys.Rev.* **D24**, 1516. Pointer basis of quantum apparatus: Into what mixture does the wave packet collapse?

Zurek, W.H. (1982), *Phys.Rev.* **D26**, 1862. Environment-induced superselection rules.

Zurek, W.H. (1992), to appear in, *Physical Origins of Time Asymmetry*, eds. J.J.Halliwell, J.Perez-Mercader and W.H.Zurek (Cambridge University Press, Cambridge). Preferred sets of states, predictability, classicality and the environment-induced decoherence.

Paper presented at the 13th Int. Conf. on General Relativity and Gravitation
Cordoba, Argentina, 1992: Part 1, Plenary Lectures

The quantum mechanics of closed systems

James B. Hartle
Department of Physics
University of California
Santa Barbara, CA 93106 USA

ABSTRACT

A pedagogical introduction is given to the quantum mechanics of closed systems, most generally the universe as a whole. Quantum mechanics aims at predicting the probabilities of alternative coarse-grained time histories of a closed system. Not every set of alternative coarse-grained histories that can be described may be consistently assigned probabilities because of quantum mechanical interference between individual histories of the set. In "Copenhagen" quantum mechanics, probabilities can be assigned to histories of a subsystem that have been "measured". In the quantum mechanics of closed systems, containing both observer and observed, probabilities are assigned to those sets of alternative histories for which there is negligible interference between individual histories as a consequence of the system's initial condition and dynamics. Such sets of histories are said to decohere. We define decoherence for closed systems in the simplified case when quantum gravity can be neglected and the initial state is pure. Typical mechanisms of decoherence that are widespread in our universe are illustrated.

Copenhagen quantum mechanics is an approximation to the more general quantum framework of closed subsystems. It is appropriate when there is an approximately isolated subsystem that is a participant in a measurement situation in which (among other things) the decoherence of alternative registrations of the apparatus can be idealized as exact.

Since the quantum mechanics of closed systems does not posit the existence of the quasiclassical domain of everyday experience, the domain of the approximate aplicability of classical physics must be explained. We describe how a quasiclassical domain described by averages of densities of approximately conserved quantities could be an emergent feature of an initial condition of the universe that implies the approximate classical behavior of spacetime on accessible scales.

I. INTRODUCTION

It is an inescapable inference from the physics of the last sixty years that we live in a quantum mechanical universe — a world in which the basic laws of physics conform to that framework for prediction we call quantum mechanics. We perhaps have little evidence of peculiarly quantum mechanical phenomena on large and even

familiar scales, but there is no evidence that the phenomena that we do see cannot be described in quantum mechanical terms and explained by quantum mechanical laws. If this inference is correct, then there must be a description of the universe as a whole and everything in it in quantum mechanical terms. The nature of this description and its observable consequences are the subject of quantum cosmology.

Our observations of the present universe on the largest scales are crude and a classical description of them is entirely adequate. Providing a quantum mechanical description of these observations alone might be an interesting intellectual challenge, but it would be unlikely to yield testable predictions differing from those of classical physics. Today, however, we have a more ambitious aim. We aim, in quantum cosmology, to provide a theory of the initial condition of the universe which will predict testable correlations among observations today. There are no realistic predictions of any kind that do not depend on this initial condition, if only very weakly. Predictions of certain observations may be testably sensitive to its details. These include the large scale homogeneity and isotropy of the universe, its approximate spatial flatness, the spectrum of density fluctuations that produced the galaxies, the homogeneity of the thermodynamic arrow of time, and the existence of classical spacetime. Recently, there has been speculation that even the coupling constants of the effective interactions of the elementary particles at accessible energy scales may be probabilistically distributed with a distribution which may depend, in part, on the initial condition of the universe (Hawking, 1983, Coleman, 1988, Giddings and Strominger, 1988). It is for such reasons that the search for a theory of the initial condition of the universe is just as necessary and just as fundamental as the search for a theory of the dynamics of the elementary particles. They may even be the same searches.

The physics of the very early universe is likely to be quantum mechanical in an essential way. The singularity theorems of classical general relativity suggest that an early era preceded ours in which even the geometry of spacetime exhibited significant quantum fluctuations. It is for a theory of the initial condition that describes this era, and all later ones, that we need to spell out how to apply quantum mechanics to cosmology. Recent years have seen much promising progress in the search for a theory of the quantum initial condition. However, it is not my purpose to review these developments here.* Rather, I shall argue that this somewhat obscure branch of astrophysics may have implications for the formulation and interpretation of quantum mechanics on day-to-day scales. My thesis will be that by looking at the universe as a whole one is led to an understanding of quantum mechanics which clarifies many of the long standing interpretative difficulties of the subject.

The Copenhagen frameworks for quantum mechanics, as they were formulated in the '30s and '40s and as they exist in most textbooks today, are inadequate for quantum cosmology. Characteristically these formulations assumed, as *external* to the framework of wave function and Schrödinger equation, the quasiclassical domain we see all about us. Bohr (1958) spoke of phenomena which could be alternatively described in classical language. In their classic text, Landau and Lifschitz (1958) formulated quantum mechanics in terms of a separate classical physics. Heisenberg and others stressed the central role of an external, essentially classical, observer.* Characteristically, these formulations assumed a possible division of the world into

* For a recent review of quantum cosmology see Halliwell (1991).

* For a clear statement of this point of view, see London and Bauer (1939).

"obsever" and "observed", assumed that "measurements" are the primary focus of scientific statements and, in effect, posited the existence of an external "quasiclassical domain". However, in a theory of the whole thing there can be no fundamental division into observer and observed. Measurements and observers cannot be fundamental notions in a theory that seeks to describe the early universe when neither existed. In a basic formulation of quantum mechanics there is no reason in general for there to be any variables that exhibit classical behavior in all circumstances. Copenhagen quantum mechanics thus needs to be generalized to provide a quantum framework for cosmology.

In a generalization of quantum mechanics which does not *posit* the existence of a quasiclassical domain, the domain of applicability of classical physics must be *explained*. For a quantum mechanical system to exhibit classical behavior there must be some restriction on its state and some coarseness in how it is described. This is clearly illustrated in the quantum mechanics of a single particle. Ehrenfest's theorem shows that generally

$$M\frac{d^2\langle x\rangle}{dt^2} = \left\langle -\frac{\partial V}{\partial x}\right\rangle . \tag{1}$$

However, only for special states, typically narrow wave packets, will this become an equation of motion for $\langle x\rangle$ of the form

$$M\frac{d^2\langle x\rangle}{dt^2} = -\frac{\partial V(\langle x\rangle)}{\partial x}. \tag{2}$$

For such special states, successive observations of position in time will exhibit the classical correlations predicted by the equation of motion (2) *provided* that these observations are coarse enough so that the properties of the state which allow (2) to replace the general relation (1) are not affected by these observations. An *exact* determination of position, for example, would yield a completely delocalized wave packet an instant later and (2) would no longer be a good approximation to (1). Thus, even for large systems, and in particular for the universe as a whole, we can expect classical behavior only for certain initial states and then only when a sufficiently coarse grained description is used.

If classical behavior is *in general* a consequence only of a certain class of states in quantum mechanics, then, as a particular case, we can expect to have classical spacetime only for certain states in quantum gravity. The classical spacetime geometry we see all about us in the late universe is not property of every state in a theory where geometry fluctuates quantum mechanically. Rather, it is traceable fundamentally to restrictions on the initial condition. Such restrictions are likely to be generous in that, as in the single particle case, many different states will exhibit classical features. The existence of classical spacetime and the applicability of classical physics are thus not likely to be very restrictive conditions on constructing a theory of the initial condition.

It was Everett who, in 1957, first suggested how to generalize the Copenhagen frameworks so as to apply quantum mechanics to cosmology.* Everett's idea was to take quantum mechanics seriously and apply it to the universe as a whole. He

* The original reference is Everett (1957). For a useful collection of reprints see DeWitt and Graham (1973).

showed how an observer could be considered part of this system and how its activities — measuring, recording, calculating probabilities, etc. — could be described within quantum mechanics. Yet the Everett analysis was not complete. It did not adequately describe within quantum mechanics the origin of the "quasiclassical domain" of familiar experience nor, in an observer independent way, the meaning of the "branching" that replaced the notion of measurement. It did not distinguish from among the vast number of choices of quantum mechanical observables that are in principle available to an observer, the particular choices that, in fact, describe the quasiclassical domain.

In this essay, I will describe joint work with Murray Gell-Mann (Gell-Mann and Hartle, 1990a, 1990b) which aims at a coherent formulation of quantum mechanics for the universe as a whole that is a framework to explain rather than posit the quasiclassical domain of everyday experience. It is an attempt at an extension, clarification, and completion of the Everett interpretation. It builds on many aspects of the, so called post-Everett development, especially the work of Zeh (1971), Zurek (1981, 1982), and Joos and Zeh (1985). At important points it coincides with the, independent, earlier work of Bob Griffiths (1984) and Roland Omnès (*e.g.*, as reviewed in Omnès, 1992).

Our work is not complete, but I hope to sketch how it might become so. It is by now a very long story but I will try to describe the important parts in simplified terms.

II. Probabilities in General and Probabilities in Quantum Mechanics

Even apart from quantum mechanics, there is no certainty in this world and therefore physics deals in probabilities. It deals most generally with the probabilities for alternative time histories of the universe. From these, conditional probabilities can be constructed that are appropriate when some features about our specific history are known and further ones are to be predicted.

To understand what probabilities mean for a single closed system, it is best to understand how they are used. We deal, first of all, with probabilities for *single* events of the *single* system. When these probabilities become sufficiently close to zero or one there is a definite prediction on which we may act. How sufficiently close to 0 or 1 the probabilities must be depends on the circumstances in which they are applied. There is no certainty that the sun will come up tomorrow at the time printed in our daily newspapers. The sun may be destroyed by a neutron star now racing across the galaxy at near light speed. The earth's rotation rate could undergo a quantum fluctuation. An error could have been made in the computer that extrapolates the motion of the earth. The printer could have made a mistake in setting the type. Our eyes may deceive us in reading the time. Yet, we watch the sunrise at the appointed time because we compute, however imperfectly, that the probability of these things happening is sufficiently low.

Various strategies can be employed to identify situations where probabilities are near zero or one. Acquiring information and considering the conditional probabilities based on it is one such strategy. Current theories of the initial condition of the universe predict almost no probabilities near zero or one without further conditions. The "no boundary" wave function of the universe, for example, does not predict the present position of the sun on the sky. However, it will predict that the conditional probability for the sun to be at the position predicted by classical celestial mechanics given a few previous positions is a number very near unity.

Another strategy to isolate probabilities near 0 or 1 is to consider ensembles of repeated observations of identical subsystems in the closed system. There are no genuinely infinite ensembles in the world so we are necessarily concerned with the probabilities for deviations of the behavior of a finite ensemble from the expected behavior of an infinite one. These are probabilities for a single feature (the deviation) of a single system (the whole ensemble).

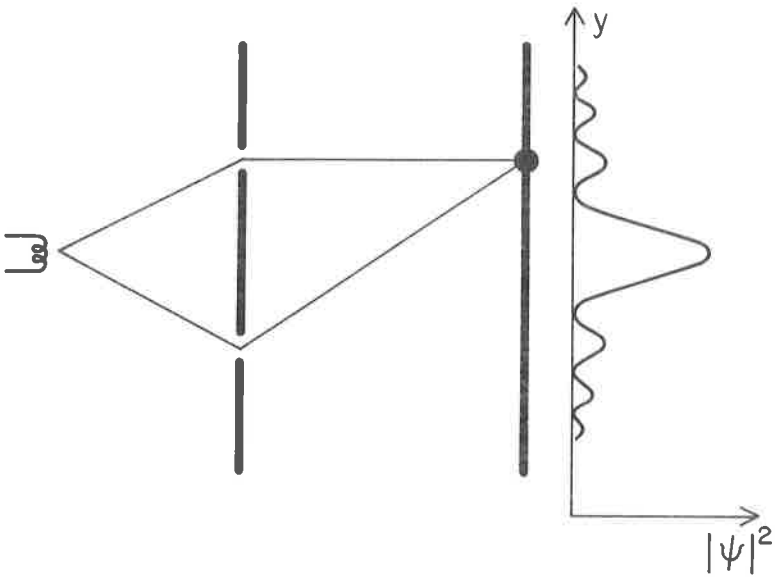

Fig. 1: The two-slit experiment. An electron gun at right emits an electron traveling towards a screen with two slits, its progress in space recapitulating its evolution in time. When precise detections are made of an ensemble of such electrons at the screen it is not possible, because of interference, to assign a probability to the alternatives of whether an individual electron went through the upper slit or the lower slit. However, if the electron interacts with apparatus that measures which slit it passed through, then these alternatives decohere and probabilities can be assigned.

The existence of large ensembles of repeated observations in identical circumstances and their ubiquity in laboratory science should not, therefore, obscure the fact that in the last analysis physics must predict probabilities for the single system that is the ensemble as a whole. Whether it is the probability of a successful marriage, the probability of the present galaxy-galaxy correlation function, or the probability of the fluctuations in an ensemble of repeated observations, we must deal with the probabilities of single events in single systems. In geology, astronomy, history, and cosmology, most predictions of interest have this character. The goal of

physical theory is, therefore, most generally to predict the probabilities of histories of single events of a single system.

Probabilities need be assigned to histories by physical theory only up to the accuracy they are used. Two theories that predict probabilities for the sun not rising tomorrow at its classically calculated time that are both well beneath the standard on which we act are equivalent for all practical purposes as far as this prediction is concerned. It is often convenient, therefore, to deal with approximate probabilities which satisfy the rules of probability theory up to the standard they are used.

The characteristic feature of a quantum mechanical theory is that not every set of alternative histories that may be described can be assigned probabilities. Nowhere is this more clearly illustrated than in the two slit experiment illustrated in Figure 1. In the usual "Copenhagen" discussion if we have not measured which of the two slits the electron passed through on its way to being detected at the screen, then we are not permitted to assign probabilities to these alternative histories. It would be inconsistent to do so since the correct probability sum rule would not be satisfied. Because of interference, the probability to arrive at y is not the sum of the probabilities to arrive at y going through the upper or lower slit:

$$p(y) \neq p_U(y) + p_L(y) \tag{3}$$

because

$$|\psi_L(y) + \psi_U(y)|^2 \neq |\psi_L(y)|^2 + |\psi_U(y)|^2 \quad . \tag{4}$$

If we *have* measured which slit the electron went through, then the interference is destroyed, the sum rule obeyed, and we *can* meaningfully assign probabilities to these alternative histories.

A rule is thus needed in quantum theory to determine which sets of alternative histories may be assigned probabilities and which may not. In Copenhagen quantum mechanics, the rule is that probabilities are assigned to histories of alternatives of a subsystem that are *measured* and not in general otherwise.

III. Probabilities for a Time Sequence of Measurements

To establish some notation, let us review in more detail the usual rules for the probabilities of time sequences of ideal measurements of subsystem using the two-slit experiment of Figure 1 as an example.

Alternatives for the electron are represented by projection operators in its Hilbert space. Thus, in the two slit experiment, the alternative that the electron passed through the lower slit is represented by the projection operator

$$P_U = \Sigma_s \int_U d^3x \, |\vec{x}, s\rangle\langle\vec{x}, s| \tag{5}$$

where $|\vec{x}, s\rangle$ is a localized state of the electron with spin component s, and the integral is over a volume around the upper slit. There is a similar projection operator P_L for the alternative that the electron goes through the lower slit. These are exclusive alternatives and they are exhaustive. These properties, as well as the requirements of being projections, are represented by the relations

$$P_L P_U = 0 \ , \quad P_U + P_L = 1, \quad P_L^2 = P_L, \quad P_U^2 = P_U \ . \tag{6}$$

There is a similarly defined set of projection operators $\{P_y\}$ representing the alternative positions of arrival at the screen.

We can now state the rule for the joint probability that the electron initially in a state $|\psi(t_0)\rangle$ at $t = t_0$ is determined by an ideal measurement at time t_1 to have passed through the upper slit and measured at time t_2 to arrive at point y on the screen. If one likes, one can imagine the case in which the electron is in a narrow wave packet in the horizontal direction with a velocity defined as sharply as possible consistent with the uncertainty principle. The joint probability is negligible unless t_1 and t_2 correspond to the times of flight to the slits and to the screen respectively.

The first step in calculating the joint probability is to evolve the state of the electron to the time t_1 of the first measurement

$$|\psi(t_1)\rangle = e^{-iH(t_1-t_0)/\hbar}|\psi(t_0)\rangle \ . \tag{7}$$

The probability that the outcome of the measurement at time t_1 is that the electron passed through the upper slit is:

$$(\text{Probability of } U) = \left\| P_U|\psi(t_1)\rangle \right\|^2 \tag{8}$$

where $\| \cdot \|$ denotes the norm of a vector in the electron's Hilbert space. If the outcome was the upper slit, and the measurement was an "ideal" one, that disturbed the electron as little as possible in making its determination, then after the measurement the state vector is reduced to

$$\frac{P_U|\psi(t_1)\rangle}{\|P_U|\psi(t_1)\rangle\|} \ . \tag{9}$$

This is evolved to the time of the next measurement

$$|\psi(t_2)\rangle = e^{-iH(t_2-t_1)/\hbar}\frac{P_U|\psi(t_1)\rangle}{\|P_U|\psi(t_1)\rangle\|} \ . \tag{10}$$

The probability of being detected at point y on the screen at time t_2 *given* that the electron passed through the upper slit is

$$(\text{Probability of } y \textit{ given } U) = \|P_y|\psi(t_2)\rangle\|^2 \ . \tag{11}$$

The *joint* probability that the electron is measured to have gone through the upper slit *and* is detected at y is the product of the conditional probability (11) with the probability (8) that the electron passed through U. The latter factor cancels the denominator in (10) so that combining all of the above equations in this section, we have

$$(\text{Probability of } y \textit{ and } U) = \left\| P_y e^{-iH(t_2-t_1)/\hbar} P_U e^{-iH(t_1-t_0)/\hbar}|\psi(t_0)\rangle \right\|^2 \ . \tag{12}$$

With Heisenberg picture projections this takes the even simpler form

$$(\text{Probability of } y \text{ and } U) = \left\| P_y(t_2)P_U(t_1)\ |\psi(t_0)\rangle \right\|^2 \ . \tag{13}$$

where, for example,

$$P_U(t) = e^{iHt/\hbar} P_U e^{-iHt/\hbar} \,. \tag{14}$$

The formula (13) is a compact and unified expression of the two laws of evolution that characterize the quantum mechanics of measured subsystems — unitary evolution in between measurements and reduction of the wave packet at a measurement.* The important thing to remember about the expression (13) is that everything in it — projections, state vectors, Hamiltonian — refer to the Hilbert space of a subsystem, in this example the Hilbert space of the electron that is measured.

In "Copenhagen" quantum mechanics, it is measurement that determines which histories of a subsystem can be assigned probabilities and formulae like (13) that determine what these probabilities are. We cannot have such rules in the quantum mechanics of closed systems. There is no fundamental division of a closed system into measured subsystem and measuring apparatus. There is no fundamental reason for the closed system to contain classically behaving measuring apparatus in all circumstances. In particular, in the early universe none of these concepts seem relevant. We need a more observer-independent, measurement-independent, quasiclassical domain-independent rule for which histories of a closed system can be assigned probabilities and what these probabilities are. The next section describes this rule.

IV. Post-Everett Quantum Mechanics

To describe the rules of post-Everett quantum mechanics, I shall make a simplifying assumption. I shall neglect gross quantum fluctuations in the geometry of spacetime, and assume a fixed background spacetime geometry which supplies a definite meaning to the notion of time. This is an excellent approximation on accessible scales for times later than 10^{-43} sec after the big bang. The familiar apparatus of Hilbert space, states, Hamiltonian, and other operators may then be applied to process of prediction. Indeed, in this context the quantum mechanics of cosmology is in no way distinguished from the quantum mechanics of a large isolated box, perhaps expanding, but containing both the observed and its observers (if any).

A set of alternative histories for a closed system is specified by giving exhaustive sets of exclusive alternatives at a sequence of times. Consider a model closed system initially in a pure state that can be described as an observer and two slit experiment, with appropriate apparatus for producing the electrons, detecting which slit they passed through, and measuring their position of arrival on the screen (Figure 2). Some alternatives for the whole system are:

1. Whether or not the observer decided to measure which slit the electron went through.
2. Whether the electron went through the upper or lower slit.
3. The alternative positions, $y_1, \cdots, y_N$, that the electron could have arrived at the screen.

This set of alternatives at a sequence of times defines a set of histories whose characteristic branching structure is shown in Figure 3. An individual history in the set is specified by some particular sequence of alternatives, *e.g.*, measured, upper, y_9.

---

* As has been noted by many authors, *e.g.*, Groenewold (1952) and Wigner (1963) among the earliest.

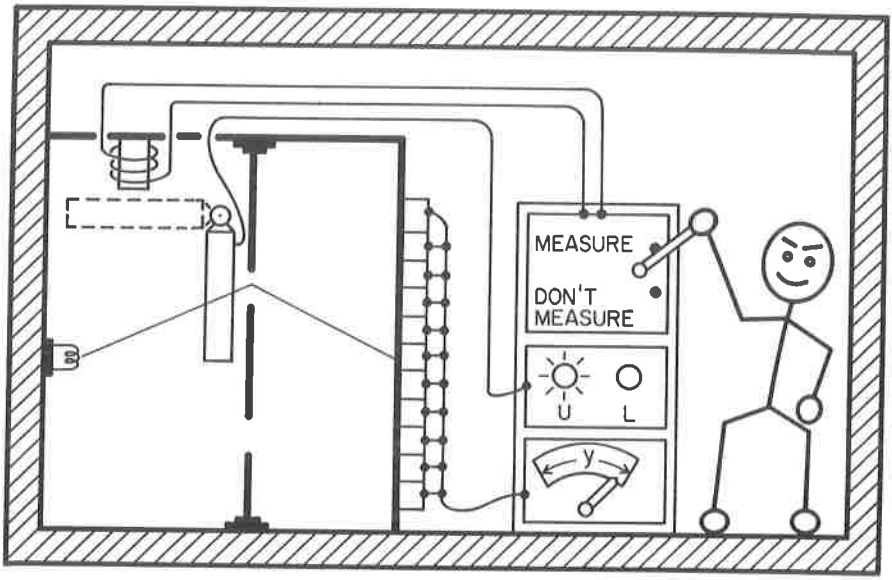

Fig. 2: A model closed quantum system containing an observer together with the necessary apparatus for carrying out a two-slit experiment. Alternatives for the system include whether the observer measured which slit the electron passed through or did not, whether the electron passed through the upper or lower slit, the alternative positions of arrival of the electron at the screen, the alternative arrival positions registered by the apparatus, the registration of these in the brain of the observer, etc., etc., etc. Each exhaustive set of exclusive alternatives is represented by an exhaustive set of orthogonal projection operators on the Hilbert space of the closed system. Time sequences of such sets of alternatives describe sets of alternative coarse-grained histories of the closed system. Quantum theory assigns probabilities to the individual alternative histories in such a set when there is negligible quantum mechanical interference between them, that is, when the set of histories decoheres.

A more refined model might consider a quantity of matter in a closed box. One could then consider alternatives such as whether the box contains a two-slit experiment or does not as well as alternative positions of atoms.

Many other sets of alternative histories are possible for the closed system. For example, we could have included alternatives describing the readouts of the apparatus that detects the position that the electron arrived on the screen. If

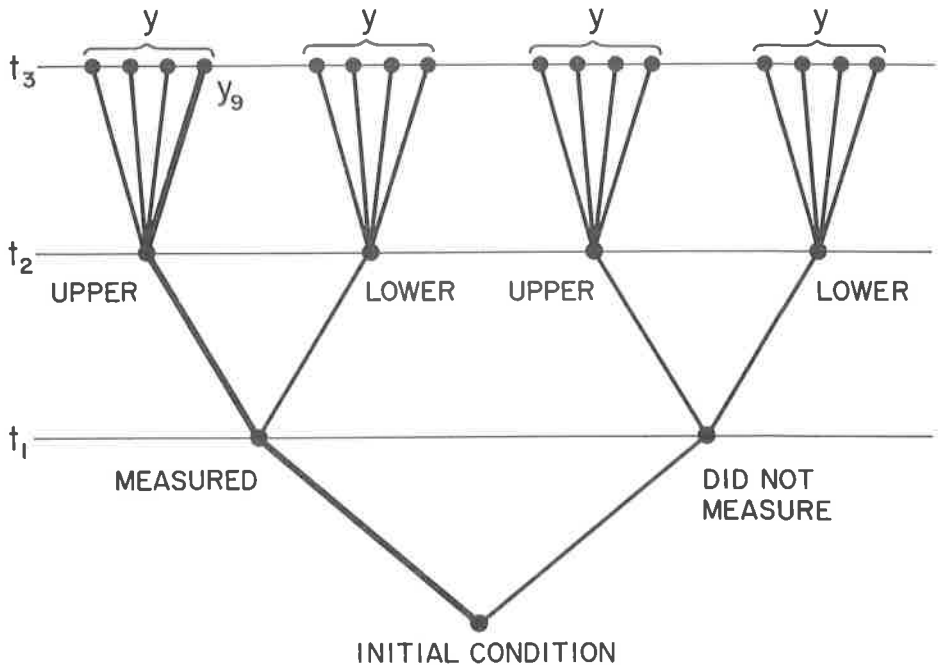

Fig. 3: Branching structure of a set of alternative histories. This figure illustrates the set of alternative histories defined by the alternatives of whether the observer decided to measure or did not decide to measure which slit the electron went through at time t_1, whether the electron went through the upper slit or through the lower slit at time t_2, and the alternative positions of arrival at the screen at time t_3. A single branch corresponding to the alternatives that the measurement was carried out, the electron went through the upper slit, and arrived at point y_9 on the screen is illustrated by the heavy line.

The illustrated set of histories does *not* decohere because there is significant quantum mechanical interference between the branch where no measurement was carried out and the electron went through the upper slit and the similar branch where it went through the lower slit. A related set of histories that does decohere can be obtained by replacing the alternatives at time t_2 by the following set of three alternatives: (a record of the decision shows a measurement was initiated and the electron went through the upper slit); (a record of the decision shows a measurement was initiated and the electron went through the lower slit); (a record of the decision shows that the measurement was not initiated). The vanishing of the interference between the alternative values of the record and the alternative configurations of apparatus ensures the decoherence of this set of alternative histories.

the initial condition corresponded to a good experiment there should be a high correlation between these alternatives and the position that the electron arrives at the screen. In a more refined model we could discuss alternatives corresponding to thoughts in the observer's brain, or to the individual positions of the atoms in the apparatus, or to the possibilities that these atoms reassemble in some completely different configuration. There are a vast number of possibilities.

Characteristically the alternatives that are of use to us as observers are very coarse grained, distinguishing only very few of the degrees of freedom of a large closed system. This is especially true if we recall that our box with observer and two-slit experiment is only an idealized model. The most general closed system is the universe itself, and, as I hope to show, the only realistic closed systems are of cosmological dimensions. Certainly, we utilize only very, very coarse-grained descriptions of the universe as a whole.

I would now like to state the rules that determine which coarse-grained sets of histories may be assigned probabilities and what those probabilities are. The essence of the rules I shall describe can be found in the work of Bob Griffiths (Griffiths, 1984). The general framework was extended by Roland Omnès (Omnès, 1992) and was independently, but later, arrived at by Murray Gell-Mann and myself (Gell-Mann and Hartle, 1990a). The idea is simple: The failure of probability sum rules due to quantum interference is the obstacle to assigning probabilities. Probabilities can be assigned to just those sets of alternative histories of a closed system for which there is negligible interference between the individual histories in the set as a consequence of the *particular* initial state the closed system has, and for which, therefore, all probability sum rules *are* satisfied. Let us now give this idea a precise expression.

Sets of alternatives at one moment of time are represented by sets of orthogonal projection operators. Employing the Heisenberg picture these can be denoted $\{P^k_{\alpha_k}(t_k)\}$. The superscript k denotes the set of alternatives being considered at time t_k (for example, the set of alternative position intervals $\{y_1, \cdots, y_N\}$ at which the electron might arrive at the screen at time t_3), α_k denotes the particular alternative in the set (for example y_9) and t_k is the time. The set of P's satisfy

$$\sum\nolimits_{\alpha_k} P^k_{\alpha_k}(t_k) = 1 \ , \quad P^k_{\alpha_k}(t_k) P^k_{\alpha'_k}(t_k) = \delta_{\alpha_k \alpha'_k} P^k_{\alpha_k}(t_k) \tag{15}$$

showing that they represent an exhaustive set of exclusive alternatives.

Sets of alternative histories are defined by giving sequences of sets of alternatives at definite moments of time, *e.g.*, $\{P^1_{\alpha_1}(t_1)\}$, $\{P^2_{\alpha_2}(t_2)\}, \cdots, \{P^n_{\alpha_n}(t_n)\}$. Different choices for $\{P^1_{\alpha_1}(t_1)\}$, $\{P^2_{\alpha_2}(t_2)\}$, etc. describe different sets of alternative histories of the closed system. An individual history in a given set corresponds to a particular sequence $(\alpha_1, \cdots, \alpha_n) \equiv \alpha$ and, for each history, there is a corresponding chain of projection operators

$$C_\alpha \equiv P^n_{\alpha_n}(t_n) \cdots P^1_{\alpha_1}(t_1) \ . \tag{16}$$

For example, in the two slit experiment in a box illustrated in Figure 2, the history in which the observer decided at time t_1 to measure which slit the electron goes through, in which the electron goes through the upper slit at time t_2, and arrives at the screen in position interval y_9 at time t_3, would be represented by the chain

$$P^3_{y_9}(t_3) P^2_U(t_2) P^1_{\text{meas}}(t_1) \tag{17}$$

in an obvious notation. The only difference between this situation and that of the "Copenhagen" quantum mechanics of measured subsystems is the following: The sets of operators $\{P^k_{\alpha_k}(t_k)\}$ defining alternatives for the closed system act on the Hilbert space of the closed system that includes the variables describing any apparatus, observers, and anything else. The operators defining alternatives in Copenhagen quantum mechanics act only on the Hilbert space of the measured subsystem.

When the initial state is pure, it can be resolved into *branches* corresponding to the individual members of any set of alternative histories. The generalization to an impure initial density matrix is not difficult (Gell-Mann and Hartle, 1990a), but for simplicity we shall assume a pure initial state throughout this article. Denote the initial state by $|\Psi\rangle$ in the Heisenberg picture. Then

$$|\Psi\rangle = \sum\nolimits_\alpha C_\alpha|\Psi\rangle = \sum_{\alpha_1,\cdots,\alpha_n} P^n_{\alpha_n}(t_n)\cdots P^1_{\alpha_1}(t_1)|\Psi\rangle \ . \tag{18}$$

This identity follows by applying the first of (15) to all the sums over α_k in turn. The vector

$$C_\alpha|\Psi\rangle \tag{19}$$

is the branch corresponding to the individual history α and (18) is the resolution of the initial state into branches.

When the branches corresponding to a set of alternative histories are sufficiently orthogonal the set of histories is said to *decohere*. More precisely a set of histories decoheres when

$$\langle\Psi|C^\dagger_{\alpha'}C_\alpha|\Psi\rangle \approx 0 \ , \quad \text{for any} \quad \alpha'_k \neq \alpha_k \ . \tag{20}$$

We shall return to the standard with which decoherence should be enforced, but first let us examine its meaning and consequences.

Decoherence means the absence of quantum mechanical interference between the individual histories of a coarse-grained set.* Probabilities can be assigned to

* The term "decoherence" is used in several different ways in the literature. Therefore, for those familiar with other work, a comment is in order to specify how we are employing the term in this simplified presentation. We have followed our previous work (Gell-Mann and Hartle, 1990a and 1990b) in using the term "decoherence" to refer to a property of a set of alternative time *histories* of a closed system. A decoherent set of histories is one for which the quantum mechanical interference between individual histories is small enough to guarantee an appropriate set of probability sum rules. Different notions of decoherence can be defined by utilizing different measures of interference. The weakest notion is just the consistency of the probability sum rules that was called "consistency" by Griffiths (1984) and Omnès (1992) and that term is used by some to refer to all measures of interference. Vanishing of the real part of (20) is a sufficient condition for the consistency of the probability sum rules called the "weak decoherence condition". We are using the stronger condition (20) because it characterizes widespread and typical mechanisms of decoherence. Eq (20) has been called the "medium decoherence condition". "Decoherence" in the context of this paper, thus, means the medium decoherence

the individual histories in a decoherent set of alternative histories because decoherence implies the probability sum rules necessary for a consistent assignment. The probability of an individual history α is

$$p(\alpha) = \|C_\alpha|\Psi\rangle\|^2 \ . \tag{21}$$

To see how decoherence implies the probability sum rules, let us consider an example in which there are just three sets of alternatives at times t_1, t_2, and t_3. A typical sum rule might be

$$\sum_{\alpha_2} p(\alpha_3, \alpha_2, \alpha_1) = p(\alpha_3, \alpha_1) \ . \tag{22}$$

We show (20) and (21) imply (22). To do that write out the left hand side of (22) using (21) and suppress the time labels for compactness.

$$\sum_{\alpha_2} p(\alpha_3, \alpha_2, \alpha_1) = \sum_{\alpha_2} \langle \Psi | P^1_{\alpha_1} P^2_{\alpha_2} P^3_{\alpha_3} P^3_{\alpha_3} P^2_{\alpha_2} P^1_{\alpha_1} | \Psi \rangle \ . \tag{23}$$

Decoherence means that the sum on the right hand side of (23) can be written with negligible error as

$$\sum_{\alpha_2} p(\alpha_3, \alpha_2, \alpha_1) \approx \sum_{\alpha'_2 \alpha_2} \left\langle \Psi | P^1_{\alpha_1} P^2_{\alpha'_2} P^3_{\alpha_3} P^3_{\alpha_3} P^2_{\alpha_2} P^1_{\alpha_1} | \Psi \right\rangle \ . \tag{24}$$

the extra terms in the sum being vanishingly small. But now, applying the first of (15) we see

$$\sum_{\alpha_2} p(\alpha_3, \alpha_2, \alpha_1) \approx \langle \Psi | P^1_{\alpha_1} P^3_{\alpha_3} P^3_{\alpha_3} P^1_{\alpha_1} | \Psi \rangle = p(\alpha_3, \alpha_1) \tag{25}$$

so that the sum rule (22) is satisfied.

Given an initial state $|\Psi\rangle$ and a Hamiltonian H, one could, in principle, identify all possible sets of decohering histories. Among these will be the exactly decohering sets where the orthogonality of the branches is exact. Indeed, trivial examples can be supplied by resolving $|\Psi\rangle$ into a sum of orthogonal vectors at time t_1, resolving those vectors into sums of further vectors such that the whole set is orthogonal at time t_2, and so on. However, such sets of exactly decohering histories will not, in general, have a simple description in terms of fundamental fields nor any connection, for example, with the quasiclassical domain of familiar experience. For this reason sets of histories that approximately decohere are of interest. As we will argue in the next two Sections, realistic mechanisms lead to the decoherence of

---

of sets of histories.

In the literature the term "decoherence" has also been used to refer to the decay in time of the off-diagonal elements of a reduced density matrix defined by tracing the full density matrix over a given set of variables (Zurek, 1991). The two notions of "decoherence of reduced density matrices" and "decoherence of histories" are not generally equivalent but also not unconnected in the sense that in particular models certain physical processes can ensure both. (See, e.g. the remarks in Section II.6.4 of (Hartle, 1991a)).

histories constituting a quasiclassical domain to an excellent approximation. When the decoherence condition (20) is approximately enforced, the probability sum rules such as (22) will only be approximately obeyed. However, as discussed earlier, these probabilities for single systems are meaningful up to the standard they are used. Approximate probabilities for which the sum rules are satisfied to a comparable standard may therefore also be employed in the process of prediction. When we speak of approximate decoherence and approximate probabilities we mean decoherence achieved and probability sum rules satisfied beyond any standard that might be conceivably contemplated for the accuracy of prediction and the comparison of theory with experiment.

Decoherent sets of histories of the universe are what we may utilize in the process of prediction in quantum mechanics, for they may be assigned probabilities. Decoherence thus generalizes and replaces the notion of "measurement", which served this role in the Copenhagen interpretations. Decoherence is a more precise, more objective, more observer-independent idea and gives a definite meaning to Everett's branches. For example, if their associated histories decohere, we may assign probabilities to various values of reasonable scale density fluctuations in the early universe whether or not anything like a "measurement" was carried out on them and certainly whether or not there was an "observer" to do it.

V. The Origins of Decoherence in Our Universe

What are the features of coarse-grained sets of histories that decohere in our universe? In seeking to answer this question it is important to keep in mind the basic aspects of the theoretical framework on which decoherence depends. Decoherence of a set of alternative histories is not a property of their operators *alone*. It depends on the relations of those operators to the initial state $|\Psi\rangle$, the Hamiltonian H, and the fundamental fields. Given these, we could, in principle, *compute* which sets of alternative histories decohere.

We are not likely to carry out a computation of all decohering sets of alternative histories for the universe, described in terms of the fundamental fields, anytime in the near future, if ever. It is therefore important to investigate specific mechanisms by which decoherence occurs. Let us begin with a very simple model due, in its essential features to Joos and Zeh (1985). We consider the two-slit example again, but this time suppose that in the neighborhood of the slits there is a gas of photons or other light particles colliding with the electrons (Figure 4). Physically it is easy to see what happens, the random uncorrelated collisions can carry away delicate phase correlations between the beams even if the trajectories of the electrons are not affected much. The interference pattern will then be destroyed and it will be possible to assign probabilities to whether the electron went through the upper slit or the lower slit.

Let us see how this picture in words is given precise meaning in mathematics. Initially, suppose the state of the entire system is a state of the electron $|\psi>$ and N distinguishable "photons" in states $|\varphi_1\rangle$, $|\varphi_2\rangle$, etc., *viz.*

$$|\Psi\rangle = |\psi\rangle|\varphi_1\rangle|\varphi_2> \cdots |\varphi_N\rangle \ . \tag{26}$$

Suppose further that $|\psi\rangle$ is a coherent superposition of a state in which the electron passes through the upper slit $|U\rangle$ and the lower slit $|L\rangle$. Explicitly:

$$|\psi\rangle = \alpha|U\rangle + \beta|L\rangle \ . \tag{27}$$

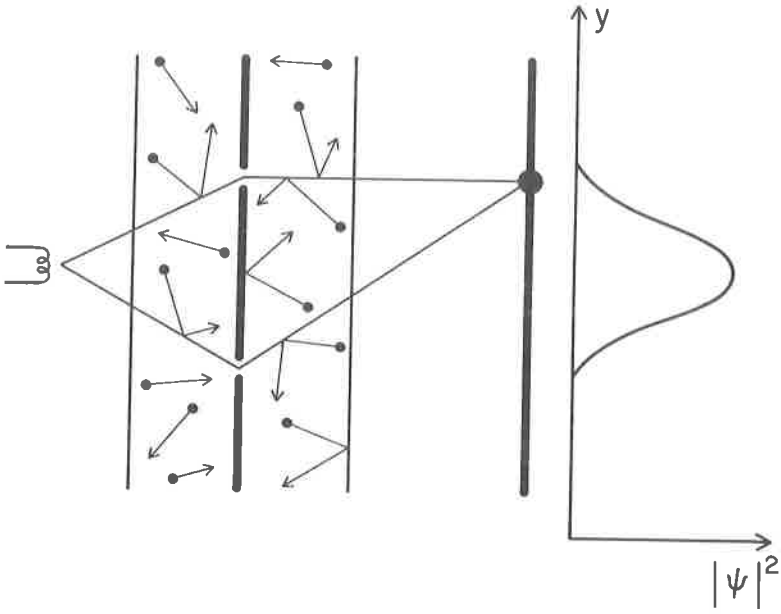

Fig. 4: The two slit experiment with an interacting gas. Near the slits light particles of a gas collide with the electrons. Even if the collisions do not affect the trajectories of the electrons very much they can still carry away the phase correlations between the histories in which the electron arrived at point y on the screen by passing through the upper slit and that in which it arrived at the same point by passing through the lower slit. A coarse graining that described only of these two alternative histories of the electron would approximately decohere as a consequence of the interactions with the gas given adequate density, cross-section, etc. Interference is destroyed and probabilities can be assigned to these alternative histories of the electron in a way that they could not be if the gas were not present (cf. Fig. 1). The lost phase information is still available in correlations between states of the gas and states of the electron. The alternative histories of the electron would not decohere in a coarse graining that included both the histories of the electron *and* operators that were sensitive to the correlations between the electrons and the gas.

This model illustrates a widely occuring mechanism by which certain types of coarse-grained sets of alternative histories decohere in the universe.

Both states are wave packets in x, so that position in x recapitulates history in time. We now ask whether the history where the electron passes through the upper slit and arrives at a detector at point y on the screen, decoheres from that in which it passes through the lower slit and arrives at point y as a consequence of the initial condition of this "universe". That is, as in Section 4, we ask whether the two

branches

$$P_y(t_2)P_U(t_1)|\Psi\rangle \quad , \quad P_y(t_2)P_L(t_1)|\Psi\rangle \tag{28}$$

are nearly orthogonal, the times of the projections being those for the nearly classical motion in x. We work this out in the Schrödinger picture where the initial state evolves, and the projections on the electron's position are applied to it at the appropriate times.

Collisions occur, but the states $|U\rangle$ and $|L\rangle$ are left more or less undisturbed. The states of the "photons", of course, are significantly affected. If the photons are dilute enough to be scattered once by the electron in its time to traverse the gas the two branches (28) will be approximately

$$\alpha P_y|U\rangle S_U|\varphi_1\rangle S_U|\varphi_2\rangle \cdots S_U|\varphi_N\rangle \ , \tag{29a}$$

and

$$\beta \ P_y|L\rangle S_L|\varphi_1\rangle S_L|\varphi_2\rangle \cdots S_L|\varphi_N\rangle \ . \tag{29b}$$

Here, S_U and S_L are the scattering matrices from an electron in the vicinity of the upper slit and the lower slit respectively. The two branches in (29) decohere because the states of the "photons" are nearly orthogonal. The overlap of the branches is proportional to

$$\langle\varphi_1|S_U^\dagger S_L|\varphi_1\rangle\langle\varphi_2|S_U^\dagger S_L|\varphi_2\rangle \cdots \langle\varphi_N|S_U^\dagger S_L \ |\varphi_N\rangle \ . \tag{30}$$

Now, the S-matrices for scattering off the upper position or the lower position can be connected to that of an electron at the orgin by a translation

$$S_U = \exp(-i\mathbf{k}\cdot\mathbf{x}_U)S \ \exp(+i\mathbf{k}\cdot\mathbf{x}_U) \ , \tag{31a}$$

$$S_L = \exp(-i\mathbf{k}\cdot\mathbf{x}_L)S \ \exp(+i\mathbf{k}\cdot\mathbf{x}_L) \ . \tag{31b}$$

Here, $\hbar\mathbf{k}$ is the momentum of a photon, $\mathbf{x}_U$ and $\mathbf{x}_L$ are the positions of the slits and S is the scattering matrix from an electron at the origin.

$$\langle\mathbf{k}'|S|\mathbf{k}\rangle = \delta^{(3)}(\mathbf{k}-\mathbf{k}') + \frac{i}{2\pi\omega_\mathbf{k}} f(\mathbf{k},\mathbf{k}')\delta(\omega_k - \omega_k') \ , \tag{32}$$

where f is the scattering amplitude and $\omega_k = |\vec{k}|$.

Consider the case where initially all the photons are in plane wave states in an interaction volume V, all having the same energy $\hbar\omega$, but with random orientations for their momenta. Suppose further that the energy is low so that the electron is not much disturbed by a scattering and low enough so the wavelength is much longer than the separation between the slits, $k|\mathbf{x}_U - \mathbf{x}_L| << 1$. It is then possible to work out the overlap. The answer according to Joos and Zeh (1985) is

$$\left(1 - \frac{(k|\mathbf{x}_U - \mathbf{x}_L|)^2}{8\pi^2 V^{2/3}}\sigma\right)^N \tag{33}$$

where σ is the effective scattering cross section and the individual terms have been averaged over incoming directions. Even if σ is small, as N becomes large this tends to zero. In this way decoherence becomes a quantitative phenomenon.

What such models convincingly show is that decoherence is frequent and widespread in the universe for histories of certain kinds of variables. Joos and Zeh calculate that a superposition of two positions of a grain of dust, 1mm apart, is decohered simply by the scattering of the cosmic background radiation on the timescale of a nanosecond. The existence of such mechanisms means that the only realistic isolated systems are of cosmological dimensions. So widespread is this kind of phenomena with the initial condition and dynamics of our universe, that we may meaningfully speak of habitually decohering variables such as the center of mass positions of massive bodies.

VI. The Copenhagen Approximation

What is the relation of the familiar Copenhagen quantum mechanics described in Section III to the more general "post-Everett" quantum mechanics of closed systems described in Sections IV and V? Copenhagen quantum mechanics predicts the probabilities of the histories of measured subsystems. Measurement situations may be described in a closed system that contains both measured subsystem and measuring apparatus. In a typical measurement situation the values of a variable not normally decohering become correlated with alternatives of the apparatus that decohere because of *its* interactions with the rest of the closed system. The correlation means that the measured alternatives decohere because the alternatives of the apparatus with which they are correlated decohere.

The recovery of the Copenhagen rule for when probabilities may be assigned is immediate. Measured quantities are correlated with decohering histories. Decohering histories can be assigned probabilities. Thus in the two-slit experiment (Figure 1), when the electron interacts with an apparatus that determines which slit it passed through, it is the decoherence of the alternative configurations of the apparatus that enables probabilities to be assigned for the electron.

There is nothing incorrect about Copenhagen quantum mechanics. Neither is it, in any sense, opposed to the post-Everett formulation of the quantum mechanics of closed systems. It is an *approximation* to the more general framework appropriate in the special cases of measurement situations and when the decoherence of alternative configurations of the apparatus may be idealized as exact and instantaneous. However, while measurement situations imply decoherence, they are only special cases of decohering histories. Probabilities may be assigned to alternative positions of the moon and to alternative values of density fluctuations near the big bang in a universe in which these alternatives decohere, whether or not they were participants in a measurement situation and certainly whether or not there was an observer registering their values.

VII. Quasiclassical Domains

As observers of the universe, we deal with coarse-grained histories that reflect our own limited sensory perceptions, extended by instruments, communication and records but in the end characterized by a large amount of ignorance. Yet, we have the impression that the universe exhibits a much finer-grained set of histories, independent of us, defining an always decohering "quasiclassical domain", to which our senses are adapted, but deal with only a small part of it. If we are preparing for a journey into a yet unseen part of the universe, we do not believe that we

need to equip ourselves with spacesuits having detectors sensitive, say, to coherent superpositions of position or other unfamiliar quantum variables. We expect that the familiar quasiclassical variables will decohere and be approximately correlated in time by classical deterministic laws in any new part of the universe we may visit just as they are here and now.

Since the post-Everett quantum mechanics of closed systems does not posit a quasiclassical domain, it must provide an explanation of this manifest fact of everyday experience. No such explanation can be provided from the dynamics of quantum theory alone. Rather, like decoherence, the existence of a quasiclassical domain in the universe must be a consequence of both initial condition of the universe and the Hamiltonian describing evolution.

Roughly speaking, a quasiclassical domain should be a set of alternative histories that decoheres according to a realistic principle of decoherence, that is maximally refined consistent with that notion of decoherence, and whose individual histories are described largely by alternative values of a limited set of quasiclassical variables at different moments of time that exhibit as much as possible patterns of classical correlation in time. To make the question of the existence of one or more quasiclassical domains into a *calculable* question in quantum cosmology we need measures of how close a set of histories comes to constituting a "quasiclassical domain". A quasiclassical domain cannot be a *completely* fine-grained description for then it would not decohere. It cannot consist *entirely* of a few "quasiclassical variables" repeated over and over because sometimes we may measure something highly quantum mechanical. Quasiclassical variables cannot be *always* correlated in time by classical laws because sometimes quantum mechanical phenomena cause deviations from classical physics. We need measures for maximality and classicality (Gell-Mann and Hartle, 1990a).

It is possible to give crude arguments for the type of habitually decohering operators we expect to occur over and over again in a set of histories defining a quasiclassical domain (Gell-Mann and Hartle, 1990a). Such habitually decohering operators are called "quasiclassical operators". In the earliest instants of the universe the operators defining spacetime on scales well above the Planck scale emerge from the quantum fog as quasiclassical. Any theory of the initial condition that does not imply this is simply inconsistent with observation in a manifest way. A background spacetime is thus defined and conservation laws arising from its symmetries have meaning. Then, where there are suitable conditions of low temperature, density, etc., various sorts of hydrodynamic variables may emerge as quasiclassical operators. These are integrals over suitably small volumes of densities of conserved or nearly conserved quantities. Examples are densities of energy, momentum, baryon number, and, in later epochs, nuclei, and even chemical species. The sizes of the volumes are limited above by maximality and are limited below by classicality because they require sufficient "inertia" resulting from their approximate conservation to enable them to resist deviations from predictability caused by their interactions with one another, by quantum spreading, and by the quantum and statistical fluctuations resulting from interactions with the rest of the universe that accomplish decoherence (Gell-Mann and Hartle, 1992). Suitable integrals of densities of approximately conserved quantities are thus candidates for habitually decohering quasiclassical operators. These "hydrodynamic variables" *are* among the principal variables of classical physics.

It would be in such ways that the classical domain of familiar experience could be an emergent property of the fundamental description of the universe,

not generally in quantum mechanics, but as a consequence of our specific initial condition and the Hamiltonian describing evolution. Whether the universe exhibits a quasiclassical domain, and, indeed, whether it exhibits more than one essentially inequivalent domain, thus become calculable questions in the quantum mechanics of closed systems.

VIII. Conclusion

Quantum mechanics is best and most fundamentally understood in the context of quantum mechanics of closed systems, most generally the universe as a whole. The founders of quantum mechanics were right in pointing out that something external to the framework of wave function and the Schrödinger equation *is* needed to interpret the theory. But it is not a postulated classical domain to which quantum mechanics does not apply. Rather it is the initial condition of the universe that, together with the action function of the elementary particles and the throws of the quantum dice since the beginning, is the likely origin of quasiclassical domain(s) within quantum theory itself.

Acknowledgements

This article is a small modification of one originally prepared for the *festshrift* for C.W. Misner, ed. by B.L. Hu, M.P. Ryan, and C.V. Vishveshwara, to be published by Cambridge University Press. The formulation of quantum mechanics described in this paper is a result of joint work with Murray Gell-Mann. It is a pleasure to thank him for the many conversations over the years and for permission to summarize aspects of our work here. Preparation of the report was supported in part by the National Science Foundation under grant PHY90-08502.

REFERENCES

In the spirit of providing a simplified introduction to the quantum mechanics of closed systems rather than a comprehensive review, no attempt has been made to provide anything more than the references that are directly relevant to the points raised in the text. These are not always the earliest nor are they the latest. More extensive, but still incomplete, lists can be found in Gell-Mann and Hartle (1990a) and Hartle (1991).

Bohr, N. (1958), *Atomic Physics and Human Knowledge*, Science Editions, New York.

Coleman, S. (1988), *Nucl. Phys.* **B310**, 643.

DeWitt, B. and Graham, R.N. eds. (1973), *The Many Worlds Interpretation of Quantum Mechanics*, Princeton University Press, Princeton.

Everett, H. (1957), *Rev. Mod. Phys.* **29**, 454.

Gell-Mann, M. and Hartle, J.B. (1990), in *Complexity, Entropy, and the Physics of Information, SFI Studies in the Sciences of Complexity*, Vol.

VIII, ed. by W. Zurek, Addison Wesley, Reading or in *Proceedings of the 3rd International Symposium on the Foundations of Quantum Mechanics in the Light of New Technology* ed. by S. Kobayashi, H. Ezawa, Y. Murayama, and S. Nomura, Physical Society of Japan, Tokyo.

_____ (1990), in the *Proceedings of the 25th International Conference on High Energy Physics, Singapore, August, 2-8, 1990*, ed. by K.K. Phua and Y. Yamaguchi (South East Asia Theoretical Physics Association and Physical Society of Japan) distributed by World Scientific, Singapore.

_____ (1992), Classical Equations for Quantum Systems. Preprint UCSBTH-91-15.

Giddings, S. and Strominger, A. (1988), *Nucl. Phys.* **B307**, 854.

Griffiths, R. (1984), *J. Stat. Phys.* **36**, 219.

Groenewold, H.J. (1952), *Proc. Akad. van Wetenschappen*, Amsterdam, Ser. B, **55**, 219.

Halliwell, J. (1991), in *Quantum Cosmology and Baby Universes: Proceedings of the 1989 Jerusalem Winter School for Theoretical Physics*, eds. S. Coleman, J.B. Hartle, T. Piran, and S. Weinberg, World Scientific, Singapore.

Hartle, J.B. (1991), *The Quantum Mechanics of Cosmology*, in *Quantum Cosmology and Baby Universes: Proceedings of the 1989 Jerusalem Winter School for Theoretical Physics*, eds. S. Coleman, J.B. Hartle, T. Piran, and S. Weinberg, World Scientific, Singapore.

Hawking, S.W. (1983), *Phys. Lett.* **B196**, 337.

Joos, E. and Zeh,H.D. (1985), *Zeit. Phys.* **B59**, 223.

Landau, L. and Lifshitz, E. (1958), *Quantum Mechanics*, Pergamon, London.

London, F. and Bauer, E. (1939), *La théorie de l'observation en mécanique quantique*, Hermann, Paris.

Omnès, R. (1992), *Rev. Mod. Phys.* **64**, 339.

Wigner, F. (1963), *Am. J. Phys.* **31**, 6.

Zeh, H.D. (1971), *Found. Phys.* **1**, 69.

Zurek, W. (1981), *Phys. Rev. D* **24**, 1516.

_____ (1982), *Phys. Rev. D* **26**, 1862.

_____ (1991), *Physics Today* **44**, 36.

Paper presented at the 13th Int. Conf. on General Relativity and Gravitation
Cordoba, Argentina, 1992: Part I, Plenary Lectures

On the mathematical theory of classical fields and general relativity

Sergiu Klainerman*

From the perspective of an analyst, like myself, the General Theory of Relativity provides an extrordinary rich and vastly virgin territory. It is the aim of my lecture to provide, first, an account of those aspects of the theory which attract me most and second a perspective of what has been accomplished so far in that respect. In trying to state our main objectives it helps to view General Relativity in the broader context of Classical Field Theory. As we know today, among all classical field theories, only the Maxwell equations and the Einstein field equations are known to have direct physical content. Others, like the Yang-Mills equations, are believed to become relevant only after they suffer the painful and still mysterious, process of quantization. In view of this fact one can describe the basic goals of the mathematical theory of classical fields as follows:

(1) Investigate the mathematical consequences of the Einstein Field Equations in so far as they describe the physical world.

(2) Investigate the mathematical properties of the other nonlinear field theories

  (a) In view of the fact that the Einstein–field equations are exceedingly complicated some of the other field theories provide a simplified testing ground for new ideas. From this point of view the final goal remains the understanding of the Einstein field equations but we hope to get important insights from the simpler nonlinear field theories. To this one might add our belief that any new ideas which lead to significant progress made on problems of nonlinear classical fields will prove useful in other areas of Mathematical Physics where nonlinear P.D.E's, in particular hyperbolic, are at the heart of the subject.

  (b) Investigate the mathematical properties of nonlinear field theories in view of their possible relevance to quantum field theory.

This final goal is most vague. One may attempt to render it slightly more meaningful by pointing out the importance of uncovering the fundamental properties of classical mechanical systems as carried out by mathematicians such as Euler, Lagrange, Hamilton, Poincare and others, to the very formulation of Quantum Mechanics.

Since according to the above scheme, General Relativity and the Einstein field equations play the dominant role in the mathematical theory of classical fields, I will concentrate my attention to it and refer to other field theories from the perspective of 2a. The mathematical structure of the Einstein–Vacuum equations, or shortly E–V, is already sufficiently complicated, I will thus restrict my attention to them.

Clearly, the aim of the mathematical theory of General Relativity is to understand the main features concerning the behaviour of general solution to the Einstein Field equations. Since general solutions can be naturaly parametrized by initial data sets one can reformulate our goal as being that of studying the main features of the evolution of general initial data sets. Now this goal is too broad since without an appropriate asymptotic restriction the evolution of an arbitrary initial data set can be very wild. A reasonable physical restriction is to consider initial data sets which look flat outside a

* Supported by the N.S.F. grant DMS-9103613

sufficiently large compact set of $\mathcal{H}$. The evolution of such initial data sets correspond to "isolated physical systems." These systems are particularly important in G.R. since, it is only for such systems that we can define the physical notions of mass, linear and angular momentum.

We thus redefine our central theme as being the study of the main features regarding the evolution of general classes of asymptotically flat initial data sets. From the perspective of (2a) we can broaden this to include all classical field theories. In this respect I will restrict myself to a discussion of field theories in Minkowski space-time.

Having thus defined our object of study I can turn my attention to what is considered to be the most fundamental mathematical question concerning the differential equations of General Relativity and the other classical Field theories. *This is to understand when and how, solutions to a classical, nonlinear, field theory can break-down.*

The break-down phenomenon can occur despite the existence of the basic conservation laws, in particular the energy, or total mass, which is positive. To understand what this means consider the comparable situation in one dimension, namely systems of ordinary differential equations which arise as the Euler-Lagrange equations of a Lagrangean with positive energy. To be more precise consider the example of the differential equation,

$$-\ddot{x} + V'(x) = 0$$

subject to the initial conditions at $t = 0$,

$$x(0) = x_0, \;\; \dot{x}(0) = x_1 \; .$$

The total energy of the system is given by the expression $\frac{1}{2}|\dot{x}|^2 - V(x)$, where $\frac{1}{2}|\dot{x}|^2$ is the kinetic energy and $-V$ the potential energy. For any reasonable physical system $-V \geq 0$. Since the total energy is conserved we immediately conclude that, for any initial conditions, the corresponding solutions exist for all time. There can be no finite time blow-up of the solutions.

The situation is very different for classical field theories. Though we have a positive energy momentum tensor which leads, in Minkowski space-time, to well defined conserved quantities, we cannot infer in general that the solutions, starting with perfectly smooth inital conditions, remain so for all time. Take as an example the field theory closest to our one dimensional example, namely the scalar wave equation in Minkowski space-time $\mathbf{M}^{n+1}$,

$$\Box\phi + V'(\phi) = 0, \qquad (N.W.E)$$

subject to the initial conditions at $t = 0$,

$$\phi(0) = f_0, \;\; \partial_t\phi(0) = f_1 \; .$$

Assume that f_0, f_1 are as regular as we want, say $f_0, f_1 \in \mathcal{C}_0^\infty(\Re^n)$. Assume also that $-V \geq 0$ so that the energy-momentum tensor satisfies the required positivity condition.

The total energy [1] at time t has the form,

$$E(t) = \int_{\Re^n} \left(\frac{1}{2}((\partial_t\phi)^2 + (\partial_1\phi)^2 \cdots + (\partial_n\phi)^2) - V(\phi) \right) dx$$

[1] which is the conserved quantity corresponding to the Killing vectorfield $T_0 = \partial_t$

One can easily check directly that $\frac{d}{dt}E(t) = 0$ and hence, if the initial data at time $t = 0$ are such that $E(0)$ is bounded, we infer that $E(t)$ is bounded for all time. Yet, unlike in the previous example we cannot conclude that the solutions remain smooth at all later times. It is easy to see that as long as ϕ remains bounded its time evolution preserves the regularity of the initial conditions in the L^2 norm . The problem, thus, is to show that ϕ stays bounded. In space dimension $n = 1$ we can conclude, from the boundedness of $E(t)$ and a simple form of the Sobolev inequalities, that ϕ is point-wise bounded. Indeed, since $-V \geq 0$, we infer that at any time t, $\int_{\Re^n} \left(\frac{1}{2}((\partial_t\phi)^2 + (\partial_1\phi)^2 \cdots + (\partial_n\phi)^2)\right) dx \leq E(0)$. On the other hand, according to the simplest version of the basic Sobolev inequalities, the sup-norm of a function in $\mathbf{R}^n$ can be bounded in terms of the square integrals of the sum of all its derivatives of order[2] $[\frac{n}{2}]$. Hence for $n = 1$,

$$\sup_x |\phi(t,x)| \leq c\left(\int_{\Re^n} \frac{1}{2}((\partial_1\phi)^2 \cdots + (\partial_n\phi)^2)dx\right)^{1/2} \leq cE(0)^{1/2} .$$

The form of the Sobolev inequality we have used above fails just a little for $n = 2$. In fact we can only estimate $(\int |\phi(t,x)|^p)^{1/p}$ for any $p < \infty$, this turns out nevertheless to be enough to conclude that the solutions remain smooth for all time[3]. For $n \geq 3$ the above form of the Sobolev inequality require a little more than $3/2$ derivatives in L^2 in order to estimate the sup norm of ϕ. The boundedness of the energy provides us with a bound of only one derivative in L^2. We have thus a gap of more than $1/2$ derivatives. The situation gets, of course, even worse in higher dimensions. We can still, nevertheless, show that ϕ remains bounded provided that $V(\phi)$ does not grow too fast as $\phi \to \infty$. For simplicity consider the case of the power potential $V(\phi) = -\frac{1}{p+1}|\phi|^{p+1}$. Prescribing to the space–time variables $t = x^0, x^1, \ldots, x^n$ the same scale L, the solution ϕ of the equation, $\Box\phi + V'(\phi) = 0$, acquires the scale $L^{-\frac{2}{p-1}}$. Therefore the total energy E has the scale L^s where s is the exponent $s = n - 2 - \frac{4}{p-1}$.

The case when the exponent s is strictly negative is called "subcritical". It is quite easy to analyze and has lead to the well–known global regularity result of Jörgens [Jö]. The case $s = 0$ is called "critical " while $s > 0$ is called "supercritical." In the supercritical regime we have no results, even for spherically symmetric solutions, despite the relatively large attention this problem has received. The critical case has been recently settled by the combined efforts of Struwe [Stru], Grillakis [Gr1] and more recently Shatah-Struwe [Sh-Stru].

Theorem. *Consider the initial value problem $\Box\phi + V'(\phi) = 0$ with initial conditions $\phi(0,x) = f_0(x), \partial_t\phi(0,x) = f_1(x)$ which, for simplicity, we may assume in $\mathcal{C}_0^\infty$. Assume that $n \leq 7$ and that the exponent $s \leq 0$. Then the equations admits unique smooth solutions globally in $\mathbf{M}^{n+1}$.*

Despite the difficulty of the problem it is widely believed that, even in the supercritical case, the solutions to (N.W.E.) remain smooth for all time.

A more interesting field theory is provided by the equations of Wave Maps defined from Minkowski space-time $\mathbf{M}^{n+1}$ with values in a Riemannian manifold $\mathcal{N}$. Relative

[2] $[\frac{n}{2}]$ denotes the smallest integer strictly greater than $\frac{n}{2}$.

[3] provided that $V(\phi)$ has polynomial growth in ϕ for large ϕ.

to standard coordinates x^α, $\alpha = 0, \ldots, n$ and local coordinates y^a, $a = 1, \ldots m$ in $\mathcal{N}$ the equations take the form,

$$\Box\phi^a + \Gamma^a_{bc}(\phi)\mathrm{m}^{\mu\nu}\partial_\mu\phi^b\partial_\nu\phi^c = 0 \qquad (W.M.)$$

where Γ^a_{bc} are the Christoffel symbols of $\mathcal{N}$. Consider the initial conditions at $x^0 = t = 0$,

$$\phi(0) = f_0, \quad \partial_t\phi(0) = f_1$$

with f_0, f_1 compactly supported smooth maps defined from $\mathbf{R}^n\mathbf{R}^n$ to $\mathcal{N}$.

Since the nonlinear terms are quadratic in the first derivatives of the map ϕ, in order to preserve the L^2 regularity of the initial conditions, we now need to have pointwise bounds not only on ϕ but also its first derivatives. The total energy, in this case, has the form,

$$E(t) = \int_{\Re^n} \left(\frac{1}{2}(|\partial_t\phi|^2 + |\partial_1\phi|^2 \cdots + |\partial_n\phi|^2)\right) dx$$

with $|\partial_\alpha\phi|^2 = h_{ab}\partial_\alpha\phi^a\partial_\alpha\phi^b$, h the Riemannian metric of $\mathcal{N}$. The law of conservation of total energy is, as before,

$$E(t) = E(0)\ .$$

This provides us with only with an L^2 bound for the derivatives of ϕ. We therefore see that the conservation law for the total energy does not suffice to control the L^∞ norm of the first derivatives of ϕ even for wave maps in $\mathbf{M}^{1+1}$. A simple remark, however, allows us to bypass the difficulty in this case (see [Sh] and also [Gu]). Indeed consider the energy momentum tensor $\mathbf{T}$. It has the form, $\mathbf{T}_{\alpha\beta} = \frac{1}{2}\Big(\langle\phi_{,\alpha}, \phi_{,\beta}\rangle - \frac{1}{2}\mathbf{g}_{\alpha\beta}(\mathbf{g}^{\mu\nu}\langle\phi_{,\mu}\phi_{,\nu}\rangle)\Big)$ with $<,>$ the scalar product in $\mathcal{N}$. Since we are in $1+1$ dimensions $\mathbf{T}$ has only the components $\mathbf{T}_{00}, \mathbf{T}_{01} = \mathbf{T}_{10}, \mathbf{T}_{11}$. Moreover, since $\mathbf{T}$ is trace-less in $\mathbf{M}^{1+1}$, we have $\mathbf{T}_{00} = \mathbf{T}_{11}$. Now recall that $\mathbf{T}$ verifies the divergence equation $\partial^\beta\mathbf{T}_{\alpha\beta} = 0$. Hence,

$$\partial_t\mathbf{T}_{00} = \partial_x\mathbf{T}_{01}$$
$$\partial_t\mathbf{T}_{01} = \partial_x\mathbf{T}_{00}$$

and therefore $\mathbf{T}_{00}$ is a solution of the linear wave equation in $\mathbf{M}^{1+1}$, $\Box\mathbf{T}_{00} = 0$. One can now easily check that $\mathbf{T}$ remains bounded for any $t > 0$ provided that $\mathbf{T}_{00}, \partial_t\mathbf{T}_{00}$ are bounded at $t = 0$. Since $\mathbf{T}_{00} = \frac{1}{2}(|\partial_t\phi|^2 + |\partial_x\phi|^2)$ we conclude that all first derivatives of the map ϕ are bounded. Therefore, in $\mathbf{M}^{1+1}$, all wave maps, which are initially smooth remain so.

The proof we have presented is typical to the sweeping simplifications which occur only in $1+1$ dimensions. The case of wave maps defined in $\mathbf{M}^{1+2}$ is already much more complicated. We can proceed as before and classify the (W.M.) according to the scale associated to the total energy E. Thus, prescribing to the space-time variables the scale L and to ϕ the scale L^0 we find that E has the scale L^s with $s = n - 2$. Consequently the (W.M.) is subcritical in $\mathbf{M}^{1+1}$, critical in $\mathbf{M}^{1+2}$ and supercritical in $\mathbf{M}^{1+n}$, $n \geq 3$. Under reasonable geometric assumptions we expect global regularity in the critical case $\mathbf{M}^{1+2}$. This conjecture has been recently checked for wave maps satisfying additional

symmetry assumptions, see [Ch-Za] in the case of spherical symmetry and [Sh-Za], [Gr2] in the equivariant case.

Of course, we also know examples of partial differential equations, for which blow-up actually occurs. This is the case of the Burger equation

$$u_t + uu_x = 0.$$

Despite the fact that the equation has, not only one but infinitely many positive conserved quantities, e.g $\int |u(t,x)|^{2k} dx$, any initial condition $u_0(x) \in C_0^\infty(\Re)$ leads to blow-up in finite time[3] Remark that the Burger equation is supercritical relative to the total energy $E = (\int |u(t,x)|^2 dx)^{1/2}$ and critical relative to the L^∞ norm of u. Indeed if we prescribe to t, x the same scale L and to u the scale L^0 then the conserved quantities $(\int |u(t,x)|^{2k} dx)^{1/2k}$ acquire the scales $L^{\frac{1}{2k}}$

There are also known examples[4] of finite time break-down of solutions for wave-maps in the supercritical case $\mathbf{M}^{3+1}$. But more important, from a physical point of view, is the well known fact that the Einstein equations also leads to singularities. It is interesting to remark in this respect that, relative to the total ADM mass, the Einstein field equations are supercritical [5].

According to the classification we have indicated above it is widely expected that in subcritical situations break-down can be ruled out. We also believe that the same holds true in most critical problems. Finally, in the supercritical cases we have a lot of evidence that break-down can in fact occur.

Though we do not expect any bad behaviour in the subcritical problems, proving that it is indeed so is not necessarily an easy task. This is the case of the beautiful result of Eardley and Moncrief [E-M] for the Yang-Mills equations[6] in $\mathbf{M}^{3+1}$. The proof required an insightful observation concerning the structure of the nonlinear terms of the Yang-Mills equations expressed in the Cronstrom gauge. As we have indicated above the question of regularity in the critical case has been solved for (N.W.E.) and is now the focus of considerable attention for (W.M.).

It is helpful to divide the general question of break-down and regularity into a sequence of simpler ones for which is easier to envision ongoing progress. Once more I will refer directly to the Einstein field equations and look at the other field theories from the perspective of (2a). The simplest of them all is,

Question 1. *Under what assumptions on the initial conditions is the Cauchy problem locally well posed ?*

Technically this is the question of local in time existence and uniqueness of the development of an initial data set. Our present technology, based on energy estimates and Sobolev inequalities, requires too much differentiability on the initial data set.

---

[3] .

[4] see [Sh]

[5] Proceeding as before for (N.W.E.) and (W.M.) we assign to the space-time metric the scale L^0 and remark that the ADM mass $E = \frac{1}{16\pi} \lim_{r\to\infty} \int_{S_r} \sum_{i,j} \Big(\partial_i g_{ij} - \partial_j g_{ii}\Big) N^j da$ has the "supercritical" scale L^1.

[6] Proceeding as before we associate to the vector potential A the scale L^{-1} and to the electromagnetic field F the scale L^{-2}. Thus the total energy at time t has the scale L^{n-4} which is subcritical for $n = 3$.

Lowering these differentiability requirements is crucial in understanding the regularity properties of the solutions to (E-V) and also the general Einstein equations in the presence of matter.

Question 2. *Under what assumptions on the initial data sets do there exist global, smooth and geodesically complete solutions of the Einstein–Vacuum equations ?*

This question is intimately connected to that of stability of the Minkowski space–time. Namely the Minkowsky space–time $\mathbf{M}^{3+1}$ is a special solution of E–V free of singularities. As mentioned above an initial data set is said to be flat if its development is diffeomorphic to $\mathbf{M}^{3+1}$. It is thus natural to ask what happens to the developments of initial data sets which are small perturbations of a flat initial data set. The answer to this question is not only important in view of the general program of studying the regularity of solutions evolving from regular Cauchy data but also in regard to the question of the structure of null infinity, of general solutions which evolve from asymptotically flat initial data sets. Indeed any A.F. initial data set can be interpreted, outside a sufficiently large relatively compact set $\mathcal{K}$, as a small perturbation of a flat initial data set. Thus the methods used in the study of the global stability of the Minkowski space-time can also be used in the study of the asymptotic properties of the development of any A.F. initial data set outside the future set of a sufficiently large set $\mathcal{K} \subset \mathcal{H}$.

The problem of the stability of the Minkowski space-time has been recently addressed in my joint work with D. Christodoulou [Ch-Kl2]. The result which we were able to prove asserts the following,

Theorem(CH-Kl). *Any S.A.F.* [7] *initial data set which satisfies, in addition, a Global Smallness Assumption, leads to a unique, smooth and geodesically complete development, solution of the Einstein-Vacuum Equations. Moreover, this development is globally asymptotically flat, by which we mean that its Riemann curvature tensor approaches zero*[8] *on any causal or space-like geodesic, as the corresponding affine parameter tends to infinity.*

The global smallness assumption requires that an appropiate L^2-norm of up to 2 derivatives of the curvature tensor of g and 3 derivatives of k are small. A more primitive version of our result requires one derivative less, in line with the statement of the local existence theorem. Further improvements of our result will depend crucially on progress made on Question 1.

Question 3. *Under what conditions do initial data sets develop black holes and singularities*?

The famous incompleteness theorem of Penrose asserts that if an initial data set of a space–time verifying the Einstein field equations (with very general assumptions on the energy–momentum tensor T) has a trapped sphere[9] then some outgoing null geodesics normal to S must be future–incomplete. Though the Penrose theorem indicates that

[7] *of order $k = 4$. We note that the precise fall-off conditions of the initial data set are in fact given in L^2-weighted norms, and thus differ slightly from those we have discussed below*

[8] *Our result gives precise information on the rate of decay of different components of the curvature tensor.*

[9] i.e. a space–like sphere S on $\mathcal{H}$ with a compact filling such that the outgoing null normal to S are everywhere converging

singularities can indeed occur for general solutions of the Einstein equations it remains entirely unclear what is the nature of these singularities and, more important, if the trapped sphere hypothesis is of any relevance in the actual process of a gravitational collapse.

At the present time we have no results concerning the formation of black holes for the (E–V) equations. Since the general problem is too hard one needs to look first at simplified situations. Unfortunately the (E–V) does not allow interesting A.F. solutions with additional symmetries, e.g. spherically symmetric. To overcome this one has to consider the general Einstein field equations coupled with some matterfield. The simplest such field is the linear scalar wave equation (S.W) with $V = 0$. This coupling allows one to obtain nontrivial dynamics even in the spherically symmetric case. In this case the group $SO(3)$ acts as an isommetry group on the space-time $(\mathcal{M}, \mathbf{g})$. The group orbits are space-like metric 2-spheres S of Gauss curvature r^{-2} where r is the area radius of S, i.e. $A(S) = 4\pi r^2$. Due to the $SO(3)$ symmetry the field equations can be reduced to quotient Q of the space–time by the group. This a 2-dimensional manifold with boundary. The boundary corresponds to the set of fixed points of the group action and forms a time-like geodesic Γ. Choosing, on Q, a pair of conjugate optical functions u, v with u constant on any future directed null curve initiating at Γ, v constant along each of the conjugate familly of null curves and such that both functions are increasing towards the future, the induced metric on Q has the form $-\Omega^2 du dv$. Remark that u, v are determined up to general transformations of the form $u \mapsto f(u)$, $v \mapsto f(v)$ with f,g arbitrary increasing functions. The reduced field equations form a system on Q for the functions r, Ω, ϕ,

$$
\begin{aligned}
r\frac{\partial^2 r}{\partial u \partial v} &= -\frac{1}{4}\Omega^2 - \frac{\partial r}{\partial u}\frac{\partial r}{\partial v} \\
\frac{\partial^2 r}{\partial u^2} &= 2\Omega^{-1}\frac{\partial r}{\partial u}\frac{\partial \Omega}{\partial u} - r\Big(\frac{\partial \phi}{\partial u}\Big)^2 \\
\frac{\partial^2 r}{\partial v^2} &= 2\Omega^{-1}\frac{\partial r}{\partial v}\frac{\partial \Omega}{\partial v} - r\Big(\frac{\partial \phi}{\partial v}\Big)^2 \\
r\frac{\partial^2 \phi}{\partial u \partial v} &= -\frac{\partial r}{\partial u}\frac{\partial \phi}{\partial v} - \frac{\partial r}{\partial v}\frac{\partial \phi}{\partial u} \qquad (E - S.W.)
\end{aligned}
$$

The above system is invariant under the transformations $u \mapsto f(u)$, $v \mapsto g(v)$ and $\Omega \mapsto (f'g')^{1/2}\Omega$.

The program of studying the spherically symmetric solutions of the coupled Einstein-scalar wave equation was initiated and carried out with remarkable success by D. Christodoulou, see [Ch2],[Ch3]. A basic ingredient of his analysis relies on the monotonicity properties of the mass function m defined by the formula $1 - \frac{2m}{r} = -4\Omega^{-2}\frac{\partial r}{\partial u}\frac{\partial r}{\partial v}$. We have $\frac{\partial m}{\partial v} \geq 0$ and $\frac{\partial m}{\partial u} \leq 0, = 0, \geq 0$ according to whether $1 - \frac{2m}{r} \leq 0, = 0, \geq 0$. In [Ch2] Christodoulou combines this montonicity properties of m together with the powerful method of characteristics, typical to 2-dimensional hyperbolic problems, to prove a deep result concerning the formation of black holes for the system (E-S.W.). He considers the evolution of regular initial conditions defined on an outgoing null curve C_0^+, starting at a point on Γ, and shows that under reasonable assumptions of the data trapped spheres must form.

Question 4. *What are the global regularity features of arbitrary devlopments of A.F. initial data sets ?*

One of the main distant goals of the mathematical theory of General Relativity is to give a correct formulation and solve what is referred to as the "Cosmic Censorship" conjecture. Loosely speaking the conjecture asserts that there are no singularities outside black holes or, more picturesque, there exist no naked singularities.

Once again the understanding of the global regularity properties of the general Einstein equations is beyond our reach. To make progress we have to either restrict our attention to solutions with spherical symmetry or to consider simpler field theories. Concerning the first approach I have to mention once more the pioneering work of Christodoulou on the coupled E–S.W. equations. Following his result on the formation of black holes, which was mentioned above, he has concentrated his efforts on the regularity of solutions outside black holes. His results show that though "naked singularities" are possible they are unstable in the sense that the initial conditions from which these could evolve form an exceptional set of strictly positive codimension in the set of all allowable initial conditions.

In what follows I will give a short description on some of the progress made recently on the questions 1,3,4 and give a short description of the proof of Theorem (Ch-Kl) mentioned in connection with question 2.

Q1. As we have mentioned below the first solution to this problem was given by Y.C. Bruhat, see [Br1], under the assumption that the initial metric g has 3 derivatives, and the tensor k has 2 derivatives locally on $L^2(\mathcal{H})$. Her proof is based on two ingredients. The first consists on an auxilliary construction of wave coordinates relative to which the E–V equations take the form of a system of nonlinear wave equations. The special choice of wave coordinates is in fact not important, they can be replaced by other more geometric structures [10]. The second ingredient is more basic as it reflects on our present, limited, techniques in dealing with the local theory for all systems of nonlinear hyperbolic equations in more than 2 space-time dimensions[11]. It combines standard energy estimates for derivatives of solutions of the hyperbolic system together with Sobolev inequalities and an iteration procedure. The procedure has a built in limitation, as it requires too many derivatives on the initial data. For E–V it means that we have to restrict ourselves to initial data sets with one derivative of the curvature tensor and two derivatives of k in $L^2_{loc}(\mathcal{H})$. This limitation is very significant when one passes from the local to the global study of the regularity properties of the system. To overcome the limitation it seems imperative to develop new analytic techniques. It helps, of course, to first address this question for a much simpler field theory. A good example is provided by the equations of Wave Maps from the Minkowski space-time $\mathbf{M}^{n+1}$ to an arbitrary Riemannian manifold $(\mathcal{N}, h)$. Recall that, relative to

[10] In [Ch-Kl2] the local existence theorem is proved relative to a maximal foliation. In other words the space-time is constructed locally together with a "time function" t whose level hypersurfaces are maximal, i.e. $trk = 0$.

[11] In 2 space-time dimensions the situation is radically different because of the method of characteristics which allows better results based on the method of characteristics. This remains true in higher dimensions if one considers only solutions with spherical symetry.

a local chart in $\mathcal{N}$, the W.M. equations take the form,

$$\Box\phi^I + \sum_{J,K} \Gamma^I_{JK}(\phi) Q(\mathbf{D}\phi^J, \mathbf{D}\phi^K), \quad I = 1, \ldots, N \tag{1}$$

where,

$$Q(\mathbf{D}\phi, \mathbf{D}\psi) = \mathbf{D}_\alpha \phi \mathbf{D}^\alpha \psi = \mathbf{m}^{\alpha\beta} \mathbf{D}_\alpha \phi \mathbf{D}_\beta \psi. \tag{2}$$

Here $\phi = (\phi^I)_{I=1,\ldots,N}$ is a vector valued function defined on $\mathbf{M}^{3+1}$. The Γ^I_{JK}'s are given arbitrary smooth functions of ϕ.
We consider the initial value problem on the hyperplane $t = x^0 = 0$,

$$\phi(0, x) = f_0(x), \quad \partial_t(0, x) = f_1(x). \tag{3}$$

The classical local existence result for equations of type (1) requires $f_0 \in H_3(\mathbf{R}^3)$, $f_1 \in H_2(\mathbf{R}^3)$. Recently, in collaboration with M. Machedon [Kl-Ma] we were able to prove the following,

Theorem (Kl-Ma). *Consider systems of nonlinear wave equations in* $\mathbf{M}^{3+1}$ *of the type (1), (2) subject to the initial conditions (3) under the assumptions* $f_0 \in H_2(\mathbf{R}^3)$, $f_1 \in H_1(\mathbf{R}^3)$. *There exists a* $T_* > 0$ *and a unique solution* ϕ *defined in the slab* $\mathcal{D}_* = [0, T_*] \times \mathbf{R}^3$ *and verifying the following,*

(i) $\sup_{\mathcal{D}_*} |\phi(t, x)| < \infty$

(ii) $\int\int_{\mathcal{D}_*} (|Q(\phi, \phi)|^2 + |\mathbf{D}Q(\phi, \phi)|^2)\, dtdx < \infty$

(iii) $\sup_{[0,T_*]} \|\phi(t, \cdot)\|_{H^2(\mathbf{R}^3)} < \infty$.

As we have mentioned above the L^2 estimates, provided by the energy method, are not enough to prove a result like this. On the other hand any attempt to replace L^2 with any other L^p is known to fail (see [L]) even for the linear problem, in $\mathbf{M}^{n+1}$,

$$\Box\phi = F; \quad \phi(0, x) = f_0(x), \partial_t(0, x) = f_1(x). \tag{4}$$

The only other possibility to get additional information is to consider space-time integrals. The best known result of this type, for the wave equation (4), is due to Strichartz [Str1, 2] who has proved the following inequality,

$$\|\phi\|_{L^q} \leq C\Big(\|f_0\|_{H^{1/2}} + \|f_1\|_{H^{-1/2}} + \|F\|_{L^q}\Big) \tag{5}$$

where L^p, L^q are space-time norms with exponents, $p = 2\frac{n+1}{n+3}$, $q = 2\frac{n+1}{n-1}$.

The novelty of our result consists of estimating, in space–time, the Lorentz invariant bilinear form $Q(\phi, \psi)$ and derive for it an estimate which is stronger than that which could be derived from the Strichartz inequality. This allows us to gain a full derivative over the classical result. The basic new tool is contained in the following

Proposition. *Let* ϕ, ψ *be solutions in* $\mathbf{M}^{3+1}$ *to the equations* $\Box\phi = F$, $\Box\psi = G$ *with initial conditions* $\phi(0, x) = f_0(x)$, $\partial_t\phi(0, x) = f_1(x)$ *and* $\psi(0, x) = g_0(x)$, $\partial_t\psi(0, x) = g_1(x)$. *Assume that* $f_0, g_0 \in H_2(\mathbf{R}^3)$, $f_1, g_1 \in H_2(\mathbf{R}^3)$ *and that the integrals* $\int_0^T \|\mathbf{D}F(t, \cdot)\|_{L^2(\mathbf{R}^3)}\, dt$, $\int_0^T \|\mathbf{D}F(t, \cdot)\|_{L^2(\mathbf{R}^3)}\, dt$ *are both finite. Then the*

space-time integral $\int_0^T \int_{\mathbf{R}^3} |\mathbf{D}Q(\phi,\psi)^2\, dtdx$ *is also bounded, with a bound depending only on the product of* $\left(\|f_0\|_{H^2(\mathbf{R}^3)} + \|f_1\|_{H^1(\mathbf{R}^3)} + \int_0^T \|\mathbf{D}F(t,\cdot)\|_{L^2(\mathbf{R}^3)}\, dt\right)$ *with* $\left(\|g_0\|_{H^2(\mathbf{R}^3)} + \|g_1\|_{H^1(\mathbf{R}^3)} + \int_0^T \|\mathbf{D}G(t,\cdot)\|_{L^2(\mathbf{R}^3)}\, dt\right)$.

Remark 1. The result of the proposition remains true if we replace the quadratic form Q by $Q_{\alpha\beta}(\phi,\psi) = \partial_\alpha\phi\partial_\beta\psi - \partial_\alpha\psi\partial_\beta\phi$. On the other hand the result is wrong if we replace the bilinear form Q with any other bilinear form in the space–time gradients of ϕ, ψ. The result of Theorem (Kl-Ma) remains true if we also allow the forms $Q_{\alpha\beta}$ but we believe is wrong if one considers in (1) general expressions quadratic in the $\mathbf{D}\phi$. This suggests that any future improvements of the classical local existence and uniqueness results require special features of the equations one is interested in.

Remark 2. The proof of Theorem (Kl-Ma) depends heavily on the fact that the nonlinear terms in (1), (2) do not depend on the top derivatives. It is reasonable to believe however that the result of the theorem can be extended to certain classes of quasilinear equations. Most important we believe that a similar result holds true for the Einstein vacuum equations, in other words the (E–V) equations should be well posed for initial data sets $(\mathcal{H}, g, k)$ for which the components curvature tensor $R(g)$ and the first derivatives of k are bounded in L^2.

Q2. In proving the result on the stability of the Minkowski space-time mentioned above we had to overcome the following major obstacles,

a. *The problem of coordinates.*
b. *The strongly nonlinear character of the equations.*
c. *The long range character of the initial data set.*
d. *The nontrivial character of the asymptotic properties of the causal structure of any small perturbation of the Minkowski space-time.*
e. *The conjunction of all of the above.*

The scope of this lecture doesn't allow me to discuss in detail each of the above difficulties and describe our strategy for overcoming them. For those interested I would like to refer to the extensive introduction of [Ch-Kl2] or the recent Bourbaki seminar [Bour]. In what follows I will only give a very short description of the main ideas in [Ch-Kl2].

The difficulty (a) is typical to General Relativity. In short one is faced with the following dilemma. To write the equations in a meaningful way one seems forced to introduce an additional structure, e.g. wave coordinates. Such a structure seems necessary even to allow the formulation of well posed Cauchy problem and a proof of a local in time existence result. Nevertheless, as the particular case of wave coordinates illustrates, the structure, if not carefully chosen, may lead, in the large, to problems of its own making. Indeeed, as pointed by Y. Chocquet–Bruhat [Br2], the "wave coordinates" are unstable in the large even when one starts with initial conditions close to flat. In [Ch-Kl2] we solve this problem by constructing our space-times together with two additional geometric stuctures. One is given by a time function t whose level hypersurfaces are maximal. The second, much more important, is given by the level hypersurfaces of an optical function . This is a special solution u of the Eikonal equation $\mathbf{g}^{\alpha\beta}\partial_\alpha u\partial_\beta u = 0$, whose level hypersurfaces are outgoing null hypersurfaces with correct asymptotic properties at null future infinity.

The other major obstacle in the study of the Einstein equations consists in their hyperbolic and strongly nonlinear character. As we have already mentioned the only

apriori bounds we have available in the study of quasilinear hyperbolic equations, in the physical space-time dimension, are those based on energy estimates. Yet the classical energy estimates are limited to proving estimates which are local in time. The difficulty has to do with the fact that, in order to control the higher energy norms of the solutions, one has to control the integral in time of their bounds in uniform norm. For this we need to control the decay of the L^∞ norms for which we have no direct information. This difficulty was overcome by us using a strategy based on two important ideas. The first has to do with the existence of the so called Bell-Robinson tensor which plays, for the Einstein field equations, a role simmilar to that played by the the energy-momentum tensor of the standard field theories of matter. The second is a technique, developed by us, of deriving all the asymptotic behaviour of a field based entirely on energy estimates. Recall that, if $\mathbf{T}$ is the energy-momentum tensor of a field theory we have,

$$\mathbf{D}^\nu \mathbf{T}_{\mu\nu} = 0 \, .$$

Let X be an arbitrary vectorfield and P be the 1–form obtained by contracting the energy–momentum tensor with X i.e. $P_\alpha = \mathbf{T}_{\alpha\beta} X^\beta$. Then, since $\mathbf{T}$ is symmetric and divergence-less,

$$D^\alpha P_\alpha = \frac{1}{2} \mathbf{T}^{\alpha\beta} \, \pi_{\alpha\beta} \tag{6}$$

where,

$$\pi_{\alpha\beta} = D_\beta X_\alpha + D_\alpha X_\beta \tag{7}$$

is the deformation tensor of X. We now integrate (6) on a lens shaped domain $\mathcal{D}$ be bounded by two space-like hypersurfaces $\mathcal{H}_0$, $\mathcal{H}_1$. In view of the divergence theorem we derive the integral identity,

$$\int_{\mathcal{H}_0} \mathbf{T}(X,T) da_g - \int_{\mathcal{H}_1} \mathbf{T}(X,T) da_g = - \int_{\mathcal{D}} \mathbf{T}^{\alpha\beta} \, \pi_{\alpha\beta} \tag{8}$$

where T is the future oriented unit normal, g the induced metric and da_g the area element of $\partial\mathcal{D}$.

In the particular case when X is Killing[12] its deformation tensor π vanishes identically and we derive the conservation law,

$$\int_{\mathcal{H}_0} \mathbf{T}(X,T) da_g = \int_{\mathcal{H}_1} \mathbf{T}(X,T) da_g \tag{9}$$

Remark that the result (9) remains true if X is a conformal Killing vector field, i.e. it generates conformal isometries, and $\mathbf{T}$ is traceless. Indeed, if X is conformal Killing, $\pi_{\alpha\beta} = \Lambda \mathbf{g}_{\alpha\beta}$ and hence $\pi_{\alpha\beta}\mathbf{T}^{\alpha\beta} = \Lambda tr(\mathbf{T}) = 0$. One can easily show that if the action integral $\mathcal{S}[\psi, \mathbf{g}]$ is invariant under conformal rescalings of the metric, i.e. $\tilde{\mathbf{g}} = \Omega^2 \mathbf{g}$, then the corresponding energy–momentum tensor is traceless. This is the case of the Yang–Mills Theory in $3+1$ dimensions.

The identities (8) and (9) are usually applied to time-like future oriented vectorfields X, in which case, in view of the positive energy condition, the integrand $\mathbf{T}(X,T)$ is positive. This is the essence of the standard energy method. The method allows us to obtain apriori bounds, typically L^2 on space-like hypersurfaces, of the field under

[12] i.e. it generates a one parameter group of isometries

consideration. Our main new idea is based on a simple extension of this method in situations when the background has symmetries or, more appropiate for the stability of the Minkowski space-time, approximate symmetries.

To illustrate the method consider the Maxwell equations in $\mathbf{M}^{3+1}$. As mentioned above the Maxwell theory is conformal invariant. This means that whenever F is a solution and X is an arbitrary conformal Killing vectorfield the Lie derivative of F with respect to X is also a solution to the Maxwell equations. Combining this fact with the standard energy method described above we can now derive L^2 estimates for various vectorfields applied to F. For example, consider the following quantities,

$$\mathbf{Q}_0(t) = \int_{\mathcal{H}_t} \mathbf{T}(F)(K_0, T_0)dx$$

$$\mathbf{Q}_1(t) = \sum_{a=1}^{3} \int_{\mathcal{H}_t} \mathbf{T}(\mathcal{L}_{\Omega(a)}F)(\overline{K_0}, T_0)dx + \int_{\mathcal{H}_t} \mathbf{T}(\mathcal{L}_S F)(\overline{K_0}, T_0)dx + \int_{\mathcal{H}_t} \mathbf{T}(\mathcal{L}_{T_0}F)(\overline{K_0}, T_0)dx$$

$$\mathbf{Q}_2(t) = \sum_{a,b=1}^{3} \int_{\mathcal{H}_t} \mathbf{T}(\mathcal{L}_{\Omega(a)}\mathcal{L}_{\Omega(b)}F)(\overline{K_0}, T_0)dx + \sum_{a=1}^{3} \int_{\mathcal{H}_t} \mathbf{T}(\mathcal{L}_S\mathcal{L}_{\Omega(a)}F)(\overline{K_0}, T_0)dx + \int_{\mathcal{H}_t} \mathbf{T}(\mathcal{L}_S\mathcal{L}_S F)(\overline{K_0}, T_0)dx + \int_{\mathcal{H}_t} \mathbf{T}(\mathcal{L}_{T_0}\mathcal{L}_{T_0}F)(\overline{K_0}, T_0)dx. \quad (10)$$

Here $\mathcal{H}_t$ are the level hypersurfaces of the time function $t = x^0$ with future directed unit normal $T_0 = \partial_0$. T_0 is also the generator of time translations. The vectorfields $\Omega(a) = \epsilon_{abc}x_b\partial_c$, $a = 1, 2, 3$ are the usual angular momentum operators. $S = x^\alpha\partial_\alpha$ is the generator of dilations while $\overline{K}_0 = K_0 + T_0$ with $K_0 = (t^2 + r^2)\partial_0 + 2tx^i\partial_i$ the generator of inverted time translations.

In view of the above discussion the quantities $\mathbf{Q}_0(t), \mathbf{Q}_1(t), \mathbf{Q}_2(t)$ are all time independent and therefore bounded for all $t \in \Re$ if $\mathbf{Q}_i(0)$, $i = 0, 1, 2$ are all bounded. These L^2 bounds can be combine with some global version of the classical Sobolev inequalities to derive very sharp uniform estimates for F.

To do this we decompose F relative to the pair of conjugate null vectors $e_+ = \partial_t + \sum_{i=1}^3 \frac{x^i}{r}\partial_i$, $e_- = \partial_t - \sum_{i=1}^3 \frac{x^i}{r}\partial_i$ and an arbitrary orthonormal frame $(e_A)_{A=1,2}$ on the 2-spheres $S_{u,v}$ obtain by the intersection of the level hypersurfaces of $u = t - r$ with those of $v = t + r$, into the vectors $\alpha_A = F(e_+, e_A)$, $\underline{\alpha}_A = F(e_-, e_A)$ and scalars $\rho = \frac{1}{2}F(e_+, e_-)$ $\sigma = \frac{1}{2}{}^*F(e_+, e_-)$. This is called the null decomposition of F. We can now state the following,

Proposition 1 [Ch-Kl1]. *Let F be a solution of (M) and assume that the initial conditions on the hyperplane $t = 0$ are such that the quantities $\mathbf{Q}_i(0)$, $i = 0, 1, 2$ are all bounded. Consider the asymptotic behaviour of F for $t > 0$. In the space-time region $r \leq 1 + \frac{t}{2}$ all components of F behave in the same way, $|F(t, x)| \leq c(1 + t)^{-\frac{5}{2}}$. Outside that region[13] we have, $|\alpha(t, x)| \leq (1 + t + r)^{-\frac{5}{2}}$, $|\rho(t, x), \sigma(t, x)| \leq (1 + t + r)^{-2}(1 + |t - r|)^{-\frac{1}{2}}$, $|\underline{\alpha}(t, x)| \leq (1 + t + r)^{-1}(1 + |t - r|)^{-\frac{3}{2}}$.*

[13] *The crucial region, where different null components of F decay in different ways, is the wave zone region where $u = t - r$ is bounded.*

We have thus been able to derive the asymptotic properties of the Maxwell-equations relying only on their geometric properties. This feature is crucial in applications to nonlinear problems. A similar method lies also at the heart of our proof of the stability of the Minkowski space-time. It consists in the derivation of L^2 estimates for the modified Lie derivatives of the curvature tensor $\mathbf{R}$ of a space-time $\mathcal{M}, \mathbf{g}$ verifying the Einstein–Vacuum equtions (E–V) relative to some modification of the vectorfields $T_0, \Omega, S, \overline{K_0}$, as above. This is based on the crucial observation that the Bianchi identities,

$$\mathbf{D}_\epsilon \mathbf{R}_{\alpha\beta\gamma\delta} + \mathbf{D}_\alpha \mathbf{R}_{\beta\epsilon\gamma\delta} + \mathbf{D}_\beta \mathbf{R}_{\epsilon\alpha\gamma\delta} = 0 \qquad (B)$$

if interpreted as equations satisfied by $\mathbf{R}$ relative to a fixed background metric $\mathbf{g}$, are conformally covariant. The role of the Bell-Robinson tensor, $\mathbf{T}_{\alpha\beta\gamma\delta} = \frac{1}{2}(\mathbf{R}_{\alpha\mu\beta\nu}\mathbf{R}_\gamma{}^\mu{}_\delta{}^\nu + {}^\star\mathbf{R}_{\alpha\mu\beta\nu}\,{}^\star\mathbf{R}_\gamma{}^\mu{}_\delta{}^\nu)$ where, ${}^\star\mathbf{R}_{\alpha\beta\gamma\delta} = \frac{1}{2}\in_{\alpha\beta\mu\nu}\mathbf{R}^{\mu\nu}{}_{\gamma\delta}$, plays the same role for (B) as that played by the energy-momentum tensor before. Indeed $\mathbf{T}$ exhibits all the properties of an energy-momentum tensor. Thus[14] $\mathbf{T}$ is symmetric relative to all pair of indices and satisfies the conservation law, $\mathbf{D}^\delta \mathbf{T}_{\alpha\beta\gamma\delta} = 0$. Moreover it verifies the positive energy condition, i.e. $\mathbf{T}(X, Y, X, Y)$ is positive whenever X, Y are future directed time-like vectors. The conformal covariance of (B) translates into the very important fact that $\mathbf{T}$ is traceless. Also, if we define the modified Lie derivative of $\mathbf{R}$ relative to a vectorfield X by the formula, $\hat{\mathbf{L}}_X \mathbf{R} = \mathbf{L}_X \mathbf{R}_{\alpha\beta\gamma\delta} - \frac{1}{2}(\pi^\mu{}_\alpha \mathbf{R}_{\mu\beta\gamma\delta} + \pi^\mu{}_\beta \mathbf{R}_{\alpha\mu\gamma\delta} + \pi^\mu{}_\gamma \mathbf{R}_{\alpha\beta\mu\delta} + \pi^\mu{}_\delta \mathbf{R}_{\alpha\beta\gamma\mu}) + \frac{3}{8} tr\pi \mathbf{R}_{\alpha\beta\gamma\delta}$, we see that the commutation of $\hat{\mathbf{L}}_X$ with (B) produces error terms depending on the vector X only through the trace-less part of its deformation tensor π. Thus, if the background metric is that of the Minkowski space-time the modified Lie derivative relative to a conformal Killing vectorfield commutes with the Bianchi equations (B). These properties of (B) allow us to define generalized energy norms for $\mathbf{R}$, as we have done previously for F. These norms allow us to control the asymptotic behaviour of $\mathbf{R}$. We illustrate this in Minkowski space-time as follows. Assume that W is a 4-covariant trace-less tensor verifying all the symmetry properties of the curvature tensor $\mathbf{R}$. Assume that W is a solution of the Bianchi equations (B) in $\mathbf{M}^{3+1}$. Let $\mathbf{Q}_0(t), \mathbf{Q}_1(t), \mathbf{Q}_2(t)$ be the quantities,

$$\mathbf{Q}_0(t) = \int_{\mathcal{H}_t} \mathbf{T}(W)(\overline{K_0}, \overline{K_0}, T_0 T_0) dx$$

$$\mathbf{Q}_1(t) = \sum_{a=1}^{3} \int_{\mathcal{H}_t} \mathbf{T}(\hat{\mathcal{L}}_{\Omega(a)} W)(\overline{K_0}, \overline{K_0}, T_0 T_0) dx + \int_{\mathcal{H}_t} \mathbf{T}(\hat{\mathcal{L}}_{T_0} W)(\overline{K_0}, \overline{K_0}, \overline{K_0} T_0) dx$$

$$\mathbf{Q}_2(t) = \sum_{a,b=1}^{3} \int_{\mathcal{H}_t} \mathbf{T}(\hat{\mathcal{L}}_{\Omega(a)} \hat{\mathcal{L}}_{\Omega(b)} W)(\overline{K_0}, \overline{K_0}, T_0, T_0) dx + \sum_{a=1}^{3} \int_{\mathcal{H}_t} \mathbf{T}(\hat{\mathcal{L}}_{\Omega(a)} \hat{\mathcal{L}}_{T_0} W)(\overline{K_0}, \overline{K_0}, \overline{K_0}, T_0) dx$$

[14] These properties hold true as a consequence of the symmetries of $\mathbf{R}$ and, in view of the E–V equations, the condition $\mathbf{R}_{\alpha\beta} = 0$.

$$+\int_{\mathcal{H}_t}\mathbf{T}(\hat{\mathcal{L}}_S\hat{\mathcal{L}}_{T_0}W)(\overline{K_0},\overline{K_0},\overline{K_0},T_0)dx$$
$$+\int_{\mathcal{H}_t}\mathbf{T}(\hat{\mathcal{L}}_{T_0}\hat{\mathcal{L}}_{T_0}W)(\overline{K_0},\overline{K_0},\overline{K_0},T_0)dx. \qquad (11)$$

with $T_0, \Omega, S, \overline{K_0}$ as in (10).

The null decomposition of W is given by $\alpha_A = W(e_A, e_+, e_B, e_+)$, $\beta_A = \frac{1}{2}W(e_A, e_+, e_-, e_+)$, $\rho = \frac{1}{4}W(e_-, e_+, e_-, e_+)$, $\sigma = \frac{1}{4}{}^*W(e_-, e_+, e_-, e_+)$, $\underline{\beta}_A = \frac{1}{2}W(e_A, e_-, e_-, e_+)$, $\underline{\alpha}_{AB} = W(e_A, e_-, e_B, e_-)$. We can now state the following,

Proposition 2. *Let W be a solution of the Bianchi equations in $\mathbf{M}^{3+1}$ and assume that the initial conditions on the hyperplane $t = 0$ are such that the quantities $\mathbf{Q}_i(0)$ are all bounded. Consider the asymptotic behaviour of W for $t > 0$. In the space-time region $r \leq 1 + \frac{t}{2}$ all components of W behave in the same way, $|W(t,x)| \leq c(1+t)^{-\frac{7}{2}}$. Outside that region we have, $|\alpha(t,x)| \leq (1+t+r)^{-\frac{7}{2}}$, $\beta \leq (1+t+r)^{-\frac{7}{2}}$, $|\rho(t,x), \sigma(t,x)| \leq (1+t+r)^{-3}(1+|t-r|)^{-\frac{1}{2}}$, $|\underline{\beta}| \leq (1+t+r)^{-2}(1+|t-r|)^{-\frac{3}{2}}$ $|\underline{\alpha}(t,x)| \leq (1+t+r)^{-1}(1+|t-r|)^{-\frac{5}{2}}$.*

Both Propositions 1,2 have been proved in [Ch-Kl1].

The implementation of a similar strategy for the actual E–V equations requires some fundamental modifications. We rely on the same quantities [15] $\mathbf{Q}_1(t), \mathbf{Q}_2(t)$ introduced in (11). The major departure from the linear theory of Prop. 2 is that the vectorfields T_0, Ω, S, K_0 are now themselves unknown and have to be constructed together with our space–time. This is due to the difficulty (d). Indeed, if the asymptotic behaviour of the causal structure of the space-time we construct would have been trivial we could have chosen as vectorfields T_0, Ω, S, K_0 the ones given to us in $\mathbf{M}^{3+1}$. As it stands, due to the mass term which appears in the Schwarzschild part of an (S.A.F.) initial data set, has the long range effect of changing the asymptotic position of the null geodesic cones relative to the maximal foliation. They are expected to diverge logarithmically from their corresponding position in flat space-time. In addition to this their asymptotic shear[16] differs drastically from that in the Minkowski space-time. This difference reflects the presence of gravitational radiation in any nontrivial perturbation of the Minkowski space-time. To take this effect into account we rely heavily on our optical function u and construct the vectorfields T_0, Ω, S, K_0 based on its properties[17]. The construction of u itself is very elaborate and requires a systematic study of the geometry of null hypersurfaces.

In linear theory the time derivatives of the $\mathbf{Q}_1, \mathbf{Q}_2$ are zero. In the case of the (E–V) equations they give rise to cubic error[18] terms which depend linearly on the traceless part of the deformation tensors of K_0, T, S, Ω , and quadratic with respect to $\mathbf{R}$ and its covariant and Lie derivatives in the direction of T, S, Ω. The crucial point of our overall strategy is to control the time integral of these error terms. This depends on the one

[15] Because of the difficulty (c) we have to avoid the use of $\mathbf{Q}_0(t)$ which would be infinite at $t = 0$.

[16] the traceless part of their null second fundamental form

[17] and also those of the maximal foliation induced by the time function t.

[18] generated each time we commute the Bianchi Identities with a one of the vectorfields used in the definition of $\mathbf{Q}_1, \mathbf{Q}_2$

hand on the asymptotic behaviour of all components of $\mathbf{R}$ and its covariant derivatives, which are themselves controlled by the basic quantities $\mathbf{Q}_1, \mathbf{Q}_2$. On the other hand it depends on the asymptotic behavior of the deformation tensors of our vectorfields, and finally, due to the general covariance of the equations, on the cancellations of the "worst possible" cubic terms. It is well known that arbitrary quadratic nonlinear perturbations of the scalar wave equation, even when derivable from a Lagrangean, could lead to formation of singularities unless a certain structural condition, which we have called the Null condition, is satisfied (see [Kl2], anad [Ch1], [Kl2]). It turns out that the appropriate, tensorial version of this structural condition is satisfied by the Einstein equations. Roughly speaking one could say that the troublesome nonlinear terms, which could have led to formation of singularities, are in fact excluded due to to the covariance and algebraic properties[19] of the Einstein equations.

Putting together all the elements discussed above requires an elaborate boot-strap argument based on the method of continuity. In the end we show that $\mathbf{Q}_1, \mathbf{Q}_2$ can never become large, if they are sufficiently small at $t = 0$. The local existence theorem then guarantees that we can extend our space-time indefinitely.

Finally it remains to note that our work has many important conclusions. Far from beeing an abstract proof of existence it provides precise information on the asymptotic nature of gravitational radiation. Some of our conclusions confirm the non-rigurous results[20] obtained by Bondi, Sachs, Penrose etc.

Some of them are however new. One of these has lead D. Christodoulou to a real experimental prediction. In [Ch4] he shows that gravitational waves generated by astronomical sources can have a nonlinear effect on laser interferometer detectors on Earth. This effect is shown to be of the same order of magnitude as the linear effects upon which all previous efforts to detect gravitational waves were based. Moreover the nonlinear effect is shown to produce a permanent displacement of test masses after the passage of a wave. It can thus alter significantly the startegy upon which the experimentalists plan to build their future detectors.

---

[19] These basic algebraic properties of the Einstein equations, which allow us to prove the above stated global existence result, are in sharp contrast with the nonlinear hyperbolic equations of classical continuum mechanics. Indeed the equation of Nonlinear Elasticity [John] and of Compressible fluids [Si], in four space and time dimensions, form singularities even for arbitrary small initial conditions.

[20] I want to point out however that our results are inconsistent with the assumption, made by Penrose [Pe1], [Pe2], concerning the smoothness of null infinity. Our results show weaker "peeling" than those implied by the Penrose requirement. It remains questionable whether there are any smooth initial conditions which lead to Ricci flat space-times for which the Penrose requirement is valid.

References

[Bour] J.P Bourguignon, "Stabilite par deformation non-lin. de la metrique de Minkowski," *Seminaire N. Bourbaki* **1990-1991** June 1991.

[Br1] C. Bruhat, "Theoreme d'existence pour certain systemes d'equations aux deriveès partielles nonlinaires," *Acta Matematica* **88**, 1952, 141–225.

[Br2] C. Bruhat, "Un theoreme d'instabilite pour certain equations hyperb. non-lineaire," *C.R. Acad. Sci. Paris* **276A**, 1973, pp. 281.

[Ch1] D. Christodoulou, "Global solutions for nonlinear hyperbolic equations for small data," *Comm. Pure Appl. Math.* **39** 1986, 267–282.

[Ch2] D. Christodoulou, "A Mathematical Theory of Gravitational Collapse," *Comm. Math. Ph.* **109**, p 613 (1987).

[Ch3] D. Christodoulou, "The Formation of Black Holes and Singularities in Spherically Symmetric Gravitational Collapse," *CP. Appl. Math.* **44**, 339-373(1991).

[Ch4] D. Christodoulou, "The Nonlinear Nature of Gravitation and Gravitational Wave Experiments," Preprint June 1991.

[Ch-Kl1] D. Christodoulou-S.Klainerman, "Asymptotic Properties of Linear Field Equations in Minkowski Space-Time," *Comm. Pure Appll. Math.* **43** 137-199(1990).

[Ch-Kl2] D. Christodoulou-D. S.Klainerman, "The nonlinear stability of the Minkowski space-time," To appear in Princeton Univ. Press .

[Ch-Za] D. Christodoulou-A. Shadi Tahvildar-Zadeh, "Regularity of Spherically Smmetric Harmonic Maps of the $2+1$ dim. Minkowski Space," Preprint.

[E-M] D. Eardley-V. Moncrief, "The Global Existence of Yang-Mills-Higgs fields in $\mathbf{M}^{3+1}$," *C.M.P.* **83**, 171-212(1982).

[Gr1] M. Grillakis, "Regularity and Asympt. Behaviour of the Wave Eq. with as critical nonlinearity," *Ann. of Math.* **132** , 485-509(1990).

[Gr2] M. Grillakis, "Classical solutions for the Equivariant Wave Maps in 1+2 dimensions," preprint.

[Gu] Chao, Gu, Hoo, "On the Cauchy PR for harmonic maps defined on 2-D Minkowski space-time," *C.P.A.M.* **33**, 1980, pp. 727-737.

[Jö] K. Jörgens, "Das Angfangswertproblem im Grossen Fur eine Klasse nichtlinearer Wellengleichungen," *Math. Z.*, **77**, 295-307(1961).

[John] F. John, "Formation of Singularities in Elastic Waves," **Lecture Notes in Phys.**, Springer–Verlag, 1984, pp. 190-214.

[Kl1] S. Klainerman, "Uniform decay estimates and the Lorentz invariance of the classical wave equation," *Comm. Pure. Appl. Math.* **38**, 1985, 321–332.

[Kl2] S. Klainerman, "Long time behaviour of solutions to nonlinear wave equations," Proc. of the I.C.M. Warsaw, 1982, pp. 1209-1215.

[Kl3] S. Klainerman, "The null condition and global existence to nonlinear wave equations," *Lect. Appl. Math.* **23**, 1986, 293–326.

[Kl4] S. Klainerman, "Remarks on the global Sobolev inequalities in Minkowski space," *Comm. Pure Appl. Math.* **40**, 1987, 111–117.

[Kl-Ma] S. Klainerman-M.Machedon, "Space–Time estimates for null forms and the local existence theorem," preprint.

[L] W. Littman "The wave operator and L^p norms," *Journal of Math. and Mech.* **12** pp. 55-68(1963).

[Pe1] R. Penrose, *Structure of Space–Time*, Batelle Rencontre, C.M. DeWitt and J.A. Wheeler, 1967.

[Pe2] R. Penrose, "Zero rest mass fields including gravitation: asymptotic behaviour," *Proc. Roy. Soc. Lond.* **A284**, 1962, 159–203.

[Sh-Stru] work in progress.

[Sh] "Weak solutions and development of singularities for the SU(2) σ-model," *Comm. P. Appl. Math.* 1988, pp. 459-469.

[Sh-Za] J. Shatah, "A. Tahvildar-Zadeh Regularity of harmonic maps from the Minkowski space-time into rotationally symmetric manifolds," preprint.

[Si] T. Sideris, "Formation of Singularities in 3-d Compressible Fluids," *Comm. Math. Phys.* **101**, 1985, 275–485.

[Str1] R.S. Strichartz, "A Priori Estimates for the Wave Equation and some Applications," *J. of Funct. Anal.***5**, 218-235(1970).

[Str1] R.S. Strichartz, "Restrictions of Fourier Transforms to Quadratic Surfaces and Decay of Solutions of Wave Eq." *Duke Math. J.* **44, No 3**, 705-714 (1977).

[Stru] M. Struwe, "Globally Regular Solutions to the Φ^5- Klein-Gordon equation."

Paper presented at the 13th Int. Conf. on General Relativity and Gravitation
Cordoba, Argentina, 1992: Part 1, Plenary Lectures

Canonical quantum gravity

Karel V. Kuchař

Department of Physics, University of Utah,
Salt Lake City, Utah 84112, U.S.A.

Abstract. This is a review of the aspirations and disappointments of the canonical quantization of geometry. I compare the two chief ways of looking at canonical gravity, geometrodynamics and connection dynamics. I capture as much of the classical theory as I can by pictorial visualization. Algebraic aspects dominate my description of the quantization program. I address the problem of observables. The reader is encouraged to follow the broad outlines and not worry about the technical details.

CLASSICAL CANONICAL GRAVITY

Dynamical laws and instantaneous laws

One of the main preoccupations of classical physics has been finding the laws governing physical data. One of the oldest schemes of quantization has been to subject such data to canonical commutation relations. I am going to review where this program leads us when it is applied to geometry.

Classical physics deals with two kinds of laws: dynamical laws, and instantaneous laws. The discovery of dynamical laws started the Newtonian revolution. The first instantaneous law was found by Gauss: at any instant of time, the divergence of the electric field is determined by the distribution of charges. In empty space, the instantaneous electric field is divergence-free.

Theorema egregium

Without knowing it, and without most of us viewing it this way, Gauss also came across the fundamental instantaneous law of general relativity: the Hamiltonian constraint. This constraint is a simple reinterpretation of the famous result Gauss obtained when studying curved surfaces embedded in a flat Euclidean space [1].

To start with, Gauss drew the key distinction between intrinsic and extrinsic properties of a surface. The intrinsic properties are not changed by bending the surface without stretching; the extrinsic ones are. The basic intrinsic property of a surface

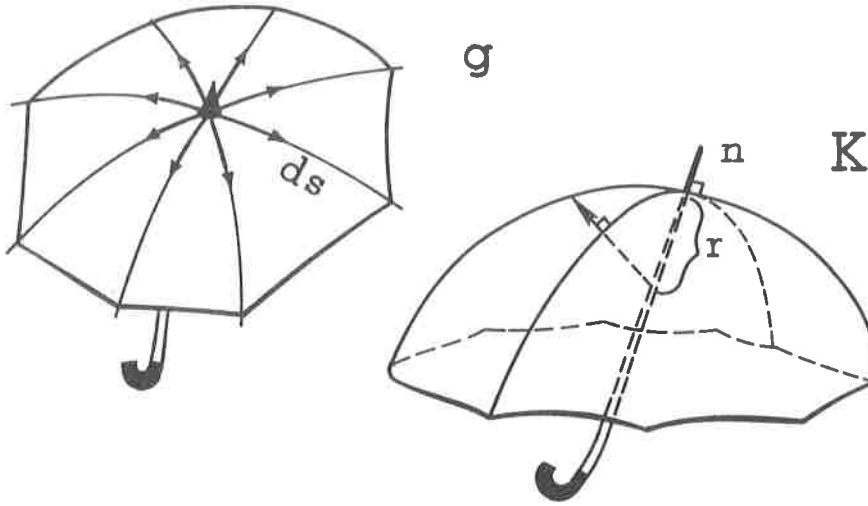

Figure 1. Intrinsic metric and extrinsic curvature.

is its *intrinsic metric*; the basic extrinsic property, its *extrinsic curvature*. These are highlighted on the umbrellas of Figure 1. The network of distances from the tip of the umbrella along the ribs is encapsulated by the familiar metric tensor

$$ds^2 = g_{ab}(x)\,dx^a dx^b. \tag{1}$$

The ribs lie in the planes passing through the shaft of the umbrella; they represent its *normal sections*. Each normal section has a radius of curvature, r, whose reciprocal value, k, is the curvature of the normal section. The extrinsic curvature, K_{ab}, of the surface is an inventory of the curvatures of all its normal sections:

$$r^{-1} = k = K_{ab}(x)\,dx^a dx^b \;/\; ds^2. \tag{2}$$

From the metric, one can derive other intrinsic objects: lengths, angles, and areas. Geodesics are completely determined by the metric. So is the *parallel transport* of a tangent vector: a vector is parallel transported from a point to a neighboring point if it keeps its angle with the geodesic segment connecting the points. In its turn, parallel transport leads to the concept of *scalar* (or Gaussian) *curvature* (Figure 2):

Take a curve enclosing the tip of the umbrella. Mark where you want to **START**, take the tangent vector to the curve, and parallel transport it along the curve back to the starting point. On the way, the tangent vector (bold) rotates clockwise with respect to the parallel transported vector (double arrow). If the umbrella were a plane (I would not like to use such an umbrella on a rainy day), the tangent vector would run all around the clock. On the bulging umbrella of Figure 2, it does not quite make it; it still has an angle $\delta\omega$ to go. As the curve is drawn tighter and tighter around the tip, the deficit angle

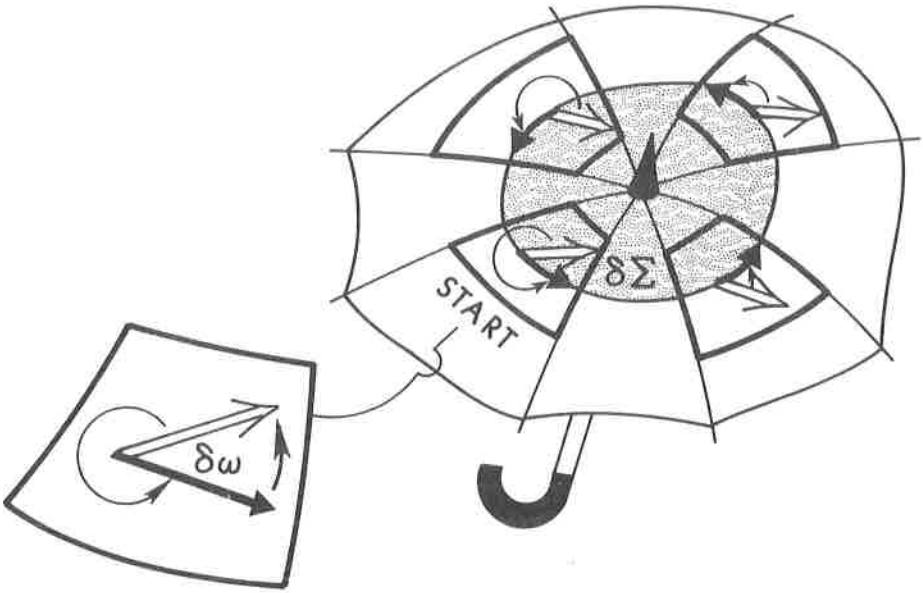

Figure 2. Scalar curvature.

$\delta\omega$ becomes proportional to the surface area $\delta\Sigma$ surrounded by the curve. The ratio of these two quantities defines the scalar curvature:

$$\delta\omega = \tfrac{1}{2} R\, \delta\Sigma\,. \tag{3}$$

(Gauss did not obtain the scalar curvature this way; by brute force, he expressed it as a function of the metric and its first and second derivatives: $R[g]$.)

Let the ribs again represent normal sections. Of all the ribs, one has the gentlest bending, k_{MIN}, and another has the steepest bending, k_{MAX}. Which rib is which, how large is k_{MAX}, and how small is k_{MIN}, depends on the wind. On a quiet day, all ribs have the same curvature $k_{\mathrm{MAX}} = k_{\mathrm{MIN}}$. The values k_{MAX} and k_{MIN} are called *principal curvatures*.

Because the principal curvatures depend on the wind, they are clearly extrinsic rather than intrinsic properties of an umbrella. However, their *product*, called the *total curvature*, remains always the same. Indeed, it is proportional to the scalar curvature which, as we have seen, is an intrinsic property of the umbrella:

$$\begin{aligned} R &= 2\ k_{\mathrm{MAX}} \cdot k_{\mathrm{MIN}} \\ \textbf{scalar curvature} &= 2\ \textbf{total curvature}\,. \end{aligned} \tag{4}$$

Gauss has found many remarkable theorems in his life, but this one he himself regarded to be remarkable: he named the result (4) *theorema egregium*.

By looking at the surface of a single umbrella, we cannot be sure if it is worn by an upright old man walking across Great Plains, or if it protects a crooked old goblin wedged into a curved space. However, if the intrinsic metric and extrinsic curvature of all

possible umbrellas are connected by Gauss' theorema egregium, we can safely conclude that the space in which they are embedded is flat and Euclidean. The flatness of space is thereby guaranteed by the match of certain intrinsic and extrinsic properties of all embedded surfaces.

Still, what has all of this to do with the assertion that dynamics is a consequence of instantaneous laws? There are no instants and hence no dynamics in a Euclidean space. To talk about instants, space must be Lorentzian. In a flat three-dimensional Lorentzian spacetime the theorema egregium still holds. The only thing we need to change is the sign. On every spacelike surface,

$$R = -2\, k_{\mathrm{MAX}} \cdot k_{\mathrm{MIN}} \,. \quad \text{(Lorentzian)} \tag{5}$$

And, conversely, if the Lorentzian theorema egregium (5) holds on every spacelike surface in a three-dimensional Lorentzian spacetime, we can be sure that the spacetime is flat.

Now, the Lorentzian theorema egregium is an instantaneous law. On the other hand, the statement that spacetime is flat is a dynamical law, albeit a very simple dynamical law, about geometry. Roughly speaking, it tells us that there is no dynamics: spacetime remains flat all the time. This argument illustrates how an instantaneous law, the Lorentzian theorema egregium, can lead to a dynamical law, that the spacetime is flat.

General relativity, I remember someone saying, does not confine us to Euclidean barracks. Even if the spacetime is empty (which, for simplicity, I shall assume for the rest of my lecture), its dynamics is quite rich. The ripples of gravitational radiation can travel around, interfere, attract each other, and amplify. They can hold themselves together in a gravitational geon. Part of the gravitational radiation can leak out, part of it may collapse and form a black hole. I find it quite surprising that all this dynamics is encoded in an almost trivial generalization of Gauss' theorema egregium:

The intrinsic geometry and the extrinsic curvature of a three-dimensional hypersurface embedded in a four-dimensional Riemannian spacetime have the same definition and the same geometric significance as those of a two-dimensional surface in a three-dimensional flat space. However, instead of two principal sections there are three, with principal (extremal) curvatures k_1, k_2, and k_3. One cannot define the total curvature T as the product of a selected couple of principal curvatures. As a true egalitarian, one takes

$$T = k_1 k_2 + k_1 k_3 + k_2 k_3 \,. \tag{6}$$

Similarly, there is no single surface on which one can determine the deficit angle $\delta\omega$ by parallel transporting the tangent vector along a curve. Instead, one chooses three perpendicular surfaces passing through the tip of a three-dimensional umbrella, and determines the three deficit angles $\delta\omega_1$, $\delta\omega_2$, and $\delta\omega_3$. The scalar curvature R has the geometric meaning

$$\tfrac{1}{2} R = \frac{\delta\omega_1}{\delta\Sigma_1} + \frac{\delta\omega_2}{\delta\Sigma_2} + \frac{\delta\omega_3}{\delta\Sigma_3} \,. \tag{7}$$

Compare now the total curvature (6) and the scalar curvature (7) of a hypersurface in an arbitrary Ricci-flat spacetime. Behold, the theorema egregium still holds:

$$R = \begin{cases} - \text{ (Lorentzian)} \\ + \text{ (Euclidean)} \end{cases} 2T \,. \tag{8}$$

Inversely, if the scalar curvature is related to the total curvature by Eq.(8) on any spacelike hypersurface, the spacetime is necessarily Ricci-flat. Therefore, the statement that the theorema egregium holds at any instant is entirely equivalent to the Einstein law of gravitation in empty space!

I am sorry that the continuation of my narrative requires some juggling of indices. The total curvature (6) is a quadratic combination of the three principal curvatures. Because each of these is a linear function of the extrinsic curvature (2), the total curvature can be expressed as a quadratic form of the extrinsic curvature:

$$T = -\tfrac{1}{2} K_{ab}\, G^{ab\,cd}\, K_{cd}\,. \tag{9}$$

The coefficient

$$G^{ab\,cd} = \tfrac{1}{2}(g^{ac}g^{bd} + g^{ad}g^{bc} - 2\,g^{ab}g^{cd}) \tag{10}$$

is called the *supermetric*. Symmetric pairs of covariant indices can be raised by the supermetric, and symmetric pairs of contravariant indices lowered by its inverse, $G_{ab\,cd}$. The contravariant version of the extrinsic curvature is

$$p^{ab} := G^{ab\,cd}\, K_{cd}\,. \tag{11}$$

The total curvature is a quadratic form of p^{ab}, and the Lorentzian theorema egregium (8) assumes the form [1]

$$H(x) := p(x) \cdot G(x;\, g) \cdot p(x) - R(x;\, g] = 0\,. \tag{12}$$

The theorema egregium is the most fundamental instantaneous law of Einstein's theory of gravitation. Gauss did not realize that the theory of curved surfaces in a flat Euclidean space (and of curved hypersurfaces in a Ricci-flat spacetime) is subject to yet another instantaneous law, closely resembling the law which he had found for electricity. As shown by Codazzi [2], the covariant divergence of the extrinsic curvature p^{ab} vanishes: [2]

$$H_a :\simeq {}_g\nabla_b\, p_a{}^b(x) = 0\,. \tag{13}$$

Canonical geometrodynamics

The stage is now ready for stating (not proving, nor even properly explaining) the sea change which the theorema egregium (12) and the Codazzi law (13) suffered a century later. Working from quite an opposite direction of variational principles and Hamiltonian dynamics, Dirac [3] and Arnowitt, Deser, and Misner [4] have shown that the intrinsic metric g_{ab} and the (densitised) extrinsic curvature p^{ab} are canonically conjugate to each other. In canonical theory, the instantaneous laws (12) and (13) are called the Hamiltonian and diffeomorphism constraints. They start playing a double role. On one hand, they *restrict* the canonical data. On the other hand, as dynamical variables on the phase space, they become capable of *evolving* the canonical data. The Poisson bracket of the data with $H_a(x)$ generates their change by a Lie derivative in the direction along the

[1] The notation $R(x;\, g]$ emphasizes that R is a function of x and a functional of g.

[2] I write $\simeq$ whenever I want to sweep a numerical factor under the rug.

hypersurface. Similarly, the Poisson bracket with $H(x)$ generates the change of the data under a normal displacement of the hypersurface. These two processes enable us to organize the embeddings by displacements which deform one embedding into another, and to correlate the data which the embeddings carry. Instead of checking the Einstein law by criss-crossing the spacetime by all possible hypersurfaces, we obtain it by an orderly Hamiltonian evolution which smoothly deforms the original hypersurface. The change of the canonical data by the generators $H_a(x)$ and $H(x)$, together with the statement that the generators, once they generated the change, are constrained to vanish, *is* the Einstein law. This is the new strange role of the instantaneous laws: they become the agents of dynamics.

The change generated by $H_a(x)$ is induced by a spatial diffeomorphism DiffΣ on a given hypersurface. This property gave the Codazzi constraint its new name — the diffeomorphism constraint. The constraint ensures that the theory is invariant under DiffΣ. In other words, canonical geometrodynamics does not depend on the intrinsic metric and the extrinsic curvature, but only on such combinations of these variables which are unaffected by spatial diffeomorphisms, i.e., only on the intrinsic and extrinsic geometries. There are *fewer* physical variables than the symbols which meet the eye.

This message can also be read backwards: by making the theory dependent on *more* variables, one can make it invariant with respect to a wider class of transformations. A good example of this process is triad dynamics.

Triad dynamics

Let us choose as our basic variables a triad E^a_i, $i = 1, 2, 3$, of orthonormal vectors [5]. These determine the intrinsic metric,

$$g^{ab} = \delta^{ij} E^a_i E^b_j \,, \tag{14}$$

but the metric determines the triad only up to an x-dependent SO(3) rotation. The rotation group SO(3) becomes a gauge group of the Einstein theory. Canonical analysis reveals that the projected extrinsic curvature

$$-K^i_a(x) = -K_{ab}(x) E^b_j \delta^{ji} \tag{15}$$

is the canonical coordinate whose conjugate momentum is the (densitised) triad E^a_i. The SO(3) rotations of the canonical variables are generated by the dynamical variable

$$G_i(x) := \epsilon_{ij}{}^k (-K^j_a(x)) E^a_k(x) = 0 \,, \tag{16}$$

which has the familiar structure of angular momentum. After generating the rotations, $G_i(x)$ is constrained to vanish. The *rotation constraint* (16) ensures that the extrinsic curvature K_{ab} related to K^i_a by Eq.(15) is symmetric.

Any vector, u^a, can be characterized by its internal components, u^i, in the orthonormal basis E^a_i:

$$u^a = u^i E^a_i . \tag{17}$$

Let us parallel transport the vector u^a from x to $x + \mathrm{d}x$; we get the double-arrow vector of Figure 3. There is a basis, $E^a_i(x + \mathrm{d}x)$, sitting at $x + \mathrm{d}x$. In this basis, I draw a

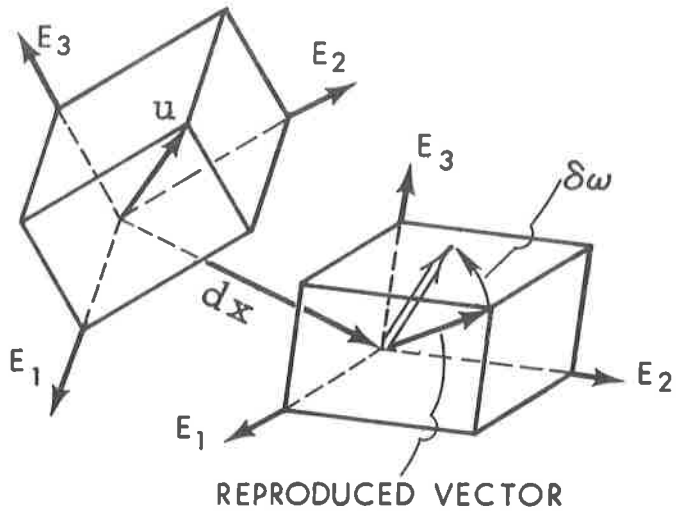

Figure 3. The SO(3) parallel transport.

vector which has the same components, u^i, as the original vector had at x. I call it the *reproduced vector*. To turn the reproduced vector into the parallel transported vector, I must rotate its internal components by an angle $\delta\omega^i$. This angle is a linear function of the displacement $\mathrm{d}x^a$:

$$\delta\omega^i = -\Gamma_a^i \mathrm{d}x^a. \tag{18}$$

The coefficient Γ_a^i tells us how the parallel transport of a vector affects its internal components. It can be expressed in terms of the triad $E_i^a(x)$ and its first derivatives. It is called the SO(3) connection.

As usual, the curvature tensor of a connection is defined by the parallel transport of a vector u along a small parallelogram with the edges $\mathrm{d}x$ and δx (Figure 4). The parallel transported vector (shown as double arrow) does not return back to its original position (shown in bold). To turn the original vector into the parallel transported vector, we must subject its internal components to a rotation:

$$\delta\omega^i = -R_{ab}{}^i \mathrm{d}x^a \delta x^b. \tag{19}$$

The coefficient $R_{ab}{}^i$ is the curvature tensor of the SO(3) connection. It can be expressed in terms of the basis vectors E_i^a, and their first and second derivatives.

The curvature tensor satisfies the *cyclic identity*. Its geometric significance is illustrated on the right in Figure 4. Take a small box with edges u, v and w. Parallel transport w along the boundary of the face u, v. The parallel transported vector does not coincide with w; it differs from it by δw. Repeat this procedure for the remaining two vectors, and obtain the differences δu and δw. While none of them in general vanishes, their sum is identically equal to zero: $\delta u + \delta v + \delta w \equiv 0$.

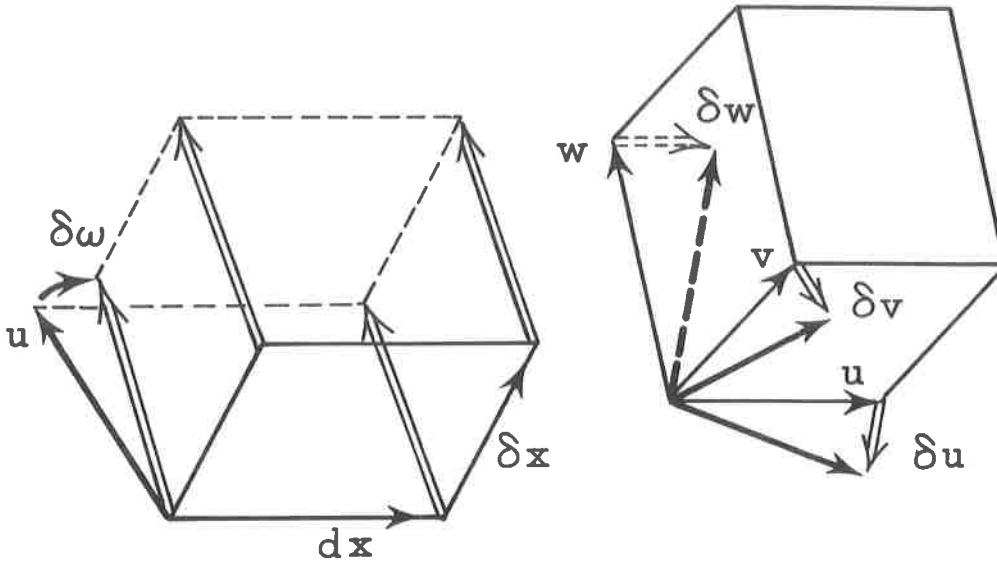

Figure 4. The SO(3) curvature tensor and the box identity.

It is easy to write down what this geometric construction yields for a small box whose edges lie in the direction of the orthonormal vectors E^a_i. We obtain

$$R_{ab}{}^{i} E^b_i \equiv 0 \,. \tag{20}$$

The SO(3) curvature tensor $R_{ab}{}^{i}[E]$ necessarily satisfies the box identity (20).

Once we know the curvature tensor, we can determine the curvature scalar as in Eq.(7). We take three mutually perpendicular curves, each of them enclosing a unit area, parallel transport their tangent vectors, determine the deficit angles, and add them together. In particular, we can choose for the curves the parallelograms spanned by the pairs of the orthonormal vectors E^a_i. In this way we learn that

$$R[E] = -R_{ab}{}^{i}\, \epsilon_i{}^{jk} E^a_j E^b_k \,. \tag{21}$$

Algebraically, $R[E]$ can be obtained by substituting the metric (14) into $R[g]$.

We can now take the Hamiltonian constraint (12) and express it in terms of the new canonical variables $-K^i_a$ and E^a_i. By using Eqs.(9)–(10) and (14), we cast the total curvature into a form in which it is quadratic both in $-K^i_a$ and in E^a_i. The curvature scalar is a concomitant (21) of the triad E^a_i. As a result, H is set forth as a functional of $-K^i_a$ and E^a_i. The diffeomorphism constraint can be handled in the same way. Neither of the constraints looks any simpler in the new variables than it did in the old ones.

In addition to the old constraints, we have the rotation constraint (16). More variables call for more constraints. There are nine entries in the triad E^a_i, while there are only six entries in the symmetric metric g_{ab}. Similarly, there are nine entries in $-K^i_a$, while there are only six in the extrinsic curvature K_{ab}. However, the physics

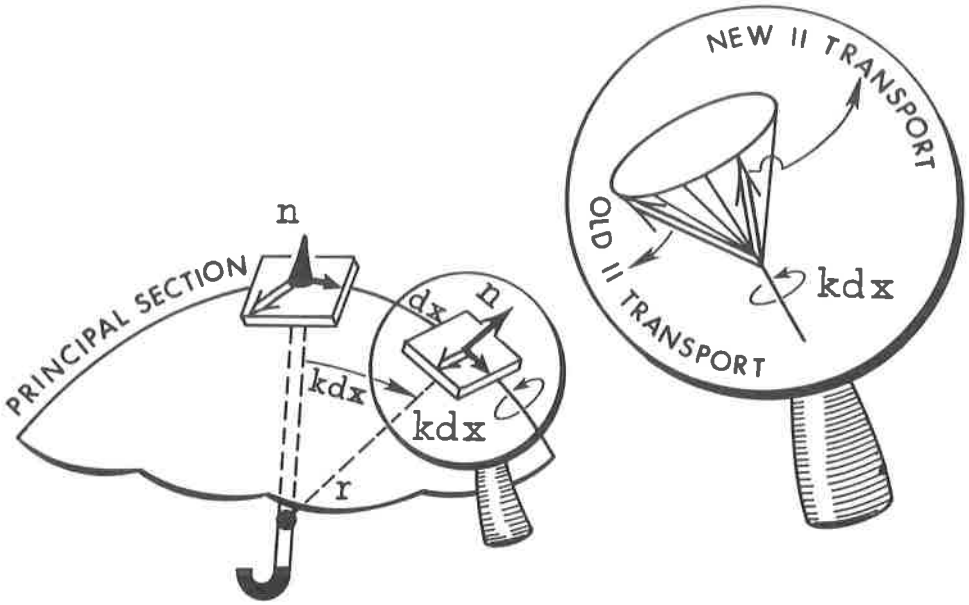

Figure 5. New parallel transport.

depends only on the old set of variables, g_{ab} and K_{ab}. The triad E^a_i enters into the Hamiltonian and diffeomorphism constraints only through the combination (14), and the extrinsic curvature given by Eq.(15) is forced to be symmetric by the rotation constraint (16). The surplus dynamics – rotations of the triad generated by the constraint (16) – is expendable. In the end, only such quantities which are unaffected by rotations (like g_{ab} and K_{ab}) physically matter. We can return to them, forget about the rotation constraint, and retrieve geometrodynamics from triad dynamics. To gain more invariance by introducing more variables is not a big deal.

Connection dynamics

To simplify the constraints, one must go one step beyond introducing the triads: one must modify the parallel transport. Figure 5 shows a three-dimensional umbrella protecting us from a storm of gravitational waves in a four-dimensional Euclidean Ricci-flat spacetime. Take a tangent vector to the umbrella (shown as a double arrow) and parallel transport it from the tip along one of the principal sections. The transported vector is again shown as a double arrow. Viewed from the center of curvature, the arc dx along which the vector is transported subtends the angle $k dx$. Give now the parallel transported vector an additional twist by the angle $k dx$ about the principal direction. The position of the vector after the twist defines the new parallel transport.

The twist does not look quite right in the two-dimensional sketch of the three-dimensional umbrella. It seems that to rotate the vector about the rib we must rotate the whole tangent plane, and destroy thereby its tangential character. To see what

is happening, we must look at the tangent plane through one of the most powerful instruments ever invented by a theoretical physicist: John A. Wheeler's dimensional magnifying glass [6]. Under the glass, the plane thickens into what it actually is, a three-dimensional tangent space, and the twist moves the vector along a cone in this space into its new position. The normal to the umbrella stays fixed because the twist takes place in the plane perpendicular to the principal section.

To transport a vector in an arbitrary direction, we must decompose the displacement dx into the three principal directions and perform the appropriate twists one after another. The new parallel transport amounts to a single rotation (18) of the reproduced vector. The angle of rotation is given by a new SO(3) connection [7, 8]

$$A_a^i = \Gamma_a^i - K_a^i \,. \qquad \text{(Euclidean)} \tag{22}$$

From the canonical standpoint, it is remarkable that Eq.(22) represents a canonical transformation: the new SO(3) connection A_a^i is a coordinate canonically conjugate to the momentum E_i^a. (To see that A_a^i is conjugate to E_i^a is trivial, because $-K_a^i$ is conjugate to E_i^a. It is more difficult to prove that $A_a^i(x)$ can serve as a field coordinate, i.e., that it has a vanishing Poisson bracket with $A_b^j(x')$.) But why should we ever want to perform the canonical transformation (22)? Cui prodest?

The new parallel transport leads, by Figure 4 and Eq.(19), to the new curvature tensor $F_{ab}{}^i[A]$. In terms of this tensor, the instantaneous laws of Euclidean Ricci-flat spacetimes take a remarkably simple form. The new curvature tensor no longer satisfies the box identity (20). Instead, the expression $F_{ab}{}^i E_i^b$ yields the supermomentum [3]

$$H_a \simeq F_{ab}{}^i[A] E_i^b \,. \tag{23}$$

The box equation (20) still holds, but no longer as an identity. It is now equivalent to the Codazzi law (13) in a Ricci-flat spacetime. Even more remarkably, the H of Gauss' theorema egregium turns out to be the scalar curvature (21) of the new parallel transport:

$$H \simeq F_{ab}{}^i[A] \tfrac{1}{2} \epsilon_i{}^{jk} E_j^a E_k^b \,. \tag{24}$$

These striking facts were discovered by Ashtekar [8]. It is quite tempting to call Eq.(24) Ashtekar's theorema egregium: In a Ricci-flat spacetime, the scalar curvature of Ashtekar's connection A vanishes on every hypersurface.

The new connection A_a^i defines the new covariant derivative ${}_A D_a$ acting on internal indices. In terms of this derivative, the rotation generator (16) takes the form

$$G_i = {}_A D_a \, E_i^a \,. \tag{25}$$

(Because E_i^a is a vector density, ${}_A D_a$ does not need to act on spatial indices to produce a scalar density.)

These are the good news. Now, for the bad news. The transition from a Euclidean to a Lorentzian spacetime enforces a change of sign in the Gauss theorema egregium. The Ashtekar theorema egregium can absorb this change of sign only at the price of introducing a complex SO(3) connection:

$$A_a^i = \Gamma_a^i - i \, K_a^i \,. \qquad \text{(Lorentzian)} \tag{26}$$

To handle this complication in canonical quantum gravity is not entirely trivial.

[3] This time, $\simeq$ sweeps under the rug not only numerical factors, but also the fact that the rotation constraint is used in rearranging the diffeomorphism constraint and the Hamiltonian constraint.

Dialogue concerning the two chief systems of canonical gravity: geometrodynamics and connection dynamics

Geometrodynamics and connection dynamics are the two chief forms of canonical gravity. Let us pause and compare them before proceeding with quantization. I suppress the turmoil of indices and highlight the two structures in a table. Let

· denote contraction in spatial indices,

o denote contraction in spatial indices *and* internal indices,

* denote internal dualization,

and the details take care of themselves. Then

GEOMETRODYNAMICS			CONNECTION DYNAMICS	
Coordinates	**Momenta**		**Coordinates**	**Momenta**
g	p	CANONICAL VARIABLES	A	E
Intrinsic metric	**Extrinsic curvature**		**SO(3) connection (mixed)**	**Triad (intrinsic)**
		GENERATORS OF		
$p \cdot G(g) \cdot p - R[g]$		⊥ EVOLUTION	$E \circ {}^*F[A] \circ E$	
${}_g\nabla \cdot p$		DiffΣ	$F[A] \circ E$	
None		ROTATIONS	${}_A D \cdot E$	

A comparison of the two columnes brings forward a number of simple observations:

1. *Variables and constraints.* Geometrodynamics works with fewer variables and fewer constraints than connection dynamics. The geometrodynamical variables are invariant under triad rotations.

2. *Connection with gauge theories.* The rotation constraint makes connection dynamics resemble an SO(3) Yang-Mills theory. In geometrodynamics, the rotation constraint is eliminated and one works with SO(3)-invariant canonical variables.

3. *Dimension.* I discussed the constraints in $3+1$ dimensions. Geometrodynamics remains virtually the same in any dimension $n+1$, $n \geq 2$. The SO(3) connection dynamics is intimately adapted to a three-dimensional space and it is not easily generalized to $n > 3$.

4. *Positivity restrictions.* The Cauchy problem works only if the hypersurfaces are spacelike, i.e., if the induced metric g is positive definite. In geometrodynamics, this puts a restriction on the domain of the configuration space. In connection dynamics, the metric is automatically positive definite, as long as the triad is non-degenerate. However, even for degenerate triads (leading to degenerate metrics) the formalism seems to make sense, and it is viable to lift the non-degeneracy restriction.

5. *Structure of the Hamiltonian constraint.* In geometrodynamics, H is a *quadratic function* of the momentum p. The supermetric $G(g)$ is ultralocal, and there is a local potential term $R[g]$. In connection dynamics, the potential is absorbed into the quadratic term; H is a *quadratic form* of the momentum E. However, the supermetric ${}^*F[A]$ is no longer ultralocal, but merely local.

6. *Polynomiality of constraints.* In connection dynamics, all constraints are low-degree polynomials in the canonical variables A and E. It was originally claimed that geometrodynamical constraints are non-polynomial in the canonical variables g and p, but a more careful look [9] reveals that a simple scaling by a power of $\det(g)$ also makes them polynomial. However, they are polynomials of a rather high degree.

7. *Reality conditions.* Geometrodynamics works with real canonical variables on a real phase space.. We have seen that in a Lorentzian spacetime the Ashtekar variable A is necessarily complex. This forces one to work with complex canonical variables, either on a complex or a real phase space. A pair A and E of canonical variables that satisfy the constraints define a real Ricci-flat spacetime only if they satisfy the *reality conditions*
$$E - \bar{E} = 0\,, \quad A + \bar{A} = \Gamma[E]\,. \tag{27}$$
These conditions are non-polynomial in E. People replace [10] the reality conditions (27) by somewhat weaker conditions that are polynomial. A simpler procedure is to scale the second condition (27) by $[\det(g)]^2$ which makes it polynomial in E and A. Whichever way one proceeds, the polynomial reality conditions are of a rather high degree.

In view of these observations, which scheme is simpler, geometrodynamics or connection dynamics? Simplicity, of course, is in the eye of the beholder, and my assessment is quite personal.

1. I believe that, on one hand, one should not make too much fuss about the count of the variables and constraints and, on the other hand, one should not overemphasize the resemblance between connection dynamics and the SO(3) gauge theories.

2. The SO(3) invariance is a simple consequence of introducing the redundant variables. The ease with which such variables are eliminated and connection dynamics

reduced back to geometrodynamics may be an indication that the achieved SO(3) invariance is not that deep. However, one should not overlook that the mixing of the extrinsic and intrinsic variables brings in a true simplification of the constraints prior to the imposition of the reality conditions.

3. Our space is three-dimensional and a theory which makes an effective use of this fact is not to be blamed. I view the simplifications which can be achieved only in three dimensions speaking for rather than against connection dynamics.

4. The positivity restrictions on the metric are quite a nuisance in quantum theory. The possibility of lifting the non-degeneracy condition on the triad without endangering the connection dynamics is a real advantage. However, when listing the achievements of quantum connection dynamics, one should bear in mind that many of these correspond to situations in which the triad and hence the metric are degenerate.

5. Trading an ultralocal supermetric and a local potential term for a local supermetric without any potential is an interesting quid pro quo. Whether such a trade-off pays off in quantum theory depends quite heavily on whether it is easier or not to turn the new Hamiltonian constraint into a well-defined operator.

6. A low-degree polynomiality is certainly an asset in quantizing a classical theory. In this respect, the constraints of connection dynamics are definitely simpler than the geometrodynamical ones.

7. One should bear in mind, however, that connection dynamics is sooner or later confronted with the task of implementing the reality conditions. These conditions are unseemly, being polynomials of such a high degree as the geometrodynamical constraints. The simplifications which the connection dynamics achieves may thus be a mere temporary advantage.

The Galilean overtones of my section heading are meant to go beyond a mere joke. Eliminating variables or constraints is like getting rid of epicycles. The presence of a potential term may be aesthetically repugnant like the use of an equant. Nevertheless, the fact remains that geometrodynamics and connection dynamics are entirely equivalent at the classical level, just as the Ptolemaic and Copernican systems are entirely equivalent at the kinematical level. The Copernican system may be aesthetically more pleasing, but its real power emerges only when one starts asking dynamical questions. Similarly, the real power of connection dynamics may emerge only when one starts quantizing the classical theory. This is the task we should now discuss.

CONSTRAINT QUANTIZATION: A PROGRAM

Canonical gravity is a system whose dynamics is entirely generated by constraints. Its quantization and interpretation presents some special difficulties. The ground rules for quantizing constrained systems were laid by Dirac [3] and refined over the years. Every major review of canonical quantum gravity [4, 11, 12] attempted to list a sequence of steps expected to lead to a satisfactory theory. People more or less agree about what

these steps are, but they do not know how to implement them: by listing the steps, they present a mere quantization program. I shall picture seven steps of a quantization program as seven gateways on a road paved with good intentions.

1. Fundamental variables

The first step is the selection of *fundamental variables*. These are classical dynamical variables that are to be turned into operators whose commutator algebra replicates the classical Poisson algebra. The fundamental variables are expected to span a vector space $\mathcal{V}$ closed under the Poisson brackets $\{\ ,\ \}$. The space $\mathcal{V}$ should be *complete* in the sense that any dynamical variable F can be approximated by an element of the free algebra $\mathcal{A}$ over $\mathcal{V}$, i.e., expressed as a sum of products of the elements of $\mathcal{V}$.

In geometrodynamics, $\mathcal{V}$ is taken to be a real vector space spanned by g, p, and the unit dynamical variable 1. In connection dynamics, $\mathcal{V}$ is taken to be a complex vector space spanned by A, E, and 1. The elements $V \in \mathcal{V}$ are expected to be mapped into operators $\hat{V} \in \hat{\mathcal{V}}$ in such a way that

$$V_3 = \{V_1, V_2\} \implies \hat{V}_3 = -i[\hat{V}_1, \hat{V}_2]\,. \tag{28}$$

In geometrodynamics, the metric g should be positive definite. The positivity conditions cannot be written as relations in $\mathcal{V}$, and their imposition is quite tricky [13]. Connection dynamics is fully equivalent to geometrodynamics only for non-degenerate triads E. People working in connection dynamics propose not to impose the condition that E be non-degenerate.

Connection dynamics has a difficulty which does not exists in geometrodynamics: not all elements of the complex vector space $\mathcal{V}$ describe a real spacetime. Ultimately, one must impose the *reality conditions* (27). However, $\mathcal{V}$ is not closed under the complex conjugation. The reality conditions thus cannot be formulated in $\mathcal{V}$ (the old SO(3) connection $\Gamma(E)$ does not lie in $\mathcal{V}$). It was proposed [12, 14] that at the level of representing the fundamental variables by operators one should simply forget about the reality conditions. These are to be taken care of much later, during the construction of a Hilbert space. Conforming to this view, I return to the issue of reality conditions at the last of my gates.

2. Dynamical variables (including constraints)

In the next step, one must decide on how to turn an arbitrary dynamical variable $F[g,p] \in \mathcal{A}$ or $F[A,E] \in \mathcal{A}$ into an operator. Such variables typically do not lie in $\mathcal{V}$, but they can be approximated by sums of products of the elements of $\mathcal{V}$. Van Hove [15] has proved that it is impossible to turn dynamical variables into operators in such a way that Eq.(28) holds for all of them. Without the guiding principle (28), the quantization of dynamical variables is subject to factor-ordering ambiguities. It is popular to dismiss these as 'mere quantum-mechanical corrections'. I do not share this view. Unless one knows how to factor order significant dynamical variables, one really does not know how to construct quantum theory. In a sense, the right factor ordering *is* the quantum theory. If one does

not set any rules about factor ordering, one can turn a classical variable $F(Q,P)$ into any quantum operator one pleases:

Let $F(Q,P)$ be a classical dynamical variable, and $G(Q,P)$ any other dynamical variable (whose dimension is that of F divided by the action). Fix some factor ordering of $\hat{F} = F(\hat{Q},\hat{P})$ and $\hat{G} = G(\hat{Q},\hat{P})$, and define the quantum variable

$$\hat{F}' := \hat{F} - i\,[\hat{Q},\hat{P}])\hat{G} = \hat{F} + \hat{G}\,. \tag{29}$$

The quantum variables $\hat{F}'$ and $\hat{F}$ have the same classical limit, namely, $F(Q,P)$, and yet they differ by an arbitrary operator $\hat{G}$. This is what a 'mere' factor ordering can do.

A particular case of dynamical variables in canonical gravity are the constraints. One should turn them into operators $\hat{H}(x)$, $\hat{H}_a(x)$ (and possibly $\hat{G}_i(x)$). In field theory, this presents a regularization problem. Moreover, in the next step of the quantization procedure one wants to impose the constraints on the states. This poses a consistency problem. Both of these problems are troublesome, but they at least impose severe restrictions on the factor ordering. On the other hand, very little is known and, even more remarkably, said about what to do with other dynamical variables.

3. Representation space $\mathcal{F}$

The operators representing the dynamical variables are expected to act on a space of states. One way of choosing this space is to rely on the Schrödinger representation: the canonically conjugare pairs of fundamental variables are taken as multiplication and differentiation operators acting on functionals of the configuration variables. Thus, in geometrodynamics $\mathcal{F}$ is taken to be a complex vector space whose elements are the functionals $\Psi[g]$ of the metric. In connection dynamics, the elements of $\mathcal{F}$ are the functionals $\Psi[A]$ of the Ashtekar connection. [4]

There is an important difference between the Schrödinger representation for unconstrained systems and the Schrödinger representation in canonical gravity: the representation space $\mathcal{F}$ is not necessarily assumed to be a Hilbert space, and (real) dynamical variables are not required to be represented by self-adjoint operators [17]. Physically, the Hilbert space structure is needed to calculate the expectation values of observables. However, prior to the imposition of constraints, the states in $\mathcal{F}$ do not necessarily describe physical states, and it does not have a good meaning to ask what is the expectation value of an observable in such a state.

The rejection of the Hilbert space structure liberates us from a straitjacket that often leads to inconsistencies [18], but it unfortunately leads to a loss of control over mathematical objects. I shall later comment on both of these aspects.

4. Space of solutions

The key idea of the Dirac constraint quantization is to turn the constraints, which I shall now collectively call $\mathbf{H}$, into operators (gate 2), and impose them as restrictions on the states:

$$\mathbf{H}\Psi = 0\,. \tag{30}$$

[4] The connection representation $\Psi[A]$ is formally related to the *loop representation*. This is discussed by Smolin in this volume, and in a recent review by Rovelli [16].

One surmises that only such states Ψ which solve the constraints can be physical. All physics is to be done on the space $\mathcal{F}_0$ of states which solve Eq.(30).

A number of remarks is appropriate. First of all, the quantum constraints should not limit the quantum states more than the classical constraints limit the classical states: they should not beget other constraints by commutation. This imposes stringent requirements on the factor ordering of the constraint functions. One can see on simple models that these requirements virtually dictate the factor ordering, and that to satisfy them the constraints cannot and should not be represented by self-adjoint operators on $\mathcal{F}$ [18]. In geometrodynamics (and in connection dynamics), a consistent factor ordering of constraints is a notorious unsolved problem. The task is seriously hampered by the field-theoretical aspects of canonical gravity, which call for regularization of the constraint operators. [5]

The absence of a Hilbert space structure on $\mathcal{F}$ helps us to make the constraints consistent. However, it also makes the quantization badly dependent on the choice of representation. One can see the problem already when solving the Schrödinger equation of a simple unconstrained system like an anharmonic oscillator [26],

$$\hat{h} = \tfrac{1}{2}\hat{P}^2 + \tfrac{1}{2}\hat{Q}^2 + \tfrac{1}{4}\hat{Q}^4. \tag{31}$$

The solution of the Schrödinger equation calls for finding the eigenfunctions of the energy operator (31). In the Q-representation, the eigenfunction equation is a differential equation of the second order. In the P-representation, it is a differential equation of the fourth order. As a differential equation, the first equation has fewer solutions than the second equation. The mismatch is removed by requiring that the solutions we seek be square integrable (in Q and in P), i.e., belong to the Hilbert space based on the Schrödinger norm.

When, as in canonical gravity, we are unwilling to impose a Hilbert-space structure on $\mathcal{F}$, the size of $\mathcal{F}$ depends on the choice of representation. Thus, in principle, the solution space $\Psi[g]$ in geometrodynamics is different from the solution space $\Psi[p]$. Similarly, the connection representation $\Psi[A]$ is not necessarily equivalent to the triad representation $\Psi[E]$. In other words, by not requiring that the representation space $\mathcal{F}$ be a Hilbert space, one affirms a strong belief in the primacy of those fundamental variables on which the representation is based.

This is only a part of a larger problem. Unless one imposes some boundary conditions on the solutions of the constraint equation (31), the solution space $\mathcal{F}_0$ may be much too big. The quadratic character of the Hamiltonian constraint in the momenta p

[5] One should find a factor ordering of the Hamiltonian and diffeomorphism constraints such that the commutator of the Hamiltonian constraints yields an expression in which the diffeomorphism constraint acts on the state function first, followed by the structure functions of the 'Dirac algebra'. It was noticed by Anderson [19] that this task cannot be accomplished if one insists on representing the constraints by self-adjoint operators on $\mathcal{F}$. A solution to the factor-ordering problem was offered by Schwinger [20] and criticized by Dirac [21]. The best, and certainly the shortest, exposition of Schwinger's solution may be found in a footnote of the paper [22] by DeWitt; this was later rediscovered by Komar [23]. DeWitt himself made a rather sweeping proposal on how to remove the problem by letting any two field operators taken at the same point formally commute [22]. Ashtekar [8] proposed a simple factor ordering of his constraints which (disregarding the regularization difficulties) satisfies the consistency requirement. Unfortunately, all these results are purely formal: Tsamis and Woodard [24], and Friedman and Jack [25] have persuasively argued that by formal manipulations of the commutator one can obtain whatever result one wants.

(or E) evokes the analogy with the Klein-Gordon constraint for the relativistic particle. There we know that the solution space of the mass-shell constraint is also too big: the physical states of a one-particle system correspond only to positive-energy solutions. In geometrodynamics (and in connection dynamics), we do not have any accepted method of cutting the basis of the solution space into half. We are thus stuck with a solution space which may be physically too big.

If, on the other hand, we start imposing boundary conditions or some other limitations on the states, we may inadvertently force the solution space to be too small. This may (though it does not need to) happen in connection dynamics when one requires that the states $\Psi[A]$ be holomorphic functions of the complex connection A. One imposes such a requirement in analogy with the Bargmann representation for the states of a harmonic oscillator [27]. The solution of the eigenvalue equation for the oscillator Hamiltonian $\hat{h}$ on the space of holomorphic functions of $Z = Q - iP$ gives automatically a correct spectrum for $\hat{h}$. This is surprising, because at this stage we do not yet have any Hilbert space. Only much later is the space of holomorphic functions turned into a Hilbert space which yields the same spectrum. This may not work so smoothly in canonical gravity. The complex connection is in some respects quite different from the complex variable Z for the harmonic oscillator. (I shall return to this point three steps later.) There is a chance that the 'preestablished harmony' between holomorphic functions and the subsequent construction of the Hilbert space no longer exists.

To summarize, without the Hilbert space structure on $\mathcal{F}$ and without boundary conditions or some auxiliary conditions on the states, we are bound to end up with a solution space $\mathcal{F}_0$ that contains many unphysical states. For this reason, I am reluctant to call $\mathcal{F}_0$ 'the physical space', and prefer to stick to a more neutral name, the space of solutions.

5. Observables

An outstanding question in the theory of constrained systems is what dynamical variables can in principle be observed. An often made proposal [28, 12] is that

- Classical 'observables' are those dynamical variables F whose Poisson brackets with the constraints weakly vanish:

$$\mathbf{H} = 0 \quad \Longrightarrow \quad \{F, \mathbf{H}\} = 0\,. \tag{32}$$

Its quantum mechanical counterpart is that

- Quantum 'observables' are those operators $\hat{F}$ that commute with the constraint operators $\hat{\mathbf{H}}$ on the space of solutions $\mathcal{F}_0$:

$$\hat{\mathbf{H}}\Psi = 0 \quad \Longrightarrow \quad [\hat{F}, \hat{\mathbf{H}}]\Psi = 0\,. \tag{33}$$

The second definition seems to be virtually forced on us if we insist that the measurement of an observable does not throw the state Ψ out of the space of solutions $\mathcal{F}_0$.

These two definitions are straightforward generalizations of the concept of an observable in gauge theories. I am going to argue that they are inappropriate for canonical gravity.

To see why the definitions (32) and (33) are natural in ordinary gauge theories, consider electrodynamics. The vector potential A describes the state of the electromagnetic field. The potentials $A_{(1)}$ and $A_{(2)}$ which lie on the same orbit of the Gauss constraint $G(x) = \nabla \cdot E(x)$ differ by a gauge transformation. They are *physically indistinguishable*: they represent two equivalent descriptions of the same physical state. One cannot observe the individual A's along the orbit, only the magnetic field B. The magnetic field remains the same if we change the vector potential by a gauge transformation:

$$\{B(x'),\, G(x)\} = 0\,. \tag{34}$$

The magnetic field is an example of an observable.

A quantum state of the electromagnetic field is described by the state functional $\Psi[A]$. This functional is the probability amplitude for finding the electromagnetic field in the state described by the vector potential A. The probability should remain the same when we change A by a gauge transformation. This is ensured by the Gauss constraint

$$\hat{G}(x)\,\Psi[A] = 0\,. \tag{35}$$

Equation (35) implies that Ψ can depend on A only via the classical observable B: $\Psi = \Psi[B]$. On an ensemble of systems described by the state functional $\Psi[B]$, we cannot measure $\hat{A}$, but only $\hat{B}$. The magnetic field operator $\hat{B}$ is a quantum observable. It satisfies the quantum counterpart of Eq.(34):

$$[\hat{B}(x'),\, \hat{G}(x)] = 0\,. \tag{36}$$

The same case which I made for the Gauss constraint G in electrodynamics can be repeated for the rotation constraint G_i and the diffeomorphism constraint H_a in canonical gravity:

Spacelike hypersurfaces in a Ricci-flat spacetime carry the induced geometry, but do not come equipped with an orthonormal triad E. The triad is a mere tool for calculating the metric (14). Two triads, $E_{(1)}$ and $E_{(2)}$, on the same orbit of the constraint (16) differ by a rotation. They both yield the same metric (14). Rotations can be thought about as a gauge, and metric as an observable. In general, the SO(3) observables are those dynamical variables which are unaffected by rotations,

$$G_i(x) = 0 \quad \Longrightarrow \quad \{F,\, G_i(x)\} = 0\,. \tag{37}$$

In the triad representation, the quantum state of the gravitational field is described by the state functional $\Psi[E]$. The rotation constraint

$$\hat{G}_i(x)\,\Psi[E] = 0 \tag{38}$$

implies that Ψ can depend on E only through the metric (14): $\Psi = \Psi[g]$.

However, the metric is not yet an observable with respect to diffeomorphisms. Two metric fields, $g_{(1)}(x)$ and $g_{(2)}(x)$, that differ only by the action of DiffΣ, i.e., which lie on the same orbit of $H_a(x)$, are physically indistinguishable. This is due to the fact that we have no direct way of observing the points $x \in \Sigma$. A dynamical variable constructed from the metric field is a true observable only if its value is unaffected by diffeomorphisms:

$$H_a(x) = 0 \quad \Longrightarrow \quad \{F,\, H_a(x)\} = 0\,. \tag{39}$$

Thus, e.g., the volume of Σ is an observable:

$$V[g] = \int_\Sigma \mathrm{d}^3x \, |\det(g(x))|^{\frac{1}{2}} . \tag{40}$$

The momentum constraint

$$\hat{H}_a(x) \, \Psi[g] = 0 \tag{41}$$

implies that the value of the state functional $\Psi[g]$ is the same for all metrics connected by DiffΣ, i.e. , that $\Psi[g]$ does not depend on the individual metrics $g(x)$, but only on the three-geometry ${}^3\mathcal{G}$.

However, the definition (32) of an observable requires yet something more. It claims that a dynamical variable F cannot be observed unless it has a vanishing Poisson bracket with the Hamiltonian constraint H. I feel that this requirement is misguided.

The action of G_i on the dynamical variables generates their change under rotations SO(3). The action of H_a on the dynamical variables generates their change under DiffΣ. Both of these actions operate in the space of the instantaneous data on a fixed hypersurface. The change of the data which they generate is unobservable. The action of H is different: it generates the dynamical change of the data from one hypersurface to another. The hypersurface itself is not directly observable, just as the points $x \in \Sigma$ are not directly observable. However, the collection of the canonical data $g_{(1)}$, $p_{(1)}$ on the first hypersurface is clearly distinguishable from the collection $g_{(2)}$, $p_{(2)}$ of the evolved data on the second hypersurface. If we could not distinguish those two sets of the data, we would never be able to observe dynamical evolution.

The same reasoning applies to quantum theory. In the Schrödinger picture, the evolution is carried by the state Ψ. The Hamiltonian constraint

$$\hat{H}(x)\Psi = 0 \tag{42}$$

plays a different role from the diffeomorphism constraint or the rotation constraint. It does not tell us that the evolved state is indistinguishable from the initial state, but rather it tells us how the state evolves. Thus, in geometrodynamics, the constraint (42) is a second-order variational differential equation for the state $\Psi[{}^3\mathcal{G}]$ of the three-geometry, called the Wheeler-DeWitt equation [6, 22]. This can be viewed as analogous to the Klein-Gordon equation for the state $\psi(x^\alpha)$ of a relativistic particle. The three-geometry ${}^3\mathcal{G}$ is considered as an internal configuration space*time* variable, similar to the argument x^α of the Klein-Gordon state. The Wheeler-DeWitt equation is supposed to describe the dynamical evolution of the state in an internal configuration spacetime.

It is this fundamental distinction between the states which are and the states which are not distinguishable that leads me to reject the definition (32) according to which 'observables' should also have a vanishing Poisson bracket with the Hamiltonian constraint. The dynamical variable F which satisfies this requirement,

$$H(x) = 0 \quad \Longrightarrow \quad \{F, \, H(x)\} = 0 \, , \tag{43}$$

must have the same value on all spacelike hypersurfaces. Therefore, it is necessarily a constant of motion. This underscores the point which I already made: If we could observe only constants of motion, we could never observe any change.

I hold that one can observe other dynamical variables, like the volume variable (40), not only constants of motion. Therefore, I shall call *observables* those dynamical variables which are invariant under SO(3) and DiffΣ, but which do not necessarily obey Eq.(43). Those observables which also satisfy Eq.(43) I shall call *perennials*. I want to argue that

- One can observe dynamical variables which are not perennial,

and that

- Perennials are often difficult to observe.

To make these two points, I do not need to deal with general relativity. Any parametrized (or already parametrized) system [29] illustrates the same point. I shall try to clarify the issues on the simplest of such systems, a parametrized free Newtonian particle moving on a line. The phase space of the system is the cotangent bundle $(T, Q; P_T, P)$ over the configuration spacetime (T, Q), and the Hamiltonian constraint amounts to the definition of the energy $-P_T$ in terms of the momentum P:

$$H := P_T + \tfrac{1}{2}P^2 = 0\,. \tag{44}$$

Perform a canonical transformation [30]

$$Q' = Q - PT, \qquad P' = P, \tag{45}$$

$$T' = T, \qquad P_{T'} = P_T + \tfrac{1}{2}P_T{}^2. \tag{46}$$

The primed canonical variables (45) are the initial data at $T = 0$. The primed time T' is identical with the Newtonian time T. The momentum $P_{T'}$ conjugate to T' coincides with the Hamiltonian constraint:

$$H := P_{T'} = 0\,. \tag{47}$$

Due to the constraint (47), any dynamical variable $G(T', Q'; P_{T'}, P')$ can be replaced by an equivalent variable $F(T', Q'; P') := G(T', Q'; 0, P')$. The variable F is a perennial if $\{F, H\} = 0$. Equation (47) enables us to conclude that perennials are simply arbitrary functions of the initial data. They cannot depend on T':

$$\textbf{Perennials} : \qquad F = F(Q'; P'). \tag{48}$$

No perennial ever changes along a dynamical trajectory. To observe change, we must observe at least one dynamical variable, like T or Q, which changes.

An opposite view has been expressed by Rovelli [31]. I interpret his paper as saying that to observe a changing dynamical variable, like Q, amounts to observing a one-parameter family

$$Q'(\tau) := Q' + P'\,\tau = Q - P\,(T - \tau), \quad \tau \in \mathsf{R} \tag{49}$$

of perennials. The perennials (49) are the values of Q at $T = \tau$. By observing the perennials $Q'(\tau_1)$ and $Q'(\tau_2)$ one can infer the change of Q from $T = \tau_1$ to $T = \tau_2$.

The problem with such a view is that one is not told how to observe τ. One way of observing τ is to watch the dynamical variable T (the hand of an ideal Newtonian clock). The value of T is τ. However, this amounts to observing a dynamical variable T which is not a perennial. An alternative is to say that one can observe τ directly. Again, one is forced to admit that one can observe an entity which is not a perennial. The third alternative is to say that because perennials are constants of motion, it does not matter when they are observed. One can observe all the perennials $Q'(\tau)$, $\tau \in \mathsf{R}$ at once, and infer 'the change of Q with T' from that instantaneous observation. Any instant is like any other, and each contains the same set $\tau \in \mathsf{R}$ of perennials from which the change is inferred. This does not make me too happy either. If all time τ is eternally present, all time is irredeemable.

My discussion was so far concerned with the epistemological status of observables. I tried to argue that the identification of observables with perennials drives one to a Parmenidean view of the world. Physicists are soundly sceptical of epistemological arguments, and I am not deluding myself that my argument is an exception. "Refutations are seldom final; in most cases, they are only a prelude to further refinements." Significantly, Bertrand Russell made this remark when closing his discussion of Parmenides [32].

So far I argued that some observables are not perennial. I must now defend my other point, namely, that perennials are often difficult to observe. In this part of the discussion, I take the attitude of physical common sense, that at any instant one can directly observe the position Q of the particle, its momentum P, and the time T on an ideal Newtonian clock, but not the position Q' which the particle had at time $T = 0$. The initial position Q', which does not change with T and is a perennial, is *inferred* from the observed data Q, P, and T by using Eq.(45). For a free particle, such an inference is easy because we know how to integrate equations of motion. However, even for such a simple system as a free particle, the inference may be hampered by experimental errors. If one determines P with an error ΔP, the error in the inferred value of Q' scales with T. If the particle moves on a circle and T is large, it is practically impossible to infer from the observations at T where on the circle the particle was at $T = 0$. For more complicated Hamiltonians, like those governing dynamics of many interacting particles, the task of inferring perennials becomes pretty hopeless. Take, e.g., a globular cluster, observe the current positions and momenta of the stars, and then try to infer what were their positions and momenta when the cluster was formed some 15 billion years ago.

In quantum theory, there is yet another reason why perennials are difficult to observe. To measure a quantum variable, one needs to design an apparatus with appropriate coupling. [6] Theoretically, it is possible to find an apparatus which measures an arbitrary quantum variable $\hat{F} = F(\hat{Q}, \hat{P})$. Experimentally, this can be done only for a small number of especially simple variables, like $\hat{F} = \hat{Q}$ or $\hat{F} = \hat{P}$. For elementary systems, like a free particle or a harmonic oscillator, the initial-data perennials (45) are linear functions of the current data $\hat{Q}$ and $\hat{P}$. Experimentalists know how to build the apparatuses for measuring such perennials. An example is the discussion of non-demolition experiments for detecting gravitational waves [33]. To circumvent the limits imposed by the uncertainty principle, one constructs an apparatus for monitoring the initial length of an oscillating bar, i.e., the perennial like Q' of Eq.(45) for a linear harmonic oscillator. However, even for such a simple system as the hydrogen atom, the initial-data perenni-

[6] My discussion takes place within the framework of von Neumann's theory of measurement. It should be rephrased in a scheme like that advocated by Hartle in the present volume.

als are complicated functions of the current data $\hat{Q}$ and $\hat{P}$. It is difficult to conceive an apparatus which would monitor such perennials at all times.

If the dynamical system is not Newtonian, i.e., if the Hamiltonian constraint is not linear in the momentum P_T conjugate to a time variable T, the practical difficulty of determining classical perennials from the current data turns into something much more serious: into an argument questioning their very existence. A classical example is an asymmetric top spinning around a fixed point in a homogeneous gravitational field. Describe the configuration of the top by the Euler angles $Q^a = (\phi, \psi, \theta)$, where θ is measured from the direction of the field. The Hamiltonian h of the top is a quadratic function of the momentum P_a. Constrain the motion of the top to be taking place with a definite energy E:

$$H := h - E = 0\,, \quad h = \tfrac{1}{2} G^{ab}(Q) P_a P_b + V(Q)\,. \tag{50}$$

The trajectory of the top in the phase space (Q^a, P_a) is generated by the Hamiltonian constraint (50). Notice that we do not ask how the top *moves* in the Newtonian time T, we are merely asking about its *trajectory*. The momentum P_T does not enter into the constraint (50); it was replaced by a constant E.

Perennials are defined as those dynamical variables $F(Q^a, P_a)$ that have a (weakly) vanishing Poisson bracket with H. Notice that T cannot be used in the construction of perennials because it no longer is a canonical variable.

One perennial is the angular momentum M_θ about the direction of the gravitational field. This perennial is linear in the momentum P_a:

$$M_\theta = M^a(Q) P_a\,. \tag{51}$$

A century ago, Poincaré asked the question [34]: Does the top have any other integrals of motion than those of vis viva and the area? In the way I formulated the problem, this translates into the question: Is there any perennial besides M_θ? The answer is *no* [35].

The configuration space (T, Q) of a parametrized free Newtonian particle is two-dimensional, and there is one Hamiltonian constraint. There are $2 \times (2 - 1) = 2$ independent perennials (45); any other perennial is their function. An n-dimensional parametrized Newtonian system should have $2(n - 1)$ independent perennials. The top is a three-dimensional system, and one would expect to find four independent perennials. However, the constraint (50) does not have the Newtonian form, and there is only one perennial, (51).

Let me briefly return from models to canonical gravity. General relativity is not a parametrized field theory whose constraints have a 'Newtonian' form (44). In particular, both in geometrodynamics and in connection dynamics, the Hamiltonian constraint is quadratic in the momenta. The supermetric has some non-trivial dependence on the canonical coordinates. In these respects, the Hamiltonian constraint resembles the constraint (50) for the top. This prompts the following remarks:

- We do not know how to construct perennials for canonical gravity.
- We do not know how to select families of perennials (similar to the family (49)) labeled by a functional time parameter (similar to τ) which would correspond to 'simple' dynamical variables as the volume observable (40) (similar to Q).

- So far, we did not find a single gravitational perennial. [7] The existence of a complete set of perennials would imply that gravity is a completely integrable theory. They are indications that it is not [36, 37]. It is likely that the gravitational perennials are rare, and it is quite possible that there are none.

Perennials in canonical gravity may have the same ontological status as unicorns —*a priori*, these are possible animals, but *a posteriori*, they are not roaming on the Earth. According to bestiaries, the unicorn is a beast of fabulous swiftness, strength, and beauty, but, alas, it can be captured only by a virgin [38]. Corrupt as we are, we better stop hunting mythical beasts.

6. Hilbert space

Once we have decided what dynamical variables can be observed, we need to know what is the statistical distribution of their observed values. In quantum mechanics, probabilities are determined by the inner product in a Hilbert space. Therefore, we need to endow the space of physical states with a Hilbert space structure.

The proposals on how to find the inner product depend on what position one takes on observables. Let me first discuss the proposal [39], which relies on identifying observables with perennials:

- Choose an inner product $\langle\Psi_1|\Psi_2\rangle$ on the solution space $\mathcal{F}_0$ such that all *real* quantum perennials are self-adjoint under it.

In geometrodynamics, the phase space is real and it is easy to say when a dynamical variable is real. I return to the reality problem in connection dynamics in the next section.

There are several problems with the above proposal. First of all, we have seen that there may not be any perennials in canonical gravity, or that at least there may not be a sufficient number (a complete set) of them. If so, the proposal on how to determine the inner product either loses its content, or becomes too weak. Secondly, even when one disregards this difficulty, one should notice that the proposal as it stands is self-contradictory. If $\hat{F}$ and $\hat{G}$ are quantum perennials, so is $\hat{F}\hat{G}$. If $\hat{F}$ and $\hat{G}$ are self-adjoint under the inner product $\langle\Psi_1|\Psi_2\rangle$, $\hat{F}\hat{G}$ is not. To remove the contradiction, one needs to find 'fundamental perennials', and approximate all other perennials by polynomials of the fundamental perennials. One can then require that only the fundamental perennials be self-adjoint, and symmetrically factor order the polynomials which define the remaining perennials. Unfortunately, the original fundamental variables g and p (or A and E) are not perennials, and we lack a guiding principle on what the fundamental perennials may be.

The third problem with the proposal is that the solution space $\mathcal{F}_0$ is probably larger than the space of physical states. We have seen that it may contain 'improper elements', 'unbounded states', and 'states with negative norms'. The definition of a perennial $\hat{F}$ requires that $\hat{F}$ commutes with the constraints on the solution space $\mathcal{F}_0$. If $\mathcal{F}_0$ is too

[7] It is not clear whether the interesting result reported at this meeting by Goldberg *et al.* can be recast into a construction of a perennial.

large, the set of perennials may be too small: Some physically significant perennials may have been excluded by the requirement that they commute with the constraints on a larger-than-physical space of solutions. Further, it may happen that those perennials which remain cannot be made self-adjoint under an inner product on the whole solution space, but only on a drastically reduced space from which the 'unphysical' states have been excluded. In brief, it seems impossible to follow step by step the 'quantization program': firstly, to find the space of solutions without having the inner product to determine which states are physical, secondly, on that space of solutions to define the perennials, and thirdly, to find the inner product on $\mathcal{F}_0$ which makes all such perennials self-adjoint. Rather, the steps should be replaced by a single jump. As I am growing older, the difficulty of replacing three steps by a single jump is becoming more and more obvious.

The second standpoint is that observables do not need to commute with the Hamiltonian constraint, but only with the gauge constraints. If so, they do not act in the space of solutions: if $\Psi \in \mathcal{F}_0$ and $\hat{F}$ is an observable, $\hat{F}\Psi \notin \mathcal{F}_0$. To proceed, one should

- abandon the space of solutions and work instead in the space of instantaneous states.

To talk about instantaneous states requires a decision about what is an instant. An instant in a relativistic spacetime is a spacelike hypersurface. However, spacelike hypersurfaces are not elements of the gravitational phase space. The task is to find an observable T (or, rather, a set of ∞^3 commuting observables, to account for ∞^3 hypersurfaces) whose value uniquely fixes a hypersurface in a Ricci-flat spacetime generated by the evolution of the classical canonical data. Such an observable is called an *internal time*. (The adjective 'internal' means 'constructed solely from the phase-space variables'.)

The Hamiltonian constraint is interpreted as an evolution equation for Ψ in T. One tries to cut down $\mathcal{F}_0$ to a linear subspace $\mathcal{F}_0' \subset \mathcal{F}_0$ whose elements are in a one-to-one correspondence with the instantaneous values of Ψ: the restrictions Ψ_T of Ψ to a fixed hypersurface T. These restrictions are the instantaneous states $\Psi_T \in \mathcal{F}_T$. The program is to find an inner product in $\mathcal{F}_T$ which is independent of T, i.e., which is conserved in internal time. The discussion centers on how different forms of the Hamiltonian constraint (the Wheeler-DeWitt form, and others) suggest what such an inner product may be. A T-independent inner product can be interpreted as an inner product in $\mathcal{F}_0'$. In general, the observables $\hat{F}$ depend on T. One requires that they be factor ordered so that, at each T, they are self-adjoint under the inner product in $\mathcal{F}_T$. The expression $\langle \Psi_T | \hat{F} | \Psi_T \rangle$ is interpreted as the mean value of $\hat{F}$ in the state Ψ_T at the internal time T.

These things are more easily said than done. The internal time proposal meets as many difficulties as the approach based on the concept of perennials. I discussed the problems of time in a recent review [40] which complements my present treatment of observables.

It is sometimes maintained that the approach based on perennials somehow avoids the problems of time. It would be great if it did, but I fear it does not. A closer look reveals that the problems of time and the problem of perennials are rather closely related. A Czech saying has it that the devil thrown out of the door returns through a window.

7. Reality conditions

The connection dynamics looks in many respects simpler than geometrodynamics, but its simplicity has been bought at a price: the SO(3) connection A is necessarily complex. One needs to ensure that the quantum theory based on such a connection describes a real gravitational field.

One can attempt to accomodate complex objects in canonical gravity in two different ways:

Complexify the Einstein theory, i.e., work with complex metrics γ on a real spacetime manifold $\mathcal{M}$. The statement that $(\mathcal{M}, \gamma)$ is Ricci-flat amounts to a system of coupled equations for the real and imaginary parts of the complex metric γ. These equations can be derived from a real action whose Lagrangian is the real part of the complex curvature scalar. Introduce the Ashtekar variables A_a^i, E_i^a for the complexified spacetime. Both A and E are now complex. The canonical form of the action leads to the Poisson brackets among these variables and their complex conjugates.

To restrict the spacetime metric to be real, one imposes the condition that its imaginary part vanishes. In the canonical version of the theory, this imposes the reality conditions (27) on A and E. The reality conditions are preserved by the constraints: when the evolution starts from real canonical data, it continues building a real spacetime. However, the Poisson brackets among the reality conditions do not vanish: to put the imaginary part of the metric and its rate of change equal to zero amounts to requiring both a canonical coordinate and its conjugate momentum to vanish. It means that the reality conditions are, in Dirac's terminology, second-class constraints [3]. Such constraints must be eliminated before quantization. Unfortunately, their elimination destroys the new variables.

An alternative is to derive the complexified equations from a holomorphic Lagrangian [41]. The corresponding canonical theory knows how to form the Poisson brackets among A and E, but the Poisson brackets involving the complex conjugates $\bar{A}$ and $\bar{E}$ are undefined. The status of the reality conditions thus remains unclear and one does not know what to do with them on quantization.

Use complex chart on a real phase space. The second option is to consider A and E as a complex chart on a real phase space $(E, -K)$. This is similar to introducing a complex chart Q and $Z = Q - iP$ on the real phase space (Q, P) of a harmonic oscillator. The proposal [12, 14] is to ignore the reality conditions in the first five steps of the quantization program. In particular, the vector space $\mathcal{V}$ spanned by the fundamental variables A and E is allowed to be complex, and so are the dynamical variables $F[A, E] \in \mathcal{A}$ and the perennials $F[A, E] \in \mathcal{A}_0$.

One knows how to complex conjugate, $\bar{F}$, the elements F of the classical spaces $\mathcal{V}$ and $\mathcal{A}$. The task is to define the corresponding operation, $\star$, on the elements $\hat{F}$ in $\hat{\mathcal{V}}$ and $\hat{\mathcal{A}}$. Ashtekar's proposal is first to define the $\star$ operation in $\hat{\mathcal{V}}$ by requiring that complex conjugate elements of $\mathcal{V}$ are carried into the $\star$- related elements of $\hat{\mathcal{V}}$:

$$F, \bar{F} \in \mathcal{V} \quad \Longrightarrow \quad \hat{F} = \hat{\bar{F}}^\star . \tag{52}$$

The $\star$ operation is then extended from $\mathcal{V}$ to $\mathcal{A}$ by using the axioms of the involution operation:

$$(a\hat{F} + b\hat{G})^\star \quad = \quad \bar{a}\hat{F}^\star + \bar{b}\hat{G}^\star ,$$

$$\begin{aligned} (\hat{F}\hat{G})^\star &= \hat{G}^\star\hat{F}^\star\,, \\ (\hat{F}^\star)^\star &= \hat{F}\,, \\ \forall \hat{F},\hat{G}\in\mathcal{A} \quad &\text{and} \quad \forall a,b\in\mathbb{C}\,. \end{aligned} \tag{53}$$

If $\hat{F}^\star = \hat{F}$, the operator $\hat{F}$ represents a real dynamical variable. If there are no constraints, this dynamical variable is an observable. The expectation value of $\hat{F}$ should be real. This objective can be achieved by requiring that the inner product $\langle\Psi_1|\Psi_2\rangle$ in $\mathcal{F}$ be such that it makes all $\star$-related operators Hermitian adjoints,

$$\hat{F} = \hat{G}^\star \implies \langle\Psi_1|\hat{F}\Psi_2\rangle = \langle\hat{G}\Psi_1|\Psi_2\rangle\,, \tag{54}$$

and hence all operators representing real variables self-adjoint. If the $\star$ operation in $\hat{\mathcal{A}}$ is determined by the $\star$ operation in $\hat{\mathcal{V}}$ as in Eqs.(52) and (53), it is sufficient to require that the condition (54) holds for all fundamental variables $\hat{F},\hat{G}\in\hat{\mathcal{V}}$.

Canonical gravity, however, is a constrained system. Ashtekar's program assumes that only perennials can be observed, and that their expectation values are obtained from an inner product on the space of solutions. To impose the reality conditions, one needs to define the $\star$ operation for perennials. This would be straightforward if the $\star$ operation from $\hat{\mathcal{A}}$ could be restricted to perennials. Unfortunately, this does not need to be the case: if $\hat{F}$ is a perennial, $\hat{F}^\star$ does not need to be a perennial (though it may be a perennial under special circumstances). The hope is that there is a 'sufficient' number of perennials $\hat{F}$ whose $\star$-adjoints $\hat{F}^\star$ are also perennials. By 'sufficient' one means that the condition (54), when imposed on these perennials, uniquely determines the inner product in $\mathcal{F}_0$.

To summarize, Ashtekar's program calls for implementing the reality conditions as requirements on the inner product in the space of solutions $\mathcal{F}_0$. Firstly, one must find a sufficient number of $\star$-adjoint perennials, and then require that these be Hermitian adjoints under the inner product.

One can ask two questions about this proposal. The first is whether it works for simple model systems. The second is whether it can reasonably be expected to work in canonical gravity.

The answer to the first question is yes. Ashtekar's proposal determines the inner product for a number of simple systems (a harmonic oscillator with complex chart, a parametrized Newtonian particle, a free relativistic particle on a flat background). It also works for $2+1$ gravity and linear field theories on a $(3+1)$-dimensional flat Lorentzian background, including Maxwell's electrodynamics and linearized gravity. With the exception of $2+1$ gravity (which does not have any field degrees of freedom) these examples are reducible to collections of harmonic oscillators.

To approach the second question, one should ask whether there are any relevant differences between the prototype of a linear harmonic oscillator and full canonical gravity. (By 'relevant' I mean relevant to the proposal on handling the reality conditions.) I feel there are two such differences:

In the harmonic oscillator problem, one works with the fundamental variables Q and $Z = Q - iP$, which are analogous to E and A in connection dynamics. The vector space $\mathcal{V}$ is spanned on Q, Z, and 1. The reality conditions are the conditions

$$\bar{Q} = Q \quad \text{and} \quad \bar{Z} = -Z + 2Q \tag{55}$$

on the dynamical variables $F = Q$, $G = Z$, and their complex conjugates $\bar{F}$ and $\bar{G}$. Both F and G, and $\bar{F}$ and $\bar{G}$ lie in $\mathcal{V}$. It is thus possible to define the $\star$ on $\mathcal{A}$ by Eqs.(52) and (53), and to impose the reality condition (54).

In connection dynamics, $\mathcal{V}$ is spanned by A, E and 1. The second reality condition (27), however, is not a condition on the elements of $\mathcal{V}$, because $\Gamma[E]$ is a non-linear functional of E. (The same remark applies to polynomial forms of reality conditions.) This prevents one from defining the $\star$ operation on $\mathcal{V}$, as in Eq.(52), and from extending it to $\mathcal{A}$, as in Eq.(53).

This difficulty can be clarified on simple models. Take a one-dimensional system with the Hamiltonian [8]

$$h := QP^2 + Q^{-2} \tag{56}$$

and introduce the complex chart (Q, Z), with

$$Z = Q^{-1} - iP, \tag{57}$$

on the real phase space (Q, P). The Hamiltonian (56) becomes polynomial in Q and Z,

$$h = -QZ^2 + 2Z. \tag{58}$$

The reality condition on Z can be written either in a non-polynomial form linear in Z, or in a polynomial form:

$$\tfrac{1}{2}(Z + \bar{Z}) = Q^{-1}, \quad \text{or} \quad Q(Z + \hat{Z}) = 2. \tag{59}$$

Whichever form we use, it is not a condition in the complex vector space $\mathcal{V}$ spanned by the fundamental variables Q and Z.

This is my first reason for believing that the harmonic oscillator is not quite representative of canonical gravity. The algorithm for handling reality conditions needs to be checked on more general models than those which have been investigated so far, like the model I have just described.

The second difference between the oscillator and canonical gravity is that the later is a parametrized theory. Ashtekar's proposal on how to handle the reality conditions depends on the existence of a sufficient number of perennials, and on the possibility to define a $\star$ operation on their algebra. I expressed my doubts that there exists a sufficient number of perennials in canonical gravity. Even if there is a sufficient number of perennials, it remains unclear whether it is possible to extend the $\star$ operation from $\hat{\mathcal{V}}$ to $\hat{\mathcal{A}}$, and then to restrict it to a suitable subset of perennials.

I do not claim that these problems are insurmountable, but I feel that they represent a major unsolved problem of connection dynamics.

8. Conclusions

Where do we stand? We certainly gained in the years a good geometric understanding of classical general relativity as a canonical dynamical system. In quantum theory, we

[8] In the sector $Q > 0$, the Hamiltonian (56) can be brought into the form $h = \frac{1}{2}p^2 + 4q^{-4}$ by the canonical transformation $q = \sqrt{2}\,Q^{\frac{1}{2}}$, $p = \sqrt{2}\,Q^{\frac{1}{2}}P$.

inherited a set of rules of thumb called Dirac constraint quantization. They were never precise, and Dirac himself never claimed they were much more than rules of thumb. People tried to make them more precise and they ended with something resembling the seven gates I described.

Let me revisit those gates and ask what steps in the quantization program have actually been accomplished. And, even more importantly, let me summarize what are the main unsolved problems.

1. Different sets of fundamental variables (not only those which I mentioned in this report) have been explored and understood. We also know how to take care of the positivity restrictions on the metric variables [13].

2. Most of the work on turning constraints into operators is formal. Both the regularization problem and consistency problem remain open. Very little is known about how to handle other dynamical variables, especially the future candidates for observables or perennials.

3. Important work has been done on clarifying the mathematical status of the states $\Psi[g]$ and $\Psi[A]$ and of the fundamental operators [13]. The connection representation has been linked to the loop representation [42]. One should note that the latter investigation has been successful only for *real* connections.

4. Because the regularization and consistency problems for the constraints have not been satisfactorily resolved, all attempts to find the states which solve the quantum constraints (30) are to a large extent formal. It is notable that connection dynamics actually exhibited a large number (indeed, infinitely many) such solutions. Most of these were obtained in the loop representation and lie outside the scope of this report [16]. When comparing this success with the lack of solutions in geometrodynamics, one should keep in mind that these solutions correspond to degenerate metrics which geometrodynamics excludes. One solution that can be written directly in the connection representation is the exponential of the Chern-Simons form [43]. Passing from particular solutions to general considerations, it is not clear what boundary or other conditions should be imposed on the solutions $\Psi \in \mathcal{F}_0$ to select the true physical states.

5. The problem of what quantities can be observed (and how they can be observed) is one of the most intriguing and important questions in quantum gravity. A widely held view (which I dispute) is that one can observe only perennials. No true perennials, classical or quantum, have so far been found, and even if they exist, finding them is difficult. I feel we should instead concentrate on formulating and proving (non?)existence theorems about perennials.
   Unlike perennials, there are many concrete examples of classical observables. It is, however, obscure what classical observables are to be represented by operators, and on what space these operators act. This is connected with the problem of time: one does not expect the time observable to be represented in quantum mechanics by an operator.

6. Another outstanding problem of canonical quantum gravity is the construction of the inner product. Quantum geometrodynamics has been unsuccessful in this task

[22, 36], and connection dynamics has hardly done more than formulate broad guidelines on how one might try to proceed. These guidelines crucially depend on the existence of perennials.
In contrast, one knows how to construct (at the formal level) the inner product for parametrized field theories [17]. Each choice of an internal time casts canonical gravity into the mold of a parametrized field theory and leads to an inner product. The procedure, however, is not without problems [40]. One which is closely related to the problem of perennials is that internal time may not exist globally [44].

7. Connection dynamics, unlike geometrodynamics, needs to take care of reality conditions. Ashtekar's proposal is to impose them as requirements which determine the inner product. Two problems arise: firstly, the necessity of finding a complete set of perennials and defining on them the $\star$ operation and, secondly, the high polynomiality of the reality conditions, which takes them out of the realm of the fundamental vector space $\mathcal{V}$.
The reality conditions are the only major problem which does not exist in geometrodynamics. The ability of connection dynamics to handle this problem will be crucial for judging its success in the Galilean contest between the two chief systems of canonical gravity.

The problem of reality conditions exemplifies the general pitfall of any quantization program. As I described it, the program resembles seven doors to the law, each of them guarded by a doorkeeper. We certainly did not sit on a stool at the side of the first door for days and years: we tried to enter the law. However, on our way through the doors we learned that their orderly sequence is deceptive. One can never be sure of passing a door before all have been passed. The entries are so interconnected that they cannot be made separately: What is a solution of the quantum constraints depends on the choice of fundamental variables and the form of the constraints. What solutions are physical depends on the inner product. What is an inner product depends on what quantities are observable. What quantities are observable may depend on what solutions are physical. More often than not we are caught in a vicious circle which calls for entering all the doors at once.

This may be frustrating, but it should have been expected. Indeed, it would be rather disappointing if one could reach a truly fundamental theory like quantum gravity by following step by step a travel guide, or its medieval predecessor, a pilgrim's itinerary to a wholy shrine. In this spirit, let me end my account of canonical quantum gravity in the dark ais les of St. Vít's cathedral of my native city of Prague, talking to a priest [45]:

> "You have studied the story more exactly and for a longer time than I have," said K. They were both silent for a while. Then K. said: "So you think that the man was not deceived?" "Don't misunderstand me," said the priest, "I am only showing you the various opinions concerning that point. You must not pay too much attention to them. The scriptures are unalterable and the comments often enough merely express the commentators' despair."

Acknowledgments

The work on this report has been partially supported by the NSF grants PHY-9207225 and INT-8901512 to the University of Utah. I want to thank Julian Barbour for his careful reading of the final draft of the paper.

References

[1] Gauss K F 1828 *Disquisitiones generales circa superficies curvas* (Gottingae: Typis Dieterichiansis)
1965 *General Investigations of Curved Surfaces of 1827 and 1825* trans J C Morehead and A M Hiltebeitel (New York: Raven Press)

[2] Codazzi D 1869 *Annali di matem. ser.2* **2** 269

[3] Dirac P A M 1964 *Lectures on Quantum Mechanics* (New York: Yeshiva University); and references therein.

[4] Arnowitt, R, Deser, S and Misner C W 1962 The dynamics of general relativity *Gravitation: An Introduction to Current Research* ed L Witten (New York: Wiley); and references therein.

[5] Deser S and Isham C J 1976 *Phys. Rev.* **D14** 2505
Nelson J E and Teitelboim C 1976 *Ann. Phys., NY* **116** 86
Henneaux M 1983 *Phys. Rev.* **D27** 986
Henneaux M, Nelson J E and Schombland C 1989 *Phys. Rev.* **D39** 437

[6] Wheeler J A 1968 Superspace and the nature of quantum geometrodynamics *Batelle Rencontres: 1967 Lectures in Mathematics and Physics* ed C DeWitt and J A Wheeler (New York: Benjamin)

[7] Sen A 1981 *J. Math. Phys.* **22** 1781

[8] Ashtekar A 1987 *Phys. Rev.* **D36** 1587

[9] Moncrief V Private communications
Tate R S 1992 *Class. Quantum Grav.* **9** 101

[10] Ashtekar A, Romano J D and Tate R S 1989 *Phys. Rev.* **D40** 2572

[11] Bergmann P G 1956 *Helv. Phys. Acta Suppl.* **4** 79
Kundt W 1966 *Canonical Quantization of Gauge Invariant Field Theories* (Berlin: Springer)
Brill D R and Gowdy R H 1970 *Rep. Progress Phys.* **33** 413
Kuchař K V 1973 Canonical quantization of gravity *Relativity, Astrophysics and Cosmology* ed W Israel (Dordrecht: Reidel)
Geroch R and Ashtekar A 1974 *Rep. Prog. Phys.* **37** 1211
Kuchař K V 1981 Canonical methods of quantization *Quantum Gravity 2: A Second Oxford Symposium* ed C J Isham *et al.* (Oxford: Clarendon)
Isham C J 1987 Quantum gravity *General Relativity and Gravitation: Proceedings of the 11th International Conference on General Relativity and Gravitation* (Cambridge: University Press)
Isham C J 1992 Conceptual and geometrical problems in canonical quantum gravity *Recent Aspects of Quantum Fields* ed H Mitter and H Gausterer (Berlin: Springer)

[12] Ashtekar A 1991 *Lectures on Non-Perturbative Canonical Gravity* (Singapore: World Scientific)

[13] Isham C J 1984 Topological and global aspects of quantum theory *Relativity, Groups and Topology II* ed B S DeWitt and R Stora (Amsterdam: North Holland)
Isham C J and Kakas A 1984 *Class. Quantum Grav.* **1** 621; 623

[14] Tate R S 1992 An algebraic approach to the quantization of constrained systems: Finite dimensional examples. PhD Thesis *Syracuse University Preprint* SU-GP-92/8-1

[15] Van Hove L 1951 *Acad.Roy. Belg. Bull. Cl. Sci.* **37** 610

[16] Rovelli C 1991 *Class. Quantum Grav.* **8** 1613

[17] Kuchař K V 1981 Canonical methods of quantization *Quantum Gravity 2: A Second Oxford Symposium* ed C J Isham *et al.* (Oxford: Clarendon)

[18] Kuchař K V 1986 *Phys. Rev.* **D34** 3031; 3044
Kuchař K V 1988 *Cont. Math.* **71** 285
Hajíček P and Kuchař K V 1990 *Phys. Rev.* **D41** 1091

[19] Anderson J 1959 *Phys. Rev.* **114** 1182
Anderson J 1962 Q-number coordinate transformations and the ordering problem in general relativity *Proceedings of the First Eastern Theoretical Physics Conference, October 26-27, 1962* ed M E Ross (New York: Gordon and Breach)

[20] Schwinger J 1963 *Phys. Rev.* **132** 1317

[21] Dirac P A M The quantization of the gravitational field *Contemporary Physics: Trieste Symposium 1968, vol. 1* ed L Fonda (Vienna: IAEA 1969)

[22] DeWitt B S 1967 *Phys. Rev.* **160** 1113

[23] Komar A 1979 *Phys. Rev.* **D19** 2908; **D20** 830

[24] Tsamis N C and Woodard R P 1987 *Phys. Rev.* **D36** 1587

[25] Friedman J L and Jack I 1988 *Phys. Rev.* **D37** 3495

[26] Isham C J 1992 Conceptual and geometrical problems in canonical quantum gravity *Recent Aspects of Quantum Fields* ed H Mitter and H Gausterer (Berlin: Springer)

[27] Bargmann V 1962 *Proc. Natl. Acad. Sci. (U.S.A.)* **48** 199
Klauder J R and Sudarshan E C G 1968 *Fundamentals of Quantum Optics* (New York: Benjamin)

[28] Newman E and Bergmann P G 1957 *Rev. Mod. Phys.* **33** 510
Bergmann P G 1961 *Rev. Mod. Phys.* **33** 510
Bergmann P G and Komar A B 1962 Observables and commutation relations *Les Theories Relativistes de la Gravitation* (Paris: CNRS)

[29] Lanczos C 1970 *The Variational Principles of Mechanics* 4th ed (Toronto: University of Toronto)
Synge J L 1960 Classical dynamics *Handbuch der Physik, vol.III/1* ed S Flügge (Berlin: Springer)

[30] Kuchař K V 1988 Canonical quantization of generally covariant systems *Highlights in Gravitation and Cosmology* ed B R Iyer *et al.* (Cambridge: University Press)

[31] Rovelli C 1991 *Phys. Rev.* **D43** 442

[32] Russell B 1945 *A History of Western Philosophy* (New York: Simon and Schuster)

[33] Thorne K S *et al.* 1979 *Sources of Gravitational Radiation* ed L Smarr (Cambridge: University Press)
Caves C M *et al.* 1980 *Rev. Mod. Phys.* **52** 341

[34] Poincaré H 1892/1893/1899 *Les méthodes nouvelles de la méchanique céleste, vols.1-3* (Paris: Gauthier-Villars); 1957 (New York: Dover)

[35] Arnold V I, Kozlov V V and Neishtadt A J 1988 Mathematical aspects of classical and celestial mechanics *Dynamical Systems III* ed V J Arnold trans A Iacob (Berlin: Springer)

[36] Kuchař K V 1981 *J. Math. Phys.* **22** 2640. Contains a proof that there is no perennial linear in the extrinsic curvature.

[37] Anderson I and Torre C G 1992 Symmetries of the Einstein Equations *Work in Progress.* It is shown that the vacuum Einstein equations do not have any generalized symmetries except spacetime diffeomorphisms and constant scalings of the metric. On the other hand, all known completely integrable field theories have non-trivial generalized symmetries.

[38] Warner M 1976 *Alone of Her Sex. The Myth and Cult of Virgin Marry* (New York: Knopf)

[39] See [12], Chapter 10: The quantization program

[40] Kuchař K V 1992 Time and interpretations of quantum gravity *Proceedings of the 4th Canadian Conference on General Relativity and Relativistic Astrophysics* ed G Kunstatter *et al.* (Singapore: World Scientific)
A complementary review of the problem of time is
Isham C J 1992 Canonical quantum gravity and the problem of time, Lectures presented at the NATO Advanced Summer Institute on 'Recent Problems in Mathematical Physics', Salamanca, June 15-27, 1992 *Imperial College Preprint* Imperial TP/91-92/25

[41] Samuel J 1987 *Pramana J. Phys* **28** L429
Jacobson T and Smolin L 1988 *Class. Quantum Grav.* **5** 583

[42] Ashtekar A and Isham C J 1992 *Phys. Lett* **274B** 393
Ashtekar A and Isham C J 1992 *Class. Quantum Grav.* **9** 1433

[43] Kodama H 1990 *Phys. Rev.* **D42** 2548

[44] Torre C G 1992 1992 *Phys. Rev.* **D46** R3231

[45] Kafka F 1946 *Der Prozess. Gessammelte Schriften Band III* (New York: Schocken)
Kafka F 1960 *The Trial* trans W and E Muir (New York: Knopf)

Paper presented at the 13th Int. Conf. on General Relativity and Gravitation
Cordoba, Argentina, 1992: Part 1, Plenary Lectures

Recent results from COBE

John C Mather†[1], Charles L Bennett†, Nancy W Boggess†, Michael G Hauser†, George F Smoot‡, and Edward L Wright§

† NASA Goddard Space Flight Center, Laboratory for Astronomy & Solar Physics, Greenbelt, MD 20771

‡ Bldg 50-351, Lawrence Berkeley Labs, Berkeley, CA 94720

§Astronomy Department, UCLA, Los Angeles, CA 90024

Abstract. The Cosmic Background Explorer (*COBE*[2]), NASA's first space mission devoted primarily to cosmology, carries three scientific instruments to make precise measurements of the spectrum and anisotropy of the cosmic microwave background (CMB) radiation on angular scales greater than 7° and to conduct a search for a diffuse cosmic infrared background (CIB) radiation with 0.7° angular resolution. The observing strategy is designed to minimize and allow determination of systematic errors that could result from spacecraft operations, the local environment of the spacecraft, and emissions from foreground astrophysical sources such as the Galaxy and the solar system. Data from the Far-InfraRed Absolute Spectrophotometer (FIRAS) show that the spectrum of the CMB is that of a blackbody of temperature T=2.73±0.06 K, with no deviation from a blackbody spectrum greater than 0.25% of the peak brightness. The first year of data from the Differential Microwave Radiometers (DMR) show statistically significant CMB anisotropy. The anisotropy is consistent with a scale invariant primordial density fluctuation spectrum and with the gravitational potential variations required to cause the observed present day structure. Infrared sky brightness measurements from the Diffuse InfraRed Background Experiment (DIRBE) provide new conservative upper limits to the CIB. Extensive modeling of solar system and galactic infrared foregrounds is required for further improvement in the CIB limits.

[1] E-mail: mather@stars.gsfc.nasa.gov (Internet).

[2] The National Aeronautics and Space Administration/Goddard Space Flight Center (NASA/GSFC) is responsible for the design, development, and operation of the *COBE*. Scientific guidance is provided by the *COBE* Science Working Group. GSFC is also responsible for the development of the analysis software and for the production of the mission data sets.

1. Introduction to the *COBE* and mission objectives

The observables of modern cosmology include the Hubble expansion of the universe; the ages of stars and clusters; the distribution and streaming motions of galaxies; the content of the universe (its mass density, composition, and the abundances of the light elements); the existence, spectrum and anisotropy of the cosmic microwave background radiation; and other potential backgrounds in the infrared, ultraviolet, x-ray, gamma-ray, etc. The purpose of the *COBE* mission is to make definitive measurements of two of these observable cosmological fossils: the cosmic microwave background (CMB) radiation and the cosmic infrared background (CIB) radiation. Since the discovery of the CMB in 1964 (Penzias & Wilson 1965), many experiments have been performed to measure the CMB spectrum and spatial anisotropies over a wide range of wavelengths and angular scales. Fewer attempts have been made to conduct a sensitive search for a CIB radiation, expected to result from the cumulative emissions of luminous objects formed after the universe cooled sufficiently to permit the first stars and galaxies to form.

In 1974 NASA issued Announcements of Opportunity (AO-6 and AO-7) for new Explorer class space missions. A proposal for a Cosmic Background Radiation Satellite was submitted by John Mather *et al.*(1974) from NASA/Goddard. The objectives of this mission were: (1) make "definitive measurement of the spectrum (of the 2.7 K CBR).. with precision of 10^{-4} around the peak... It will also look for the emission from cold dust clouds and from infrared galaxies"; (2) "measure the large scale isotropy of the background radiation... to a precision of 10^{-5}... Measurements at several wavelengths are required in order to distinguish anisotropy in the background radiation itself from anisotropy due to discrete sources"; and (3) "... search for diffuse radiation in the 5-30 micron wavelength range, expected to arise from interplanetary dust, interstellar dust, and, in particular, from the integrated luminosity of very early galaxies. The experiment is designed to separate these contributions by their spectral and directional properties." Additional proposals were also submitted for large angular scale microwave isotropy experiments by Sam Gulkis *et al.*(1974) from JPL and by Luis Alvarez *et al.*(1974) from UC Berkeley. NASA selected six investigators from these proposals and formed the core of what was to become the COBE Science Working Group, as shown in Table 1.

To achieve the full benefit of space observations, a goal of the mission and instrument design was that *COBE* measurements would be limited ultimately by our ability to identify and model the various components of the astrophysical foreground emission, and discriminate between them and the cosmological emission. This goal drove the design of the mission strategy, the spacecraft and operations, and the choice of instruments. Basic elements in the mission strategy were the requirements for highly redundant full sky coverage and for sufficient time in orbit to achieve necessary sensitivity and evaluate potential sources of systematic errors in the observations. The instruments were designed to measure specific attributes of the cosmological backgrounds and also, through their complementary spectral coverage, to enable the modeling and subtraction of foreground emissions. Early descriptions of the mission concept have been given by Mather (1982, 1987) and by Gulkis *et al.*(1990).

The three scientific instruments are the Far Infrared Absolute Spectrophotometer (FIRAS), the Differential Microwave Radiometers (DMR), and the Diffuse Infrared Background Experiment (DIRBE). The FIRAS objective is to make a precision mea-

Table 1. *COBE* Science Working Group

Bennett, C. L.	NASA-GSFC	DMR Deputy Principal Investigator
Boggess, N. W.	NASA-GSFC	*COBE* Deputy Project Scientist
Cheng, E. S.	NASA-GSFC	*COBE* Deputy Project Scientist
Dwek, E.	NASA-GSFC	
Gulkis, S.	NASA-JPL	
Hauser, M. G.	NASA-GSFC	DIRBE Principal Investigator
Janssen, M.	NASA-JPL	
Kelsall, T.	NASA-GSFC	DIRBE Deputy Principal Investigator
Lubin, P. M.	U.C.S.B.	
Mather, J. C.	NASA-GSFC	FIRAS Principal Investigator & COBE Project Scientist
Meyer, S. S.	M.I.T.	
Moseley, S. H.	NASA-GSFC	
Murdock, T. L.	Gen. Res. Corp.	
Shafer, R. A.	NASA-GSFC	FIRAS Deputy Principal Investigator
Silverberg, R. F.	NASA-GSFC	
Smoot, G. F.	LBL & UCB	DMR Principal Investigator
Weiss, R.	M.I.T.	*COBE* SWG Chairman
Wilkinson, D. T.	Princeton	
Wright, E. L.	U.C.L.A.	

surement of the spectrum of the CMB from 1 cm to 100 μm. The DMR objective is to search for CMB anisotropies on angular scales larger than 7° at frequencies of 31.5, 53, and 90 GHz. The DIRBE objective is to search for a CIB by making absolute brightness measurements of the diffuse infrared radiation in 10 photometric bands from 1 to 240 μm and polarimetric measurements from 1 to 3.5 μm. The FIRAS and DIRBE instruments are located inside a 650 liter superfluid liquid helium dewar.

2. *COBE* mission design & implementation

A full description of the *COBE* mission by Boggess *et al.*(1992) is summarized here. Many papers giving overviews, implications, and additional detailed information about the *COBE* have been presented by Mather *et al.*(1990b), Mather *et al.*(1991b), Mather (1991), Janssen & Gulkis (1992), Wright (1990), Wright (1991a), Hauser (1991a), Hauser (1991b), Smoot *et al.*(1991c), Smoot (1991), Bennett (1991), and Boggess (1991).

The need to control and measure potential systematic errors led to the requirements for an all-sky survey and a minimum time in orbit of six months. The instruments required temperature stability to maintain gain and offset stability, and a high level of cleanliness to reduce the entry of stray light and thermal emission from particulates. The control of systematic errors in the measurement of the CMB anisotropy and the

need for measuring the interplanetary dust cloud at different solar elongation angles for subsequent modeling required that the satellite rotate.

In near-Earth orbit, the Sun and Earth are the primary continuous sources of thermal emission and it was necessary to ensure that neither the instruments nor the dewar were exposed to their radiation. A circular Sun-synchronous orbit satisfied these requirements. An inclination of 99° and an altitude of 900 km were chosen so that the orbital plane precesses 360° in one year due to the Earth's gravitational quadrupole moment. The 900 km altitude is a good compromise between contamination from the Earth's residual atmosphere, which increases at lower altitude, and interference due to charged particles in the Earth's radiation belts at higher altitudes. A 6 PM ascending node was chosen for the *COBE* orbital plane; this node follows the terminator (the boundary between sunlight and darkness on the Earth) throughout the year. By maintaining the spacecraft spin axis at about 94° from the Sun and close to the local zenith, it is possible to keep the Sun and Earth below the plane of the instrument apertures for most of the year. However, since the Earth's axis is tilted 23.5° from the ecliptic pole, the angle between the plane of the *COBE*'s orbit and the ecliptic plane varies through the seasons from -14.5° to +32.5°. As a consequence, the combination of the tilt of the Earth's axis, the orbit inclination, and the offset of the spacecraft spin axis from the Sun brings the Earth limb above the instrument aperture plane for up to 20 minutes per orbit near the June solstice. During this period the limb of the Earth rises a few degrees above the aperture plane for part of each orbit, while on the opposite side of the orbit the spacecraft goes into the Earth's shadow. In the nominal *COBE* orbit the spacecraft's central axis scans the full sky, though not with uniform coverage, every six months. The orbital period is 103 minutes, giving 14 orbits per day.

A 3-axis attitude control system was implemented by using a pair of inertia wheels (yaw angular momentum wheels), with their axes oriented along the spacecraft spin axis. These wheels carry an angular momentum opposite that due to the spacecraft rotation to create a nearly zero net angular momentum system. The spacecraft orientation is controlled by three reaction wheels with spin axes 120° apart in the plane perpendicular to the spacecraft spin axis and by electromagnetic coils (torquer bars) that interact with the Earth's magnetic field. Earth and Sun sensors (one of each on each of the three transverse control axes) provide control signals to point the spin axis away from the Earth and at least 90° from the Sun. Rate damping and fine resolution attitude sensing are provided by six gyros, one on each transverse control axis and three on the spin axis. Coarse attitude parameters are calculated by using telemetered data from the attitude control sensors to produce attitude solutions good to 4 arcmin (1 σ). A fine aspect is determined by using gyro data to interpolate between the positions of known stars detected in the short wavelength bands of the DIRBE instrument. The fine aspect solution has an accuracy of 1.5 arcmin (1 σ) and is now used in the analysis of data from all three instruments (Kumar *et al.*1991).

The FIRAS instrument, located inside the dewar, points along the spin axis with its 7° field of view. The three pairs of DMR receivers are spaced 120° apart around the aperture plane of the dewar. Each radiometer channel measures the difference in sky signal from a pair of horn antennas defining 7° fields of view separated by 60°, each beam being 30° from the spin axis. The spin causes a short-term interchange of the two

beams associated with a single differential radiometer and thereby gives a modulation of the differential sky signal at the spin rate. The 0.8 rpm spin rate was chosen to be fast enough to reduce the noise and systematic errors that could otherwise arise from radiometer gain and offset instabilities. The DIRBE, also located inside the dewar, views 30° from the spin axis. The spin allows DIRBE to measure the emission and scattering by the interplanetary dust cloud over a range of solar elongation angles for each celestial direction, which aids in the discrimination and subsequent modelling of zodiacal radiation. DMR and DIRBE trace out a pattern of epicycles that enable them to scan half of the sky every day and obtain multiple measurements for each pixel of the sky.

The dewar is a 650 liter superfluid helium cryostat that kept the FIRAS and DIRBE instruments cooled to ~1.6 K. A deployable dewar aperture cover protected the cryogen and permitted calibration and performance testing of the cryogenic instruments prior to launch. A contamination shield attached to the inside of the dewar cover protected the DIRBE primary mirror from particulate or gaseous contamination until ejection of the dewar cover in orbit. It also protected DIRBE from emission from warm parts of the cryostat during ground testing. The deployed conical Sun-Earth shield protects the scientific instruments from direct solar and terrestrial radiation and provides thermal isolation for the dewar. The shield also provides the instruments isolation from Earth-based radio frequency interference (RFI) and from the spacecraft transmitting antenna. The shield was designed to be flexible and was folded to fit within the Delta rocket fairing for launch. Contamination covers attached to the Sun-Earth shield were placed over the DMR horn antennas and were pulled away in orbit by the deployment of the shield. The deployed solar arrays provide the nominal spacecraft and instrument power load of 542 Watts.

The *COBE* has two omnidirectional antennas, one to communicate with the Tracking and Data Relay Satellite System (TDRSS), and the other to transmit data stored on tape recorders directly to the ground. The antennas are located on a mast at the bottom of the spacecraft deployed after launch. The *COBE* has a command and data handling system that stores and decodes the commands received from the ground, collects data from the instruments and spacecraft at the rate of 4 kbps, and prepares data for transmission to the ground. The on-board tape recorders and data system allow 24 hours of data to be transmitted to the Wallops Flight Facility in a single 9 minute pass. The data rate allocations for DIRBE, FIRAS, and DMR are 1716, 1362 and 250 bps, respectively. The remainder of the telemetry is assigned to spacecraft subsystems.

The *COBE*, as initially proposed, was to have been launched by a Delta rocket. However, once the design was underway, the Shuttle was adopted as the NASA standard launch vehicle. After the Challenger accident occurred in 1986, ending plans for Shuttle launches from the West Coast, the spacecraft was redesigned to fit within the weight and size constraints of the Delta. The final *COBE* satellite had a total mass of 2,270 kg, a length of 5.49 m, and a diameter of 2.44 m with Sun-Earth shield and solar panels folded (8.53 m with the solar panels deployed).

Ground testing of the *COBE* was necessary to demonstrate that the individual subsystems, and ultimately the entire spacecraft and instrument assembly in its flight configuration, could satisfy both the sensitivity and systematic error requirements. System

level tests were performed: to simulate the space environments, including vacuum and temperature; to determine susceptibility to vibration, acoustic excitation, and acoustic shock; to quantify electromagnetic interference (EMI) self-compatibility and RFI susceptibility; to determine the interaction between instruments and spacecraft; to simulate the thermal and power conditions that would occur during eclipse periods; to test the deployables and moving parts (Sun-Earth shield, antenna boom, solar panels, dewar cover, FIRAS external calibrator and moving mirror transport, and DIRBE shutter and chopper); and characterize and calibrate the instruments.

Technical papers on various aspects of the *COBE* have been published. These include papers on contamination control (Barney 1991); test facility requirements for thermal balance tests (Milam 1991); design of the dewar (Hopkins & Castles 1985); optical alignments (Sampler 1990); thermal performance of the dewar prior to launch and in orbit (Hopkins & Payne 1987, Volz & Ryschkewitsch 1990a, Volz *et al.*1990, Volz *et al.*1991a & 1991b, Volz & Dipirro 1992); thermal design of the cryogenic optical assembly (Mosier 1991a, 1991b); cryogenic cool-down tests (Coladonato *et al.*1990), and attitude control (Bromberg & Croft 1985).

3. *COBE* operations in orbit

The *COBE* was launched aboard Delta rocket No. 189 at 1434 UT on November 18, 1989 from the Western Space and Missile Center at Vandenberg Air Force Base, California. The DMR receivers began operating the day after launch. The dewar cover was ejected three days after launch, and the FIRAS and DIRBE instruments began obtaining data on the same day. During the first month in orbit, various tests were undertaken to evaluate the performance of the instruments and spacecraft, and to optimize instrument parameters.

The *COBE* operated in a routine survey mode. The three instruments completed their first full sky coverage by mid-June 1990, and returned high quality data until the depletion of the liquid helium at 0936 UT September 21, 1990. The FIRAS, which had surveyed the sky 1.6 times, ceased operating when the helium ran out, but the DMR is still operating normally in all of its six channels. By November 1991 (over one year after helium depletion) the dewar temperature at the DIRBE detectors was about 50 K. The six longest wavelength bands were turned off in September 1990, but the four short wavelength bands of the DIRBE continue to acquire data at reduced sensitivity. The detector system responsivity in the short wavelength bands decreased about an order of magnitude following cryogen depletion (largely due to the change in load resistance). However, sky maps of the large scale interplanetary dust signals are of adequate quality to permit searching for evidence of temporal changes on annual time scales.

In flight, the helium temperature inside the main cryogen tank was 1.40 K and the temperature of the inner surface of the Sun-Earth shield was 180 K. As expected, the Earth limb rose a few degrees above the Sun-Earth shield for a part of every orbit during a three month period starting in May. At these times, the Earth's radiation produced thermal transients in the instruments and adversely affected data for a portion of each orbit. Some of these data are still usable after careful calibration. One of the gyros for a transverse control axis failed electrically on the fourth day after launch. On September

7, 1991, one of the three gyros on the spin axis failed, but no data were lost and satellite operations continue in the nominal orbit.

4. Spectral results from FIRAS

4.1. The spectrum of the primeval radiation

The discovery of the cosmic microwave background radiation by Penzias & Wilson (1965) provided strong evidence for Big Bang cosmology. Radiation produced in the very early universe was frequently scattered until about 300,000 years after the Big Bang. At this point, the "recombination", the characteristic energy in the universe fell to the point where previously free electrons could combine with nuclei to form neutral atoms. The 2.7 K radiation we see today has been traveling to us unimpeded since that time. The rapid production and destruction of photons within the first year after the Big Bang forced the radiation to have a Planck (blackbody) spectrum. Any mechanism that injected energy into the Universe (e.g. a particle decay) between a year after the Big Bang and $\sim$ 2000 years after the Big Bang would give rise to a radiation spectrum characterized by a non-zero chemical potential. Thus there would be a Bose-Einstein spectral distortion with the photon occupation number

$$N(\epsilon) \sim \frac{1}{e^{(\epsilon-\mu)/kT} - 1} \tag{1}$$

where μ is the chemical potential and ϵ is the photon energy. A Compton distortion is usually parameterized in terms of a Compton y-parameter,

$$y = \frac{\sigma_T}{m_e c^2} \int n_e k (T_e - T_{CMB}) c dt \tag{2}$$

where σ_T is the Thomson scattering cross-section and the integral is the electron pressure along the line of sight. A Compton distortion of the spectrum can become important when $(1+z)dy/dz > 1$, which occurs $\sim$ 2000 years after the Big Bang. The thermodynamic temperature distortion observed at a frequency ν is

$$\frac{\delta T}{T} \approx y \left(x \frac{e^x + 1}{e^x - 1} - 4 \right) \tag{3}$$

where $x = h\nu/kT_{CMB}$, h is the Planck constant, and k is the Boltzmann constant (Sunyaev & Zeldovich 1980). After recombination it becomes nearly impossible to distort the CMB spectrum short of reionizing the universe. Thus a perfect Planck CMB spectrum would support the prediction of the simplest Big Bang model of the universe, while spectral distortions would indicate the existence of more complicated releases of energy.

4.2. The FIRAS instrument

The FIRAS instrument is a polarizing Michelson interferometer (Mather 1982, Mather *et al.*1991a) with two separate spectral channels. The low frequency channel, extending from 0.5 mm to 1 cm, was designed to obtain a precision comparison between the CMB spectrum and a Planckian calibration spectrum. The objective was to attain, in each

5% wide spectral element and each 7° pixel, an accuracy and sensitivity of $\nu I_\nu \cong 10^{-9}$ W m^{-2} sr^{-1}, which is 0.1% of the peak brightness of a 2.7 K blackbody. The high frequency channel, with a useful spectral range from 0.12 mm to 0.5 mm, was designed to measure the emission from dust and gas in our galaxy and to remove the effect of galactic radiation on the measurements of the CMB made in the low frequency channel.

The FIRAS uses a multimode flared horn (Mather, Toral, & Hemmati 1986) with a 7° beam. The instrument directly measures the difference between the sky signal in its beam and that from a temperature-controlled internal reference body. The best apodized spectral resolution is 0.2 cm^{-1} (6 GHz). The in-orbit absolute calibration of FIRAS was accomplished by inserting an external blackbody calibrator periodically into the mouth of the horn. The calibrator is a precision temperature-controlled blackbody, with an emissivity greater than 0.999. The FIRAS uses bolometric detectors (Mather 1981, 1984a,b) in both bands.

In ten months of cryogenic operation the FIRAS obtained over two million interferograms. This complete data set is now undergoing careful analysis.

4.3. FIRAS results

Analysis of the FIRAS data to date confirm the prediction of the simplest Big Bang model that the CMB must have a thermal spectrum. Initial results based on only nine minutes of data showed that there is no deviation from a blackbody spectrum $B_\nu(T)$ as large as 1% of the peak brightness (Mather *et al.*1990a, 1991a) over the spectral range from 500 μm to 1 cm. The temperature of the CMB in the direction of the north galactic pole is 2.735 ± 0.060 K, where 60 mK is the initial conservative uncertainty in the calibration of the thermometry of the absolute calibrator. These data also ruled out the existence of a hot smooth intergalactic medium that could emit more than 3% of the observed x-ray background. The thermal character of the CMB spectrum was subsequently confirmed by Gush, Halpern, & Wishnow (1990), who obtained virtually the same temperature over the spectral range 2-30 cm^{-1}. Neither mean CMB temperature quoted above are corrected for the dipole distortion. These experiments found no submillimeter excess as previously reported by Matsumoto *et al.*(1988b).

More recently, Shafer *et al.*(1991) and Cheng *et al.*(1991a) have examined FIRAS spectra in a direction known previously to be very low in interstellar material ($l = 142°$, $b = 55°$). In this direction, known as Baade's Hole, the temperature is 2.730 ± 0.060 K and there is no deviation from a blackbody spectrum greater than 0.25 % of the peak brightness. The lack of deviations from a Planck spectrum translate to a limit on a chemical potential (see eqn 1) of $|\mu/kT| < 0.005$ (95% CL) and a limit on the Compton y-parameter (eqn 2) of $y < 0.0004$ (95% CL). These results rule out a hot smooth intergalactic medium that could emit more than 1% of the observed x-ray background.

The dipole anisotropy of the CMB, presumed due to our peculiar motion relative to the Hubble flow, can be seen clearly in the FIRAS data, and is consistent with previous results (Cheng *et al.*1990). The FIRAS data show for the first time that the difference in spectra between the poles of the dipole is that expected from two Doppler-shifted blackbody curves. This result also indicates that the stability of the FIRAS instrument is better than one part in 5000 over long time scales. The dipole amplitude measured by FIRAS is 3.31 ± 0.05 mK in the direction (l,b)=($266° \pm 1°$, $47.5° \pm 0.5°$).

FIRAS results also include the first nearly all-sky, unbiased, far infrared survey of the galactic emission at wavelengths greater than 120 μm (Wright *et al.*1991). Wright *et al.*present a map of the dust emission across the sky from the *COBE* FIRAS experiment. They write the absolute Galactic emission intensity in the form $I(\nu, l, b) = g(\nu)G(l, b)$, where $g(\nu)$ is the mean spectrum of the emission and $G(l, b)$ is a dimensionless map. The dust component of the mean spectrum is given as

$$g_d(\nu) = 0.00016 \left(\bar{\nu}/30\mathrm{cm}^{-1}\right)^{1.65} B_\nu(23.3\ \mathrm{K}), \tag{4}$$

or,

$$g_d(\nu) = 0.00022 \left(\bar{\nu}/30\mathrm{cm}^{-1}\right)^{2} [B_\nu(20.4\ \mathrm{K}) + 6.7 B_\nu(4.77\ \mathrm{K})] \tag{5}$$

where $\bar{\nu}$ is the frequency in units of cm^{-1}, and $B_\nu(T)$ is the Planck function. The total far infrared luminosity of the Galaxy is inferred to be $(1.8 \pm 0.6) \times 10^{10} \mathrm{L}_\odot$ (Wright *et al.*1991).

Wright *et al.*report that spectral lines from interstellar [C I], [C II], [N II], and CO are detected in the mean galactic spectrum, $g(\nu)$. The lines of [C II] at 158 μm and [N II] at 205.3 μm were sufficiently strong to be mapped. This is the first observation of the 205.3 μm line. Wright *et al.*interpret the [C II] line as coming from photodissociation regions and the [N II] lines as partially arising from a diffuse warm ionized medium and partially arising from dense H II regions. Petuchowski & Bennett (1992) agree with this conclusion and further elaborate on it by apportioning the [C II] and [N II] transition line intensities among various morphologies of the interstellar medium. Petuchowski & Bennett (in preparation) have conducted observations on NASA's Kuiper Airborne Observatory to measure the scale height of the 205.3 μm [N II] line with a much higher angular resolution ($\sim$1 arcmin) than FIRAS.

5. DMR: microwave anisotropy measurements and interpretations

Primordial gravitational potential fluctuations at the surface of last scattering give rise to the distribution and motions of galaxies and to large angular scale fluctuations in the CMB (Sachs & Wolfe 1967). In inflationary models of cosmology (Guth 1981, Linde 1982, Albrecht & Steinhardt 1982) the gravitational energy fluctuations arise from quantum mechanical fluctuations from 10^{-35} seconds after the Big Bang that inflate to become classical fluctuations with a nearly scale invariant power spectrum (Bardeen, Steinhardt, & Turner 1983, Guth & Pi 1982, Hawking 1982, Starobinskii 1982). Inflation theories are not yet sufficiently constrained to be able to make accurate predictions of the amplitude of current energy fluctuations. Rather, estimates of the energy fluctuations inferred from observations are a constraint on the theories.

The large angular scale CMB temperature anisotropy ΔT and gravitational potential fluctuations at the surface of last scattering $\Delta\Phi$ are simply related by $3\Delta T/T = \Delta\Phi/c^2$ for adiabatic fluctuations in a universe with no cosmological constant ($\Lambda_0 = 0$). On much smaller scales than the DMR measures ($\theta < 4°$), i.e. on scales in causal contact with one another after the universe became matter dominated, the gravitational potential fluctuations are affected by the growth of structures through gravitational instability (e.g. Bond & Efstathiou 1987). Usually the mass density is written as

$$\rho(\vec{x}, t) = \overline{\rho}(1 + \delta(\vec{x}, t)), \tag{6}$$

where $\delta(\vec{x},t)$ describes the spatial density fluctuations. Density fluctuations are often considered in terms of their spectrum by comoving wavenumber, k. The Fourier relations are

$$\delta(\vec{k}) = \int \frac{d\vec{x}}{(2\pi)^3} e^{-i\vec{k}\cdot\vec{x}} \delta(\vec{x}), \qquad \delta(\vec{x}) = \int d\vec{k} e^{i\vec{k}\cdot\vec{x}} \delta(\vec{k}). \tag{7}$$

Newtonian gravitational potential fluctuations are related to density fluctuations through the Poisson equation, $\nabla^2\Phi = 4\pi G\rho$, and its Fourier transform $\Phi_k = -4\pi G\bar{\rho}a^2\delta_k/k^2$, where a is the cosmological scale factor relating the physical scale size r to the comoving scale size x, $r = a(t)x$.

A "transfer function", $T(k)$, relates the initial primordial density fluctuations at the epoch t_i to those observed at the present epoch, t_o: $\delta(k,t_o) \propto T(k)\delta(k,t_i)$. By convention, on large angular scales (small k) $T(k) = 1$ for Λ_0=0 and $\Omega_0 = 1$ cosmologies. That is, the fluctuations on the largest angular scales are primordial and unaffected by any physical evolution since there was never sufficient time for causal contact on this scale. The statistics of the primordial density fluctuations, $\delta(k,t_i)$, are described by a primordial power spectrum, the Peebles-Harrison-Zeldovich (Peebles & Yu 1970, Harrison 1970, Zeldovich 1972) spectrum,

$$P(k,t_i) = \left\langle |\delta(k,t_i)|^2 \right\rangle = Ak^n \tag{8}$$

where the angle brackets represent a spatial average over a large volume of the universe. For this spectrum the rms Sachs-Wolfe potential fluctuations (and the resulting CMB temperature anisotropies) as a function of angle θ are $\Delta\Phi_{rms}/c^2 = 3\Delta T_{rms}/T \propto \theta^{(1-n)/2}$ for $\Omega_0 = 1$ and $\theta \gtrsim 4°$ (Sachs & Wolfe 1967). Note that $P(k,t_o) = T^2(k)P(k,t_i) = Ak^nT^2(k)$. The scale invariant value $n = 1$ gives gravitational potential fluctuations with an rms amplitude that is independent of scale size and a large angular scale CMB temperature fluctuation spectrum that is approximately independent of the separation angle. For $\Omega_0 < 1$ and $\Lambda_0 = 0$ the above expression is still approximately true for angles $\theta < \Omega_0/(1-\Omega_0)^{1/2}$.

Measurements of the abundances of the light elements together with nucleosynthesis calculations imply that $0.011 \leq \Omega_B h^2 \leq 0.037$ (Walker *et al.*1991, Olive *et al.*1990), where Ω_B is the fraction of the critical mass density ($\rho_c = 3H_o^2/8\pi G = 1.88h^2 \times 10^{-29}$ gm cm^{-3}) in baryons and $h = H_0/100$ km s^{-1} Mpc^{-1}. Inflation requires that $\Omega_0 + \Lambda_0/3H_0^2 = 1$ so that either $\Lambda_0 \neq 0$, or inflation theory is incorrect, or most of the mass in the universe is yet to be detected nonbaryonic material. It is useful to assume that this nonbaryonic material does not interact with light. This simultaneously explains why it is not seen, and allows it to begin clustering while the universe was radiation dominated, earlier than is possible for the baryonic matter. The nonbaryonic material is broadly categorized as "hot" or "cold" dark matter, depending on whether it was or was not relativistic when the universe became matter-dominated. A neutrino with mass is a favorite hot dark matter candidate.

Bond & Efstathiou (1984) calculated the transfer function of the "standard cold dark matter model" assuming $T_{0,CMB} = 2.7$ K, $\Omega_B \ll \Omega_{CDM}$, and three massless neutrinos with $T_{0,\nu} = 1.9$ K, giving

$$T(k) = \frac{1}{[1 + (ak + (bk)^{3/2} + (ck)^2)^\nu]^{1/\nu}} \tag{9}$$

where $a = 6.4/(\Omega_0 h^2)$ Mpc, $b = 3.0/(\Omega_0 h^2)$ Mpc, $c = 1.7/(\Omega_0 h^2)$ Mpc, and $\nu = 1.13$. The scale size corresponding to the time when CDM and radiation have equal energy densities is $10/(\Omega_0 h^2)$ Mpc. Holtzman (1989) presents the results of calculations of $T(k)$ for 94 cosmological models.

A successful model of cosmology and the evolution of structure must match the amplitude and spectrum of density fluctuations from the galaxy scale to the horizon scale. Several observables have been derived from galaxy surveys, including the two-point correlation function, the amplitude of its integral, the rms mass fluctuation in a fixed radius sphere, and rms galaxy streaming velocities. The two point correlation function is defined by $\xi(x) = \langle \delta\rho(\vec{x}' + \vec{x})\delta\rho(\vec{x}')/\bar{\rho}^2 \rangle$. $\xi(x)$ is simply the Fourier conjugate of the power spectral density

$$P(k) = \int \frac{d\vec{x}}{(2\pi)^3} e^{-i\vec{k}\cdot\vec{x}} \xi(|\vec{x}|). \tag{10}$$

The integral of the two point correlation function is $J_3(R) \equiv \int_0^R \xi(x) x^2 dx$ (Peebles 1981). Based on the CfA redshift survey Davis & Peebles (1983) find $J_3(10/h \text{ Mpc}) \approx 277 h^{-3}$ Mpc^3 and $J_3(25/h \text{ Mpc}) \approx 780 h^{-3}$Mpc^3. J_3 relates directly to the power spectral density of fluctuations according to

$$J_3(R) = 4\pi \int \frac{P(k)}{k} [\sin kR - kr \cos kR]\, dk \tag{11}$$

where the J_3 definition assumed a spatial top-hat sampling or window function that results in the k-space weighting function $3(kR)^{-3}(\sin kR - kR \cos kR)$. Galaxies are not necessarily distributed in the same way as the mass density fluctuations. In general a linear proportionality is assumed, $(\delta\rho/\rho)_g = b(\delta\rho/\rho)_\rho$ where the constant b is the "biasing factor". The two point correlation function then scales as $\xi_g = b^2 \xi_\rho$. The rms density fluctuation in a sphere of radius 8/h Mpc is $\sigma_8 \approx 1/b$ (Peebles 1982). The radius 8/h Mpc is chosen because the rms fluctuation of the galaxy distribution is unity at this radius. Note that, if the biasing factor is a constant (independent of angular scale), ratios of the rms density fluctuations over different scale sizes, such as σ_8/σ_{25} are independent of b. Kaiser *et al.*(1990) find that $b/\Omega_0^{0.6} = 1.16 \pm 0.21$ based on redshift surveys of IRAS galaxies so it is likely that $b < 2$. The rms of the peculiar velocities of galaxies averaged over a sphere of radius R is another observable that relates directly to the presently observed spectral power density by

$$\left\langle v^2(R) \right\rangle \approx 36\pi H_0^2 \Omega_0^{8/7} \int P(k) \frac{(\sin kR - kR\cos kR)^2}{(kR)^6} dk. \tag{12}$$

Bertschinger *et al.*(1990) derive average galaxy velocities in a sphere of given radii centered on the local group of v(R=4000 km s^{-1})= 388 $\pm$ 67 km s^{-1} and v(R=6000 km s^{-1})= 327 $\pm$ 82 km s^{-1} Thus measurements of ξ, J_3, σ_8, and $\langle v^2 \rangle$ help to determine $P(k)$ today.

Peacock (1991) finds that the power spectrum that is a best-fit to several independent observations of galaxy clustering is

$$P(k) = \frac{1}{4\pi k^3} \frac{(k/k_0)^\alpha}{1 + (k/k_c)^{-\beta}} \tag{13}$$

where $\alpha = 1.6$, $\beta = 2.4$, $k_0 = 0.19$ h Mpc^{-1}, and $k_c \approx 0.025$ h Mpc^{-1}. More recently, in light of the new *COBE* results summarized below, Peacock prefers $k_c \approx 0.033$ h Mpc^{-1} (private communication).

Hence, measurements of large scale (i.e. primordial) CMB anisotropies can provide the observational link between the production of gravitational potential fluctuations in the early universe and the observed galaxy distributions and velocities today. Large scale CMB anisotropy measurements provide both the amplitude and the power spectrum of the primordial fluctuations . Large scale anisotropy measurements are usually expressed in terms of a multipole expansion and a correlation function. The multipole expansion of the CMB temperature as a function of sky location is

$$T(\theta,\phi) = \sum_l \sum_{m=-l}^{l} a_{lm} Y_{lm}(\theta,\phi). \tag{14}$$

where $Y_{lm}(\theta,\phi)$ are the spherical harmonic functions. Since DMR is a differential experiment, as are almost all anisotropy experiments, the $l = 0$ monopole term is not observed. (It is observed by FIRAS.) The $l = 1$ dipole term is also dropped since it is dominated by the Doppler effect due to our local peculiar velocity and not by cosmic perturbations. Thus the $l = 2$ quadrupole term is the first term of interest. We are at liberty to select any coordinate system we choose. Since galactic emission dominates the sky signal, we choose galactic coordinates, with the usual galactic coordinate angles angles l and b. We define $Q(l,b)$ to be the $l = 2$ term of equation (14) and rewrite the five $Y_{l=2,m}$ components:

$$\begin{aligned} Q(l,b) &= Q_1(3\sin^2 b - 1)/2 + Q_2 \sin 2b \cos l + Q_3 \sin 2b \sin l + \\ &\quad Q_4 \cos^2 b \cos 2l + Q_5 \cos^2 b \sin 2l \end{aligned} \tag{15}$$

where the rms quadrupole amplitude is

$$Q_{rms}^2 = \frac{1}{4\pi}\int_{4\pi} Q^2(l,b)d\Omega = \frac{4}{15}(\frac{3}{4}Q_1^2 + Q_2^2 + Q_3^2 + Q_4^2 + Q_5^2). \tag{16}$$

There is a small kinematic quadrupole, $Q_{rms} = 1.2$ μK, from the second order terms in the relativistic Doppler expansion (Peebles & Wilkinson 1968), for which $(Q_1, Q_2, Q_3, Q_4, Q_5)$=(0.9, -0.2, -2.0, -0.9, 0.2) μK.

The measured correlation function determines the parameters of the fluctuation power spectrum. The correlation function is

$$C(\alpha) = \sum_{l>1} \Delta T_l^2 W(l)^2 P_l(cos(\alpha)), \tag{17}$$

where P_l are Legendre polynomials, and a 3.2° rms Gaussian beam gives a weighting $W(l) = \exp[-\frac{1}{2}(l(l+1)/17.8^2)]$ and

$$\Delta T_l^2 = \frac{1}{4\pi}\sum_m |a_{lm}|^2 \tag{18}$$

are the rotationally-invariant rms multipole moments. As with the spherical harmonic expansion, the $l = 0$ is excluded from the correlation function since it is not measured by differential instruments, and the $l = 1$ term is excluded because it is contaminated

by the kinematic dipole. The $l = 2$ quadrupole term is sometimes excluded since the quadrupole has only $2l + 1 = 5$ degrees of freedom, and thus has an intrinsically high statistical, or "cosmic" variance, independent of the measurement. For a power law primordial fluctuation spectrum the predicted moments, as a function of spectral index $n < 3$, are given by Bond & Efstathiou (1987):

$$< \Delta T_l^2 >= (Q_{rms})^2 \frac{(2l+1)}{5} \frac{\Gamma(l + (n-1)/2)\Gamma((9-n)/2)}{\Gamma(l + (5-n)/2)\Gamma((3+n)/2)}. \quad (19)$$

For $n = 1$ this simplifies to

$$< \Delta T_l^2 >= (Q_{rms})^2 \frac{6}{5} \frac{2l+1}{l(l+1)}. \quad (20)$$

Smoot *et al.*(1991b) presented preliminary DMR results based on six months of data. Smoot *et al.*(1992) describe results based upon the first year of DMR data, Bennett *et al.*(1992a) describe the calibration procedures, Kogut *et al.*(1992) discuss the treatment of systematic errors, and Bennett *et al.*(1992b) discuss the separation of cosmic and Galactic signals. Wright *et al.*(1992) compare these data to other measurements and to models of structure formation through gravitational instability. Previously published large-angular-scale anisotropy measurements include Fixsen *et al.*(1983), Lubin *et al.*(1985), Klypin *et al.*(1987), and Meyer *et al.*(1991). Some excellent reviews of CMB anisotropy and cosmological perturbation theory include Bertschinger (1992), Efstathiou (1990), Kolb & Turner (1990), Peebles (1971, 1980), and Wilkinson (1986).

5.1. The DMR instrument and data processing

The *COBE* DMR instrument is described by Smoot *et al.* (1990). DMR operates at three frequencies: 31.5, 53 and 90 GHz (wavelengths 9.5, 5.7, and 3.3 mm), chosen to be near the minimum in Galactic emission and near the CMB maximum. Wright et al. (1990) have used the FIRAS and DMR data to show that the ratio of the galactic emission to that of the CMB reaches a minimum between 60 and 90 GHz. There are two nearly independent channels, A and B, at each frequency. The orbit and pointing of the *COBE* result in a complete survey of the sky every six months while shielding the DMR from terrestrial and solar radiation (Boggess *et al.*1992).

The DMR measures the difference in power received between regions of the sky separated by 60°. For each radiometer channel a baseline is subtracted and the data are calibrated. Data are rejected when the limb of the Earth is higher than 1° below the Sun/Earth shield plane, when the Moon is within 25° of a beam center, when any datum deviates from the daily mean by more than 5σ, or when the spacecraft telemetry or attitude solution is of poor quality. Small corrections are applied to remove the estimated emission from the Moon and Jupiter in the remaining data. Corrections are also applied to remove the Doppler effects from the spacecraft's velocity about the Earth and the Earth's velocity about the solar system barycenter. A least-squares minimization is used to fit the data to spherical harmonic expansions and to make sky maps with 6144 nearly equal area pixels using a sparse matrix technique (Torres *et al.* 1989, Janssen & Gulkis 1992). The DMR instrument is sensitive to external magnetic fields. Extra equations are included in the sparse matrix to allow these magnetic susceptibilities to

be fit separately as a linear function of the Earth's field and the radiometer orientation. The magnetic corrections are on the scale of 10 to 100 μK in the time-ordered data. Residual uncertainties in the individual radiometer channel maps, after correction, are typically 2 μK and never more than 8.5 μK.

Kogut *et al.*(1992) have searched the DMR data for evidence of residual systematic effects. The largest such effect is the instrument response to an external magnetic field. Data binned by the position of the Earth relative to the spacecraft show no evidence for contamination by the Earth's emission at the noise limit (47 μK at 95% CL). The contribution of the Earth's emission to the maps is estimated to be less than 2 μK. The time-ordered data with antenna beam centers more than 25° away from the Moon are corrected to an estimated accuracy of 10% (4 μK) of the lunar flux. The estimated residual effect on the maps is less than 1 μK. Kogut *et al.*list upper limits for the effects of variations in calibration and instrument baselines, solar and solar system emissions, RFI, and data analysis errors. The quadrature sum of all systematic uncertainties in a typical map, after corrections, is < 8.5 μK for rms sky fluctuations, < 3 μK for the quadrupole and higher-order multipole moments, and < 30 μK^2 for the correlation function (all limits 95% CL).

5.2. The DMR anisotropy

The DMR maps are dominated by the dipole anisotropy and the emission from the Galactic plane. The dipole anisotropy ($\Delta T/T \approx 10^{-3}$) is seen consistently in all channels with a thermodynamic temperature amplitude 3.36 ± 0.1 mK in the direction $l = 264.7° \pm 0.8°$, $b = 48.2° \pm 0.5°$, consistent with the FIRAS results, above. Our motion with respect to the CMB (a blackbody radiation field) is assumed to produce the dipole anisotropy, so the dipole and associated ≈ 1.2 μK rms kinematic quadrupole are removed from the maps.

The DMR instrument noise and the intrinsic fluctuations on the sky are independent and thus add in quadrature to give the total observed signal variance

$$\sigma^2_{obs} = \sigma^2_{DMR} + \sigma^2_{Sky}. \tag{21}$$

The σ_{obs} is estimated from the two channel (A+B)/2 sum maps, and the (A-B)/2 difference maps provide an estimate of σ_{DMR}, yielding the sky variance $\sigma_{Sky}(10°) = 30 \pm 5$ μK for $|b| > 20°$. The observations are made with a 7° beam, and the resulting maps are smoothed with an additional 7° Gaussian function, resulting in the effective 10° angular resolution.

The correlation function, C(α), is the average product of temperatures separated by angle α. It is calculated for each map by rejecting all pixels within the Galactic latitude band $|b| < 20°$, removing the mean, dipole, and quadrupole from the remaining pixels by a least squares fit, multiplying all possible pixel pair temperatures, and averaging the results into 2.6° bins. Bennett *et al.*(1992b) conclude that the galactic contribution to the correlation signal is small for $|b| > 15°$. This is consistent with the fact that the correlation function and rms sky fluctuation are insensitive to the Galactic latitude cut angles so long as $|b| < 15°$ is excluded. The DMR correlation functions exhibit temperature anisotropy on all angular scales greater than the beam size (7°) and differ significantly ($> 7\sigma$) from the flat correlation function due to receiver noise alone.

All six channels show a statistically significant quadrupole signal. A comparison of the fitted quadrupoles between channels and frequencies, and between the first and second six months of data, shows that individual quadrupole components, Q_i, typically differ from map to map by ≈ 10 μK with comparable uncertainty. Determination of the cosmic quadrupole is linked to its separation from Galactic emission (Bennett *et al.*1992b), summarized below. Discrete extragalactic sources individually contribute less than 2 μK in the DMR beam and the expected temperature variations are less than 1 μK (Franceschini *et al.* 1989).

5.3. Separation of galactic signals & the cosmic quadrupole

The DMR anisotropy maps are sufficiently sensitive and free from systematic errors that our knowledge of Galactic emission is a limiting factor in interpreting the measurements of the 1-year DMR maps. The detected signals expressed in thermodynamic temperature are nearly constant amplitude: the rms fluctuations on a 10° scale are proportional to $\nu^{-0.3\pm1}$ and the quadrupole and correlation functions $\propto \nu^{-0.2\pm1}$. The flat spectral index of the DMR anisotropy, without correction for Galactic emissions, is consistent with a cosmic origin and inconsistent with an origin from a single Galactic component. However, from this fact alone we are unable to rule out a correlated superposition of dust, synchrotron, and free-free emission and thus more detailed galactic emission models are required. Bennett *et al.*(1992b) constructed preliminary models of microwave emission from our Galaxy based on *COBE* and other data for the purpose of distinguishing cosmic and Galactic signals.

Four emission components are important at microwave wavelengths. CMB anisotropies are assumed to produce differences in the measured antenna temperature according to $\Delta T_A = \Delta T\, x^2 e^x/(e^x-1)^2$, where $x = h\nu/kT$, T is thermodynamic temperature. Synchrotron emission arises from relativistic electrons accelerated by magnetic fields. Free-free emission occurs when free electrons are accelerated by interactions with ions. Thermal emission from dust is also important at microwave wavelengths.

The brightest pixels in the DMR maps are T_A=5.9 $\pm$ 0.4 mK at 31.5 GHz, 1.9 $\pm$ 0.2 mK at 53 GHz, both at (l,b)=(337°, -1°), and 1.3$\pm$0.2 mK (348°,+1°) at 90 GHz. Galactic plane emission would have to be removed to better than 1% to reveal cosmologically interesting fluctuations in the CMB at low Galactic latitudes, so our preliminary models concentrate on $|b|>10°$.

The intensity of synchrotron radiation is given by the integral $I(\nu) = \iint P(\nu,\vec{B},E)\, N(E,\vec{l})\, dE\, dl$ where $P(\nu,\vec{B},E)$ is the power emitted at a frequency ν by a single electron of energy E in a magnetic field $\vec{B}$, and $N(E,\vec{l})dE$ is the number of relativistic electrons of energy E per unit volume at the position $\vec{l}$ along the line of sight. Since $N(E,\vec{l})$ and $\vec{B}(\vec{l})$ are not known for every position in the Galaxy, approximations must be made to model synchrotron emission. Bennett *et al.*(1992b) use the local electron spectrum in conjunction with radio data to approximate the synchrotron integral.

Free-free emission is characterized by an antenna temperature that depends only on the emission measure and electron temperature along the line of sight for each frequency, with a spectral index $\beta_{ff} = 2 + [10.48 + 1.5 \ln (T_e/8000\ \mathrm{K}) - \ln \nu_{\mathrm{GHz}}]^{-1}$. β is the spectral index of the antenna temperature, $T_A \propto \nu^{-\beta}$, which relates to the flux

density, $S(\nu)$, by $S(\nu) \propto 2kT_A(\nu)\nu^2/c^2$. With $T_e = 8000$ K (Reynolds 1985), the ratio of the 53 GHz free-free antenna temperature to Hα intensity can be expressed as $T_A(\mu\mathrm{K})/I(\mathrm{Rayleigh}) = 0.83\ \mu\mathrm{K}/0.44$ Rayleigh $= 2\ \mu$K/Rayleigh, almost independent of the electron temperature. Reynolds (1984, 1992) has observed several high-latitude lines of sight with a 0.8° beam and reports that the high latitude Hα diffuse Galactic distribution is $I(\mathrm{R}) \approx 1.2 \csc|\mathrm{b}|$ for $|b| > 15°$. Deviations from a cosecant-law are larger than the $\approx$15% Hα measurement uncertainties (Reynolds 1992).

Another estimate of the high latitude emission measure comes from the *COBE* FIRAS measurement of the N^+ ground state transition at 205 μm (Wright *et al.*1991). The observed intensity of this line is $I(\mathrm{N}^+) \approx (7 \pm 2) \times 10^{-8} \csc|\mathrm{b}|$ erg s^{-1} cm^{-2}sr^{-1} for $|b| > 15°$ compared with $I(\mathrm{N}^+) \approx 2.4\times10^{-8}\csc|\mathrm{b}|$ erg s^{-1} cm^{-2}sr^{-1} predicted by Reynolds (1992) for the diffuse component. In doing Hα background measurements, Reynolds picks locations that are free from discrete sources, while the *COBE* observations do not exclude sources. The *COBE* observations may also include emission with velocities outside of Reynolds' Hα passband. If we assume that the excess observed N^+ arises from the full sky (unbiased) sampling by FIRAS and any Hα bandwidth exclusions by Reynolds, then we can use the ratio of measured-to-predicted N^+ to correct the diffuse free-free predictions for these effects. We deduce from the above that a correction factor of $\sim$3 is required; the free-free emission is then approximately $T_A(\mu\mathrm{K}) \approx 7 \csc|b|$ for $|b| > 15°$ at 53 GHz. This prediction is only approximate since it depends on assumptions of the chemical abundance of nitrogen and its fractional population in the N^+ state. Our factor of $\sim$3 to convert from diffuse to full sky-averaged Hα intensity is consistent with the Reynolds (1992) a priori estimate of a factor of $\sim$2. Unfortunately, there exists no full sky survey of free-free emission to the sensitivity required by the *COBE* DMR, so the DMR 31.5 GHz map must serve this purpose. We compare this with the cosecant-law predictions discussed above and find good agreement.

Along each line of sight the observed dust emission is the sum over the emission from each dust grain. Since the grain temperatures, emissivities, and spatial distributions are not known, it is not possible to make *a priori* full sky dust emission models. We must rely, instead, on full sky measurements of the dust intensity at higher frequencies and extrapolate these with empirical spectral fits. The dust antenna temperature is $T_A = 33\ G(l,b)\ R_\nu^d\ \mu$K , where R_ν^d=(0.43, 1.0, 2.3) at (31.5, 53, 90 GHz) and $G(l,b)$ is the dimensionless map of the dust distribution from FIRAS, discussed earlier. $G(l,b) \approx 1$ in the plane of the Galaxy and ranges from 0.03 to 0.1 at high Galactic latitudes.

Bennett *et al.*(1992b) present three approaches to modeling the Galactic emission signal in the DMR maps. A "subtraction technique" makes use of external data to *subtract* Galactic emission maps from the DMR maps. The dust and synchrotron emission models, described above, were subtracted from the DMR maps. Since no map exists of the ionized component of our Galaxy, the 31 GHz residual map is used to subtract the free-free emission from the 53 and 90 GHz DMR maps. The angular autocorrelation functions of the individual emission components show that the Galactic components have different angular correlations than the residual cosmic signal. In fact, the cosmic correlation function is largely unaffected by the Galactic model subtraction. A "fitting technique" directly *fits* the DMR maps pixel-by-pixel. The results of the fit are maps of free-free and Planckian emission. A "combination technique" uses a linear *combina-*

tion of the DMR 31.5, 53, and 90 GHz maps to minimize the Galactic emission without recourse to Galactic models or other data. The subtraction, fitting, and combination techniques produce consistent results. The combination technique is independent of the fitting and subtraction techniques, aside from the use of DMR data and the assumed free-free spectral index. Bennett *et al.*conclude that no known Galactic emission component or superposition of components can account for most of the observed anisotropy signal. In the absence of significant extragalactic source signals or systematic errors, as argued above, this signal must be intrinsic to the CMB radiation.

DMR maps, with the modeled Galactic emission removed, are fit for a quadrupole distribution. Bennett *et al.*derive a cosmic quadrupole, corrected for the expected kinematic quadrupole, of $Q_{rms} = 13 \pm 4$ μK, $(\Delta T/T)_Q = (4.8 \pm 1.5) \times 10^{-6}$, for $|b| > 10°$. When Galactic emission is removed from the DMR data, the residual fluctuations are virtually unaffected and therefore they are not dominated by any known Galactic emission component(s).

5.4. Interpretation of the DMR anisotropy

The anisotropy detected by the DMR is interpreted as being a direct result of primordial fluctuations in the gravitational potential. Assuming a power spectral density of density fluctuations of the form $P(k) = Ak^n$, the best-fit results are $n = 1.1 \pm 0.5$ with $Q_{rms-PS} = 16 \pm 4$ μK. Q_{rms-PS} is the rms quadrupole amplitude resulting from this power spectrum fit, i.e. making use of fluctuation information from all observed angular scales, as opposed to the Q_{rms} derived from a direct quadrupole fit. Forcing the spectral index to $n = 1$ gives $Q_{rms-PS} = 16.7 \pm 4$ μK and increases the χ^2 from 79 to 81 for 68 degrees of freedom. Interpreted as a power-law spectrum of primordial fluctuations with a Gaussian distribution, the ΔT_l^2 in each horizon have a χ^2 distribution of $2l + 1$ degrees of freedom, giving a cosmic variance for observations within a single horizon volume in the universe of $2 < \Delta T_l^2 >^2 /(2l + 1)$. Best fit values are $n = 1.15^{+0.45}_{-0.65}$ and $Q_{rms-PS} = 16.3 \pm 4.6$ μK including the cosmic variance, with a χ^2 of 53. Cross-correlation of the 53 GHz and 90 GHz maps are consistent with a power law spectra with index $n = 1 \pm 0.6$ and amplitude $Q_{rms-PS} = 17 \pm 5\mu$K, including cosmic variance.

The observed cosmic quadrupole from the maps [$Q_{rms} = 13 \pm 4\mu$K from Bennett *et al.*1992b (see above)] is slightly below the mean value predicted by the higher-order moments deduced from the correlation function ($Q_{rms-PS} = 16 \pm 4\mu$K). This is a likely consequence of cosmic variance: the mode of the χ^2 distribution is lower than the mean. A map quadrupole value of 13 μK or lower would be expected to occur 35% of the time for an $n = 1$ universe with $Q_{rms-PS} = 16$ μK. The results above exclude the quadrupole before computing C(α). Including the quadrupole in computing C(α) increases the χ^2, raises n to 1.5, and decreases Q_{rms-PS} to 14μK.

The measured parameters [$\sigma_{Sky}(10°)$, Q_{rms}, Q_{rms-PS}, $C(\alpha)$, and n] are consistent with a Peebles-Harrison-Zeldovich (scale invariant) spectrum of perturbations, which predicts $Q_{rms} = (1^{+0.3}_{-0.4})Q_{rms-PS}$ and $\sigma_{Sky}(10°) = (2.0 \pm 0.2)Q_{rms-PS}$. The theoretical 68% CL errors take into account the cosmic variance due to the statistical fluctuations in perturbations for our observable portion of the Universe. The minimum Q_{rms} for models with an initial Peebles-Harrison-Zel'dovich perturbations, normalized to the local large-

scale galaxy streaming velocities, is predicted to be 12 μK, independent of the Hubble constant and the nature of dark matter (Gorski 1991, Schaefer 1991).

These observations are consistent with inflationary cosmology models. The natural interpretation of the DMR signal is the observation of very large (presently >>100 Mpc) structures in the Universe which are little changed from their primordial state (t << 1 sec). These structures are part of a power law spectrum of small amplitude gravitational potential fluctuations that on smaller length scales are sources of the large scale structure observed in the Universe today. The DMR data provide strong support for gravitational instability theories (Wright *et al.*1992). Wright *et al.*compare the 94 cosmological models for which Holtzman (1989) has computed a transfer function, $T(k)$, with the DMR anisotropy results. None of the Holtzman isocurvature models are compatible with the DMR anisotropy amplitude for a biasing factor $b < 4$. Wright *et al.*find that three Holtzman models fit the observational data (galaxy clustering, galaxy streaming velocity, and CMB quadrupole amplitude) reasonably well. These models are described below.

A model with vacuum energy density with $\Omega_{vac} = \Lambda_0/3H_0^2 = 0.8$, $H_0 = 100$ km s^{-1} Mpc^{-1}, $\Omega_B = 0.02$, $\Omega_{CDM} = 0.18$ is an excellent fit to the observational data (see, e.g. Gorski, Silk & Vittorio 1992; Efstathiou, Bond & White 1992, and Peebles 1984, 1991).

A "mixed dark matter" (MDM) model that fits the data uses both hot dark matter (a massive neutrino with $\Omega_{HDM} = 0.3$) and cold dark matter ($\Omega_{CDM} = 0.6$) with baryonic dark matter $\Omega_B = 0.1$ and $H_0 = 50$ km s^{-1} Mpc^{-1}. See, for example, Efstathiou, Bond & White (1992), Davis, Summers & Schlegel (1992), and van Dalen & Schaefer (1992) for further recent discussions of mixed dark matter models.

An open universe model with $\Omega_0 = 0.2$, $\Omega_B = 0.02$, and $\Omega_{CDM} = 0.18$ for $H_0 = 100$ km s^{-1} Mpc^{-1} satisfies the observations, except perhaps for the galaxy rms peculiar velocities, but is in conflict with the inflation model and theoretical prejudices for $\Omega_0 = 1$ (see Gouda & Sugiyama 1992, Peebles 1984, 1991).

The unbiased standard cold dark matter is in conflict with galaxy clustering data, even without the constraint of the COBE data (e.g. Vogeley *et al.*1992, Loveday *et al.*1992). Hogan (1991, 1992, 1993) and Hoyle & Burbidge (1992) interpret the *COBE*-DMR results in terms of models where the temperature anisotropies do not arise from gravitational potential fluctuations on the surface of last scattering. Bennett & Rhie (1992) interpret the DMR data in terms of global monopoles and textures.

In summary, the *COBE* detection of CMB temperature anisotropy has added another important observational piece of knowledge to the cosmic puzzle. There is not yet a clear favorite among the models that attempt to account for all of the pieces, nor is there likely to be one without further observational information.

6. The DIRBE experiment

6.1. The cosmic infrared background radiation

The Diffuse Infrared Background Experiment (DIRBE) is the first space experiment designed primarily to measure the CIB radiation. The aim of the DIRBE is to conduct a definitive search for an isotropic CIB radiation, within the constraints imposed by the local astrophysical foregrounds.

Cosmological motivations for searching for an extragalactic infrared background have been discussed in the literature for several decades (early papers include Partridge & Peebles 1967; Low & Tucker 1968; Peebles 1969; Harwit 1970; Kaufman 1976). Both the cosmic redshift and reprocessing of short-wavelength radiation to longer wavelengths by dust act to shift the short-wavelength emissions of cosmic sources toward or into the infrared. Hence, the wide spectral range from 1 to 1000 μm is expected to contain much of the energy released since the formation of luminous objects, and could potentially contain a total radiant energy density comparable to that of the CMB.

The CIB radiation has received relatively little attention in the theoretical literature compared to that devoted to the CMB (Negroponte 1986), which has a central significance to Big Bang cosmology and quite distinctive and definite predictions as to its character. However, advances in infrared instrumentation, and especially the introduction of cryogenically cooled infrared instruments on space missions, have stimulated increasing attention to prediction of the character of the CIB radiation (Fabbri and Melchiorri 1979; Bond, Carr, and Hogan 1986; McDowell 1986; Fabbri *et al.*1987; Fabbri 1988; Bond, Carr, and Hogan 1991). Measurement of the spectral intensity and anisotropy of the CIB radiation would provide important new insights into intriguing issues such as the amount of matter undergoing luminous episodes in the pregalactic Universe, the nature and evolution of such luminosity sources, the nature and distribution of cosmic dust, and the density and luminosity evolution of infrared-bright galaxies.

Observing the CIB radiation is a formidable task. Bright foregrounds from the atmosphere of the Earth, from interplanetary dust scattering of sunlight and emission of absorbed sunlight, and from stellar and interstellar emissions of our own Galaxy dominate the diffuse sky brightness in the infrared. Even when measurements are made from space with cryogenically cooled instruments, the local astrophysical foregrounds strongly constrain our ability to measure and discriminate an extragalactic infrared background. Furthermore, since the absolute brightness of the CIB radiation is of paramount interest for cosmology, such measurements must be done relative to a well established absolute flux reference, with instruments which strongly exclude or permit discrimination of all stray sources of radiation or offset signals which could mimic a cosmic signal.

Hauser (1991b) lists recent experiments capable of making absolute sky brightness measurements in the infrared (for a compilation including some earlier measurements, see Negroponte 1986). Instruments or detector channels designed specifically to measure that part of the spectrum dominated by the CMB radiation have been excluded. Murdock & Price (1985) flew an absolute radiometer with strong stray light rejection on a sounding rocket in 1980 and 1981. Their primary objective was measuring scattering and emission from interplanetary dust, and no attempt was made to extract an extragalactic component. Matsumoto *et al.*(1988a) flew a near-infrared experiment on a rocket in 1984. They have reported possible evidence for an isotropic residual near 2 μm, perhaps in a line feature, for which they cannot account in their models of emission from the interplanetary medium and the Galaxy. This group has flown a modified instrument early in 1990 to investigate further this result (Noda *et al.*1992). The IRAS sky survey instrument, though not specifically designed for absolute background measurements, was, within the limits of long term stability, capable of good relative total sky brightness measurements, and so is included in this list. Uncertainties in the IRAS

absolute calibration have impeded efforts to extract an estimate of the CIB radiation (Rowan- Robinson 1986). The FIRAS high frequency channel (100 to 500 μm), with its all-sky coverage, excellent stray light rejection, absolute calibration, and high sensitivity, also promises to be an important instrument for CIB radiation studies. Quantitative comparison of the measurements from the experiments discussed above, and a summary of current CIB radiation limits are discussed further below.

6.2. The DIRBE instrument

The experimental approach is to obtain absolute brightness maps of the full sky in 10 photometric bands (J[1.2], K[2.3], L[3.4], and M[4.9]; the four IRAS bands at 12, 25, 60, and 100 μm; and 140 and 240 μm bands). To facilitate discrimination of the bright foreground contribution from interplanetary dust, linear polarization is also measured in the J, K, and L bands, and all celestial directions are observed hundreds of times at all accessible angles from the Sun in the range 64° to 124°. The instrument rms sensitivity per field of view in 10 months is $\lambda I(\lambda)$ = (1.0, 0.9, 0.6, 0.5, 0.3, 0.4, 0.4, 0.1, 11.0, 4.0) $\times 10^{-9}$ W m^{-2} sr^{-1}, respectively for the ten wavelength bands listed above. These levels are generally well below both estimated CIB radiation contributions (e.g., Bond, Carr, Hogan 1986) and the total infrared sky brightness.

The DIRBE instrument is an absolute radiometer, utilizing an off-axis Gregorian telescope with a 19-cm diameter primary mirror. Since the DIRBE was designed to make an absolute measurement of the spectrum and angular distribution of the diffuse infrared background it must have extremely strong rejection of stray light. The optical configuration (Magner 1987) has strong rejection of stray light from the Sun, Earth limb, Moon or other off-axis celestial radiation, or parts of the COBE payload (Evans, 1983; Evans and Breault 1983). Stray light rejection features include both a secondary field stop and a Lyot stop, super-polished primary and secondary mirrors, a reflective fore-baffle, extensive internal baffling, and a complete light-tight enclosure of the instrument within the COBE dewar. Additional protection is provided by the Sun and Earth shade surrounding the COBE dewar, which prevents direct illumination of the DIRBE aperture by these strong local sources. The DIRBE instrument, which was maintained at a temperature below 2 K within the dewar as long as the helium was present, measures absolute brightness by chopping between the sky signal and a zero flux internal reference at 32 Hz using a tuning fork chopper. The synchronously demodulated signal is averaged for 0.125 second before transmission to the ground. Instrumental offsets are measured by closing a cold shutter located at the prime focus. All spectral bands view the same instantaneous field-of-view, 0.7° x 0.7°, oriented at 30° from the spacecraft spin axis. This allows the DIRBE to modulate the angle from the Sun by 60° during each rotation, and to sample fully 50% of the celestial sphere each day. Four highly-reproducible internal radiative reference sources can be used to stimulate all detectors when the shutter is closed to monitor the stability and linearity of the instrument response. The highly redundant sky sampling and frequent response checks provide precise photometric closure over the sky for the duration of the mission. Calibration of the photometric scale is obtained from observations of isolated bright celestial sources. Careful measurements of the beam shape in pre-flight system testing and during the mission using scans across bright point sources allow conversion of point-source calibrations to surface brightness

calibrations.

The data obtained during the helium temperature phase of the mission are of excellent photometric quality, showing good sensitivity, stability, linearity, and stray light immunity. Few artifacts are apparent other than those induced by energetic particles in the South Atlantic Anomaly and variations in instrument temperature. Both of these effects will be removed in final data processing. Strong rejection of off-axis radiation sources is confirmed by the absence of response to the Moon (which saturates the response in all detectors when in the field of view) until it comes within about 3° of the field of view. The sensitivity per field of view, listed above, is based on noise measured with the shutter closed and response determined from measurements of known celestial sources. The noise when the shutter was open is somewhat above the shutter-closed values due to discrete source confusion. The nuclear radiation environment in orbit caused very little response change (<1%) in all detectors except the Ge:Ga photoconductors used at 60 and 100 μm. Thermal and radiative annealing procedures applied to these detectors following passages through the South Atlantic Anomaly will allow response correction to about 1% at these wavelengths. It is expected that fully reduced DIRBE sky maps will have photometric consistency over the sky better than 2% at each wavelength, nearest neighbor band-to-band (color) brightness accuracy of 3% or better, and absolute intensity scale accuracy better than 20%.

6.3. DIRBE results

Preliminary results of the DIRBE experiment have been described by Hauser *et al.*(1991), Hauser (1991a, 1991b). Qualitatively, the initial DIRBE sky maps show the expected character of the infrared sky. For example, at 1.2 μm stellar emission from the galactic plane and from isolated high latitude stars is prominent. Zodiacal scattered light from interplanetary dust is also prominent. These two components continue to dominate out to 3.4 μm, though both become fainter as wavelength increases. A composite of the 1.2, 2.3, and 3.4 μm images was presented by Mather et al. (1990b). Because extinction at these wavelengths is far less than in visible light, the disk and bulge stellar populations of the Milky Way are dramatically apparent in this image. At 12 and 25 μm, emission from the interplanetary dust dominates the sky brightness. As with the scattered zodiacal light, the sky brightness is strongly dependent upon ecliptic latitude and solar elongation angle. At wavelengths of 60 μm and longer, emission from the interstellar medium dominates the galactic brightness, and the interplanetary dust emission becomes progressively less apparent. The patchy infrared cirrus noted in IRAS data (Low *et al.*1984) is evident at all wavelengths longer than 25 μm. Weiland *et al.*(1991) gave a preliminary description of the emission from the interplanetary dust bands, and Murdock *et al.*(1991) have used the DIRBE data to carry out a preliminary examination of the ecliptic pole emission.

The DIRBE data will clearly be a valuable new resource for studies of the interplanetary medium and Galaxy as well as the search for the CIB radiation.

In searching for the extragalactic infrared background, the most favorable conditions are directions and wavelengths of least foreground brightness. In general, because of the strong interplanetary dust foreground and the relatively modest gradient of that foreground over the sky, the infrared sky is faintest at high ecliptic latitude. A prelimi-

nary DIRBE spectrum of the sky brightness toward the south ecliptic pole was presented by Hauser *et al.*(1991), and is reproduced in Table 2.

Table 2. Cosmic Infrared Background Limits

Reference	λ (μm)	λI_λ (10^{-7} W m^{-2} sr^{-1})
DIRBE	1.2	8.3 ± 3.3
(South Ecliptic Pole)	2.3	3.5 ± 1.4
	3.4	1.5 ± 0.6
	4.9	3.7 ± 1.5
	12.	29. ± 12
	22.	21. ± 8
	55.	2.3 ± 1
	96.	1.2 ± 0.5
	151.	1.3 ± 0.7
	241.	0.7 ± 0.4

This table shows the strong foreground from starlight and scattered sunlight at the shortest wavelengths, a relative minimum at 3.4 μm, emission dominated by interplanetary dust peaking around 12 μm, and generally falling brightness from there out to submillimeter wavelengths.

To meet the cosmological objective of measuring the CIB radiation, the foreground light from interplanetary and galactic sources must be discriminated from the total observed infrared sky brightness. This task requires extensive careful correlation studies and modelling, which in the case of the DIRBE investigation is in progress. A conservative upper limit on extragalactic light is the total observed brightness in a relatively dark direction. The sky brightness at the south ecliptic pole is a fair representation of the best current limits from the DIRBE. The faintest foregrounds occur at 3.4 μm, in the minimum between interplanetary dust scattering of sunlight and re-emission of absorbed sunlight by the same dust, and longward of 100 μm, where interstellar dust emission begins to decrease. Through careful modelling, we hope to be able to discriminate isotropic residuals at a level as small as 1 percent of the foregrounds. These near-infrared and submillimeter windows will allow the most sensitive search for, or limits upon, the elusive cosmic infrared background.

These data are to be compared with the theoretical estimates of contributions to the CIB radiation from pregalactic and protogalactic sources in a dust free universe (Bond, Carr, and Hogan 1986, Carr 1988). The present conservative observational limits are beginning to constrain some of the theoretical models at short infrared wavelengths, though in a dusty universe energy from these sources can be redistributed farther into the infrared. If the foreground components of emission can confidently be identified, the current *COBE* measurements will seriously constrain (or identify) the CIB radiation across the infrared spectrum. However, the spectral decade from about 6 to 60 μm will

have relatively weak limits until measurements are made from outside the interplanetary dust cloud.

The CIB radiation promises to enhance our understanding of the epoch between decoupling and galaxy formation. The high quality and extensive new measurements of the absolute infrared sky brightness obtained with the DIRBE and FIRAS experiments on the *COBE* mission promise to allow a definitive search for this elusive background, limited primarily by the difficulty of distinguishing it from bright astrophysical foregrounds.

7. *COBE* data products and plans

Extensive data products from the *COBE* mission consisting of calibrated maps and spectra with associated documentation are planned. The *COBE* databases have been described by White & Mather (1991). An overview of the *COBE* software system has been given by Cheng (1991). All *COBE* data processing and software development for analysis take place at the Cosmology Data Analysis Center (CDAC) in Greenbelt, MD, a facility developed by the *COBE* project for that purpose. This facility, and the software tools developed there, will become available to the scientific community when the data products are released.

Initial data products are planned for release in mid 1993. Galactic plane maps, including the nuclear bulge, will be available at all 10 DIRBE wavelengths and the high frequency FIRAS band. Full sky maps from all six DMR radiometers will also be available.

Full sky maps from all three *COBE* instruments, spanning four decades of wavelength, are planned for release in mid-1994. These data gathered by the *COBE*'s three instruments will constitute a comprehensive data set unprecedented in scope and sensitivity for studies of cosmology, and large scale galactic and solar system science.

Acknowledgments

The authors gratefully acknowledge the contributions to this report by their colleagues on the *COBE* Science Working Group and the other participants in the *COBE* Project. Many people have made essential contributions to the success of *COBE* in all its stages, from conception and approval through hardware and software development, launch, flight operations, and data processing. To all these people, in government, universities, and industry, we extend our thanks and gratitude. In particular, we thank the large number of people at the GSFC who brought this challenging in-house project to fruition.

References

Albrecht A and Steinhardt P J 1982 *Phys. Rev. Lett.* **48** 1220

Alvarez L W, Buffington A, Gorenstein, M V, Mast, T S, Muller, R A, Orth, C D, Smoot, G S, Thornton, D D and Welch, W J 1974 *Observational Cosmology: The Isotropy of the Primordial Black Body Radiation*, UCBSSL 556/75 Proposal to NASA

Bardeen J M, Steinhardt P J and Turner M S 1983 *Phys. Rev.* D **28** 679

Barney R D 1991 *Illuminating Eng. Soc.* **34** 34

Bennett C L 1991 *Highlights Astron.* in press

Bennett C L *et al.* 1992a *ApJ* **391** 466

——1992b *ApJ* **396** L7

Bennett D P and Rhie S H 1992, preprint

Bertschinger E, Dekel A, Faber S M, Dressler A and Burstein D 1990 *ApJ* **364** 370

Bertschinger E 1992, in *Current Topics in Astrofundamental Physics*, eds. N Sanchez & Z Zichini, in press and in *New Insights Into The Universe*, eds. V J Martinez, M Portilla & D Saez, in press

Boggess N W 1991 *Highlights Astron.* in press

Boggess N *et al.* 1992 *ApJ* **397** 420

Bond J R, Carr B J and Hogan C J 1986 *ApJ* **306** 428

——1991 *ApJ* **367** 420

Bond J R and Efstathiou J 1984 *ApJ* **285** L45

——1987 *MNRAS* **226** 655

Bromberg B W and Croft J 1985 *Adv.Astron.Sci.* **57** 217

Carr B J 1988, *Comets to Cosmology*, ed. A Lawrence, (Springer Verlag:London), 265

Cheng E S 1991, *1st Annual Conf. on Astronomical Data Analysis Software and Systems*, NOAO, Tucson, AZ

Cheng E S *et al.* 1990 *Bull.Am.Phys.Soc.* **35** 971

——1991 *BAAS* **23** 896

Coladonato R J, Irish S M and Mosier C L 1990, *Third Air Force/NASA Symp. on Recent Advances in Multidisciplinary Analysis and Optimization*, (Hayward, CA: Anamet), 370

Davis M and Peebles P J E 1983 *ApJ* **267** 465

Davis M, Summers F J and Schlegel D 1992, *Nature*, in press

Efstathiou G 1990, in *Physics of the Early Universe*, eds. J A Peacock, A F Heavens and A T Davies (Edinburgh Univ. Press: Edinburgh, Scotland), p. 361

Efstathiou G, Bond J R and White S D M 1992, *MNRAS*, in press

Evans D C 1983 *SPIE Proc.* **384** 82

Evans D C and Breault R P 1983 *SPIE Proc.* **384** 90

Fabbri R 1988 *ApJ* **334** 6

Fabbri R and Melchiorri F 1979 *AA* **78** 376

Fabbri R, Andreani P, Melchiorri F and Nisini B 1987 *ApJ* **315** 12

Fixsen D J, Cheng E S and Wilkinson D T 1983 *Phys.Rev.Lett.* **50** 620

Franceschini A, Toffolatti L, Danese and De Zotti G 1989 *ApJ* **344** 35

Gorski K 1991 *ApJ* **370** L5

Gorski K M, Silk J and Vittorio N 1992 *Phys. Rev. Lett.* **68** 733

Gulkis S, Carpenter R L, Estabrook F B, Janssen M A, Johnston E J, Reid M S, Stelzried C T and Wahlquist, H D 1974 *Cosmic Microwave Background Radiation Proposal*, NASA/JPL Proposal

Gulkis S, Lubin P M, Meyer S S and Silverberg R F 1990 *Sci.Amer.* **262** 132

Gush H P, Halpern M and Wishnow E H 1990 *Phys. Rev. Lett.* **65** 537

Guth A 1981 *Phys.Rev.D* **23** 347

Guth A and Pi Y-S 1982 *Phys.Rev.Lett.* **49** 1110

Harrison E R 1970 *Phys.Rev. D* **1** 2726

Harwit M 1970 *Rivista del Nuovo Cimento II*, **253**

Hauser M G 1991a *Proc. Infrared Astronomy and ISO* Les Houches, June in press

——1991b *Highlights Astron.* in press

Hauser M G *et al.*1991 *After the First Three Minutes* eds. S S Holt, C L Bennett and V Trimble, (New York:AIP Conf. Proc. 222), 161

Hawking S 1982 *Phys.Lett.* **115B** 295

Hogan C J 1991 *Nature* **350** 469

——1992 submitted to *ApJ Letters*

——1993 *ApJ*, in press

Holtzman J A 1989 *ApJS* **71** 1

Hopkins R A and Castles S H 1985 *Proc. SPIE* **509** 207

Hopkins R A and Payne D A 1987 *Adv.Cryog.Eng.* **33** 925

Hoyle F and Burbidge G 1992 preprint

Janssen M A and Gulkis S 1992, *The Infrared and Submillimetre Sky After COBE*, eds M Signore and C Dupraz, (Dordrecht: Kluwer), 391

Kaiser N, Efstathiou G, Ellis R, Frenk C, Lawrence A, Rowan-Robinson M and Saunders W 1990 *MNRAS* **252** 1

Kaufman M 1976 *Ap.Sp.Sci.* **40** 369

Kogut A *et al.* 1992 *ApJ* in press

Kolb E and Turner M S 1987 *The Early Universe*, (Addison Wesley:Redwood City, CA)

Klypin A A, Sazhin M V, Strukov I A and Skulachev D P 1987 *Sov.Astr.Letters* **13** 104

Kumar V K, Freedman I, Wright E L and Patt F S, 1991, *Flight Mechanics and Estimation Theory Symposium*

Linde A 1982 *Phys.Lett.* **108B** 389

Loveday J, Efstathiou G, Peterson B A and Maddox S J 1992 preprint

Low F J and Tucker W H 1968 *Phys.Rev.Lett.* **22** 1538

Low F J *et al.* 1984 *ApJ* **278** L19

Lubin P, Villela T, Epstein G and Smoot G 1985 *ApJ* **298** L1

Magner T J 1987 *Opt.Eng.* **26** 264

Mather J C 1981 *Applied Optics* **20** 3992

——1982 *Opt.Eng.* **21** 769

——1984a *Applied Optics* **23** 584

——1984b *Applied Optics* **23** 3181

——1987 *Proc. 13th Texas Symp. Rel. Astrophys.*, (World Scientific Pub.:Singapore) p. 232

——1991 *Highlights Astron.*, in press

Mather J, Thaddeus P, Weiss R, Muehlner D, Wilkinson D T, Hauser M G and Silverberg R F 1974 *Cosmological Background Radiation Satellite*, NASA/Goddard Proposal

Mather J C, Toral M and Hemmati H 1986 *Applied Optics* **25** 2826

Mather J C *et al.* 1990a *ApJ* **354** L37

Mather J C *et al.* 1990b *IAU Colloq. 123, Observatories in Earth Orbit and Beyond*, ed. Y. Kondo, (Boston: Kluwer), 9

——1991a *After the First Three Minutes*, eds. S S Holt, C L Bennett and V Trimble, (New York: AIP Conf Proc 222), 43

——1991b *Adv.Sp.Res.* **11** 181

Matsumoto T, Akiba M and Murakami H 1988a *ApJ* **332** 575

Matsumoto T, Hayakawa S, Matsuo H, Murakami H, Sato S, Lange A E, and Richards P L 1988b *ApJ* **329** 567

McDowell J C 1986 *MNRAS* **223** 763

Meyer S S, Page L and Cheng, E S 1991 *ApJ* **371** L7

Milam L J 1991 *Illum.Eng.Soc.J.* **34** 27

Mosier C L 1991a "Thermal Design, Testing, and Analysis of the Cosmic Background Explorer's External Calibrator" in AIAA 29th Aerospace Sciences Conference

——1991b "Thermal Design of the Cosmic Background Explorer Cryogenic Optical Assembly" in AIAA 29th Aerospace Sciences Conference

Murdock T L and Price S D 1985 *AJ* **90** 375

Murdock T L, Burdick S V, Hauser M G, Kelsall T, Moseley S H, Silverberg R F, Toller G N, Stemwedel S W and Freudenreich H T 1991 *BAAS* **23** 1313

Negroponte J 1986 *MNRAS* **222** 19

Noda M, Christov V V, Matsuhara H, Matsumoto T, Matsuura S, Noguchi K, Sato S and Murakami H 1992 *ApJ* **391** 456

Olive K A, Schramm D N, Steigman G and Walker T P 1990 *Phys. Lett. B* **236** 454

Partridge R B and Peebles P J E 1967 *ApJ* **148** 377

Peebles P J E 1969 *Phil.Trans.Royal Soc.London, A* **264** 279

——1971 *Physical Cosmology*, (Princeton Univ. Press: Princeton, NJ)

——1980 *The Large-Scale Structure of the Universe*, (Princeton Univ. Press: Princeton, NJ)

——1981 *ApJ* **248** 885

——1982 *ApJ* **263** L1

——1984 *ApJ* **284** 439

——1991 *After the First Three Minutes*, eds. S S Holt, C L Bennett and V Trimble, (New York: AIP Conf Proc 222), 3

Peebles P J E and Wilkinson D T 1968 *Phys. Rev.* **174** 2168

Peebles P J E and Yu J T 1970 *ApJ* **162** 815

Penzias A A and Wilson R W 1965 *ApJ* **142** 419

Petuchowski S J and Bennett C L 1992 *ApJ* in press

Reynolds R J 1984 *ApJ* **282** 191

——1985 *ApJ* **294** 256

——1992 *ApJ* **392** L35

Rowan-Robinson M 1986 *MNRAS* **219** 737

Sachs R K and Wolfe A M 1967 *ApJ* **147** 73

Sampler H P 1990 *Proc. SPIE* **1340** 417

Shafer R A *et al.* 1991 *Bull.Am.Phys.Soc.* **36** 1398

Schaefer R K 1991 *After the First Three Minutes*, eds S S Holt, C L Bennett and V Trimble, (New York: AIP Conf. Proc. 222), 119

Smoot G F 1991 *Highlights Astron* in press

Smoot G F *et al.* 1990 *ApJ* **360** 685

——1991a *Adv.Space Res.* **11** 193

——1991b *ApJ* **371** L1

——1991c *After the First Three Minutes*, eds S S Holt, C L Bennett and V Trimble, (New York: AIP Conf. Proc. 222), 95

——1992 *ApJ* **396** L1

Starobinskii A A 1982 *Phys.Lett.* **117B** 175

Sunyaev R A and Zeldovich Ya B 1980 *Ann. Rev. Astr. Ap.* **18** 537

Torres S, *et al.* 1989 *Data Analysis in Astronomy* eds V di gesu, L Scarsi and M C Maccarone, Erice, June 20-27, Plenum Press

van Dalen A and Schaefer R K 1992 *ApJ*, in press

Vogeley M S, Park C, Geller M J and Huchra J P 1992 *ApJ* **391** L5

Volz S M and Ryschkewitsch M G 1990a *Superfluid Helium Heat Transfer*, eds J P Kelly and W J Schneider, (New York: AME), **134**, 23

Volz S M, Dipirro M J, Castles S H, Rhee M S, Ryschkewitsch M G and Hopkins R 1990 *Proc. Intl. Symp. Optical and Opto-electronic Appl. Sci. and Eng.*, (San Diego: SPIE), 268

Volz S M, Dipirro M J, Castles S H, Ryschkewitsch M G and Hopkins R 1991a *Adv.Cry.Eng.* **35B** 1703

Volz S M, Dipirro M J, Ryschkewitsch M G and Hopkins R 1991b *Adv.Cry.Eng.* **37A** 1183

Volz S M and DiPirro M J 1992 *Cryogenics* **32** 77

Vrtilek J M and Hauser M G 1992 in preparation

Walker T P, Steigman G, Schramm D N, Olive K A and Kang H-S 1991 *ApJ* **376** 51

Weiland J L, Hauser M G, Kelsall T, Moseley S H, Silverberg R F, Odegard N P, Spiesman W J, Stemwedel S W and Freudenreich H T 1991 *BAAS* **23** 1313

White R A and Mather J C 1991 *Astrophysics and Space Sciences Library*, eds M A Albrecht and D Egret, (Dordrecht: Kluwer),**171**, 29

Wilkinson D T 1986 *Science* **232** 1517

Wright E L 1990 *Ann. NY Acad. Sci. Proc.* **647** 190

——1991a *The Infrared and Submillimetre Sky After COBE*, eds M Signore and C Dupraz, (Dordrecht:Kluwer), 231

——1991b *ApJ* **375** 608

Wright E L *et al.* 1990 *BAAS* **22** 874

——1991 *ApJ* **381** 200

——1992 *ApJ* **396** L13

Zeldovich Ya B 1972 *MNRAS* **160** 1

Paper presented at the 13th Int. Conf. on General Relativity and Gravitation
Cordoba, Argentina, 1992: Part 1, Plenary Lectures

Gravity and quantum mechanics

Roger Penrose
Mathematical Institute
Oxford UK

0 Introduction

At a conference in honour of John S. Bell, held in October 1991, the following quotation, from a paper by Bell and Nauenberg (1966), was presented (in a talk by Dipankar Home):

> "... the quantum mechanical description will be superseded. In this it is like all theories made by man. But to an unusual extent its ultimate fate is apparent in its internal structure. It carries in itself the seeds of its own destruction."

I was struck by a similarity of sentiment, as expressed here, with one that I have myself often expressed, but now in relation to general relativity; e.g. (Penrose 1991):

> "... in a clear sense, general relativity predicts its own downfall as a complete description of the structure of space-time"

There is, indeed, a remarkable parallel, in this regard, between these two great physical theories. Both theories are now known to be exceptionally accurate, within the range of phenomena to which they are applied; yet both present us with profound difficulties. In the case of general relativity, the profound problem is that of *space-time singularities*, whose presence in Einstein's theory is an implication of the theory itself (cf. e.g. Hawking and Penrose 1970). In the case of quantum theory, the difficulty is the so-called *measurement problem*, which still has no really satisfactory solution. Many different viewpoints are expressed in relation to the measurement problem, often accompanied by claims of some kind of solution - but where the "solution" proposed satisfies none of those holding to an opposing viewpoint.

Despite the fact that *both* theories have their profound difficulties, the normal attitude to them, amongst contemporary physicists, is very different in the two cases. The standard reply to the question of the space-time singularities of classical general relativity is that that theory should be modified by applying to it, in some appropriate way, the rules of standard quantum theory - or, if this fails to work, then the classical Einstein theory itself should be changed (as would be the case with Kaluza-Klein-type theories, supergravity, and the classical limit of superstring theory) so as to force it into a more amenable shape for the consistent application of quantum procedures. When it comes to quantum theory's own problems, on the other hand, the normal attitude among contemporary physicists seems to be that the problem ought just to go away - if we only understood the theory itself properly. There is little suggestion that the very rules of quantum mechanics might in any way be in need of modification, or that quantum mechanics might itself derive any benefit from general relativity. (Isolated expressions of

dissent from this general viewpoint have, however, been put forward from time to time; cf. Károlyházy 1966, 1974, Károlyházy, Frenkel, and Lukács 1986, Komar 1969, Diosi 1989, Penrose 1981, 1989.)

My own attitude to these problems is a much more even-handed one than that normally adopted. I believe that the sought-for union between general relativity and quantum theory will involve as much change in the structure of quantum theory as in general relativity. I believe that *both* quantum theory and general relativity are (superb) approximations to some hitherto undiscovered new theory, where each would be valid in some appropriate limiting sense. The solution to the singularity problem and to the measurement problem would then both find their resolutions within this new theory; indeed, the solutions to these two problems should arise, accordingly, as two sides of but a single coin.

1 Quantum theory's fundamental problem

Since the singularity problem of general relativity is now generally accepted as providing a limit to that theory's classical applicability, I shall concentrate here on the central problem of quantum theory. I think that a very profound remark concerning different people's attitudes to quantum theory was made to me some years ago in a dinner-table comment by Bob Wald:

"If you really believe in quantum mechanics, then you can't take it seriously."

This expresses the fact that it is only those who dissent from the standard "Copenhagen" view who are prepared to regard the state-vector $|\psi>$ as actually representing (even an approximation to) physical *reality*. Niels Bohr, on the other hand, was one of the strongest proponents of the idea that $|\psi>$ was to be regarded merely as a calculational tool, an expression only of our knowledge about a physical system, and to be used simply for the mathematical calculation of probabilities with regard to the various results of "measurements" that might be performed on a system. Bohr, indeed, was someone who really did believe in quantum mechanics, and so was unable to take $|\psi>$ seriously as a description of actual physical reality.

Those who *do* take $|\psi>$ seriously as an objective description of the physical world - although possibly only a provisional or approximate one, perhaps to be superseded if quantum theory is someday replaced by an even more accurate theory - seem to me to fall into two camps. There are those who believe that the present theory of the way that $|\psi>$ evolves, namely *unitary evolution* **U**, must be preserved at all costs, and that the phenomenon of *state-vector reduction* **R** is some kind of illusion; on the other hand there are those who believe that **U** applies only to an approximate (though highly accurate degree) and that **R** represents a real physical phenomenon that effectively interrupts, from time to time, the continuous evolution according to **U** (such as might be entailed by some modification of the Schrödinger equation). In my own view, it is those in this latter camp (and I would count myself among their number) who are taking the formalism of quantum mechanics most "seriously" of all, because they believe that *both* the main ingredients of the theory, indeed **R** as well as **U**, are to be taken seriously as describing something objectively real about the evolution of the world. Among those in the latter camp would be such as Károlyházy 1974, 1986 and his co-workers, Pearle 1985, 1989, Ghirardi, Rimini, and Weber 1986, Diosi 1989 moreover, I would count John Bell as essentially being in this camp.

Belonging to the former camp within the group who "take $|\psi>$ seriously" would be those who follow a "many-worlds" viewpoint with regard to the quantum state. Accordingly, $|\psi>$ always evolves according to **U**, but all the different outcomes of a "measurement" must co-exist in different "worlds", each being perceived by a separate copy of any observer. The difficulties with this viewpoint lie, to my own mind, not so much in its lack of economy and in its extreme stretching of one's notions as to what "reality" should encompass, but in its incompleteness with regard to its descriptions as to what a conscious observer would actually perceive about the world and about the probability values that such an observer would assign to different perceived events. In short, in itself, it provides no solution to the "Hilbert space basis problem" or to the problem of why squared moduli of amplitudes should actually become probabilities. Also in this former camp might be those who hold to some form of "decoherence" explanation of R, although the standard explanations of this kind, described by John Bell as FAPP ("for all practical purposes") explanations, could really be satisfying only to those who do *not* take $|\psi>$ seriously as providing an actual description of the physical world. With regard to those "decoherence" viewpoints, such as those put forward by Griffiths (1984), Gell-Mann, and Hartle (1990), which adopt a path-integral-type picture of reality, I would regard them as being to some extent *modifications* of standard quantum theory (which would place them in the latter camp), but in any case, as not providing a real resolution of the measurement problem. For they accept their own versions of **U** and **R** as things with distinct mathematical descriptions, and they do not tell us under what actual physical circumstances a "measurement" would be deemed as taking place.

2 Two sides to the state-reduction phenomenon

Consider a simple (thought) experiment, where a photon source (the lamp L) and a detector (the photo-cell P) are placed at opposite ends of a hall, with suitable paraboloidal or ellipsoidal mirrors placed so as to ensure that virtually every photon emitted by the source would reach the detector, provided that there is no obstruction on the line joining them (fig. 1). Let us now imagine that there is a half-silvered mirror M placed mid-way between them on this line, tilted at 45° to the line. We are to take it that any photon moving along this line in *either* direction, when reflected off the mirror, would be absorbed at a point of the hall wall (at B if the photon comes from the direction of L, and at A if the photon were to come from P). Suppose that from time to time photons are emitted by L and that from time to time P (assumed to be a 100% efficient detector) registers the reception of a photon. Assume, also, that every time L emits a photon, it registers this fact (again with 100% efficiency), so that all the emission and reception events are clear-cut *measurements*.

Suppose, now, that it is given that L has emitted a photon. We ask: what is the probability that P receives one? Clearly the answer is $\frac{1}{2}$. This is a consequence of the standard quantum rule whereby probabilities are obtained by squaring the moduli of complex amplitudes. The photon wave function splits into two components, each of which has an amplitude $\frac{1}{\sqrt{2}}$ (times a possible phase factor). One component reaches the detector at P, and the other reaches the wall at B. The squared modulus of each is $\frac{1}{2}$, so the respective probabilities of reaching P or B are each $\frac{1}{2}$.

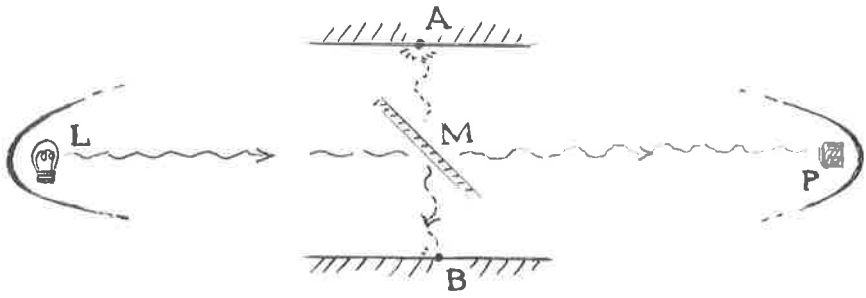

Fig.1 Photons, aimed towards the detector P, are emitted from time to time by the source L. Between the two is a half-silvered mirror M which partially deflects the photons to the absorbing wall at B. A photon ejected from the wall at A also could reach P. The probability that P receives a photon, given that L emits one, is governed by the quantum-mechanical squared-modulus rule, whereas the probability that L has emitted one given that P receives one is determined by the second law of thermodynamics.

Let us now ask a time-reversed kind of question. Suppose that we are given that P has received a photon. We now ask: what is the probability that L had emitted one? Now the answer is certainly not $\frac{1}{2}$, but we have a probability of essentially 1 that the photon came from L and a probability of essentially 0 that it came from A, which is the only other possibility. It is virtually certain that the photon came from the lamp L rather than from the wall at A. However, had we tried to use the "squared modulus" rule for calculating these probabilities, we should have indeed obtained $\frac{1}{2}$ for each of them. This discrepancy has nothing to do with the fact that it is usual to evolve wave functions into the future rather than into the past. Precisely the same answers are obtained whichever way we evolve. The conclusion is that the "squared modulus" rule simply does not work if we try to apply it to obtain retrospective probabilities. There is no real reason why it should. The miracle is that such a simple and elegant rule indeed works in the future time-direction!

If we were to ask what rules indeed govern probabilities in the past time direction, then we are forced to consider such matters as the second law of thermodynamics. This law is certainly playing a role here, because for a photon to jump out of the wall at A in order to be reflected off the mirror and arrive at the detector P, a severe violation of the second law would be needed. Basically it is the second law that is responsible for a virtual vanishing of the probability of a photon emerging from the wall at A.

It is clear from all this that there is no necessity for the probabilities in the past and future directions to match up in a particular experimental situation, such as this. However, there are strong reasons for expecting that if we were to consider, in some appropriate sense, the totality of *all* possible "experimental situations", then there would indeed be a past-future balance for the probabilities, taken as a whole. Basically, we must balance the physics responsible for the second law of thermodynamics against the physics responsible for the quantum-mechanical probabilities - in order to show that these two areas of physics must actually be two different aspects of the *same* physics.

3 The role of the Weyl curvature hypothesis

I do not wish to repeat the entire argument here, since it is one that I have given many times before (see, particularly, Penrose 1981, 1989). The essential point is that the second law arises ultimately from the enormous constraint on the space-time geometry that was operative at the big bang singularity. This initial constraint (which for a 10^{80}-baryon universe would amount to a restriction of the phase-space down to a region whose volume is about $\exp(-10^{123})$ of that of the entire phase space) starts the universe off with a very tiny entropy, as compared with what it "might have been", and the entropy has been rising ever since, in accordance with the second law. The simplest way to impose such a constraint is to take it as a geometrical restriction: the initial Weyl curvature is to be zero - or at least to be much smaller, in some appropriate sense, than it would have been for a generic big bang. This restriction, which is taken to apply only to *initial*, and not to final space-time singularities (so that we can derive the time-asymmetric second law), is referred to as the *Weyl curvature hypothesis* (WCH). (In recent work, R.P.A.C.Newman has shown that a form of WCH due to K.P.Tod can be used to derive an initial Friedmann-Robertson-Walker structure for the early universe, assuming an appropriate perfect-fluid state.)

Thus it is WCH (or something closely of this nature) that we seem to have to balance against the "squared modulus" rule of the **R** part of quantum mechanics. Moreover, if we take it that this singularity structure, as implied by WCH, is a feature of the correct quantum gravity theory - or rather, of the putative correct *new* theory out of which classical general relativity and standard quantum mechanics must both emerge as appropriate limiting cases - then we appear to conclude that it is this new theory that must also be responsible for the probabilities involved in **R**. In other words, the modification of quantum theory that would be needed in order for us to be able to understand **R** as a real physical process must be a modification that operates at the level where gravitational effects begin to become quantum-mechanically important.

In fact the connection can be made more explicit if we consider the "Hawking box" thought experiment (described in Penrose 1981, 1989). In this situation, the phase-space volume gain that occurs with experiments like that discussed in the previous section, where in effect there are more possible outputs than there are inputs (here: P and B allowed as outputs, whereas L is allowed as an imput but A is effectively forbidden), is balanced against a phase-space loss that occurs in black holes. There is a net loss because black holes are allowed by WCH, but their time-reverses - the white holes - are forbidden by it. The black hole singularity is a future singularity which absorbs information, whereas a white hole's singularity would be a past singularity, like the big bang, but with an infinite rather than zero Weyl curvature. The absorbing of information by a black hole's singularity is what is responsible for the Hawking effect, according to Hawking's original derivation (fig. 2). It should be mentioned, however, that there are some opposing points of view, according to which it is argued that if the entire history of a black hole's formation and ultimate disappearance (or possible non-disappearance) due to its Hawking evaporation is taken into account, then information is not actually lost. My own considered opinion is that such information restoration is not plausible, particularly if one is to believe that something of the nature of WCH must hold, so that white-hole singularities, with their potential for creating new phase-space volume, are not permitted.

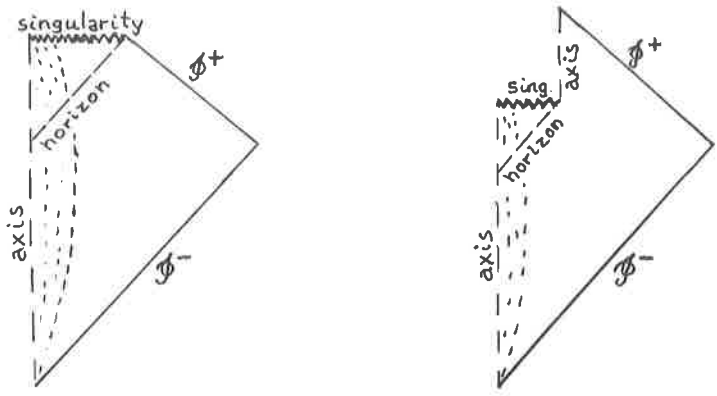

Fig.2 Hawking evaporation. The conformal diagram on the left gives the geometrical background for Hawking's original derivation of the presence of Hawking radiation - as an effect of a loss of information in the Black hole's singularity. The diagram on the right takes into account the final disappearance of the hole from the back-reaction of this radiation on the space-time geometry.

4 A gravitational origin for R?

An implication of the preceding discussion is that we should look for some criterion for the onset of quantum state-vector reduction which is of a gravitational character. There are, indeed, other reasons for suspecting that the standard quantum formalism might not apply without change in situations where the curved-space features of general relativity begin to become significant. For example, the normal ideas of energy, momentum, and angular momentum, in a quantum context, relate these physical concepts directly to symmetries of the space-time manifold, whereas in a general-relativistic setting, such symmetries would normally be absent. This leads to certain severe difficulties with regard to quantization, since it is indeed the quantum rules for energy, momentum, and angular momentum that provide the initial guide to quantization in a normal flat-space setting. (We recall that the passage from a classical to a quantum description - i.e. from a symplectic manifold to Hilbert space - is not a well-defined mathematical procedure, except in the presence of further structure such as that supplied by space-time symmetries; cf. Woodhouse 1980.) Even the notion of positive/negative energy (i.e. frequency) splitting, vital for the setting up of quantum field theory, is not well-defined in a general curved space-time.

These difficulties occur even for the problem of quantizing within a curved space-time background. As is well known, the problems that arise when it is the curvature of space-time itself that has to be subjected to the laws of quantum mechanics are of a much more serious nature. One must ask, in particular, how it is possible to interpret a physical system in which different space-time geometries are being subjected to quantum linear superposition; indeed, even worse than that, one must ask how is it possible to interpret *other* physical objects that try to inhabit such a curious quantum-superposed background? When there is no natural correspondence even between the individual

points in the different classical space-times that are to be superposed, then it is hard to see how interference between the different physical states associated with the different geometries can be understood.

What has this to do with state-vector reduction **R**? Let us consider a type of situation like a "Schrödinger's cat", in which one strives to produce a state in which a pair of macroscopically distinguishable alternatives are linearly superposed. For example, in

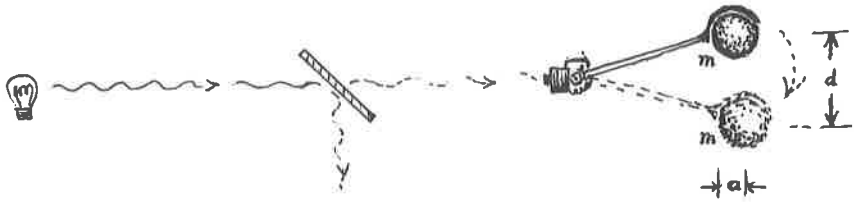

Fig.3 A photon impinges on a half-silvered mirror, so that only the transmitted part of its wave function is received by a device which if activated would move a macroscopic spherical lump of mass m and radius a through a distance d. Is there stage at which linear superposition between the two possible positions of the lump fails, and the lump actually becomes localized in one position or the other?

Fig. 3 we have a situation in which a photon impinges upon a half-silvered mirror, and the photon state becomes a linear superposition of being transmitted through it and reflected by it. The transmitted part of the photon's wave function activates (or would activate) a device which moves a macroscopic spherical lump from one location to another. So long as Schrödinger evolution **U** holds good, the "location" of the lump becomes a quantum superposition of its being in the original position with its being in the displaced position. As soon as **R** comes into effect, we are allowed to consider that the lump is in either one position or the other - and a "measurement" has been performed. The idea here is that this is an entirely objective physical process which occurs whenever the mass of the lump is large enough or the distance it moves is far enough. In particular, it has nothing to do with whether or not a conscious observer may happen to have actually "observed" the movement or otherwise of the lump. (In this, I am imagining that the device that detects the photon and moves the lump is itself "small" enough that it can be treated entirely quantum mechanically, and it is only the lump that registers the measurement. For example, in an extreme case, we might simply imagine that the lump is poised sufficiently unstably that the mere impact of the photon would be adequate to cause it to move off significantly.)

How would such a situation be treated according to the standard **U** procedure of quantum mechanics? After the photon has encountered the mirror its state would have to be considered as a non-local system, with the two parts of its wave function in two very different locations. One of these parts then becomes entangled with the device and finally with the lump. We thus have a quantum state which involves a linear superposition of two quite dfferent positions for the lump. Now the lump will have its gravitational field, which must also be involved in this superposition. Thus, the state also involves a superposition of two different gravitational fields - i.e., according to Einstein's theory, with two different space-time geometries! The question is: is there a point at which the two geometries become sufficiently different from each other that the rules of

quantum mechanics must change, and rather than forcing the different geometries into superposition, Nature chooses between one or the other of them to effect the reduction procedure **R**?

5 The weak-field gravitational symplectic integral

As a guide to establishing a plausibe criterion, in accordance with the foregoing general ideas, we consider a certain integral (Fierz 1940), over a spacelike Cauchy hypersurface Σ, which determines the symplectic structure on the function-space of weak vacuum gravitational fields (i.e. massless fields of spin 2 in flat space). This integral (up to a real factor) is:

$$\{K_{(1)}, K_{(2)}\} = \int_\Sigma (\Psi_{(1)ABCD}\bar{\eta}_{(2)A'}^{BCD} + \bar{\Psi}_{(1)A'B'C'D'}\eta_{(2)A}^{B'C'D'})d^3x^{AA'},$$

where the fields (labeled by (1) and (2), respectively) are described by linearized curvatures in Minkowski space

$$K_{abcd} = \Psi_{ABCD}\varepsilon_{A'B'}\varepsilon_{C'D'} + \varepsilon_{AB}\varepsilon_{CD}\bar{\Psi}_{A'B'C'D'},$$

where we have a (Dirac) chain of potentials

$$\begin{aligned} \nabla_{BB'}\eta_A^{B'C'D'} &= \chi_{AB}^{C'D'}, & \nabla^{AA'}\eta_A^{B'C'D'} &= 0, \\ \nabla_{BB'}\chi_{AB}^{C'D'} &= \gamma_{ABC}^{D'}, & \nabla^{AA'}\chi_{AB}^{C'D'} &= 0, \\ \nabla_{BB'}\gamma_{ABC}^{D'} &= \Psi_{ABCD}, & \nabla^{AA'}\gamma_{ABC}^{D'} &= 0, \\ & & \nabla^{AA'}\Psi_{ABCD} &= 0, \end{aligned}$$

all of Ψ_{ABCD}, $\gamma_{ABC}^{D'}$, $\chi_{AB}^{C'D'}$, $\eta_A^{B'C'D'}$ being symmetric in their unprimed and primed indices separately. (See Penrose and Rindler 1984 for the relevant spinor notation.) In vacuum, the divergence of the integrand vanishes, showing that (with suitable fall-off at infinity), the integral is independent of the choice of Σ. The symplectic form $\{,\}$ is closely related to the scalar product $<|>$. We have $< K \mid K >= i\{K, \mathbf{J}K\}$, where **J** multiplies the positive-frequency part of K by i and the negative frequency part by -i. Integrating by parts, we find, schematically, (for fields falling off suitably at infinity):

$$\int \Psi_{(1)}\bar{\eta}_{(2)} = -\int \gamma_{(1)}\bar{\chi}_{(2)} = \int \chi_{(1)}\bar{\gamma}_{(2)} = -\int \eta_{(1)}\bar{\Psi}_{(2)},$$

whence

$$\{K_{(1)}, K_{(2)}\} = -\{K_{(2)}, K_{(1)}\}.$$

The quantity $\gamma...$ describes the linearized spin-coefficients or, equivalently, the linearized Christoffel symbols; moreover, the Hermitian part of $\chi...$ describes the linearized metric.

We can use this symplectic integral as a measure of the *difference* between two weak (vacuum) gravitational fields. When this difference reaches order unity, all quantities being measured in *absolute units* ($G = c = \hbar = 1$), we can say that the two fields are "significantly different". The idea is, roughly, that when two (weak) gravitational fields are judged as being significantly different according to this criterion, then quantum linear superposition between them cannot be maintained, and the state reduces to one or the other of them - so **R** has been effected!

6 A gravitational criterion for the onset of R

How can we apply this kind of idea to the situation considered in **4.**, as illustrated in Fig. 3? A difficulty is that the symplectic integral of **5.** applies (and is divergence-free) only in the vacuum region, whereas we need it in situations where the lumps are actually present. Any spacelike Cauchy hypersurface Σ must intersect the lump (in both its alternative locations). One point to note, however, is that there is actually no need for Σ' to be spacelike everywhere. It just has to be topologically deformable to a Cauchy hypersurface. As it stands, this is no direct help, but it indicates that we need not feel restricted to spacelike hypersurfaces in applying the integral expression.

What seems suggestive, however, is to perform the integral over a *timelike* hypersurface Σ' in a region between the two locations of the lump after they have separated from one another, as indicated

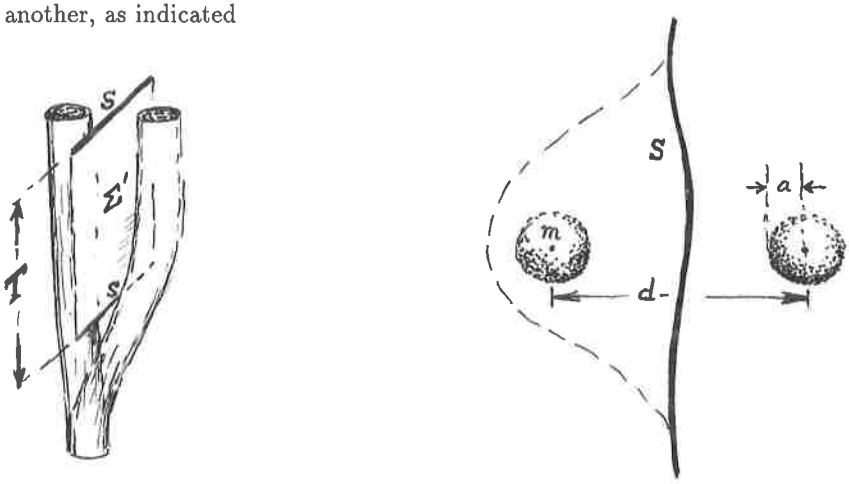

Fig.4 The two alternative locations of the lump, as depicted in Fig.3, are indicated, on the left, in a space-time diagram. The symplectic integral is performed over a portion of timelike hypersurface, suggesting that a criterion for reduction to take place is when this integral becomes of order unity. The spatial configuration is depicted on the right.

in Fig. 4. The hypersurface Σ' is bounded between two times, and let us call the time-interval between them T. A tentative suggestion might now be that the system reduces in a timescale T such that the symplectic integral, applied between the gravitational fields of the two possible positions of the lump, is of order unity.

The situation is basically a Newtonian one, assuming that, compared with the speed of light, the lump moves slowly, and the gravitational escape velocity at its surface is also tiny. In the Newtonian limit, for an essentially static situation, and taking a hypersurface Σ' that is also static - i.e. the product of a spacelike 2- surface S with an interval of the time axis (of duration T) - our symplectic integral becomes (some simple constant multiple of)

$$T \int_S (\phi_{(2)} \vec{\nabla} \phi_{(1)} - \phi_{(1)} \vec{\nabla} \phi_{(2)}) \bullet d\vec{x}.$$

Here ϕ denotes the Newtonian gravitational potential due to the lump. The parenthetic suffix specifies the lump location. We note that the integrand is divergence-free in

vacuum, so the integral remains unchanged under topological deformations of S through regions of matter-free space. Where lump matter is present (density ρ), we make use of

$$\vec{\nabla} \bullet (\phi_{(2)}\vec{\nabla}\phi_{(1)} - \phi_{(1)}\vec{\nabla}\phi_{(2)}) = -4\pi\phi_{(2)}\rho_{(1)} + 4\pi\phi_{(1)}\rho_{(2)};$$

thus, moving S across one of the two possible locations for the lump, we find that the value of our integral is $4\pi T$ times the gravitational energy of one instance of the lump in the gravitational field of the other. This would suggest that the time that it takes for the state to reduce is a simple multiple of the reciprocal of this energy.

This energy is

$$m^2/d,$$

where the lump has mass m, and the spatial displacement between the centres of its two alternative positions is d. However, the reciprocal of such a measure is not really a satisfactory suggestion as to the time it takes for reduction to occur in this situation, because for large displacements the measure gets smaller and tends to zero. Thus the suggested "time" would get longer, making it *less* likely for **R** to occur, the greater the displacement, rather than *more* likely as one would expect. Moreover, when the two instances of the lump overlap, there would be no way of locating Σ' so that it lies entirely in the vacuum region.

This suggests that the reduction time should be something a little different from this, but perhaps belonging to the same order of ideas. In fact a slight modification of this expression (arrived at some ten weeks following the Cordoba meeting) does give something reasonable. The expression m^2/d is the gravitational energy gained (in absolute units) if one moves one instance of the lump in from infinity, in the gravitational field of the other, until it reaches the required separation d. If we consider, instead, the more relevant energy of moving one instance of the lump away from the other, starting from coincidence, until they reach the required separation, then, taking the lump to have radius a, we obtain something of the order of m^2/a for this energy. Although the energy now indeed increases as the separation increases, the additional energy in moving from the contact position to all the way out to infinity is of the same order as that in moving from coincidence to the contact position. Thus, as far as orders of magnitude are concerned, one can ignore the contribution due to the displacement d, and take the reduction time to be of the order of

$$a/m^2.$$

It is reassuring that this gives very "reasonable" answers in certain simple situations. For example, in the case of a nucleon, where we take a to be 10^{-13}cm, which in absolute units is about 10^{20}, and m to be about 10^{-20}, we get a reduction time of 10^{60}, which is about a Hubble time. If we consider a droplet of water of radius 10^{-5}cm, we get a reduction time of about a day; if of radius 10^{-4}cm, the reduction time, according to this scheme is roughly a second; if of radius 10^{-3}cm, then about 10^{-5} of a second. So far, this seems to be quite plausible, but clearly more work is needed to see whether the idea will survive more stringent examination.

I am grateful to many colleagues for valuable comments, most particularly Abhay Ashtekar and Ted Newman.

References

Bell, J.S. and Nauenberg, M. (1966) The moral aspect of quantum mechanics. In *Preludes in Theoretical Physics* (eds. A. De Shalit, H. Feshbach, and L. Van Hove; North-Holland, Amsterdam; pp.279- 86). Reprinted in: Bell, J. S. (1987) *Speakable and Unspeakable in Quantum Mechanics.* (Cambridge Univ. Press, Cambridge).

Diosi, L. (1989) *Phys. Rev.* A40, 1165.

Fierz, M. (1940). Uber den Drehimpuls von Teichlen mit Ruhemasse null und beliebigem Spin. *Helv. Phys. Acta* 13, 45-60.

Gell-Mann, M. and Hartle, J.B. (1990) in *Complexity, Entropy, and the Physics of Information, SFI Studies in the Science of Complexity, Vol. VIII* (ed. W. Zurek; Addison Wesley, Reading).

Ghirardi, G.C., Rimini, A., and Weber, T. (1986). Unified dynamics for microscopic and macroscopic systems. *Phys. Rev.* D34, 470

Griffiths, R. (1984) *J. Stat. Phys* 36, 219.

Hawking, S.W. and Penrose, R. (1970) The singularities of gravitational collapse and cosmology, *Proc. Roy. Soc. (London)* A314, 529-548.

Károlyházy, F. (1966) *Nuovo Cim.* A42, 390.

Károlyházy, F. (1974) Gravitation and quantum mechanics of macroscopic bodies, *Magyar Fizikai Polyoirat* 12, 24.

Károlyházy, F., Frenkel, A. and Lukács, B. (1986) On the possible role of gravity on the reduction of the wave function in *Quantum Concepts in Space and Time* eds. R.Penrose and C.J.Isham (Oxford University Press, Oxford).

Komar, A.B. (1969) Qualitative features of quantized gravitation *Int. J. Theor. Phys.* 2, 157-60.

Pearle, P. (1985) 'Models for reduction', in *Quantum Concepts in Space and Time*, eds., C.J. Isham and R. Penrose, (Oxford Univ. Press, Oxford).

Pearle, P. (1989) Combining stochastic dynamical state-vector reduction with spontaneous localization. *Phys. Rev.* A39, 2277- 89.

Penrose, R. (1981) Time-asymmetry and quantum gravity in *Quantum Gravity 2*, eds D.W.Sciama, R.Penrose and C.J.Isham (Oxford University Press, Oxford).

Penrose, R. (1989) *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*, (Oxford Univ. Press, Oxford).

Penrose, R (1991) What does the big bang tell us about quantum gravity? *Mem. S.A.It.* 62, No.3, pp.607-614.

Penrose, R. and Rindler, W. (1984) *Spinors and Space-Time*, Vol. 1: Two-Spinor Calculus and Relativistic Fields (Cambridge University Press, Cambridge).

Woodhouse, N.M.J. (1980) *Geometric Quantization* (Clarendon Press, Oxford).

Paper presented at the 13th Int. Conf. on General Relativity and Gravitation
Cordoba, Argentina, 1992: Part 1, Plenary Lectures

Sources of gravitational waves and their detectability

B.F. Schutz

Department of Physics and Astronomy, University of Wales College of Cardiff, Cardiff, UK

Abstract. With the funding of the LIGO project in the US and the likely funding of VIRGO in Europe, we can now anticipate the detection and astrophysical study of gravitational waves before the end of this decade. I review the position of detector development today and the plans for the large-scale interferometers. I then survey likely sources of gravitational waves: supernovae (gravitational collapse), coalescences of compact-object binary systems, pulsars, and a stochastic background. I try to make realistic assessments of the uncertainties in our models of these sources, and in our estimates of their event rates; and I point out what information we may be able to extract from observations when they are made. I pay particular attention to the uncertainties that allow the possibility that even first-stage interferometric detectors might be able to detect gravitational waves from gravitational collapse or coalescing binaries. I also point out that, before observers will be able to extract maximum information from a coalescing binary signal, relativists will have to make more progress in understanding the two-body problem in general relativity.

1. Introduction

1.1. Gravitational wave astronomy

After many decades of developing techniques for gravitational wave detection and of studying possible sources of gravitational waves, we may finally be standing on the threshold of gravitation wave astronomy. The approval last year of the LIGO project by the US Congress, and the expected approval later this year of the VIRGO project in Europe, at last marks the beginning of the construction of instruments whose sensitivity should exceed by a large margin the targets that theoretical studies of sources have set for detecting waves that pass through the Earth a few times per year.

It will, nevertheless, be several years before these big interferometers begin operating, and it will probably be several years more before they begin returning observations

with the regularity that will be needed to extract reliable astronomical information from them. In the meantime, bar detectors of increasing sensitivity will be watching for rare events of unusual strength, and it could well happen that they will turn in the first detections of gravitational waves before the interferometers get going.

In this review I will discuss our present understanding of possible sources of gravitational waves, in the context of efforts to detect them. First I will review the most important recent detector developments. Then I will focus on the most promising sources, namely supernova explosions, coalescing compact-object binary systems, and a cosmological background. Because the interferometers will be built in two stages, I will speculate as well on what they might be capable of detecting even with their lower, first-stage sensitivity.

1.2. Gravitational wave uncertainties

The most important *intrinsic* characteristics of a gravitational wave are the total energy it carries and its frequency, or its spectral bandwidth. In addition, for reliable predictions about the likelihood of detecting gravitational waves we need estimates of the distance to the source and the number of sources out to that distance. These are bound up with the sensitivity of detectors, so our review begins with a survey of the types of detectors and their expected characteristics.

Not surprisingly, the astrophysics of almost all potential gravitational wave sources is sufficiently uncertain that we have to allow for wide ranges of variability in at least some of the characteristics needed for our predictions. This is not a reason to be discouraged; on the contrary, only by making gravitational wave observations can we fill in the gaps that present observational techniques leave in our understanding of many astronomical systems.

The one exception to this "rule" of uncertainty is the coalescing binary: the radiation emitted from the orbital motion of two neutron stars or black holes as they spiral ever closer together from a tight binary orbit. We are sufficiently sure of the strength of the waves and of the number of sources in a given volume of space to have great confidence that future interferometric detectors like LIGO will see them. Interestingly, this source presents a challenge to relativists at this time: can we solve the two-body problem with sufficient accuracy to allow us to extract the wealth of information that the signal contains? I will return to this question at the end.

There is at least one aspect of our predictions about gravitational waves that is not a major source of uncertainty: relativity theory itself. The Binary Pulsar system, PSR1913+16, discovered by Hulse & Taylor[1] in 1975, has confirmed the predictions of general relativity to such an accuracy that it is now necessary to take into account the acceleration of the system caused by the Galaxy before arriving at the best value of the period change attributable to gravitational radiation.[3] This is described in much more detail in the article by Taylor in this volume.

1.3. Energy carried by a gravitational wave

The amplitude of the gravitational waves that pass the Earth regularly (once a month or more often) is likely to be well below 10^{-21}. The smallness of this number is a measure of the weakness of its interaction with the matter it passes through. This is at once a

benefit and a curse: a curse because it makes the waves hard to detect, a benefit because one can be sure that the radiation has not been significantly affected by anything it may have passed through since leaving its source. Unlike photons, which are easily scattered and reprocessed, gravitational waves carry their original information intact.

The weakness of the effect of gravitational waves on our detectors should not make us overlook the fact that they carry enormous amounts of energy. Using the Isaacson stress-energy tensor (see [2]), one can easily calculate that the energy flux in a gravitational wave of amplitude h and frequency f is

$$\mathcal{F}_{\rm gw} = 3.2 \times 10^{-3} \left[\frac{f}{1\ \rm kHz}\right]^2 \left[\frac{h}{10^{-22}}\right]^2\ \rm W\ m^{-2}. \tag{1}$$

In astronomers' language, a 1 kHz wave with amplitude 10^{-22} is as bright as a star of apparent magnitude -13, which is twice as bright as the full moon! This may last only a millisecond, but since the wave originates in a distant galaxy, it is clear that enormous energies are involved.

By integrating Eq. (1) over a sphere of radius r for a time τ, we find the relation between the amplitude h and the energy E carried by the wave during τ:

$$h = 1.4 \times 10^{-21} \left[\frac{E}{0.01\ M_\odot\ c^2}\right]^{1/2} \left[\frac{f}{1\ \rm kHz}\right]^{-1} \left[\frac{\tau}{1\ \rm ms}\right]^{-1/2} \left[\frac{r}{15\ \rm Mpc}\right]^{-1}. \tag{2}$$

2. Detector developments

2.1. Types of detector

The two principal types of detector under development are bar detectors and laser interferometers. In addition, a number of other techniques have been used, such as ranging to interplanetary spacecraft and searching for irregularities in timing of millisecond pulsars.

Bar detectors use the fundamental longitudinal normal mode of a cylindrical bar to detect waves. Although bars have their greatest sensitivity near the frequency of this mode, which is usually chosen to be near 1 kHz, they do not act as resonant detectors: impulsive gravitational waves do not last long enough to excite a resonance. Instead, bars use the high Q of this mode effectively to reduce the noise due to the thermal vibrations of the bar. For this reason, they have bandwidths of some tens of Hertz, much greater than the sub-milliHertz bandwidth of the resonant mode itself. In principle they can have even larger bandwidths, but this has not yet been achieved in practice. The narrow bandwidth affects another characteristic of bars: their poor time-resolution (worse than 0.1 s) makes them unable to discriminate between the time-of-arrival of a wave at one detector and another, so that networks of bars are relatively poor at determining the direction, and therefore the intrinsic amplitude, of gravitational waves.

Bars can be made more sensitive by reducing their physical temperature, and this is the major route for progress at the moment. Novel designs of bars, such as spherical ones, may allow further improvements. But somewhere near a sensitivity of 10^{-21}, bars encounter the quantum limit, where the energy left in the bar by the gravitational wave is less than the energy of one phonon of the longitudinal mode. This can be circumvented in principle, but again there are no practical ways as yet. Bars are therefore now, and

for the foreseeable future, narrow-band detectors which can in principle reach to about 10^{-21}. Despite this limitation, bars have one great advantage over interferometers: they exist and will be taking data regularly during the next few years, while interferometers are under construction.

Interferometers are described more fully in K. Danzmann's article in this volume. They do not involve any natural frequency, so they are intrinsically broad-band detectors. Their bandwidth extends from perhaps 10 Hz up to several kHz. This means that a network of 3 or more interferometers can deduce the direction to a source by measuring the time-delays among the various detectors. Optical tricks can be used to reduce their bandwidth if desired, such as for high-sensitivity observations of a particular pulsar. Interferometers are not troubled by the quantum limit at sensitivities of 10^{-22} or so, because they can take advantage of the length effect: the tidal forces in a gravitational wave increase in proportion to the size of the apparatus, so that a detector 4 km in size will experience 100 times the displacement of its end masses that a 40 m detector experiences.

In the near future, the only existing interferometers are various prototypes that are rarely used for data-taking, being required for technical development instead. In five years or so, once the kilometer-scale interferometers begin operating, they will give us our first broad-band look at the universe in gravitational waves, with a sensitivity comparable to the best that bars can achieve. Within a few years after that, they should be operating at a sensitivity that ought to guarantee a multitude of detections.

Detectors on the ground cannot hope to observe at really low frequencies, even though much of the radiation in the sub-milliHertz region is easy to predict, coming as it does from ordinary binary systems in our Galaxy. Ground vibration and even the Newtonian gravitational fluctuations produced by movements of atmospheric masses will mask signals at low frequency. Present space-based detection methods involve ranging to interplanetary spacecraft[4] and monitoring millisecond pulsars (see the article by Taylor). Both methods are improving, and will be our only access to this frequency window until dedicated interferometers are placed in space.

2.2. Current ground-based detectors

2.2.1. Cryogenic bar detectors. The best currently operating bar detectors are cryogenic, cooled to liquid helium temperatures (4.2 K). The sites currently operating include

- LSU. The cryogenic bar at Louisiana State University takes data continuously (you can even tune in to the data system on the Internet!). Its noise level is well below 10^{-18}.

- Rome. The University of Rome bar (called Explorer) is situated at CERN. It recently ran for 2 years, and will soon be back on the air. Its sensitivity is comparable to that of the LSU bar. The Rome and LSU bars have run in coincidence and the groups are currently analyzing 6 months of data.

- Perth. This cryogenic bar, made of niobium, is planned to be in operation by the end of 1992 at the University of Western Australia. Its sensitivity will be comparable to that of LSU and Rome. Three-way coincidence experiments are expected.

2.2.2. Interferometric detectors. Laser interferometers are presently being used as development tools for the larger-scale interferometers of the future. They occasionally have made observing runs. The main ones include:

- Garching. A 30 m prototype at the Max Planck Institute for Quantum Optics is currently being modified. Its broadband sensitivity at 1 kHz is a few times 10^{-18}.
- Glasgow. The 10 m prototype at the University of Glasgow has a sensitivity similar to that of Garching. The Garching and Glasgow instruments took 100 hours' data in coincidence with each other, the first such observing run between two interferometers. Analysis of these data sets is being done in my research group at Cardiff.
- Caltech. At the California Institute of Technology, the LIGO project operates a 40 m prototype, the largest in the world at present. It will soon be rebuilt, after which its sensitivity may well better that of Garching and Glasgow. There are no plans at present for long observation periods.
- MIT. The LIGO project also operates a small prototype at the Massachusetts Institute of Technology. This was one of the first to be constructed, and is used for testing interferometer techniques.
- Tokyo. There is a 10 m prototype at ISAS, near Tokyo. Its sensitivity is about 10^{-17}. It will soon to be replaced by a 100 m instrument, which is nearing completion. The 10 m detector took data over several days that partly overlapped with the Glasgow-Garching run. These data will soon be analysed for coincidences at Cardiff as well.

2.3. Future ground-based detectors

2.3.1. Ultracryogenic bars. The next generation of bars will be cooled to well below liquid helium temperatures by dilution refrigeration. The two such bars that are now under construction are:

- Rome. This bar, called Nautilus, has already been cooled (without its instrumentation) to below 90 mK. It is expected to begin observations later this year, and to approach a sensitivity of 10^{-19} within another year. With improvements in transducers, cryogenics, and other systems, it could in principle reach about 10^{-20}.
- Stanford. The group at Stanford University operated a cryogenic bar until it was destroyed by an earthquake. They are currently building an ultracryogenic detector similar to that of Rome. First tests of it are expected later this year. Coincident observing runs between these two bars offer the hope of seeing any large supernova-like event in our Galaxy.

2.3.2. Large-scale laser interferometers. There is general agreement that laser interferometers need to go to the scale of a few kilometers in order to achieve a sensitivity of 10^{-22} with foreseeable technology. Several projects have been proposed along these lines, and are at various stages of approval:

- LIGO. This project, a collaboration between the California Institute of Technology and the Massachusetts Institute of Technology, was approved last year by the US Congress. It expects to build two detectors, each 4 km long, at sites in Hanford (Washington) and in Louisiana. The Hanford installation would also contain a half-length detector. These detectors should reach a broadband sensitivity of 10^{-21} at Stage 1 (1997–8?) and can anticipate going beyond that to 10^{-22} with present designs. As technology improves, there is no fundamental barrier to further improvements beyond that.

- VIRGO. This Italian-French collaboration plans to build a detector of 3 km arm-length near Pisa. It has been approved by France and is now awaiting approval by Italy. Its capability will be similar to that of each LIGO detector, but with an emphasis on reaching very low frequencies (10 Hz or below). Coincident observations by the two LIGO and the VIRGO detectors could determine the direction to a source to within a degree or so.

- GEO. This is a collaboration between Germany (Garching and the University of Hannover) and the UK (Glasgow and Cardiff). It is stalled at the moment by funding problems in both countries, after having been approved in the UK in 1989. The plan is to build a 3 km detector near Hannover, with capabilities similar to each LIGO detector. The close separation of the two European detectors is attractive for doing cross-correlation searches for a stochastic background of gravitational waves (see below).

- AIGO. This Australian project is still in the planning stage, looking for international partners. Good sites exist in Australia, and its long baseline from the other detectors makes it attractive for improving the directional resolution of the network.

3. Gravitational collapse

We turn now to a discussion of possible sources of gravitational waves. Informed by our survey of the available ground-based detectors, we shall concentrate on sources at frequencies above about 10 Hz. Historically, the most important source that drove the development of bar detectors was the supernova, or more generally gravitational collapse.

3.1. What we know and what we don't know

Gravitational collapse is attractive as a gravitational wave source because it involves a considerable mass compressed to a high density on a very short timescale. Collapses that produce neutron stars lead to supernovae of Type II, where the formation of the star creates a rebound of infalling material that then blows away the original envelope of the star. It is less clear that collapses that lead to black holes will produce bright supernovae, since these events might involve rebounds that do not succeed in blowing away the envelope; this might subsequently fall in and convert the neutron star into a black hole. Moreover, many supernovae (Type Ia, particularly) do not appear to involve gravitational collapse. So there will not be a one-to-one relation between supernovae and strong bursts of gravitational waves.

Moreover, despite decades of theoretical study of supernovae of type II, it is still difficult to assess their likely importance as sources of gravitational waves. This is because optical and gravitational wave observations of supernovae measure effectively orthogonal quantities. Optical observations see the expanding cloud of ejecta; this is material that never reached a particularly high density, and which is expanding roughly spherically. Gravitational waves do not come from spherical motions; they measure the asymmetries in the very core of the collapse. Making a link between these two has proved difficult, not least because modelling three-dimensional supernovae with a full nuclear reaction network, hydrodynamics, and radiation transport is a job that exceeds the capacity of supercomputers available today.

There is also contradictory observational evidence on the subject. If large asymmetries do develop in collapse, they are likely to be related to rotation (see below). But young pulsars, which were presumably formed in such events, do not show dynamically rapid rotation rates. The Crab pulsar, for example, rotates at 30 times per second, which is very much slower than its breakup speed of 1–2 kHz. This agues that rotation is not dynamically important, at least in the majority of collapses that produce pulsars.

On the other hand, there is general agreement that there must be some asymmetries, because pulsars seem to be given a "kick" of 100–200 km/s when they are formed, in addition to any velocity they acquire due to binary breakup. There is even a case which may show a velocity of 1600 km/s.[5] While this is again only a moderate speed (the rotational speed of the Crab pulsar at its surface is almost 2000 km/s), if it comes from an asymmetric emission of radiation (neutrinos or gravitational waves) then the energy carried away by the asymmetric radiation must be a sizeable fraction (perhaps 10% or more) of the energy released by the explosion.

In the absence of detailed modelling, the best one can do is to calculate the radiation amplitude at the Earth of a gravitational wave produced in a distant galaxy by a collapse that converts a certain amount of energy into gravitational waves. Assuming that the duration of the burst is the timescale of the rebound, *i.e.* about a millisecond, then Eq. (2) implies that the amplitude of a burst of energy E occurring at a distance r is

$$h_{\rm collapse} = 10^{-21} \left[\frac{E}{10^{-2} M_\odot c^2} \right] \left[\frac{r}{15 \text{ Mpc}} \right]. \tag{3}$$

The strongest possible burst would emit the whole binding energy of a neutron star, about 0.1 $M_\odot\, c^2$. This would produce an amplitude of 3×10^{-18} if it occurred in our Galaxy, and 3×10^{-21} in the Virgo cluster. A more moderate, and plausible, amount would be 0.0 $M_\odot\, c^2$, which would give an amplitude of about 4×10^{-22} at 40 Mpc.

These numbers are very interesting in view of our discussion of detectors. Present day bars and interferometers could in principle see a strong burst in our Galaxy. The second-stage large interferometers could see a moderate burst twice as far away as the Virgo cluster. This is a volume of space containing many thousands of galaxies, in which hundreds of supernovae occur each year. Even if a small fraction of them should produce bursts as large as this, the second- stage interferometers should see a few per year. We will return below to what the first-stage interferometers might see.

3.2. Scenarios for collapse leading to strong radiation

If even a small fraction of gravitational collapse events do lead to strong gravitational waves, they will probably be driven into asymmetric collapse by rotation. Rotation has

two effects. The first is the obvious one, causing the collapsing core to form a oblate figure of rotation. The second effect develops only at sufficiently high rotation rates, and probably only for strong differential rotation: the development of non-axisymmetric "bar-mode" instabilities that cause the core to assume a tumbling tri-axial ellipsoidal shape.[6]

The radiation produced by the first effect is likely to be very small. Numerical simulations of axisymmetric gravitational collapse produce very little energy, perhaps $10^{-6}\ M_\odot\ c^2$.[7, 8, 9, 10]. There is also evidence that rotational effects slow the collapse and thereby lower the dominant frequency at which the radiation comes out. If this frequency turns out to be well below 1 kHz, then bar detectors will not see such events.

If there is enough rotation to lead to non-axisymmetry, then the resulting configuration must shed its excess rotational energy. Since the instability sets in when the total rotational energy is about 25% of the binding energy of the neutron star, it is reasonable to expect that this may lead to radiated energy of the order of $0.01 M_\odot\ c^2$ or more. The timescale and frequency of the radiation will depend in detail on the dynamics of the collapse.

If rotation does dominate collapse, then it might be possible to explain the absence of young, rapidly rotating pulsars in one of the following two ways:

- A rapidly rotating collapse eventually produces a black hole. Perhaps the outgoing shock is too weak to blow away the envelope, or perhaps the collapsed ellipsoidal core, with its rotational support, exceeds the upper mass limit of a slowly rotating neutron star.

- The collapsed core may have so much angular momentum that it fissions, expelling a small part of itself that carries off substantial angular momentum. The recoil might explain high-velocity pulsars.

Better hydrodynamical calculations even in Newtonian gravity would shed considerable light on these questions.

3.3. Supernovae and Stage-1 Interferometers

The usual assumption about supernovae is that they produce a burst of radiation in a timescale characteristic of the bounce, about 1 ms. This would be a broad-band burst at about 1 kHz. It is possible, however, that considerable radiation from a collapse event emerges at a frequency well below 1 kHz, particularly if rotation is involved. As an illustration, we construct the following simplified and optimistic scenario, in which first-stage interferometers could see a good number of events:

- Suppose the energy emerges at 100 Hz as a pulse lasting 10 ms. Then the amplitude of the signal increases by a factor of 3.

- Suppose the detector is optimized in recycling mode for detecting pulses at 100 Hz instead of 1 kHz. Then the noise goes down by a factor of 10.

The result is that the signal-to-noise ratio goes up by 30. Events that radiate $0.01 M_\odot\ c^2$ of energy become visible hundreds of megaparsecs away, so hundreds of thousands of supernovae per year become potential sources.

There is no reason in principle that the real situation should not be closer to this optimistic scenario than to the conventional one. Although the rotational frequency of the collapsed core must be in the kHz range for instabilities to set in, the non-axisymmetric deformation that emits the radiation may be a CFS-unstable bar mode,[11, 12] which could have a significantly lower frequency $f_{\rm CFS}$. In such as case, the supernova would produce a long, low-amplitude wavetrain at a frequency $f_{\rm CFS} < 1000$ Hz rather than a structureless burst of duration $1/f_{\rm CFS}$, but *the signal-to-noise ratio will go up by the same factor of* $(1000\ {\rm Hz}/f_{\rm CFS})^{(3/2)}$ *provided we have by then good enough models of the waveform to perform matched filtering.*[13]

This scenario requires that detector builders decide to control their noise sources down to lower frequencies, and then to optimize the interferometers for the frequency $f_{\rm CFS}$ rather than 1 kHz. These circumstances are not part of the present plans for the first stage detectors. But the potential payoff of doing so should be kept in mind. Even if detectors are not optimized for lower frequencies, matched filters should still be applied to the low-frequency data to look for such events; in this case the potential gain in signal-to-noise is still a factor of $1000\ {\rm Hz}/f_{\rm CFS}$.

Unfortunately, this optimistic scenario does not help bar detectors, since a low-frequency gravitational collapse burst would be outside their bandwidth.

3.4. Astrophysical payoffs of detecting supernovae

In addition to the satisfaction of finding a supernova, there are many astrophysical reasons for wanting to detect them. If bar detectors register a collapse event in our Galaxy, then:

- a coincident observation of neutrinos associated with the event would confirm it and test models of collapse, and may help give the direction to the event if it is not seen optically;
- this would define at least one type of supernova that can be a strong gravitational wave emitter.

If two interferometers detect a collapse event in a distant galaxy, then:

- they will give waveform information that will greatly constrain collapse models and the nuclear physics of neutron stars (the waveform information will always be available because the detection threshold will be at a signal-to-noise ratio of 7 or so, which means that any detection will yield extra information too);
- they may, from the characteristic frequencies of the waveform, be able to measure the mass of the compact object and therefore identify it as a neutron star or black hole.

Finally, if three or more interferometers detect the event, then they can do the above plus:

- they will fix the direction to the event (to better than a degree) and be able to measure its intrinsic amplitude h and polarization;

- the direction information may lead to an identification of the galaxy or cluster of galaxies in which the event occurred, from which one could infer a distance and thence, from the amplitude h, estimate the total energy released in gravitational waves;

- if the detection can be confirmed and its direction analyzed quickly, then optical astronomers can be alerted to look in that direction for a new supernova — it is rare for astronomers to have the opportunity to see a supernova before it reaches its maximum brightness.

4. Coalescing compact-object binaries

The first suggestion that the orbital radiation emitted just before two binary neutron stars coalesce would be an interesting source of gravitational radiation was, remarkably, made by Freeman Dyson[14] before the discovery of pulsars, at a time when the existence of neutron stars was speculative. This paper correctly estimated the principal features of the radiation and its sources — the timescale of a few seconds, the waveform of increasing frequency and amplitude that we now call a "chirp", even that the distance to a typical source would be 100 Mpc — but erred in considering that these events would be detectable by the first bar detectors that were then under construction by Joseph Weber. In fact, no bar detectors so far designed have the frequency bandwidth that would be necessary to detect a chirp.

Subsequent papers[15, 16, 17] developed the basic theory of such systems, with the view that the actual coalescence event could be detectable by bars, a possibility which still exists. But the present intensive interest in coalescing binaries arose from the realization[18] that the planned interferometers would have a more reliable chance of seeing these events than of seeing supernovae. In this section I review the main characteristics of these sources, the prospects of seeing them with first-stage and second-stage detectors, and the considerable astrophysical information that we can expect to extract from the waveforms, provided relativists can make progress on the 2-body problem in the next few years.

4.1. Basics

The Binary Pulsar PSR 1913+16 referred to earlier will, in about 10^8 years, evolve to a point when the stars are separated by about 250 km. By that time the orbit will have circularized and the system will be emitting radiation with a frequency of 70 Hz. From then until the stars begin to merge, the system will evolve in a predictable way, the frequency of the radiation increasing as the stars gradually spiral together. Over a period of some 5–6 s, the frequency will increase from 70 Hz to nearly 1 kHz. Theoretical calculations by Clark and Eardley[16] and more recent numerical simulations by Oohara and Nakamura[19] have shown us that the stars do not begin to merge until they are so close together that the radiation has a frequency near 1 kHz. Although LIGO-type detectors may be able to detect this radiation from about 10 Hz, their best sensitivity will be above 70–100 Hz, so estimates of detection range are most conservatively made by assuming the detection begins at about this frequency.

The general character of the orbit is well represented if we simply take the radiation to be governed by the quadrupole formula. This tells us that there are of order $N \sim 630$ periods of gravitational waves until the stars begin to coalesce. If we can construct a detailed waveform that follows the radiation over this whole period, then we should be able to use it as a matched filter to improve the signal-to-noise ratio of a detection by about $\sqrt{N} \sim 25$ over what it would be for a broad-band burst of the same amplitude and central frequency. Put another way, if we know the waveform we can obtain the same signal-to-noise ratio as we would have for a broadband burst of the same total energy (integrated over the whole signal) at the same frequency.[13] Since the orbital motion radiates some $5 \times 10^{-3} M_\odot\, c^2$ of energy, these sources will be as detectable as a moderate supernova burst *at a low frequency, say 100 Hz.*

Our discussion above of low-frequency bursts from supernovae shows that this makes coalescing binaries some 30 times more detectable than a conventional kilo-Hertz supernova burst of the same energy. It therefore becomes possible to contemplate detecting these events out as far as 800 Mpc, which is a redshift of $z = 0.26(H_0/100 \text{ km s}^{-1} \text{ Mpc}^{-1}$.

There are also likely to be a certain number of binaries containing a neutron star and a black hole of, say, $10M_\odot$. These radiate with an amplitude approximately 5 times as large as that from two neutron stars. Although they execute fewer orbits before coalescence, such systems are still visible almost 3 times further away. And systems consisting of two black holes can be seen more than 10 times further away, which means that they are visible essentially everywhere in the Universe.

This great distance is the key to the interest in these sources. Although the coalescence events are rare in any galaxy, the volume of space we can observe is very large, and the number of events turns out to be very significant. Moreover, as we shall see, the observations themselves can carry very interesting cosmological information.

4.2. Event rate

Since coalescence events are rare and have never been observed, we can only infer a likely event rate from observations of their progenitors, binary systems containing two neutron stars in an orbit close enough to decay in a Hubble timescale. Such systems may be observable in our Galaxy if one of the neutron stars is a pulsar. The Binary Pulsar PSR1913+16 was our first example, and two others have been discovered since: PSR2127+11C (in the globular cluster M15) and PSR1534+12. Searches currently underway are very likely to turn up more.

Based on a careful analysis of the selection effects in the pulsar surveys that have taken place so far, two independent studies[20, 21] have reached broadly similar conclusions about the rate of coalescence in our Galaxy. Its most probable value is about one neutron-star coalescence per 10^7 years. When we extrapolate this to a distance of 800 Mpc, containing some 2×10^9 galaxies, we arrive at about 200 coalescences per year. Perhaps 10% of these will give strong enough signals to be detected[22], so the detection rate should be some tens per year. This estimate is to be regarded as an observational lower bound to the event rate, and it is uncertain by a factor of 10 either way. (This is a considerable improvement on the best earlier estimate, by Clark, *et al*,[17], which I estimated to be uncertain by a factor of 100.[24]) It could be increased by any of the following:

- If new searches turn up new progenitor systems, the estimated fraction of pulsars in binaries will increase and so will the predicted event rate.

- The event rate may also increase because of another remarkable conclusion of the two studies just cited. Because the formation of a black hole in a supernova explosion retains more of the mass of the original star, it has a smaller probability of disrupting a binary system. This means that the number of binaries containing a black hole and a neutron star may be much larger than one would at first suppose: it may be that one compact-object binary in three contains a black hole. If so, then pulsar surveys now underway should find such a system in our Galaxy. Since such systems are detectable at greater distances, they may even dominate the detection rate, raising it by a factor of 2 or more above the estimate we have just made.

- We have considered only the sensitivity of detectors above about 70 Hz. Depending on the noise performance at low frequencies, it may be possible to gain at least a factor of 2 in signal-to- noise ratio by going down to 40 Hz or lower.[23] This would double the range and multiply the detection rate by 8. Improving sensitivity at low frequency also offers more information about stars from the waveform, but only if the two-body problem in relativity can be solved more completely than it has been so far. I will return to this challenge below.

If one is optimistic about all these possibilities, then the expected detection rate in stage-two detectors could be pushed up from two per month to as many as 40 per day!

4.3. Coalescing binaries and first-stage detectors

More interesting, perhaps, is the possibility that stage-one detectors could see coalescing binaries. This possibility is usually discounted, because these detectors are not expected to be optimized for observing as low as 70 Hz, and this, coupled with a 10-times poorer sensitivity, leaves them with an expected event rate of about one per 10^5 years!

But if a stage-one detector is built with noise controlled down to 40 Hz, and is optimized for observing at 70 Hz, then its range is only a factor of 5 less than the conservative assumption we made for the stage-two detector above. Then if the event rate is actually 10 times higher, and if black holes contribute a further factor of 2, the expected detection rate even with a stage-one detector could be a few per year. Again, to exploit this possibility, detector designers and builders would have to give priority to low-frequency performance at the first stage of construction.

4.4. Information to be learned from observing them

The amount of information that one can in principle extract from coalescing binary observations is enormous. As for supernovae, we consider what different arrays of detectors can do. Coalescing binaries are essentially invisible to bar detectors because they emit very little of their total energy in the bandwidth of realistic detectors. If two laser interferometers were to observe a coalescing binary, they could do the following.

- They would automatically measure what we call the "chirp mass" of the binary, which is the following combination of the reduced mass μ and total mass M of the system:

$$\mathcal{M} = \mu^{3/5} M^{2/5}. \tag{4}$$

This is the only parameter that governs the quadrupolar emission of radiation, and so it determines the overall acceleration of the frequency of the chirp signal. Determining the value of this parameter is essential to extracting the signal from the interferometer noise.

- Provided sensitivity at or below 70 Hz is adequate, or if the system is near and the signal is sufficiently strong, they would be able to measure the individual component masses from post- Newtonian effects in the orbit. The dominant effect is a slowly accumulating phase error in the signal, as post-Newtonian corrections to the orbit grow secularly alongside the quadrupolar secular terms. With good low-frequency sensitivity, a stage-two detector should be able to measure the masses of hundreds of neutron stars and black holes over a few years' observing. The improvement in our understanding of neutron star formation, evolution, and equation of state will be enormous.

- Again, for the strongest systems they should be able to measure other parameters, such as the spins of the neutron stars. These have a marginal but accumulating effect on the orbital phase.

- They should observe black holes coalescing. Provided such events occur more than once a year out to quasar distances, second-stage detectors should see them. If numerical calculations of black-hole coalescence can be performed by the time observations occur, the observations will provide a unique test of strong-field gravity theory.

If three or more interferometric detectors detect a coalescing binary signal, then they can determine its intrinsic amplitude and direction, as well as secondary parameters such as the orientation of the plane of the binary system. Compact-object binaries are so easy to model that a knowledge of the chirp mass $\mathcal{M}$, the intrinsic amplitude, and the orientation of the binary are enough to tell us exactly how far away the system is. Such simple and reliable "standard candles" are rare in astronomy, and offer a wealth of new possibilities. Two such possibilities are:

- The determination of the Hubble constant.[25] This can happen in two ways.

  - We may be lucky enough to observe an event whose position can be determined precisely enough to locate it in a particular galaxy. This requires more precision than the $\pm 1^\circ$ accuracy we expect for typical events. This may happen if the event is much closer and therefore stronger, or if it is also observed with other instruments (such as gamma-ray detectors, as described in the next section). Then the redshift of the galaxy can be measured, and its gravitational-wave distance then determines the Hubble constant.
  - More likely, individual events that occur within 100–150 Mpc will fall within error boxes containing a few clusters of galaxies. Since about 50% of galaxies are members of identifiable clusters, there is a good chance that, by measuring the redshifts of all the clusters and obtaining *candidate* values of the Hubble constant from each, one will find the correct value of H_0 among the candidates. Then if a number of different events are treated in the same way,

the correct value will repeat often, while the incorrect candidates will be distributed randomly. After a few (ten or so) such events, the real value of the Hubble constant will emerge. Depending on the event rate, this could take anywhere from 1 year to a decade to accomplish.

- Test the cosmological mass distribution for super-clustering on the 200+ Mpc scale. Given positions and distances to these systems, and given that they probably sample the visible mass distribution fairly well, one can expect to accumulate statistics on the homogeneity of the mass distribution on scale that we have little information about at present.

4.5. Gamma-ray bursts and coalescing binaries

Among the most puzzling astronomical phenomena at present are gamma-ray bursts. Originally these were thought to originate in neutron stars in the Galaxy, as fairly low-energy and relatively benign events. But the recent announcement[26] that the BATSE instrument on the Compton Observatory has detected hundreds of bursts, and that they are distributed perfectly isotropically on the celestial sky, has cast grave doubt on such models, which would have predicted a concentration towards the galactic plane. A currently popular model is that they are associated with the coalescence event of two neutron stars or a neutron star and a black hole.

Although all models are uncertain at present, it this one turns out to be correct, then gamma ray bursts should also be accompanied by gravitational wave signals. Nicholson and I[27] have studied the consequences of joint observations between a gamma-ray burst detector and two or three gravitational wave detectors. The principal advantage is that the gravitational wave detectors can lower their threshold for detection, since they need only discriminate between noise and real events in a narrow window of time (perhaps a second or less) before the gamma burst. This improves the range of the detectors and of course markedly affects the detected event rate. The more distant events that are now detectable have, of course, lower signal-to-noise, so it is harder to extract information from them. In particular, distance determinations become very inaccurate outside of perhaps 500 Mpc. Nevertheless, the added statistics on the chirp mass, and the ability to correlate gravitational waveforms with the individual characteristics of the gamma bursts is likely to be of great value in making models.

4.6. Challenge to relativity

When I discussed above the information that can be extracted from the signals from coalescing binaries, I qualified the discussion with the proviso that the two-body problem should have been adequately solved by the time observations are made. The reason is that, to identify a signal buried in detector noise that is of much higher amplitude, and to measure its parameters, one must perform matched filtering of the data stream: one must match the pattern expected from a source to the actual incoming wave. To do this, one needs good theoretical predictions of the waveform over a long period of time, and the prediction must keep in phase with the actual wave all the way.

The problem is that, although we have solutions of the two-body problem including terms of post- Newtonian order, including radiation reaction, there is doubt[28] that the

post- Newtonian hierarchy of approximations will produce a convergent, or at least well-behaved, approximation for the evolution of the orbital phase as a function of time over the several minutes that systems can be observed if they are picked up at a few tens of Hz. In fact, the problem gets worse for systems that can be observed first at lower frequencies, that is when they are more nearly Newtonian. This counter-intuitive problem is worth describing.

The post-Newtonian hierarchy of approximations is an asymptotic approximation in a small parameter (essentially $\Phi \sim GM/rc^2$) that is uniform for a fixed number of orbits.[29] But observing coalescing binaries means following them as they decay, on a timescale proportional to $(v/c)^{-4} \sim \Phi^{-2}$. As the system becomes more Newtonian, so that $\Phi \to 0$, any fixed number of orbits occupies a time proportional to $1/v \sim \Phi^{-1/2}$, and therefore represents a fraction of the overall decay time that gets smaller as $\Phi^{3/2} \sim (v/c)^3$. It follows that, the more closely Newtonian the system is when it is first observed, the *worse* the post-Newtonian approximation will be for predicting what the orbit does all the way to coalescence.

If, despite this, the post-Newtonian approximations do provide a convergent description of the orbit, then it is just a matter of hard work to develop them out to post-post-Newtonian order, which might be sufficient for most observations. But if, as seems likely, they do not, then relativists have a challenge: find another way to give a reasonably complete solution. Numerical calculations for widely separated bodies seem prohibitively expensive. Is there another analytic method, involving expansions in a new small parameter, that will produce computationally feasible approximations that converge?

The reward will be signal templates that can extract the maximum information from the observations. It should be emphasized that the mere detection of the signal, and the measurement of the simple chirp mass, does not necessarily require a solution of the whole approximation problem: filters can be designed to pick up most signals even if they do not fit the whole wavetrain with a single analytic expression. But to measure physically interesting things requires filters that translate physical parameters into the observed waveforms.

5. Pulsars and the stochastic background

I will briefly discuss two other possible sources of gravitational waves: pulsars and a cosmological background. These are not as strongly emphasized in discussions of gravitational wave sources because they are more speculative, but they would nevertheless be very interesting if they are observed. It is arguable, in fact, that a positive detection of a cosmological background would be the most important observation that interferometers could make.

5.1. Pulsars old and new

A number of pulsars spin fast enough for any gravitational waves emitted by them to be of a high enough frequency (twice the pulsar frequency) to be detected by ground-based detectors. Indeed, searches for radiation from the Crab pulsar have been performed with special bar detectors built to resonate at 60 Hz.[30] Such radiation would arise in

non-axisymmetric mass distributions, small lumps or irregularities in the crust of the star.

Irregularities are certainly present, if only because the stars possess non-axisymmetric magnetic fields, but they need not be large enough to produce observable amplitudes. The only observational limits we have on most pulsars come from spindown: gravitational waves must not carry away more energy than can be accounted for by the spindown of the pulsar. In fact, the gravitational wave luminosity of pulsars is likely to be much smaller than this, since the dominant loss of energy is almost certainly in the form of electromagnetic and particle fluxes.

The bounds that spindown places on the radiation from the Crab and Vela pulsars still leaves plenty of room for detectable amplitudes, and second-stage interferometers will certainly look for it. Millisecond pulsars, on the other hand, tend to be older and to be spinning down less rapidly, so their radiation limits are more stringent. A recently-discovered nearby millisecond pulsar[31] might offer some hope of eventual detection.

5.2. All-sky search for pulsars

There are many more radio-quiet ("dead") pulsars than there are active ones: perhaps one star in 10 in our Galaxy is a neutron star. So the nearest is only a few parsecs away, and it might well be a source of gravitational waves even though its radio emission has ceased.

Finding it, however, means performing an all-sky, all-frequency survey of the sky in gravitational waves. The sensitivity of such a search, it turns out, will be limited by available computing power in the foreseeable future.[32] The reason is that, provided a data set longer than a few hours is used, it will be necessary to remove the Doppler effects of the Earth's motion from the signal, and these effects depend on the pulsar's position on the sky. The data set must therefore be searched separately for each of a large number of small patches on the sky. This will limit data sets to a couple of weeks in the near future.

5.3. Stochastic background and the early Universe

While pulsars are relatively well-known objects, sources of a stochastic cosmological background of gravitational waves are a good deal more exotic. The leading candidates are cosmic strings. If they are massive enough to act as seeds for galaxy formation, then they must also produce gravitational waves as they oscillate and decay.[33] Other candidates are bubble collisions in extended inflation scenarios and quantum fluctuations in the fields that drive inflation.[34]

A stochastic background looks just like noise in a single detector. The only way to detect it is to look for correlated noise between two different detectors. The detectors must be separated enough so that other sources of noise — ground vibration, for example — are uncorrelated. But if the detectors are separated by too much, they will also not be correlated in their response to the background: if the time it takes a random wave to travel from one detector to the other is a good fraction of the period of the wave, then the responses of the detectors will not be correlated at a given time. For observing at 50 Hz, separations less than 1500 km are ideal. The two European detectors are well placed for this, while the two LIGO sites are rather further apart.

Bar detectors can also look for background radiation, and they have done so.[35] But their narrow bandwidth and limited sensitivity allows only very weak limits to be set at present.

The strength of the background is conveniently expressed by a quantity called Ω_{gw}, which is the fraction of the closure density contributed by the energy of the background per decade of frequency. In a bandwidth of about 50 Hz about 50 Hz, stage-two LIGO-type detectors could in principle reach below $\Omega_{gw} = 10^{-9}$, which would be enough to eliminate the cosmic string model or detect its background. Current limits from pulsar timing are around 10^7 at very low frequencies (periods of several years) — see the article by Taylor in this volume.

Although the existence of such a background is very speculative, a detection would open up an entirely new window on the very early Universe; the radiation would be coming to us directly from an epoch to which we have no other possible direct access. Performing a correlation experiment is arguably one of the most important goals of interferometer development, and it has a high priority in the detector groups.

Acknowledgements

It is a pleasure to acknowledge the assistance of many colleagues, whose conversations and collaboration were invaluable: C. Cutler, J.L. Friedman, J. Hough, G. Jones, A. Lyne, D. Nicholson, J. Shuttleworth, and K.S. Thorne.

References

[1] Hulse, R A, and Taylor, J H, "Discovery of a pulsar in a binary system", *Astrophys J*, **195**, L51-3 (1975)

[2] Misner, C W, Thorne, K S, and Wheeler, J L, *Gravitation* (Freeman & Co, San Francisco, 1973)

[3] Damour, T, and Taylor, J H, "On the Orbital Period Change of the Binary Pulsar PSR-1913+16", *Astrophys J*, **366**, 501–511 (1991)

[4] Armstrong, J W, Estabrook, F B, and Wahlquist, H D, *Astrophys J*, **318**, 536 (1987)

[5] Manchester, R N, Kaspi, V M, Johnston, S, Lyne, A G, and D'Amico, N, "PSR 1758-24 and G5 4-1 2, a remarkable pulsar-supernova remnant association", *Mon Not R astr Soc*, **253**, 7p–10p (1991)

[6] Chandrasekhar, S, *Ellipsoidal figures of equilibrium* (Yale University Press, New Haven, 1969)

[7] Stark, R F, and Piran, T, "Gravitational Radiation from rotating gravitational collapse", *in* Michelson, P F, Hu, En-Ke, and Pizzella, G, eds, *Experimental Gravitational Physics* (World Scientific, Singapore, 1988) 268–275

[8] Evans, C R, "An Approach for Calculating Axisymmetric Gravitational Collapse", *in* Centrella, J M, ed, *Dynamical Spacetimes and Numerical Relativity* (Cambridge University Press, Cambridge, England, 1986), 3–39

[9] Moenchmeyer, R, Schaefer, G, Mueller, E, and Kates, R E, "Gravitational-Waves from the Collapse of Rotating Stellar Cores" *Astron Astrophys*, **246**, 417–440 (1991)

[10] Finn, L S, "Detectability of Gravitational Radiation from Stellar-Core Collapse", *Ann N Y Acad Sci*, **631**, 156–172 (1991)

[11] Chandrasekhar, S, *Phys Rev Lett*, **24**, 611 (1970)

[12] Friedman, J L, and Schutz, B F, "Secular Instability of Rotating Newtonian Stars", *Astrophys J*, **222**, 281–296 (1978)

[13] Schutz, B F, "Gravitational Wave Sources and Their Detectability", *Class Quant Grav*, **6**, 1761–1780 (1989)

[14] Dyson, F J, "Gravitational Machines", *in* Cameron, A G W, ed, *Interstellar Communication: the Search for Extraterrestrial Life*, (W A Benjamin, New York, 1963) 115–120

[15] Forward, R L, and Berman, D, "Gravitational-Radiation Detection Range for Binary Stellar Systems", *Phys Rev Lett*, **18**, 1071–1074 (1967)

[16] Clark, J P A, and Eardley, D M, *Astrophys J*, **215**, 315 (1977)

[17] Clark, J P A, van den Heuvel, E P J, and Sutantyo, W, *Astron & Astrophys*, **72**, 120 (1979)

[18] Thorne, K S, "Gravitational Radiation", *in* Hawking, S W, and Israel, W, eds, *300 Years of Gravitation* (Cambridge University Press, Cambridge, 1987) 330–458 (1987)

[19] Nakamura, T, and Oohara, K, *Ann NY Acad Sci*, **647**, 597–604 (1991)

[20] Phinney, E S, "The rate of neutron star binary mergers in the Universe: minimal predictions for gravity wave detectors", *Astrophys J*, **380**, L17 (1991)

[21] Narayan, R, Piran, T, and Shemi, A, "Neutron star and black hole binaries in the Galaxy", *Astrophys J*, **379**, L17 (1991)

[22] Tinto, M, "Coincidence Probabilities for Networks of Laser Interferometric Detectors Observing Coalescing Compact Binaries", *in* Schutz, B F, ed, *Gravitational Wave Data Analysis* (Kluwer, Dordrecht, 1989), 299–313

[23] Dhurandhar, S V, Krolak, A, and Lobo, J A, "Detection of gravitational waves from a coalescing binary system - effect of thermal noise on efficiency of the detector", *Mon Not R astr Soc*, **237**, 333–340 (1989)

[24] Schutz, B F, "Sources of Gravitational Radiation", *in* Schutz, B F, ed, *Gravitational Wave Data Analysis* (Kluwer, Dordrecht, 1989) 3–18

[25] Schutz, B F, "Determining the Hubble Constant from Gravitational Wave Observations", *Nature*, **323**, 310–311 (1986)

[26] Meegan, C A, Fishman, G J, Wilson, R B, and Paciesas, W S, "Spatial distribution of gamma-ray bursts observed by BATSE", *Nature*, **355**, 143–145 (1992)

[27] Schutz, B F, and Nicholson, D, to be published

[28] Cutler, C, Apostolatos, T A, Bildsten, L, Finn, L S, Flanagan, E E, Kennefick, D, Markovic, D M, Ori, A, Poisson, E, Sussman, G J, and Thorne, K S, "The last three minutes: issues in the measurement of coalescing compact binaries by LIGO", *Phys Rev Lett*, , submitted (1992)

[29] Futamase, T, and Schutz, B F, "The Newtonian and Post-Newtonian Approximations are Astymptotic to General Relativity", *Phys Rev D*, **28**, 2363–2237 (1983)

[30] Owa, S, Fujimoto, M K, Hirakawa, K, Morimoto, K, Suzuki, T, and Tsubono, K, "Search for gravitational radiation from the crab pulsar using cryogenic detectors", *in* CHINA87 Michelson, P F, Hu, En-Ke, and Pizzella, G, eds, *Experimental Gravitational Physics* (World Scientific, Singapore, 1991) 397–402

[31] A Lyne, private communication

[32] Schutz, B F, "Data Processing Analysis and Storage for Interferometric Antennas", *in* Blair, D G, ed, *The Detection of Gravitational Waves* (Cambridge University Press, Cambridge England, 1991) 406–452

[33] Allen, B, and Shellard, E P S, "Gravitational radiation from cosmic strings" *Phys Rev D*, **45**, 1898–1912 (1992)

[34] Turner, M, and Wilczek, F, *Phys Rev Lett*, **65**, 3080 (1990)

[35] Michelson, P F, "On Detecting Stochastic Background Gravitational Radiation with Terrestrial Detectors", *Mon Not R astr Soc*, **227**, 933–941 (1987)

Paper presented at the 13th Int. Conf. on General Relativity and Gravitation
Cordoba, Argentina, 1992: Part 1, Plenary Lectures

Computer calculations of collisions, black holes, and naked singularities

Stuart L Shapiro and Saul A Teukolsky

Center for Radiophysics and Space Research, Cornell University, Ithaca, NY 14850, USA.

Abstract. We describe a method for the numerical solution of Einstein's equations for the dynamical evolution of a *collisionless* gas of particles in general relativity. The gravitational field can be arbitrarily strong and particle velocities can approach the speed of light. The computational method uses the tools of numerical relativity and N-body particle simulation to follow the full nonlinear behavior of these systems. Specifically, we solve the Vlasov equation in general relativity by particle simulation. The gravitational field is integrated using the $3 + 1$ formalism of Arnowitt, Deser, and Misner. One application of our method is the study of head-on collisions of relativistic clusters. We have constructed and followed the evolution of three classes of initial configurations: spheres of particles at rest; spheres of particles boosted towards each other; and spheres of particles in circular orbits about their respective centers. In the first two cases, the spheres implode towards their centers and may form black holes before colliding. These scenarios thus can be used to study the head-on collision of two black holes. In the third case the clusters are initially in equilibrium and cannot implode. In this case collision from rest leads either to coalescence and virialization, or collapse to a black hole. This scenario is the collisionless analog of colliding neutron stars in relativistic hydrodynamics. Our method also provides a new tool for studying the cosmic censorship hypothesis and the possibility of naked singularities. The formation of a naked singularity during the collapse of a finite object would pose a serious difficulty for the theory of general relativity. The hoop conjecture suggests that this possibility will never happen provided the object is sufficiently compact ($\lesssim M$) in all of its spatial dimensions. But what about the collapse of a long, nonrotating, prolate object to a thin spindle? Such collapse leads to a strong singularity in Newtonian gravitation. Using our numerical code to evolve collisionless gas spheroids in full general relativity, we find that in all cases the spheroids collapse to singularities. When the spheroids are sufficiently compact the singularities are hidden inside black holes. However, when the spheroids are sufficiently large there are no apparent horizons. These results lend support to the hoop conjecture and appear to demonstrate that naked singularities can form in asymptotically flat spacetimes.

1. Introduction

For several years we have been using the tools of numerical relativity to explore the dynamical behavior of collisionless matter. One of our goals is to learn how to solve Einstein's equations on the computer, and collisionless matter provides a particularly simple matter source. The dynamical equations for collisionless particles are simply the geodesic equations, which are ODEs and are considerably easier to solve than the equations for hydrodynamic matter, which are PDEs. In contrast to hydrodynamic matter, collisionless matter is not subject to shocks or other discontinuities which present computational difficulties unrelated to solving Einstein's equations.

Basically we are solving relativistic Vlasov equation for the matter coupled to Einstein's equations for the gravitational field. The method is mean-field particle simulation scheme: the distribution function is sampled by a large number of particles to get the stress-energy tensor $T_{\mu\nu}$. This stress-energy tensor is the source for the the field equations. Once they have been solved for the metric, the particles are moved for a small time step along geodesics of this background metric, and $T_{\mu\nu}$ is recalculated. The whole procedure is repeated to follow a complete evolution.

A second goal of our study is to explore the dynamical fate and evolution of relativistic star clusters, and their possible astrophysical importance. For example, we have shown how a sufficiently relativistic star cluster collapses to form a supermassive black hole. Since supermassive black holes are likely to be the engines that power quasars and AGNs, the question of their origin is an important unsolved problem that our work addresses. In this paper, we will not deal with this and other astrophysical issues related to our work. Instead, we will focus on some of the fundamental issues connected with general relativity, rather than with astrophysical applications.

The simplest problems to tackle first are those with lots of symmetry and thus fewer degrees of freedom. Even in spherical symmetry, the dynamical evolution of a star cluster is not trivial. One must choose a good set of coordinates that allow the evolution to be calculated without the appearance of coordinate singularities. More importantly, when there are black holes, one wants coordinates that avoid the accompanying physical singularities, while continuing to follow the evolution outside the black holes. There are no recipes for choosing good coordinates in general. However, the spherical problem is essentially solved (Shapiro and Teukolsky 1985a,b,c, 1986, 1988). One can start with an arbitrary distribution of particle velocities and positions and follow the cluster's evolution, even for cases in which the velocities are close to the speed of light and the gravitational fields become arbitrarily strong. Black hole formation and the growth of the event horizon can be tracked with reasonable accuracy.

Fully three-dimensional problems with no symmetry whatsoever are beyond the reach of present computational resources and algorithmic developments. The current focus is on handling nonspherical problems in axisymmetry. These problems allow one to study two new qualitative features absent in spherical symmetry: rotation and the production of gravitational waves. Also, as we shall see below, they allow us to investigate some fundamental issues of general relativity including cosmic censorship and the formation of naked singularities.

2. Cluster collisions

As a first example of the kind of problem that is now amenable to solution, we will consider the head-on collision of relativistic star clusters. Collision experiments are a time-honored means of probing the nature of an interaction in physics, and here we discuss numerical collisions that we have performed to study the gravitational interaction.

To follow the head-on collision of two clusters, we solve the field equations in $3+1$ form following Arnowitt, Deser, and Misner (1962). We use maximal time slicing and isotropic spatial coordinates in axisymmetry. The metric is

$$ds^2 = -\alpha^2 dt^2 + A^2(dr + \beta^r dt)^2 + A^2 r^2(d\theta + \beta^\theta dt)^2 + B^2 r^2 \sin^2\theta d\phi^2. \qquad (1)$$

The key equations are presented in Shapiro and Teukolsky (1992a).

We have considered three different collision scenarios (Shapiro and Teukolsky 1992a):

(1) the collision from rest of spheres of particles with no initial internal motion;

(2) the collision of the same spheres initially boosted towards each other with a uniform velocity;

(3) the collision from rest of equilibrium spheres of particles, i.e., spheres in which the internal particle motions maintain dynamical equilibrium when the spheres are widely separated.

In the first two cases the nonequilibrium spheres collapse towards their own centers and may form black holes before they collide. Thus these scenarios provide a way of studying the head-on collision of two identical nonrotating black holes. This treatment is an alternative means of analyzing this important problem, which has only been considered previously (Smarr 1979) using the vacuum Einstein equations ("topological" black holes). We thus demonstrate how a code designed to handle matter in general relativity can be used to treat black holes, without having to develop special-purpose routines to handle the boundary conditions for vacuum black holes.

The third case is the cluster analog of the head-on collision of two neutron stars (Smarr and Wilson 1979; Gilden and Shapiro 1984; Evans 1987; Kochanek and Evans 1989). Collisions of equilibrium clusters have previously been studied only in Newtonian gravitation (see, e.g., Kochanek *et al* 1990). In a strong gravitational field the collision can lead to black hole formation. However, stable equilibrium clusters do not implode on their own centers. Thus a black hole can only form when the clusters interact closely.

2.1. Boosted spheres

By way of illustration, consider how case (2) can be used to study the head-on collision of two black holes. Figure 1 shows snapshots from the collision of two identical nonequilibrium spherical clusters boosted towards each other with a velocity $v = 0.12c$ (as measured in the frame of a "normal" observer; see Shapiro and Teukolsky 1992a). The spheres each have a radius $a = 0.8M$ and are separated by $z_0 = 1.4M$, so there are no apparent horizons initially. The velocity was chosen so that the collapse of the clusters individually gives rise to two distinct black holes well before the systems merge to form a single black hole. At $t/M = 6.5$, a "disjoint" apparent horizon forms around each of the spheres. The spheres then implode rapidly toward their centers. However, they continue

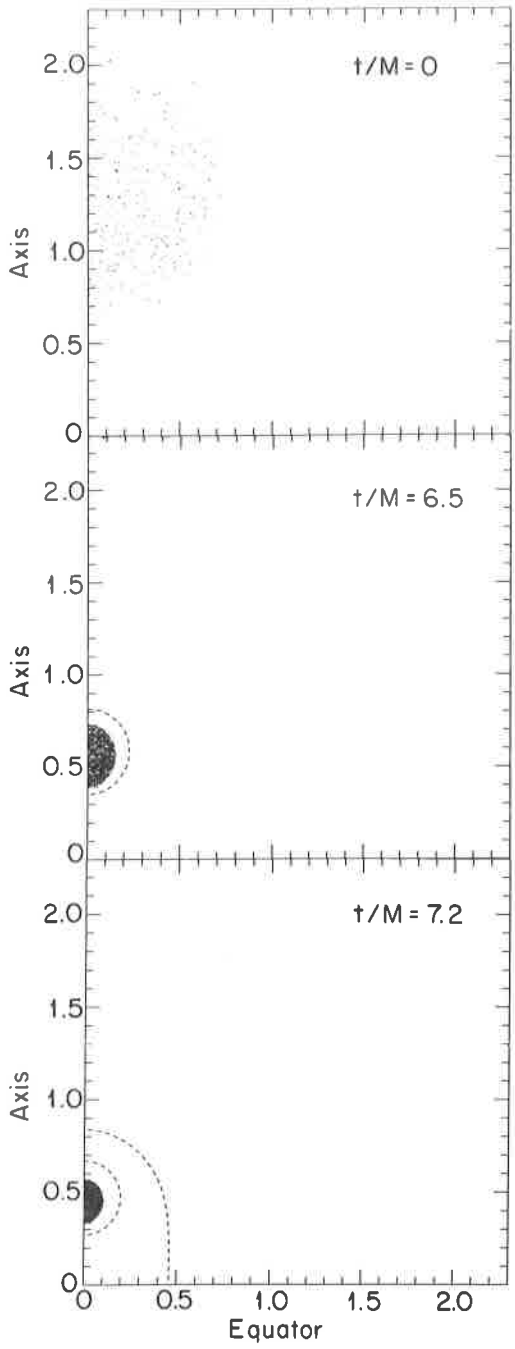

Figure 1. Snapshots of particle positions at selected times for the collision of two spheres of particles boosted towards each other with a velocity $v = 0.12$. The initial radius of each sphere is $a = 0.8M$, and their centers are at $z_0 = \pm 1.4M$. The spheres implode on their own centers as they fall toward each other, forming disjoint horizons at $t \simeq 6M$. Apparent horizons are shown as dashed lines. Because of the boost, the clusters can get close enough for a common horizon to form as well as the disjoint horizons. The area of each disjoint horizon is $\mathcal{A}/4\pi M^2 = 1.1$ at the end of the integration, while the area of the common horizon is $\mathcal{A}/16\pi M^2 = 0.95$. The polar circumference of the disjoint horizon is $\mathcal{C}_{\rm pole}/2\pi M = 1.1$. The circumferences of the common horizon are $\mathcal{C}_{\rm eq}/4\pi M = 0.70$ and $\mathcal{C}_{\rm pole}/4\pi M = 1.1$.

to move toward each other, and by the end of the simulation a "common" apparent horizon encloses the two clusters. As in other cases considered, the amount of gravitational radiation emitted here is small, $\ll 1\%M$. With the computational resources used for this simulation, the precise amount of radiation produced is not accurately determined.

We are forced to terminate the integrations when the field gradients near the cluster centers become so large that we can no longer maintain accuracy. This is the same problem that limits integrations of spherical collapse to black holes with maximal time slicing and isotropic coordinates (Shapiro and Teukolsky 1986). By the time the lapse has fallen exponentially to $\sim 10^{-2}$ at the cluster center, the code becomes inaccurate. Thus these coordinates are not adequate to follow the entire collision and coalescence.

2.2. Equilibrium spheres

Working with equilibrium spheres allows us to study collisions without the clusters imploding on their own centers before making contact. This is the collisionless analog of the head-on collision of neutron stars in hydrodynamics.

In Newtonian theory the head-on collision from rest of two clusters leads to violent relaxation, coalescence and virialization to a new equilibrium state (see, e.g., Kochanek *et al* 1990). In general relativity, however, there is the possibility that the collision will be followed by collapse to a black hole. To anticipate when this might happen, note first that a single spherical cluster is dynamically unstable to collapse whenever its radius in Schwarzschild coordinates is less than $6M$, corresponding to an isotropic radius of $4.95M$. Numerical examples of this phenomenon are given in Shapiro and Teukolsky (1985a). Two well-separated spheres of mass $M/2$ are thus individually unstable if $a < 2.47M$. If these spheres were suddenly superposed, the resulting sphere would be unstable if the original a was less than $4.95M$. In fact, the final sphere would be unstable for even larger values of a because the original internal velocities that maintained equilibrium were those appropriate to a mass $M/2$ rather than M. We thus expect to find for a value of $a \gtrsim 5M$ a transition between collisions leading to black hole formation and those leading to virialization to a merged equilibrium.

In Figures 2 – 4 we consider a sequence of collisions from rest between equilibrium clusters with successively smaller values of a/M. Figure 2 shows the evolution of spheres of radius $a = 64M$ separated by $z_0 = 256M$. This case is essentially Newtonian. Following coalescence and violent relaxation, the two clusters achieve virial equilibrium. To monitor the approach to virial equilibrium, we computed the ratio $2T/|W|$, where T is the Newtonian kinetic energy and W is the Newtonian gravitational potential energy. In equilibrium, this ratio is precisely unity. After some initial large oscillations, the ratio had fallen to 1.07 by the time we stopped the integration. As expected, violent relaxation produces equilibrium in just a few dynamical timescales.

Figure 3 shows a more relativistic encounter, with $a = 9M$ and $z_0 = 36M$. The outcome is qualitatively similar to the Newtonian case. The virial ratio is no longer applicable when relativistic effects are important. However, the quantity

$$E_0 = -\sum_j (1 + u_0^j) \tag{2}$$

becomes constant in time as the spacetime becomes stationary. Here u_0^j is the time component of the 4-velocity of the jth particle. Note that E_0 reduces to the total energy

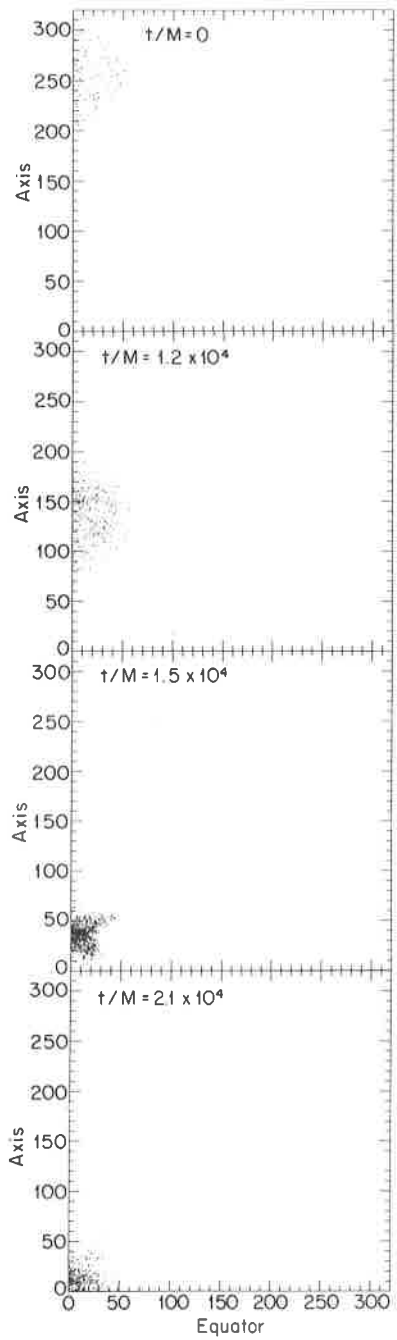

Figure 2. Snapshots of particle positions at selected times for the collision of two equilibrium spheres of particles whose centers are initially at rest. The initial radius of each sphere is $a = 64M$, and their centers are at $z_0 = \pm 256M$, so that this case is essentially in the Newtonian regime. Following collision and coalescence, the spheres virialize.

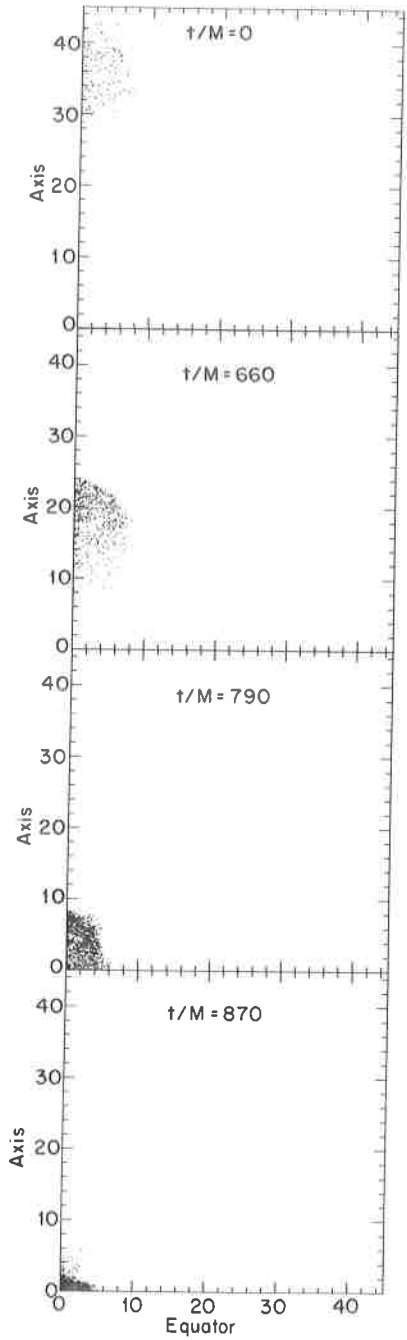

Figure 3. Snapshots of particle positions at selected times for the collision of two equilibrium spheres of particles whose centers are initially at rest. The initial radius of each sphere is $a = 9M$, and their centers are at $z_0 = \pm 36M$. While this case is more relativistic than in Figure 2, the behavior is qualitatively the same.

in the Newtonian limit. We find numerically that E_0 is conserved over several orbital periods towards the end of our integrations, indicating that we have reached a steady state.

By contrast, Figure 4 shows the evolution when $a = 7M$ and $z_0 = 28M$. Now the merger results in collapse to a black hole. Thus the critical radius at which the transition from stable merger to catastrophic collapse occurs lies in the range $7M \lesssim a \lesssim 9M$.

The real significance of this numerical experiment is that we expect analogous behavior for the head-on collision of neutron stars. In particular, the collision of two neutron stars, each of which is below the maximum mass limit, can result in a merged configuration that is too massive to be stable. The collapse may be postponed, however, if the fluid is sufficiently shock heated that thermal pressure provides an appreciable fraction of the total pressure support. This behavior has already been demonstrated for the collapse of single spherical stars above the maximum mass (Shapiro and Teukolsky 1980). These stars can exist in stable equilibrium as long as their temperature remains high. Eventually, after neutrino cooling, any neutron star must collapse if its total mass exceeds the cold mass limit. For collisions between $1.4M_\odot$ neutron stars, however, thermal pressure may be unable to impede dynamical collapse. Just as for cluster collisions, where there is a critical value of a/M, there should be a critical mass for neutron star collisions above which collapse will occur even on a dynamical timescale.

3. Cosmic censorship and naked singularities

It is well-known that general relativity admits solutions with singularities, and that such solutions can be produced by the gravitational collapse of nonsingular, asymptotically flat initial data. The *cosmic censorship hypothesis* (Penrose 1969) states that such singularities will always be clothed by event horizons and hence can never be visible from the outside (no naked singularities). If cosmic censorship holds, then there is no problem with predicting the future evolution outside the event horizon. If it does not hold, then the formation of a naked singularity during collapse would be a crisis for general relativity theory. In this situation, one cannot say anything precise about the future evolution of any region of space containing the singularity since new information could emerge from it in a completely arbitrary way.

Are there guarantees that an event horizon will always hide a naked singularity? No definitive theorems exist. Proving the validity of cosmic censorship is perhaps the most outstanding problem in the theory of general relativity. Until recently, possible counter-examples (see, e.g. Goldwirth *et al* 1989) have all been restricted to spherical symmetry and typically involve shell crossing, shell focusing, or self-similarity. Are these singularities an accident of spherical symmetry?

Very little is known about nonspherical collapse in general relativity. In the absence of concrete theorems, Thorne (1972) has proposed the *hoop conjecture*: Black holes with horizons form when and only when a mass M gets compacted into a region whose circumference in *every* direction is $\mathcal{C} \lesssim 4\pi M$. If the hoop conjecture is correct, aspherical collapse with one or two dimensions appreciably larger than the others might then lead to naked singularities.

For example, consider the Lin-Mestel-Shu instability (Lin *et al* 1965) for the collapse of a nonrotating, homogeneous spheroid of collisionless matter in Newtonian gravity. Such a configuration remains homogeneous and spheroidal during collapse. If the

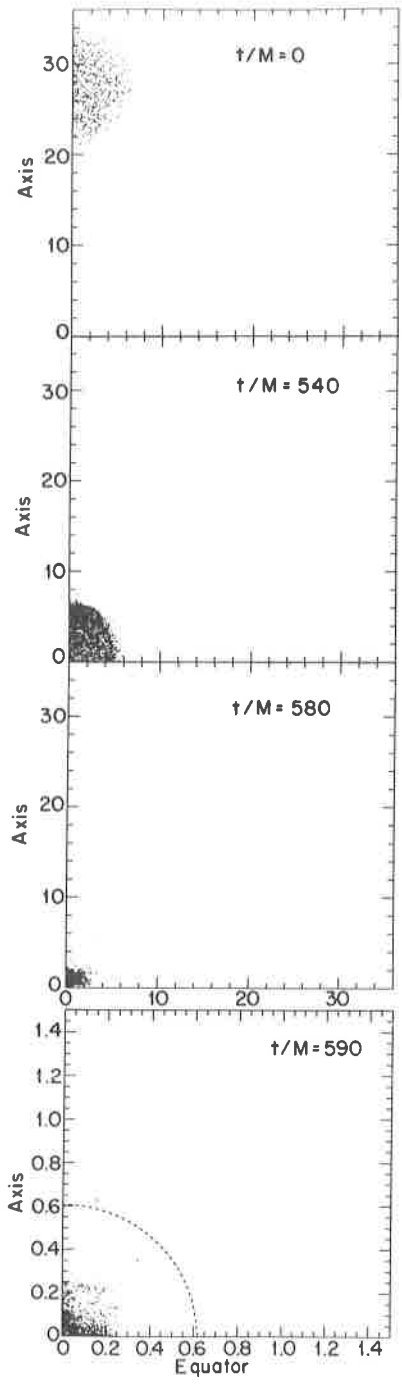

Figure 4. Snapshots of particle positions at selected times for the collision of two equilibrium spheres of particles whose centers are initially at rest. The initial radius of each sphere is $a = 7M$, and their centers are at $z_0 = \pm 28M$. Now the spheres are sufficiently relativistic that the collision is followed by catastrophic collapse to a black hole. The common horizon shown in the blowup at $t/M = 590$ has an area $\mathcal{A}/16\pi M^2 = 1.1$, while its circumferences are $\mathcal{C}_{eq}/4\pi M = 1.1$ and $\mathcal{C}_{pole}/4\pi M = 1.0$.

spheroid is slightly oblate, the configuration collapses to a pancake, while if the spheroid is slightly prolate, it collapses to a spindle. While in both cases the density becomes infinite, the formation of a spindle during prolate collapse is particularly worrisome. The gravitational potential, gravitational force, tidal force, kinetic and potential energies all blow up. This behavior is far more serious than mere shell-crossing, where the density alone becomes momentarily infinite. For collisionless matter, prolate evolution is forced to terminate at the singular spindle state. For oblate evolution the matter simply passes through the pancake state, but then becomes prolate and also evolves to a spindle singularity.

Does this Newtonian example have any relevance to general relativity? We already know that *infinite* cylinders do collapse to singularities in general relativity, and, in accord with the hoop conjecture, are not hidden by event horizons (Thorne 1972; Misner *et al* 1973). But what about *finite* configurations in asymptotically flat spacetimes?

Previously, we constructed an analytic sequence of momentarily static, prolate and oblate collisionless spheroids in full general relativity (Nakamura *et al* 1988). We found that in the limit of large eccentricity the solutions all become singular. In agreement with the hoop conjecture, extended spheroids have no apparent horizons. Can these singularities arise from the collapse of nonsingular initial data? To answer this, we have performed fully relativistic dynamical calculations of the collapse of these spheroids, starting from nonsingular initial configurations (Shapiro and Teukolsky 1991*a*, *b*).

We find that the collapse of a prolate spheroid with sufficiently large semi-major axis leads to a spindle singularity without an apparent horizon. Our numerical computations suggest that the hoop conjecture is valid, but that cosmic censorship does not hold because a naked singularity may form in nonspherical relativistic collapse.

3.1. Collapse of collisionless spheroids

We followed the collapse of nonrotating prolate and oblate spheroids of various initial sizes and eccentricities. The matter particles are instantaneously at rest at $t = 0$ and the configurations give exact solutions of the relativistic initial-value equations (Nakamura *et al* 1988). In the Newtonian limit, these spheroids reduce to homogeneous spheroids. When they are large (size $\gg M$ in all directions) we confirm that their evolution is Newtonian (Lin *et al* 1965; Shapiro and Teukolsky 1987).

We constructed a number of geometric probes to diagnose the evolving spacetime. We tracked the Brill mass and outgoing radiation energy flux to monitor mass-energy conservation. To confirm the formation of a black hole, we probed the spacetime for the appearance of an apparent horizon and computed its area and shape when it was present. To measure the growth of a singularity, we computed the Riemann invariant $I \equiv R_{\alpha\beta\gamma\delta}R^{\alpha\beta\gamma\delta}$ at every spatial grid point. To test the hoop conjecture, we computed the minimum equatorial and polar circumferences outside the matter. Because of the highly curved geometry, these minimum circumferences were not always along the matter surface, but usually further out.

Typical simulations were performed with a spatial grid of 100 radial and 32 angular zones, and with 6000 test particles. A key feature enabling us to snuggle close to singularities was that the angular grid could fan and the radial grid could contract to follow the matter.

Figure 5 shows the fate of a typical prolate configuration that collapses from a highly compact and relativistic initial state to a black hole. Note that in isotropic coordinates a Schwarzschild black hole on the initial time slice would have radius $r = 0.5M$, corresponding to a Schwarzschild radius $r_s = 2M$. Figure 6 depicts the outcome of prolate collapse with the same initial eccentricity but from a larger semi-major axis. Here the configuration collapses to a spindle singularity at the pole without the appearance of an apparent horizon. (We searched for both a single global horizon centered on the origin as well as a small disjoint horizon around the singularity in each hemisphere.) The spindle consists of a concentration of matter near the axis at $r \approx 5M$. Figure 7 shows the growth of the Riemann invariant I at $r = 6.1M$ on the axis, just outside the matter. Before the formation of the singularity, the typical size of I at any exterior radius r on the axis is $\sim M^2/r^6 \ll 1$. With the formation of the spindle singularity, the value of I rises without bound in the region near the pole. The maximum value of I determined by our code is limited only by the resolution of the angular grid: the better we resolve the spindle the larger the value of I we can attain before the singularity causes the code (and spacetime!) to break down. Unlike shell-crossing singularities, where I blows up in the matter interior whenever the matter density is momentarily infinite, the singularity also extends *outside* the matter beyond the pole at $r = 5.8M$ (Fig. 8). In fact, the peak value of I occurs in the vacuum at $r \approx 6.1M$. Here the *exterior* tidal gravitational field is blowing up, which is not the case for shell crossing. *The absence of an apparent horizon suggests that the spindle is a naked singularity.*

When our simulation terminates, I along the axis falls to half its peak value at $r \approx 4.5M$ inside the matter and $r \approx 6.7M$ outside the matter. The singularity is not a point. Rather it is an extended region which includes the matter spindle, but grows most rapidly in the vacuum exterior above the pole. A $t =$ constant slice has a spatial metric $ds^2 = A^2dr^2 + A^2r^2d\theta^2 + B^2r^2\sin^2\theta\, d\phi^2$. In flat space $A = B = 1$. At the termination of the simulation these quantities have a modest maximum value $A \approx B \approx 1.7$, which occurs at the origin. They decrease monotonically outwards, reaching unity at large distances. However, it is their second derivatives that contribute to I and these blow up. While A and B steadily grow with time, I diverges much more rapidly. The behavior is similar to the logarithmic divergence of the metric in the analytic prolate sequence of Nakamura *et al* (1988). We emphasize that the above characterization of the singularity and the behavior of the metric is dependent on the time slicing and may be different for other choices of time coordinate. In principle the spindle singularity might first occur at the center rather than the pole with a different time slicing.

The absence of an apparent horizon does not necessarily imply the absence of a global event horizon, although the converse is true. Recently Wald and Iyer (1991) have emphasized this point by showing that even Schwarzschild spacetime can be sliced with nonspherical slices that approach arbitrarily close to the singularity without any trapped surfaces. Because such singularities cause our numerical integrations to terminate, we cannot map out a spacetime arbitrarily far into the future, which would be necessary to completely rule out the formation of an event horizon. However, we do not think this is at all likely: for collapse from an initially compact state (Fig. 5), outward null geodesics turn around near the singularity. For collapse from large radius, by contrast, (Fig. 6) outward null geodesics are still propagating freely away from the vicinity of the singularity up to the time our integrations terminate. It is an interesting question for future research whether any time slicing can be found which will be more effective in

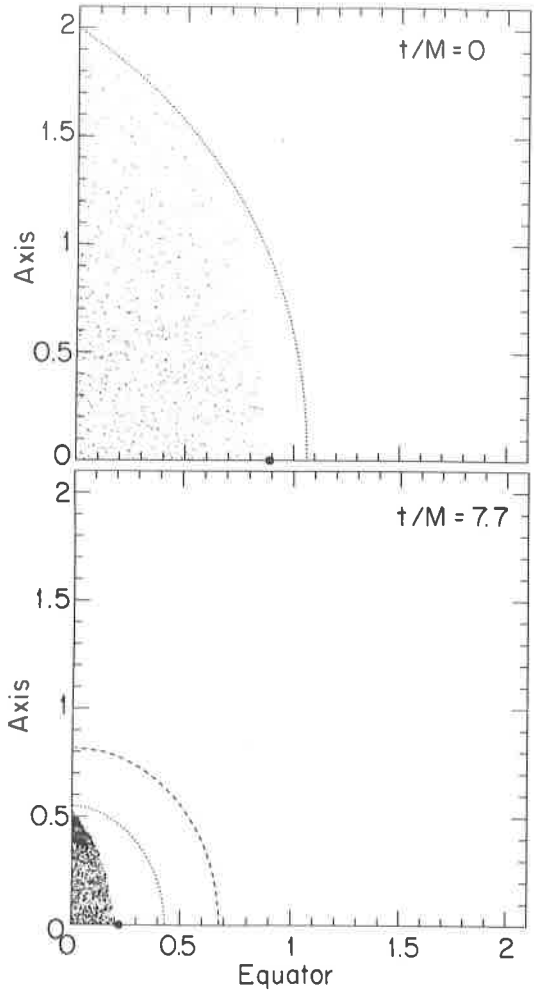

Figure 5. Snapshots of the particle positions at initial and late times for prolate collapse. The positions (in units of M) are projected onto a meridional plane. Initiallythe semi-major axis of the spheroid is $2M$ and the eccentricity is 0.9. The collapse proceeds nonhomologously and terminates with the formation of a spindle singularity on the axis. However, an apparent horizon (dashed line)forms to cover the singularity. At $t/M = 7.7$ its area is $\mathcal{A}/16\pi M^2 = 0.98$, close to the asymptotic theoretical limit of 1. Its polar and equatorial circumferences at that time are $\mathcal{C}^{\rm AH}_{\rm pole}/4\pi M = 1.03$ and $\mathcal{C}^{\rm AH}_{\rm eq}/4\pi M = 0.91$, At later times these circumferences become equal and approach the expected theoretical value 1. The minimum exterior polar circumference is shown by a dotted line when it does not coincide with the matter surface. Likewise, the minimum equatorial circumference, which is a circle, is indicated by a solid dot. Here $\mathcal{C}^{\rm min}_{\rm eq}/4\pi M = 0.59$ and $\mathcal{C}^{\rm min}_{\rm pole}/4\pi M = 0.99$. The formation of a black hole is thus consistent with the hoop conjecture.

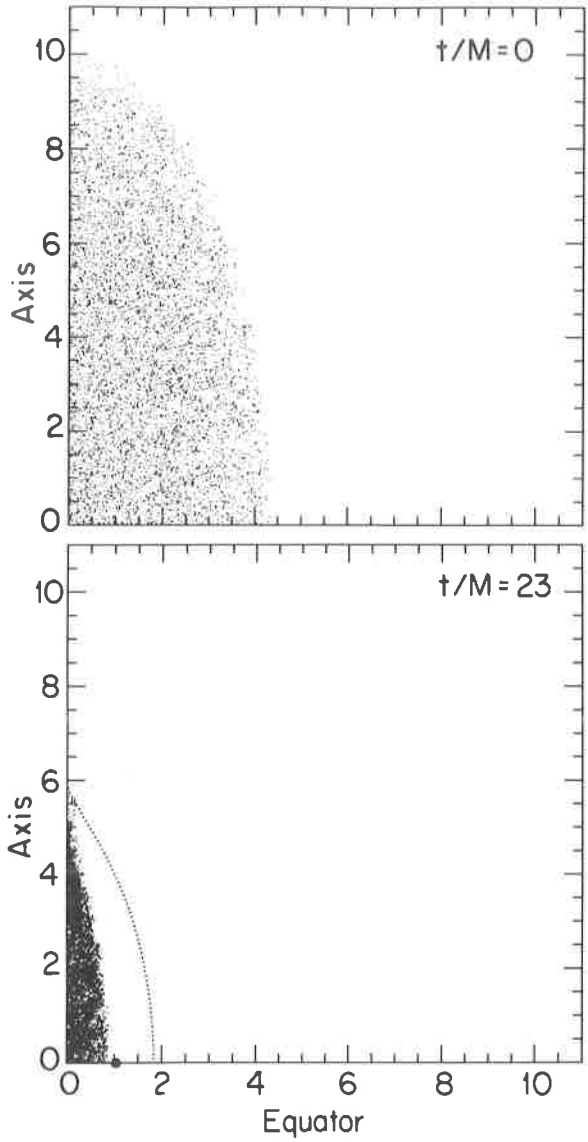

Figure 6. Snapshots of the particle positions at the initial and final times for prolate collapse with the same initial eccentricity as Figure 5 but with initial semi-major axis equal to $10M$. The collapse proceeds as in Figure 5, and terminates with the formation of a spindle singularity on the axis at $t/M = 23$. The minimum polar circumference is $\mathcal{C}^{\min}_{\text{pole}}/4\pi M = 2.8$. There is no apparent horizon, in agreement with the hoop conjecture. This is a good candidate for a naked singularity, which would violate the cosmic censorship hypothesis.

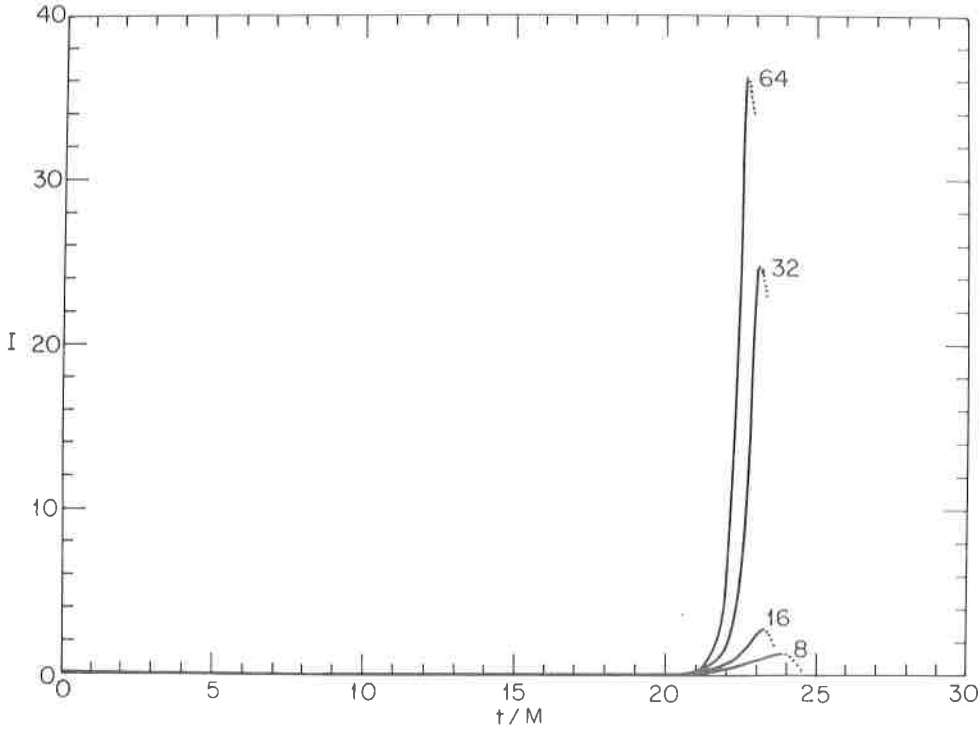

Figure 7. Growth of the Riemann invariant I (in units of M^{-4}) versus time for the collapse shown in Figure 6. The simulation was repeated with various angular grid resolutions. Each curve is labeled by the number of angular zones used. We use dots to show where the singularity has caused the code to become inaccurate.

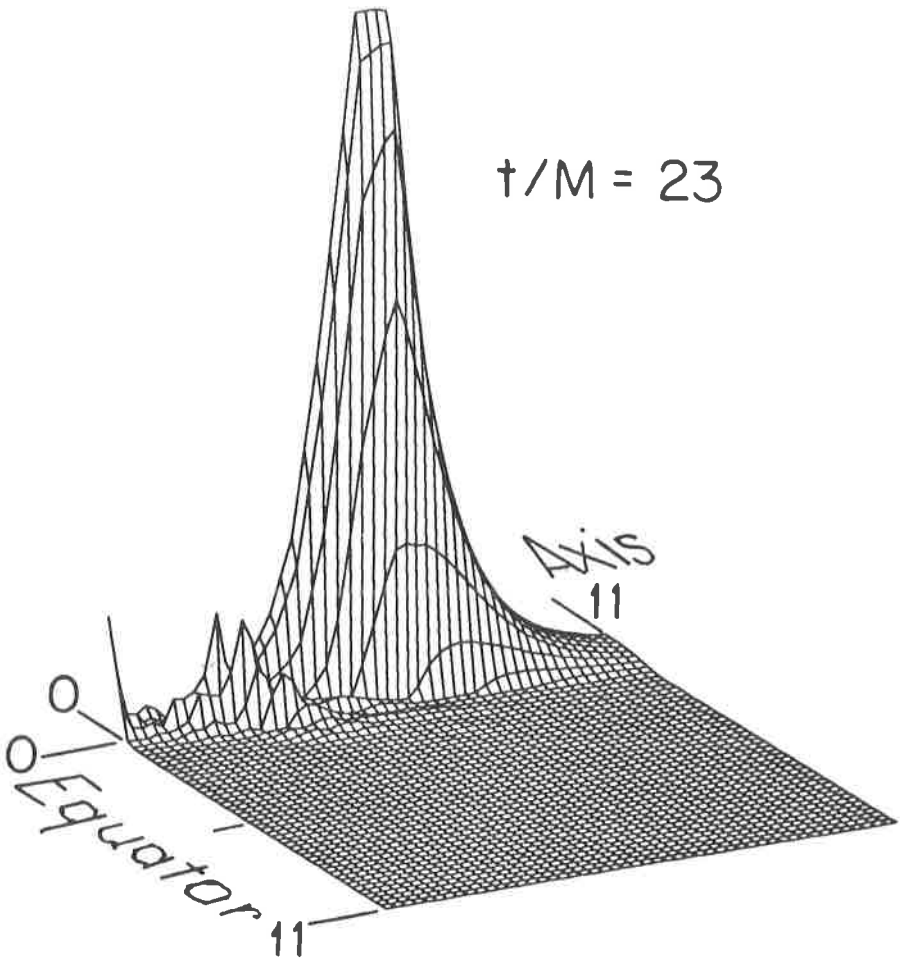

Figure 8. Profile of I in a meridional plane for the collapse shown in Figure 6. For the case of 32 angular zones shown here, the peak value of I is $24/M^4$ and occurs on the axis just outside the matter.

snuggling up to the singularity without actually hitting it.[1] Such a slicing would enable one to confirm that all outward null geodesics propagate to large distances.

Further evidence for the nakedness of the singularity is the similarity of the spindle singularity to the infinite cylinder naked singularity. In both cases the proper length of a given segment of matter along the axis grows slowly, while its proper circumference and surface area shrink to zero much more rapidly. Also, the singularity is an extended region along the axis and not just a point.

We have also followed the collapse of an initially oblate configuration with the same initial eccentricity and semi-major axis as Figure 6. Following pancaking, it overshoots, becomes prolate and forms a black hole. At the time our integrations terminate, we find that $\mathcal{C}^{\rm min}_{\rm pole} = \mathcal{C}^{\rm min}_{\rm eq} = 0.85(4\pi M)$.

All of the above results are consistent with the hoop conjecture. When black holes form, the minimum polar and equatorial circumferences satisfy $\mathcal{C}^{\rm min} \lesssim 4\pi M$. Conversely, when naked singularities form the minimum polar circumference is much bigger than this value. In all cases where an apparent horizon forms, its area satisfies to within numerical accuracy $\mathcal{A} \leq 16\pi M^2$, as required theoretically (see, e.g., Eardley 1975). In every case we find that gravitational radiation carries away a negligible fraction ($\ll 1\%$) of the total mass-energy by the time a black hole or naked singularity forms.

4. Conclusions

Numerical simulations of relativistic star clusters are a useful tool for probing the nature of spacetimes with strong gravitational fields. We have shown two examples where this technique has proven to be fruitful. One is the exploration of black hole and matter collisions. The other investigates cosmic censorship and the possible formation of naked singularities.

In the cosmic censorship investigation, we have presented numerical evidence that the hoop conjecture is a valid criterion for the formation of black holes during nonspherical gravitational collapse. We have also found numerical candidates for the formation of naked singularities from nonsingular initial configurations. These examples are in contrast with any cases of singularities which may arise during spherical collapse. There the exterior spacetime is always the Schwarzschild metric and the Riemann invariant is always exactly $48M^2/r_s^6$, which is finite outside the matter. In spherical collapse the singularities can thus only occur inside the matter. Here the singularities extend above the pole into the vacuum exterior. These examples suggest that the unqualified cosmic censorship hypothesis cannot be valid.

While the matter treated here has kinetic pressure, it is collisionless, not fluid. We do not regard the collisional properties of the matter as crucial: First, the formation of naked singularities should not depend on the particular details of the fundamental interactions affecting matter at high densities. The gravitational field equations alone should be sufficient to rule out naked singularities, at least in the vacuum exterior, for true cosmic censorship. Second, collisional effects may even accelerate the formation of singularities via relativistic "pressure regeneration" (Misner *et al* 1973) There is at least

[1] Maximal slicing apparently does not hold back the formation of prolate spindle singularities. For prolate spheroids the Newtonian potential diverges only logarithmically as the eccentricity $\to 1$, which may explain why α does not plummet precipitously near a spindle.

one tentative numerical example of prolate *fluid* collapse (adiabatic index $\Gamma \rightarrow 2$) that appears to be evolving to a singular state without the formation of an apparent horizon (Nakamura and Sato 1982; Nakamura *et al* 1987).

It is not impossible that naked singularities qualitatively similar to the ones here may even arise in vacuum spacetimes. Since the sequence of momentarily static spheroids (Nakamura *et al* 1988) proved to be a predictor of the singularities found in the dynamical calculations, we have been motivated to seek similar sequences of pure vacuum inital data. We have constructed two sequences characterized by long prolate concentrations of mass-energy: linear strings of black holes, and Brill waves with characteristic widths much less than their lengths (Abrahams *et al* 1992). We find once again that the surrounding gravitational tidal field diverges for limiting members of these sequences, but that no common apparent horizons occur when the configurations are sufficiently long. It would be interesting to employ these solutions as initial data in dynamical evolutions.

The collisionless matter simulations described so far have no angular momentum. The presence of angular momentum could prevent an infinitesimally thin spindle singularity from forming on the axis. Recently, Apostolatos and Thorne (1992) have shown that, as in Newtonian theory, an infinitesimal amount of rotation is sufficient to prevent the formation of a singularity in the relativistic collapse of an *infinite* dust cylinder. This still leaves open the possibility that collapsing configurations of finite size can collapse to singularities in the presence of rotation. Recall that a small amount of angular momentum does not prevent the formation of a singularity when a Kerr black hole forms.

To explore this question, we have recently used our code to study the collapse of rotating collisionless spheroids (Shapiro and Teukolsky 1992b). The spheroids are initially prolate and consist of equal numbers of co- and counter-rotating particles, as in the infinite case treated by Apostolatos and Thorne (1992). Although individual particles are rotating, the spheroid has no net angular momentum. This restriction greatly simplifies the spacetime—the metric still has the same form as Equation (1). We find that rotation significantly modifies the evolution when it is sufficiently large. Imploding configurations with appreciable rotation ultimately collapse to black holes. However, for small enough angular momentum, our simulations cannot at present distinguish rotating from nonrotating collapse: spindle singularities appear to arise without apparent horizons. Hence it is possible that even spheroids with some angular momentum may form naked singularities.

Acknowledgments

This research was supported in part by NSF grants AST 91-19475 and PHY 90-07834 and NASA grant NAGW-2364 at Cornell University. Computations were performed on the Cornell National Supercomputer Facility.

References

Abrahams A M, Heiderich K, Shapiro S L and Teukolsky S A 1992 *Phys. Rev.* D in press

Apostolatos T A and Thorne K S 1992 *Phys. Rev.* D, in press

Arnowitt R, Deser S and Misner C W 1962 In *Gravitation: an Introduction to Current Research* (ed. L Witten) (New York: John Wiley) p 227

Eardley D M 1975 *Phys. Rev.* D **12** 3072

Evans C R 1987 In *Proc. 13th Texas Symposium on Relativistic Astrophysics* (ed. M P Ulmer) (Singapore: World Scientific) p 152

Gilden D L and Shapiro S L 1984 *Astrophys. J.* **287** 728

Goldwirth D S, Ori A and Piran T 1989 In *Frontiers in Numerical Relativity* (ed. C R Evans, L S Finn and D W Hobill) (Cambridge: Cambridge University Press) p 414

Kochanek C S and Evans C R 1989 In *Frontiers in Numerical Relativity* (ed. C R Evans, L S Finn and D Hobill) (Cambridge: Cambridge University Press) p 297

Kochanek C S, Shapiro S L, Teukolsky S A and Chernoff D F 1990 *Astrophys. J.* **358** 81

Lin C C, Mestel L and Shu, F H 1965 *Astrophys. J.* **142** 1431

Misner C W, Thorne K S and Wheeler J A 1973 *Gravitation* (San Francisco: Freeman)

Nakamura T, Oohara K and Kojima Y 1987 *Prog. Theor. Phys. Suppl.* **90** 57

Nakamura T and Sato H 1982 *Prog. Theor. Phys.* **68** 1396

Nakamura T, Shapiro S L and Teukolsky S A 1988 *Phys. Rev.* D **38** 2972

Penrose R 1969 *Rivista del Nuovo Cimento* **1** (Numero Special) 252

Shapiro S L and Teukolsky S A 1980 *Astrophys. J.* **235** 199

——1985a *Astrophys. J.* **298** 34

——1985b *Astrophys. J.* **298** 58

——1985c *Astrophys. J. Lett.* **292** L41

——1986 *Astrophys. J.* **307** 575

——1987 *Astrophys. J.* **318** 542

——1988 *Science* **241** 421

——1991a *Phys. Rev. Lett.* **66** 994

——1991b *Amer. Sci.* **79** 330

——1992a *Phys. Rev.* D **45** 2739

——1992b *Phys. Rev.* D **45** 2206

Smarr L L 1979 in *Sources of Gravitational Radiation* (ed. L L Smarr) (Cambridge: Cambridge University Press) p 245

Smarr L L and Wilson J R 1979 as reported in Wilson J R in *Sources of Gravitational Radiation* (ed. L L Smarr) (Cambridge: Cambridge University Press) p 443

Thorne K S 1972 In *Magic Without Magic: John Archibald Wheeler* (ed. J. Klauder) (San Francisco: Freeman) p 1

Wald R M and Iyer V 1991 *Phys. Rev.* D **44** R3719

Paper presented at the 13th Int. Conf. on General Relativity and Gravitation
Cordoba, Argentina, 1992: Part 1, Plenary Lectures

What can we learn from the study of non-perturbative quantum general relativity?

Lee Smolin

Department of Physics, Syracuse University, Syracuse, New York U.S.A.

February 14, 1993

Abstract

I attempt to answer the question of the title by giving an annotated list of the major results achieved, over the last six years, in the program to construct quantum general relativity using the Ashtekar variables and the loop representation. A summary of the key open problems is also included.

Contents

1. Introduction

Most of what we know about nature at the present time is contained within the realms of either quantum field theory or general relativity. Each of these is a beautiful, powerful and profound theory. However neither can, because of the existence of the other, be said to constitute the basis for a general theory of physics. Thus, while Newtonian physics has been overthrown, it has not been replaced; and it cannot be until we can invent a synthesis of these two great theoretical edifices that can serve as a single foundation for our understanding of all of nature. To do this is the problem of quantum gravity. As such, the problem of quantum gravity is very different from other problems in which we seek to apply a well defined theoretical structure to a new phenomena. It is more open, and more difficult.

Many people who work on this problem complain about its difficulty, and about its distance from experiment. However, I think both complaints are based on a misunderstanding of the nature of the problem. After all, the last time the progress of science required a transformation of this scale in our basic understanding of nature, it took more than 140 years from the publication of Copernicus's *Revolutionibus* to the publication of Newton's *Principia* [1]. could As far as experiment is concerned, there is a large amount of data about fundamental physics that is presently unexplained, including the undetermined parameters of the standard model of particle physics, the horizon and flatness problem of cosmology and the problem of explaining the formation of structure in the universe. There are good reasons to believe that a quantum theory of gravity will have something important to contribute to the solution of each of these problems.

Many different approaches to this problem have been pursued. This is proper, as there is no way of really knowing from what direction the solution will come. Moreover, as the philosopher of science Paul Feyerabend reminds us, science functions best when the level of consensus remains near the minimum forced on us by the experimental data[2]. In this contribution, I want to discuss only one approach to the problem. This approach has been particularly active during the last six years and it has achieved a certain amount of progress, which I want to summarize here. There are also big unsolved problems, which I will also be mentioning.

The approach I want to describe can be characterized by a list of questions that those who pursue it are seeking to answer. These may be stated as follows:

1) We take it as given that any quantum theory of gravity that has a chance of being a correct description of reality must be non- perturbative. This means that the theory cannot be based on an expansion around a single classical background, in which it is the deviations from the background that are quantized. Instead, all of the geometrical quantities that describe the geometry of spacetime must be treated as quantum operators. This means that very few of the techniques of conventional quantum field theory can be directly applied to it. Can we invent a new approach to quantum field theory that applies to the cases of field theories defined on differential manifolds without fixed metric structures?

2) Specifically, it is diffeomorphism invariance that expresses the absence of a non-dynamical background geometry in classical general relativity. As such, its implementation in the quantum theory is the central problem of constructing any quantum theory of gravity. We would then like to understand how the requirement of exact diffeomorphism

invariance requires us to modify the standard quantum field theory techniques such as regularization, renormalization, operator product expansions, and so forth.

3) We are not trying to construct *the* quantum theory of gravity. We are trying to construct *a* theory which is consistent with quantum mechanics and general relativity. For this reason we study quantum general relativity because we have its exact nonperturbative formulation, classically. String theory is, in principle, an attractive program for the unification of physics. However, as long as it lacks a nonperturbative formulation it cannot be the basis of an exploration of the problems of nonperturbative approaches to quantum gravity. We will not be disappointed if, in the end, quantum general relativity is not the theory of nature. But we aim to decide cleanly whether or not there is a consistent mathematical theory, with a sensible physical interpretation, that could be given that name.

4) Certainly it may be true that a complete solution to the problems of quantum gravity and quantum cosmology will require also a solution to the problem of the unification of gravitation with the other fundamental interactions. This could happen through a "microphysics $\rightarrow$ macrophysics" approach like string theory or a "macrophysics $\rightarrow$ microphysics" approach, one example of which is proposed in [3]. At the same time, it may also be the case that important aspects of the problem of quantum gravity can be solved independently of what matter fields gravitation is coupled to. This will be the case to the extent that qualitatively new physics emerges from the construction of diffeomorphism invariant quantum field theories. We would like to see if the study of quantum general relativity at the nonpertubative level can lead to the discovery of such phenomena.

5) In any such approach, one of the key questions is what constitutes an observable. This is because, in a diffeomorphism invariant theory, coordinates and clocks have no *a-priori* meaning; any meaningful observable must express some relationship between physical fields, rather than being defined with respect to a background geometry or an external observer. One aspect of this problem is the problem of time in quantum cosmology[1]. These problems arise already at the classical level, but there they can be solved[46, 33, 34, 35, ?]. We then need to ask: Can we put what we know about the problem of observables and time in classical general relativity together with the technical developments in diffeomorphism invariant quantum field theory to learn how to construct operators that correspond to physically meaningful quantities in a quantum theory of gravity and cosmology?

While these are the main questions that have guided us in the work I will be describing below, the fact that we have been able to make any progress is due primarily to two technical developments. These are the Ashtekar formulation of general relativity [6], and the loop representation of quantum field theories [7, 8].

The Ashtekar variables and the loop representation have been the subject of a number of reviews, including three published within the last two years[9, 10, 11]. For this reason, I will not repeat here what the reader can easily find in those reviews or in the original papers. Instead, I will devote myself to describing, in as concise a manner as possible, the basic results that have been, so far, achieved by this program. As my aim is to be brief, I do not give many technical details, except when describing results

[1] Good discussion of this problem are in [4, 46]. The author's point of view about the problem of time in quantum cosmology is described in [5].

that have not so far appeared elsewhere.

I should point out that while I have tried to write this paper so it can be read by someone who is not an expert on quantum gravity, I do not include an introduction to the basics of canonical general relativity and the methods of canonical quantization. These can be found in other papers in this proceedings, such as the paper of Kuchar[12], as well as in numerous other places, among them[9, 10, 11].

I begin in the next section by asking how quantum field theory must be modified so that it can make sense in the absence of a background metric. The core of the paper is in the following two sections, where I give, correspondingly, an annotated list of major results and a list of key open problems. In the conclusion I try to provide an answer to the question in the title[2].

2. Basic ideas of the approach

A good place to begin is with the question of how to construct diffeomorphism invariant quantum field theories. The key problem is how to get rid of the dependence on the background metric that underlie the standard formulations of quantum field theory in Minkowski spacetime. Here are four ways in which Minkowski spacetime quantum field theories depend on the background metric:

A) The Fock spaces of linearized field theories are constructed by associating operators with the solutions of free field theories on a background spacetime.

B) The definition of the vacuum and the definition of the creation and annihilation operators depend on the splitting of the solutions of the linearized field equations into positive and negative frequency parts. This splitting is Poincare invariant but, as we know from our experience with quantum field theories in curved spacetime, it is not invariant under any larger group of transformations and depends also on the background metric.

C) The regularization and renormalization procedures necessary to make sense out of operator products and interactions in quantum field theory all depend explicitly on the background metric. The background metric is used in the definition of the cutoff scale, in the separation of terms in the operator product expansion and to measure how fast the cutoffs are removed in taking the limits that define renormalized operator products.

D) The inner products of conventional quantum field theories are defined by the requirement of Poincare invariance with respect to a given flat background metric.

We propose to replace each of these constructions with alternatives that do not depend on any background metric. The way in which we go about doing it defines the approach we are calling non- perturbative quantum general relativity. These alternative constructions are:

A') We construct background independent quantum field theories by constructing new representations of algebras of observables that are unitarily inequivalent to the

[2] I want to mention that this is not intended as a review of all of the developments associated with the Ashtekar variables. Particularly, there are several important developments connected with the classical theory that most likely have implications for the quantum theory, such as the Capovilla-Dell-Jacobson[64] formalism and the classical solutions of the diffeomorphism constraints of Newman and Rovelli[65]. However, as my focus is the quantum theory, and as their implications for that have yet to be developed I do not mention them, as well as a number of other very interesting developments, here.

Fock representations. It is here that the loop representation plays the key role. These representations carry unbroken unitary representations of the diffeomorphism group; this makes possible the exact solutions of the constraints that impose diffeomophism invariance on the quantum states.

B') We replace the splitting into positive and negative frequency parts by a splitting into self-dual and antiself- dual parts. This latter splitting is defined only in special cases such as that of connection fields in four spacetime dimensions. It has the great virtue that the connection and curvature of any four dimensional spacetime can be split into self-dual and antiself-dual parts; as a result, the self-dual representation, in which states are functions of the self-dual part of the field, can be defined in a nonperturbative quantization. This stands in contrast to a positive frequency representation, which can only be defined with respect to a fixed background metric and a correspondingly restricted time coordinate.

This is one of the main motivations for the use of the Ashtekar variables as the basis for the quantization.

At the linearized level, the self-dual part of a connection (whether or the electromagnetic, Yang-Mills or linearized gravitational field) consists of the positive frequency part of the left handed helicity, plus the negative frequency part of the right handed helicity. The negative frequency part must then be quantized in a kind of "anti- Bargmann" representation. It was not entirely obvious that such a quantization exists, and one of the key results of the program is that self-dual representations exist at the linearized level[13, 14].

C') As I will describe below, the process of renormalization, through which the product of local operators is defined to be another local operator, is necessarily dependent on a background metric (or at least on a background volume element [11, 15].) There is then no local background independent renormalization procedure for local operators. At the same time, we have found that there are background independent regularization procedures for certain non- local observables. Their use results in finite and background independent operators[15, 11]. Further, one can make an argument that any operator constructed from functions of local operators through a regularization procedure which is diffeomorphism invariant in the limit that the regulator is removed will necessarily be finite[11, 29].

D') Without a background metric, there is no symmetry principle that can guide the selection of the inner product. We propose to base the selection on an alternative principle: a complete set of real physical observables must be represented by self-adjoint operators[9, 10]. The proposal depends on the construction of a set of physical observables, realized as well defined operators on the space of physical states, for which the classically corresponding observables are known. We can then use the reality conditions satisfied by the corresponding classical observables to posit hermiticity relations for the physical operators. The inner product is then to be determined by the condition that it realize these hermiticity conditions.

Clearly, what I have just stated is a strategy, but it is not a completely defined procedure. Crucial questions such as which observables and how many observables are necessary are left unspecified. In a number of special cases, including Maxwell theory[16, 13] and linearized gravity[14], in 2+1 gravity[17, 18, 9] and other finite dimensional examples [19], this principle can be implemented and leads to the physically correct

inner products. However, for reasons that will become clear later, we have not yet been able to test this idea in the case of full quantum gravity.

3. The main results of nonperturbative quantum gravity

I would like now to state, concisely and with a minimum of technicalities, exactly what the main results of this approach are, to date. To frame the discussion, let me begin by recalling the main elements of canonical quantization of general relativity in the Ashtekar formalism, following the method of Dirac[20]. We begin with a phase space, which is taken to be the space of pairs of complex $SU(2)$ connections, A_a^i and conjugate electric fields $\tilde{E}^{ai}$ on a three manifold Σ of fixed topology. For everything that follows it will be crucial that $\tilde{E}^{ai}$ is a vector density field. This is necessary so that the Poisson bracket relation,

$$\{A_b^j(y), \tilde{E}^{ai}(x)\} = \delta_b^a \delta^{ij} \delta^3(x,y) \tag{1}$$

makes sense, because the delta function must, in the absence of a background metric, be a density.

To quantize the system, we proceed to construct the state space in three steps, which we call kinematic, diffeomorphism invariant and physical. To define the kinematical state space, $\mathcal{V}_{kin}$, which is the starting point for the quantization, one must first pick out of the algebra of functions on the phase space a subalgebra that one wants to have represented exactly by quantum operators. A choice is necessarily involved at this stage; because of the existence of problems of ordering and the regularization of operator products, it is impossible that the whole algebra of functions on the phase space be identically represented as an algebra of operators. We shall call the subalgebra chosen $\mathcal{A}$. The one condition it should satisfy is that its elements should coordinatize the phase space of interest.

Once we choose $\mathcal{A}$, the kinematical state space is then constructed by finding a linear space $\mathcal{V}_{kin}$ on which there acts an algebra $\hat{\mathcal{A}}$, that is isomorphic to $\mathcal{A}$, at least up to terms that vanish in the limit that $\hbar \to 0$.

The diffeomorphism invariant and physical state spaces, which will be called $\mathcal{V}_{diffeo}$ and $\mathcal{V}_{phys}$ are then constructed as follows. One constructs operators in $\mathcal{V}_{kin}$ that correspond to the classical diffeomorphism and Hamiltonian constraints. (These will be denoted $\hat{D}(v)$ and $\hat{H}(N)$, respectively, where v^a and N are smearing functions that are, respectively, a vector field and an inverse density field.) $\mathcal{S}_{diffeo}$ is then defined to be the subspace of states in $\mathcal{V}_{kin}$ that are annihilated by $\hat{D}(v)$. $\mathcal{V}_{phys}$ is defined to be the subspace of those states that are annihilated both by $\hat{D}(v)$ and $\hat{H}(N)$.

This process is called Dirac quantization[20]. Unfortunately Dirac, while setting out the procedure in one of his very readable little books, failed to include two details that must arise in any field theoretic application of the method. The first is that, at least in all representations known to this time, the Hamiltonian constraint involves operator products and must be regularized. The second is the question of how to choose the inner product.

The key point is that it is not sufficient to give an inner product at the kinematical level, because, unless one is very lucky, the solutions to the quantum constraint equations will be non-normalizable with respect to the kinematical inner product. (Thus, $\mathcal{V}_{diffeo}$

and $\mathcal{V}_{phys}$ are, in general, subspaces of $\mathcal{V}_{kin}$ as vector spaces, and not as Hilbert spaces.) The physical inner product must be chosen just on the space $\mathcal{V}_{phys}$ (and similarly for the diffeomorphism invariant states).

Let us now begin the quantization along these lines. Conventionally, one choices to quantize using the algebra (1). This is the starting point of the construction of Fock representations. However, to construct the non-Fock representations we will be interested in, we take a different subalgebra as a starting point. The main idea behind this choice is to implement the $SU(2)$ gauge invariance explicitly by choosing an algebra of gauge invariant functions. To do this we take the configuration variable is taken to be the holonomies,

$$T[\gamma] \equiv \frac{1}{2} Tr P e^{G \int_\gamma A}. \tag{2}$$

For a conjugate variables we would like to take a family of functions linear in the conjugate momentum. The requirement of gauge invariance then requires us to take functions that are also dependent on loops. To construct them, consider a loop β on which we have a preferred point $\beta(s)$. We may then define the function,

$$T^a[\beta](s) \equiv \frac{1}{2} Tr \left[\tilde{E}^a(\beta(s)) P e^{G \int_s dt \dot{\beta}^a(t) A_a(t)} \right] \tag{3}$$

defined by tracing $\tilde{E}$ at the point $\beta(s)$ against the holonomy around β starting and ending at $\beta(s)$. To complete the definition of the conjugate variable, let us consider a rubber band, β defined by one parameter family of loops $\beta_u(s)$ with $0 \geq u \geq 1$. We may then define

$$E[\beta] \equiv \int \int du ds \epsilon_{abc} \frac{d\beta^a_u(s)}{ds} \frac{d\beta^b_u(s)}{du} T^c[\beta_u](s) \tag{4}$$

The two observables (2) and (4) have a very pretty algebra. Let us define

$$I[\gamma, \beta] = \int d\gamma^a(t) \int d^2\beta^{bc}(s,u) \delta^3(\gamma(t), \beta(s,u)) \epsilon_{abc} \tag{5}$$

to be the intersection number of the loop γ with the two dimensional surface β. The intersection number is equal to an integer; it counts, with the sign reflecting the orientation, the number of times that the loop intersects the strip. A simple calculation than yields,

$$\{T[\gamma], E[\beta]\} = GI[\gamma, \beta] \left[T[\gamma \circ \beta_{u^*}] - T[\gamma \circ \beta^{-1}_{u^*}]. \right]. \tag{6}$$

Here, $\alpha \circ \beta$ refers to the loop made by combining α and β and β_{u^*} is the element of the one parameter family that goes through the point of the rubber band where γ intersects it.

This algebra is called the loop-strip algebra, or the loop algebra for short.

It is important to comment on the presence of Newton's constant, G, in the definitions of the holonomies in these observables. Because the frame field $\tilde{E}^{ai}$ should be dimensionless, the quantity it is conjugate to, the Ashtekar connection, cannot have the usual dimensions of a connection of inverse length if the Poisson brackets (1) are to hold. Instead, it is GA^i_a that has dimensions of inverse length. As a result, there is a G in (6). We will see later that this plays a key role in several results described below.

With this background, I now give an annotated list of the major results so far which have been achieved in the direction of nonperturbative quantum general relativity. The results will be organized according to the three levels of Dirac quantization: kinematical, diffeomorphism invariant and physical.

3.1. Existence of the loop representations at the kinematical level

A loop representation is a quantization of a gauge theory based on the loop algebra given by (6). The loop algebra bears a relationship to the canonical algebra (1) which is somewhat analogous to that the Weyl algebra bears the simple Heisenberg algebra. There are representations of the loop algebra which are also representations of the canonical algebra. Among them is the Fock representation. However, there are also representations of the loop algebra which in which there exist no operator linear in A_a^i, which are inequivalent to the Fock representation. Among these are the representations which are of interest to non-perturbative quantum gravity[22, 21].

A loop representation may be characterized as follows. We introduce a basis of bra states, labeled by loops, such that any ket state $|\Psi>$ may be written as

$$\Psi[\gamma] =< \gamma|\Psi > \tag{7}$$

Note that this notation does not assume the existence of an inner product. The elements of the ket space are functions $\Psi[\gamma]$ which live in some space of functions, the complete specification of which is necessary to define the representation. The bra states, $< \gamma|$ are a linear space dual to the ket space, whose action is defined by (7). For any loop representation these ket-states are constrained to satisfy

$$\sum_i c_i < \gamma_i| = 0 \tag{8}$$

whenever the holonomies satisfy

$$\sum_i c_i Tr P e^{\int_{\gamma_i} A} = 0 \tag{9}$$

for all connections A_a^i. This is the way that the identities satisfied by holonomies-the Mandelstam identities-are imposed on the representation. The loops γ may then be taken to be piecewise differentiable loops. It is convenient also to use a convention in which a loop can refer to a set of loops, if we assume that the trace of the holonomy of a set of loops is taken to be the product of the traces of the holonomies of the individual loops. Indeed, by a loop or a set of loop we really always mean equivalence classes of loops under the relation (8)[3].

The action of the operators that represent the loop variables (2) and (4) are then defined by their action on the bra states:

$$< \gamma|\hat{T}[\alpha] =< \alpha \cup \gamma| \tag{10}$$

$$< \gamma|\hat{E}[\beta] = \hbar G \left(< \gamma \circ \beta_{u^*}| - < \gamma \circ \beta_u^{-1}| \right) \tag{11}$$

In these equations the $\cup$ represents union in a set of loops while the $\circ$ represents forming a new loop by a product of two loops. It is not hard to verify that these operators satisfy the algebra (6) (with the $\hbar$ inserted. It is, in fact, interesting to note that $l_p^2 = \hbar G$ appears in the algebra even at the kinematical level. This fact (and the fact that it is the Planck area and not the Planck length that appears) plays a crucial role below.

[3] As has been emphasized by Gambini and Trias and their coworkers, the space of these equivalence classes of loops actually forms a group[8]. The basis for their alternative formulation of the loop representation is the representation theory of this group.

Finally, we may note that the fact that the group is $SL(2,C)$ (or some real subgroup of it) appears in the form of the right hand side of (6), together with the equivalence relations generated by (8). Connections based on other Lie groups also have loop representations; they differ only in the form of this action and in the equivalence relations generated by the holonomies.

3.2. Applications of the loop representations to linearized field theories: Fock representations

An example of the forgoing is the loop representations of abelian connections. The loop representation actually first appeared in a quantization of the free Maxwell field given by Gambini and Trias in 1981 [8]. The loop quantization of free Maxwell theory was also treated later in [16, 13]. In these papers it is shown that one can construct a space of states $\Psi[\gamma]$ which is isomorphic to the Fock space of free photon states. The whole of free quantum electrodynamics, including the Hamiltonian, creation and annihilation operators and inner product may be written simply in the loop representation.

Indeed, there are several different representations of the Fock space as functions of loops. These correspond to the different ways to write the Fock space as functions of the connection. The two well known connection representations of free field theories are the Schroedinger representation, in which the states are functions of the real connection, A_a and the positive frequency, or Bargmann representation, in which the states are functions of the positive frequency part of the connection.

In the loop representation that corresponds to the positive frequency connections the ground state is simply written as

$$\Psi_0[\gamma] =< \gamma|0 >= 1 \tag{12}$$

The state of one photon of momentum $\vec{p}$ and polarization $\vec{\epsilon}$ is written,

$$\Psi_{\vec{p},\vec{\epsilon}}[\gamma] =< \gamma|\vec{p},\vec{\epsilon} >= \oint ds e^{p_b \cdot \gamma^b(s)} \dot{\gamma}^a(s)\epsilon_a \equiv F[\vec{p},\vec{\epsilon}] \tag{13}$$

The $F[\vec{p},\vec{\epsilon}]$'s play a dual role in the formalism. First, they provide a useful set of coordinates for the loops modulo the relations (8), with the abelian holonomy[4] Second, the multiphoton states are written as polynomials of the $F[\vec{p},\vec{\epsilon}]$'s. The Fock space is then defined to be the space of functions of loops that are analytic functions of the loop coordinates $F[\vec{p},\vec{\epsilon}]$.

The Hamiltonian and inner product of quantum Maxwell theory are then written as operators on loop space as follows,

$$\hat{H} = \sum_{\epsilon} \int d^3p|p|F[\vec{p},\vec{\epsilon}]\frac{\delta}{F[\vec{p},\vec{\epsilon}]} \tag{14}$$

$$< \Phi|\chi >= \int [dF]\bar{\Phi}\chi e^{-\int \frac{d^3p}{|p|}|F[\vec{p},\vec{\epsilon}]|^2} \tag{15}$$

It is interesting to note that in every case where the construction of the loop representation has been completed, it has been found that there exists a transform that

[4] Their extension to a set of coordinates on the space of loops modulo the relations that arise from non- abelian holonomies are the basis of a new approach to the loop representation of Gambini and his collaborators[23].

connects it to the appropriate connection representation in which the states are functions of the connection. The general form for this transform is

$$\Psi[\gamma] = \int d\mu[A] T[\gamma, A] \psi[A] \tag{16}$$

where $d\mu[A]$ is an appropriate measure on the space of connections mod gauge transformations.

3.3. Existence of self-dual representations for both connection and loop representations

The key discovery of Ashtekar is that in full general relativity the self-dual part of the connection may be considered as a configuration variable of the theory, i.e. every component at ever point commutes with every other one. Because of this, we would like to base the quantization of the full theory on the self-dual representation, in which the observables, (2) and (4), whose algebra we quantize, are taken to be functions of the self-dual part of the connection. If this is to be a successful route to the quantum theory it would be most convenient if the linearized theory can also be quantized in a representation in which the observables are functions of the self-dual part of the linearized connection. This, however, gives rise to the following question: At the linearized level, the self-dual part of the connection is the positive frequency part of the left handed helicity component, plus the negative frequency part of the right handed helicity component. This means that the right handed component must be quantized in a kind of negative frequency, or anti-Bargmann representation. Thus, the first question that must be asked is whether there exist such anti-Bargmann representations, which are diagonal in the annihilation operator rather than in the creation operator. At first sight this seems to be impossible; consider for example the equation that the annihilation operator annihilates the ground state. In such a representation it must read

$$< \bar{z}|\hat{a}|0 >= \bar{z}\psi_0(\bar{z}) = 0. \tag{17}$$

In fact, such representations exist, but their expression requires distributions rather than holomorphic functions, as in the usual Bargmann representation[13]. Thus, in the sense of distributions, the solution to (17) is

$$\psi_0(\bar{z}) = \delta(\bar{z}). \tag{18}$$

Once this obstacle is overcome, it is straightforward to construct the full negative frequency representation, for the harmonic oscillator and for any linear field theory[13] whose configuration variable is a connection. Among these is linearized gravity[9], which may be quantized in the loop representation[14].

The existence of the loop representation for linearized general relativity not only serves as a confirmation of the basic program of nonperturbative quantum gravity, it may expected to play a key role in the physical interpretation of the exact theory.

3.4. Non-Fock representations of the loop algebra

The Fock representations are important for the construction of linearized field theories, but they are inappropriate as a starting point for the construction of diffeomorphism invariant theories. This is because they cannot carry unbroken representations of the

diffeomorphism group for the simple reason that the diffeomorphisms are broken by the existence of the background metric. Since the diffeomorphisms cannot be represented one cannot use them as a starting point to construct diffeomorphism invariant states.

The key question is then, do there exist representations of the kinematical observable algebra that carry unbroken representations of the spatial diffeomorphism group? I do not know the answer for the case of the canonical algebra (1). But, for the loop algebras of the form of (6) (and their generalizations to other groups) there do exist representations with this property[22, 21].

These representations are based on the use of the discrete measure on the space of loops[5]

To construct this representation it will be useful to introduce a set of basis states, which we call characteristic states. For a loop α which contains no intersections, we may define the state, such that, for γ also nonintersecting,

$$\chi_\alpha[\gamma] =< \gamma|\alpha >= 1 \quad \text{if} \quad \alpha = \gamma \text{and} \quad \text{otherwise} \quad \text{vanishes.} \tag{19}$$

Here, the equality is always meant in the sense of the equivalence relations (8). There is one more case, which is if $\gamma \neq \alpha$, but contains intersections. In this case $\chi_\alpha[\gamma]$ does not necessarily vanish, its actual value is determined by requiring that it be an eigenstate of the operators defined in (25) and (30), below [11]. There are also basis states associated with intersecting loops [11].

We will denote these characteristic states abstractly by $|\alpha >$, making use of the standard Dirac formalism in which bras represent elements of function spaces and kets are linear maps from those function spaces to the complex numbers.

Let us then consider the linear space, $\mathcal{V}_{discrete}$, which consists of states of the form,

$$\Psi[\gamma] = \sum_I c_I \chi_{\alpha_I}[\gamma] \tag{20}$$

where we require that

$$\sum_I |c_I|^2 < \infty \tag{21}$$

Here the sum is over any countable set of loops in Σ. It can be easily verified that the formal definitions of the loop operators, (10) and (11), are well defined when acting on the states in $\mathcal{V}_{discrete}$.

We can impose an inner product on $\mathcal{V}_{discrete}$ by defining

$$< \alpha|^\dagger = |\alpha > \tag{22}$$

Note that this inner product does not realize the kinematical relativity conditions for the $T[\gamma]$ observable[6], but it does realize them (at least formally) for functions of $\tilde{E}^{ai}$ only. Like any kinematical inner product, it is useful only as a mathematical device.

With any choice of inner product which makes the characteristic states normalizable, the discrete representation is unitarily inequivalent to Fock space. There is a

[5] recall that by loops I always mean loops modulo the relations (8).

[6] For the reader unfamiliar with the reality conditions, they are the conditions that the three metric and its time derivative both turn out to be real when computed in the Ashtekar formalism. They imply that A_a^i is a complex connection, which, together with its complex conjugate, satisfies a certain polynomial condition. The result is that the loop operators are not real.

possibility that it may be of use for nonperturbative treatments of gauge theories, because it implements, in the continuum, the quantization of the electric flux. This is a possibility that needs further development. At the present time its use comes from its application to diffeomorphism invariant quantum field theories. This is because an exact, unbroken, unitary representation of the diffeomorphism group can be defined on it as,

$$\hat{U}(\phi)|\gamma>=|\phi^{-1}\circ\gamma> \tag{23}$$

where ϕ is any diffeomorphism. This means that the generator of diffeomorphisms is well defined in this representation,

$$\hat{D}(v)\Psi[\gamma]=\frac{d}{dt}\hat{U}(\phi_t)\Psi[\gamma], \tag{24}$$

where ϕ_t is a one parameter group of diffeomorphisms generated by the vector field v^a. $\hat{D}(v)$ may be shown from these definitions to satisfy the algebra of vector fields on Σ.

3.5. Classification of the loop representations in the real case

I have described two different representations of the loop algebra (6). In one case, in which we require that both both A_a^i and $\tilde{E}^{ai}$ are real, the representations of the $SU(2)$ loop algebra have been completely classified. This was done by Ashtekar and Isham [21], who make use of the fact that in this case the loop algebra (6) is a star algebra. The classification can then be done using some of the technology of the representation theory of star algebras developed by Gel'fand and collaborators.

3.6. Application of the loop representation to quantum Yang- Mills theory

As I mentioned above, the loop representation may be applied to non-abelian gauge theories[24]. The loop representation may be developed in the context of the lattice regularization, where the loop states provide a gauge invariant basis for the state space. This formulation has been the starting point for several works in which new numerical approaches to lattice gauge theory, both with and without fermions have been explored. These works involve approximation procedures which are based on the fact that in the loop basis almost all the matrix elements of both the Hamiltonian and inner product are zero . As a result, sparse matrix and cluster techniques may be applied[25, 26]. For example, in $2+1$ dimensions extensive numerical calculations have been done for both $SU(2)$ [25] and $SU(3)$ [26], which showed that results for the ground state energy and mass gap (as functions of the coupling constant), obtained previously by Monte Carlo simulations and are reproduced accurately. Furthermore, in $3+1$ dimensions numerical work has been, and is being, done for the case of QED with fermions[27].

In addition to these numerical approaches, there have been some very interesting analytical work done on non-abelian gauge theories in the loop representation, by Loll[28], Rovelli[29] and others.

3.7. Nonexistence of local operators in non-perturbative quantum gravity

For the remainder of this section, I will confine myself to the applications of the loop representation to quantum gravity. I begin with several results about observables and the classical limit at the kinematical level. First, of all, it is not trivial to construct quantum

observables at the kinematical level because such observables must be invariant under the Yang-Mills gauge transformations, and any such observables that involve the frame fields involve operator products. Thus, regularization is an issue even at the kinematical level.

As the kinematical level is meant to be a stepping stone to the diffeomorphism invariant and physical levels, we will be interested only in regularization procedures that do not introduce extra background dependence into the final definitions of the operators. Any regularization procedure depends on additional structure such as background metrics or coordinate systems as these are needed to specify how the point splitting is done or define the cutoffs. What we must then require is that when finite operators are finally produced as a result of the process they have no dependence on these structures.

We have discovered that this requirement seems to rule out the conventional renormalization procedures of Poincare invariant quantum field theories[15, 11]. Although this was discovered through a painful process in which many possible approaches were tried and discarded, the reason for this can be stated very simply. Local operators are distribution valued and distributions are, in the absence of a background metric, densities of weight one. A renormalization procedure is a procedure by which a product of two local operators is defined to be a third local operator. It is thus a procedure for multiplying two distributions to get a third distribution. However, there is a problem with the density weights, because the product of two distributions should have density weight two, but a local operator will have only density weight one. The result is that any such renormalization procedure must introduce an additional scalar density so that the density weights on the left and right hand sides of the product match. What we found was that in any procedure we tried, such a density always appeared which was a function of the background structures introduced in the regularization and renormalization procedures.

Now, in Minkowski spacetime, or even in quantum field theory in a curved spacetime, there is a preferred density which is given by the determinant of the background metric. In these cases the ambiguity may be reduced to one free renormalization constant. This is the reason for the existence of a free renormalization scale in the conventional renormalizations of operator products.

However, in nonperturbative quantum gravity there is no preferred density and the ambiguity of a scale in the renormalization of a quantum field theory in a classical background becomes an ambiguity up to a density. The result is that it is very difficult to imagine how a renormalization procedure could be constructed for operator products in this context that did not lead to a breaking of diffeomorphism invariance.

This means, in particular, that when using a frame field formalism such as the Ashtekar variables there is no operator to measure the metric at a point. This is because the basic variable is the frame field, $\tilde{E}^{ai}$, which is related to the metric through $\tilde{q}^{ab} = \tilde{E}^{ai}\tilde{E}^{b}_{i}$.

3.8. Existence of finite, background independent non-local operators at the kinematical level

One might think that as a result of the situation I've just described it is impossible to define meaningful observables that measure the spatial metric in non-perturbative quantum gravity. Fortunately, this is not the case, because there are non-local observables that

are equivalent to the metric in the sense that a complete measurement of them allows the metric to be reconstructed. We have found that it is possible to construct quantum operators that correspond to some of these non-local observables and that these operators are finite and background independent, when constructed by means of the right regularization procedure[15, 11]. Thus, as these operators don't need to be renormalized, they escape the difficulty I described in the previous paragraph.

The idea behind the construction of these operators is very simple: If there is no unambiguous procedure for multiplying two distributions to get a third distribution, we may construct unambiguous procedures that define the *square root of the product of two distributions.*

I will mention here three examples of such observables[15, 11]. First, given any one form ω, we may define the integral of its norm as follows,

$$Q[\omega, \tilde{E}^{ai}] = \int_\Sigma \sqrt{\tilde{E}^{ai}\omega_a \tilde{E}^{bi}\omega_b} \tag{25}$$

This observable can be regulated through a modified point splitting procedure. I will not describe it here, the details are given in [11]. As may be expected, the hardest part of the construction is taking the operator square root. The result is easiest to express in terms of the bras $< \alpha|$. For the case of a non-intersecting loop, α, the bra is, for every ω, an eigenstate of the operator corresponding to $Q[\omega, \tilde{E}^{ai}]$. The action of the operator is given by,

$$< \alpha|\hat{Q}[\omega] = \frac{l^2_{Planck}}{2} \oint_\alpha ds \ |\dot{\alpha}^a \omega_a(\alpha(s))| \ < \alpha|. \tag{26}$$

A second nonlocal operator that can be defined as a quantum operator is the area of any surface. Given a surface $\mathcal{S}$, there is a function on the kinematical phase space that is its area, it is given by,

$$\mathcal{A}[\mathcal{S}, q] = \int_\mathcal{S} \sqrt{\tilde{q}^{ab} n_a n_b} \tag{27}$$

where n^a is the unit normal of the surface. The problem is how to turn this into an operator when there is no operator for the metric $\tilde{q}^{ab}$? There is a solution, which is the following. Let me represent the surface by a distributional one form, $\pi^\mathcal{S}_a$ which is given by,

$$\pi^\mathcal{S}_a(x) = \int d^2 S^{bc}(\sigma) \ delta^3(x, \mathcal{S}(\sigma))\epsilon_{abc}, \tag{28}$$

where σ are coordinates on the surface and ϵ_{abc} is the inverse of the Levi-Civita density. I then can consider the expression

$$A(\mathcal{S}, \tilde{E}) = Q(\pi, \tilde{E}) \tag{29}$$

It is not difficult to show that this is equal to the area of the surface given by the metric by (27). To show this, we demonstrate that an equivalent expression, when the frame fields are smooth, is given by

$$A(\pi, \tilde{E}) = \lim_{N\to\infty} \sum_{N=1}^{N} \sqrt{A^2_{approx}[\mathcal{R}_i]} \tag{30}$$

where space has been partitioned into N regions $\mathcal{R}_i$ such that in the limit $N \to \infty$ the regions all shrink to points. Here, the observable that is measured on each region is defined by,

$$A^2_{approx}[\mathcal{R}] \equiv \int_{\mathcal{R}} d^3x \int_{\mathcal{R}} d^3y T^{ab}(x,y)\pi_a(x)\pi_b(y) \tag{31}$$

Here $T^{ab}(x,y)$ is a loop operator that is quadratic in the frame fields $\tilde{E}^{ai}$. It is constructed in the following way. Pick a background Euclidean metric and use it to define, for every two points, x and y, in the spatial manifold Σ, a circle, γ_{xy}, such that $\gamma_{xy}(0) = x$ and $\gamma_{xy}(\pi) = y$ and such that in the limit that y approaches x, the circles shrink to the point x. Then, define

$$T^{ab}(x,y) = \frac{1}{2} Tr \left[(\mathcal{P} \exp G \int_y^x A_a d\gamma^a_{xy}) \tilde{E}^a(x) (\mathcal{P} \exp G \int_x^y A_a d\gamma^a_{xy}) \tilde{E}^b(y) \right]. \tag{32}$$

To show the equivalence between (29) and (30), we start with (29) and regulate it by means of a point splitting procedure by introducing, with respect to the background euclidean coordinate system, a set of test fields $f_\epsilon(x,y)$ by

$$f_\epsilon(x,y) = \frac{1}{\epsilon^3} \theta[\frac{\epsilon}{2} - |x^1 - y^1|] \theta[\frac{\epsilon}{2} - |x^2 - y^2|] \theta[\frac{\epsilon}{2} - |x^3 - y^3|]. \tag{33}$$

In these coordinates

$$\lim_{\epsilon \to 0} f_\epsilon(x,y) = \delta^3(x,y) \tag{34}$$

We can then write

$$A(\pi, \tilde{E}) = Q(\pi, \tilde{E}) = \lim_{\epsilon \to 0} \int d^3x \sqrt{\int d^3y \int d^3z T^{ab}(y,z)\pi_a(y)\pi_b(z) f_\epsilon(y,x) f_\epsilon(z,x)} \tag{35}$$

When the expression inside the square root is slowly varying in x we can re-express it in the following way. We divide space into regions $\mathcal{R}_i$ which are cubes of volume ϵ^3 centered on the points $x_i = (n\epsilon, m\epsilon, p\epsilon)$ for n, m, p integers. We then write,

$$\begin{aligned} A(\pi, \tilde{E}) &= \lim_{\epsilon \to 0} \sum_i \epsilon^3 \left[\int d^3y \int d^3z T^{ab}(y,z)\pi_a(y)\pi_b(z) f_\epsilon(y,x_i) f_\epsilon(z,x_i) \right]^{\frac{1}{2}} \\ &= \lim_{N \to \infty} \sum_{N=1}^{N} \sqrt{A^2_{approx}[\mathcal{R}_i]} \end{aligned} \tag{36}$$

If we now plug into these expressions the distributional form (28) it is straightforward to show that

$$A(\pi_{\mathcal{S}}, \tilde{E}) = \int_{\mathcal{S}} \sqrt{h} \tag{37}$$

where h is the determinant of the metric of the two surface, which is given by $h = \tilde{q}^{ab} n_a n_b$ where n^a is the unit normal of the surface.

It is not difficult to show that that starting from the expressions (30) and (31) we may construct a quantum operator for the area of a surface $\mathcal{S}$ in the loop representation. We can show that the expression (30) is equivalent to an expression in which the surface is partitioned into N subsurfaces $\mathcal{S}_I$, $I = 1, ..., N$. We then write

$$A_{\mathcal{S}} = \lim_{N \to \infty} A^2_{\rm appr}[\mathcal{S}_I]. \tag{38}$$

where $A^2_{\rm appr}[\mathcal{S}_I]$ denotes an approximate expression for the area of the subsurface, which is defined by

$$A^2_{\rm appr}[\mathcal{S}_I] \equiv \int_{\mathcal{S}_I} d^2 S^{bc}(x) \epsilon_{abc} \int_{\mathcal{S}_I} d^2 S^{\prime b' c'}(x') \epsilon_{a'b'c'} T^{aa'}(x, x') \tag{39}$$

This last expression may be written as a quantum operator, by writing an operator for $T^{aa'}(x, x')$. This can be done, but, as the action is a bit complicated, I do not give it here. It may be found in [7, 9, 10, 11]. The result is that the limit (38) may be taken on any loop state, leading to a final expression that is finite and independent of the background structure that went into the definition of the loop operator. The result, for non-intersecting loops α is[11, 15],

$$< \alpha | \hat{A}[\mathcal{S}] = \frac{l^2_{Planck}}{2} I^+[\mathcal{S}, \alpha] < \alpha |. \tag{40}$$

Here $I^+[\mathcal{S}, \alpha]$ is the positive, unoriented, intersection number, which counts (independent of orientation) the number of intersections of the loop with the surface.

If the loop α has intersections the action of the operator is more complicated, but it is still finite and background independent. Details are given in [11].

The third observable that can be constructed in this way is the volume of any region. It is described in [11]

3.9. Quantization of areas and volumes in the discrete representation

The result (40) says that the bras in the loop representation are, at least for intersecting loops, eigenstates of the operator that measures areas. This does not mean that, in general, there are normalizable states that are eigenstates, this can only be the case if the inner product is defined in such a way that there is a state which is the hermitian conjugate of the bra $< \alpha |$ which is a normalizable state and if the area operator is hermitian in that inner product.

In general these conditions will not be satisfied, for example, there are no normalizable eigenstates of the area operator in the Fock representation of linearized quantum gravity. But in the context of discrete representations an inner product can be defined by the imposing the condition that the inner product be chosen such that the area operator is hermitian so that its eigenstates comprise an orthonormal basis. For nonintersecting loops this is given by (22), for intersecting loops it is more complicated [11].

With such an inner product, we may say that area is quantized in the discrete representation, because the spectrum of the operator that measures area is discrete. This spectrum consists, first of all, of the eigenvalues $N l^2_{Planck}/2$, for every nonnegative integer N. There is also another discrete sequence of eigenvalues corresponding to eigenstates that are labeled by intersecting loops, these are described in [11].

In the same representation, the volume operator turns out also to have a discrete spectrum. The basic action of the volume operator in this representation turns out to be to annihilate the states $|\alpha >$ associated with nonintersecting loops α and to rearrange the routings through the intersections of the intersecting loops. That is, with each intersecting loop, α, one can associate a finite dimensional subspace of the state space which is spanned by the loops which have the same support as α but differ as to how the loops are routed through the intersection points. The action of the volume operator

is then to induce a finite dimensional matrix in each such subspace that rearranges the routings then multiplies by l^3_{Planck}. Its non-zero eigenvalues are given by l^3_{Planck} times the eigenvalues of these finite dimensional matrices.

3.10. The correspondence principle: Existence of states which approximate classical metrics at large scales

We are used to describing the classical limit of quantum theories in situations in which there is a background metric against which to measure distance intervals. It is not a completely trivial problem to understand what it means to take the classical limit in a non-perturbative quantum theory of gravity, in which there is no background metric. To do this we need to first understand two simple points. First, in pure quantum general relativity the classical limit is a limit of large distances. This is because the theory has only one dimensional parameter, the Planck length, $l_{Planck} = \sqrt{\hbar G/c^3}$. This obviously goes to zero as $\hbar \rightarrow 0$. It is perhaps also significant that it is the Planck area that is proportional to $\hbar$, this perhaps is the reason why length intervals are not defined in the quantum theory, while areas are both defined and are quantized.

The second thing to be understood is that without a classical metric we don't know what distance and area means and so we cannot tell which intervals are small or large compared to the Planck scale.

Because of these two points, it is easiest to express the classical limit in a way that may seem backwards, as follows[15, 10, 11]. Given any classical metric h_{ab}, whose curvatures are small compared to the Planck scale, we seek a quantum state $|\Psi>$ which has the property that it is an eigenstate of the operators $\hat{Q}[\omega]$ and $A[\mathcal{S}]$, and where, for every one form ω which is slowly varying with respect to the metric h_{ab} and every surface which has small extrinsic curvatures, again with respect to h_{ab} (where, again these area measured with respect to the Planck scale) we have

$$\hat{Q}[\omega]|\Psi>=(Q(\omega,h)+O(l_{Planck}|\bigtriangledown\omega|))\,|\Psi>, \tag{41}$$

and

$$\hat{A}[\mathcal{S}]|\Psi>=\left(A[\mathcal{S},h]+O\left(\frac{l^2_{Planck}}{A[\mathcal{S},h]}\right)\right)|\Psi>. \tag{42}$$

Thus, the eigenvalues are required to give back the corresponding values for the metric h_{ab} up to terms that are small measured in Planck units. When these conditions are satisfied we say that $|\Psi>$ is a semiclassical state that approximates the metric h_{ab}.

In the loop representation we call states that have this property weaves, because it can be satisfied by loop states $|\Delta>$, where the multiloop Δ consists of many small loops which are arranged so that, through every surface, $\mathcal{S}$, as described above, approximately one line of a loop pierces $\mathcal{S}$ per half Planck area of the surface, measured in the metric h_{ab}. It is easy to give examples of such states; a particularly simple one, for the case that the metric h_{ab} is flat, is constructed as follows [15].

We use h_{ab}, to introduce a random distribution of points on $\Sigma = R^3$ with density n. This means that in any given volume V there are $nV(1+\mathcal{O}(1/\sqrt{nV}))$ points. We center a circle of radius $a=(1/n)^{\frac{1}{3}}$ at each of these points, with a random orientation. Again, the notion of a random orientation is defined with respect to h_{ab}. We call this whole collection of circles Δ.

It is now straightforward to show that $|\Delta>$ is an eigenstate of $\hat{Q}[\omega]$ and $A[\mathcal{S}]$. However the conditions (41) and (42) are only satisfied if the density n is chosen so that[15],

$$a = \sqrt{\pi/2}\, l_p. \tag{43}$$

3.11. The necessity of discrete structure at the Planck scale

This last result (43) means that if we require that the state $|\Delta>$ approximate the classical metric h_{ab}, when we measure it with operators that average the metric information over scales that are large in Planck units, it is necessary that the state have discrete structure at the Planck scale, where, in both cases, what we mean by the Planck scale is determined by h_{ab}. This is a direct consequence of the fact that we were able to construct non-local operators to measure the metric information that are finite and background independent. This result can be generalized by considering families of loops that generalize Δ by being described by more parameters. In each case it is found that there is one combination of parameters that is fixed to be a certain exact multiple of the Planck scale. The other combinations of parameters with dimensions length are also restricted to be on the order of the Planck scale, so that the requirements on the orders of the errors in (41) and (42) are satisfied[15].

This means that in nonperturbative quantum gravity, at least in the formulation I am describing here, it is possible to have states that are semiclassical on large scales. However, there are no states that are semiclassical on the Planck scale.

This completes my discussion of the kinematical level of the theory. I now turn to results concerning diffeomorphism invariant states.

3.12. Complete solution of the diffeomorphism constraints

We have defined the action of diffeomorphisms on states in the loop representation by (23) and (24). Using these, we define the space of diffeomorphism invariant states, $\mathcal{V}_{diffeo}$, to be those loop states that satisfy,

$$\Psi[\alpha] = \hat{U}(\phi)\Psi[\alpha] = \Psi[\phi^{-1} \circ \alpha] \tag{44}$$

for all elements of the connected component[7] of the diffeomorphism group of Σ.

The definition (44) of diffeomorphism invariant states may be compared with a similar condition in the metric representation, which says that the quantum states are functions of the three geometry, which are defined to be diffeomorphism equivalence classes of three metrics. The difference is that the diffeomorphism equivalence classes of loops are countable[8][30] and a great deal is known about their classification. If we denote by $\{\alpha\}$ the diffeomorphism equivalence class of the loop α (also known as its knot or link class) the condition (44) means that[7]

$$\Psi[\alpha] = \Psi[\{\alpha\}]. \tag{45}$$

[7] There are two ways to treat the large diffeomorphisms: as symmetries or as part of the gauge group. I do not discuss this issue here.

[8] Assuming certain mild conditions on the finiteness of the components and the intersections.

Because the knot classes are countable the space of diffeomorphism invariant states, $\mathcal{V}_{diffeo}$, has a countable basis, which are the characteristic states of the knot classes. These are,

$$\Psi_{\{\gamma\}}[\{\alpha\}] = \delta_{\{\alpha\}\{\gamma\}}. \tag{46}$$

We can make $\mathcal{V}_{diffeo}$ a Hilbert space by imposing an inner product. The simplest possibility is one in which these characteristic states are orthogonal, that is if we chose a basis of bra states $< \{\alpha|$ such that $\Psi[\{\alpha\}] =< \{\alpha\}|\Psi >$ and we chose the inner product so that $< \{\alpha\}|^\dagger = |\Psi_{\{\alpha\}} >$, we have

$$< \{\alpha\}|\Psi_{\{\gamma\}} >= \delta_{\{\alpha\}\{\gamma\}} \tag{47}$$

This is almost certainly not the right inner product for general relativity, because it corresponds to a reality condition in which the connection is real. However, it may be useful as a technical device to bound limits in certain calculations.

It should be mentioned there is a diffeomorphism invariant theory for which (47) is the physical inner product. This is the Husain-Kuchar model[31], which is a limit of general relativity in which the speed of light has been taken to infinity[32]. In the classical version of this theory, all physical evolution has been frozen, and all solutions are static. Because of this the theory has no Hamiltonian constraint-it has only the gauge and diffeomorphism constraints.

The Husain-Kuchar model is a very interesting model because it is a three plus one dimensional diffeomorphism invariant theory that has an infinite number of physical degrees of freedom. It is solved to the extent that the exact state space and inner product have been constructed. The theory needs to be completed by the construction of a sufficient number of diffeomorphism invariant observables, represented by operators on $\mathcal{V}_{diffeo}$, on which the interpretation can be grounded. As there are only the spatial diffeomorphism invariant constraints, this problem is significantly easier than in the case of the full theory. The Husain-Kuchar model is a very useful laboratory to study those problems of the interpretation of diffeomorphism invariant quantum field theory that are *not* related to the problems of time and time reparametrization invariant observables.

3.13. Some finite diffeomorphism invariant operators

In fact, a small number of diffeomorphism invariant observables can be directly written down. I will describe here one of them, which is closely related to the area observable I described in subsection 3.8 above. The idea is to introduce a dynamical field whose configuration can define a surface. The area of that surface will then be a diffeomorphism invariant quantity[9].

One way to do this is to couple an antisymmetric tensor gauge field to gravity[36]. This is a two form, $C_{ab} = -C_{ba}$ subject to a gauge transformation generated by a one form Λ_a by,

$$\delta C_{ab} = d\Lambda_{ab}. \tag{48}$$

It's field strength is a three form which is denoted $W_{abc} = dC_{abc}$. In the Hamiltonian theory its conjugate momenta is given by $\pi^{ab} = -\pi^{ba}$ so that

$$\{C_{ab}(x), \pi^{cd}(y)\} = \delta_a^{[c}\delta_b^{d]}\delta^3(y, x). \tag{49}$$

[9] The idea of using matter fields to define physical and diffeomorphism invariant observables is an old idea, which goes back at least to a paper of DeWitt [33]. It has been recently revived [34, 35].

The gauge transform (48) is then generated by the constraint

$$G = \partial_c \pi^{cd} = 0 \tag{50}$$

Any two dimensional surface $\mathcal{S}$ defines a distributional configuration of the π^{ab} by equation (28), where we now want to understand the field π_a in that equation as being the dynamical field dual to π^{bc} by $\pi_a = \epsilon_{abc}\pi^{bc}$. These distributional configurations are solutions to the gauge constraint (50) and I thus know of no reason it cannot be considered to be an allowed configuration of the classical field.

We can then interpret equation (29) as the definition of a diffeomorphism invariant observable by reading it as a function of a dynamical π_a field and the gravitational field. It has the interpretation that when the π^{ab} field defines a surface through a distributional configuration by equation (28), it gives the area of that surface.

This observable can be promoted to an operator if we also quantize the C_{ab} field. This can be done by constructing a surface representation to represent it, completely analogous to the loop representation. To do this we introduce a surface observable,

$$T[\mathcal{S}] = e^{k\int_S C} \tag{51}$$

associated to every closed surface $\mathcal{S}$. The k is a free constant with dimensions of inverse action[36]. The algebra we will quantize is then the surface algebra

$$\{T[\mathcal{S}], \pi^{bc}(x)\} = k\int d^2\mathcal{S}^{bc}(\sigma)\delta^3(x, \mathcal{S}(\sigma))T[\mathcal{S}] \tag{52}$$

We can then construct a representation of this algebra in which states are functions of surfaces $\Psi[\mathcal{S}]$. We then define the representation by[36],

$$\hat{T}[\mathcal{S}']\Psi[\mathcal{S}] = \Psi[\mathcal{S}' \cup \mathcal{S}] \tag{53}$$

and

$$\hat{\pi}^{ab}(x)\Psi[\mathcal{S}] = \hbar k\int d^2\mathcal{S}^{ab}(\sigma)\delta^3(x, \mathcal{S}(\sigma))\Psi[\mathcal{S}]. \tag{54}$$

To represent the coupled C_{ab}-gravity system we take the direct product of this state space with the loop representation for quantum gravity. The states are then functions, $\Psi[\alpha, \mathcal{S}]$, of loops and surfaces. We may introduce a set of bra's, $< \alpha, \mathcal{S}|$, labeled by loops and surfaces so that $\Psi[\alpha, \mathcal{S}] =< \alpha, \mathcal{S}|\Psi >$.

We may then impose the diffeomorphism constraints, suitably extended to the coupled Einstein-C_{ab} system[36]. I will not give the details here, the result is that the diffeomorphism invariant states may be constructed and they are functions of the diffeomorphism equivalence classes of loops and surfaces. Denoting these classes by $\{\alpha, \mathcal{S}\}$, the diffeomorphism invariant state space then consists of functions of the form

$$\Psi[\{\alpha, \mathcal{S}\}] =< \{\alpha, \mathcal{S}\}|\Psi > . \tag{55}$$

We then want to express the area observable (37) as a diffeomorphism invariant operator and show that it does indeed measure areas. It is straightforward to show that the bras at the kinematical level, $< \alpha, \mathcal{S}|$, are, for nonintersecting loops α, eigenstates of the operator $\hat{A}$. This operator may be constructed by using the expressions (30) and (31) as the definition of a regularization procedure, in the usual way[11]. As the regularization

breaks diffeomorphism invariance, this calculation must be done at the kinematical level. A straightforward calculation shows that

$$< \alpha, \mathcal{S}|\hat{A}^2_{approx}[\mathcal{R}] = (\frac{\hbar k l^2_{Planck}}{2})^2 I[\alpha, \mathcal{S} \cap \mathcal{R}]^2 < \alpha, \mathcal{S}| \tag{56}$$

where $\mathcal{S} \cap \mathcal{R}$ means the part of the surface that lies inside the region. It then follows from (30) that

$$< \alpha, \mathcal{S}|\hat{A} = \frac{\hbar k l^2_{Planck}}{2} I^+[\alpha, \mathcal{S}] < \alpha, \mathcal{S}| \tag{57}$$

This may be compared with (40), we see that the only difference is that now that the surface is dynamical it is specified by the state and not by the operator. In addition, we see that if we want agreement between the units measured by this and the kinematical area operator we must pick $k = 1/\hbar$.

The action of $\hat{A}$ can be lifted to the space of diffeomorphism invariant states, giving us,

$$< \{\alpha, \mathcal{S}\}|\hat{A} = \frac{\hbar k l^2_{Planck}}{2} I^+[\{\alpha, \mathcal{S}\}] < \{\alpha, \mathcal{S}\}| \tag{58}$$

We have thus defined a diffeomorphism invariant operator that assigns to the surface an area which is given by $\hbar k l^2_{Planck}/2$ times the number of intersections of the loop with the surface. Thus, we see that the same techniques that gave us finite and background independent kinematical operators work to give us finite operators acting on diffeomorphism invariant states.

If we use the inner product (47) of the Husain-Kuchar model on the space of diffeomorphism invariant states, suitably extended to include the coupling to the C_{ab} field [36], we see that $\hat{A}$ is a hermitian operator and that its spectrum is quantized. Thus, we see that the technology we have been developing allows us to derive a prediction from a $3+1$ dimensional diffeomorphism invariant quantum field theory. This is that the area of any two dimensional surface is quantized in units of the Planck area. While that theory is a model, which corresponds classically only to a limit of full general relativity, this is an encouraging result. Moreover, it does not seem impossible that with the addition of structure corresponding to clocks it will be possible to extend the $\hat{A}$ observable to the full physical case, and that we will find that this prediction stands in full quantum general relativity.

It is known that a few additional diffeomorphism invariant operators can be constructed in this way, in either the case of pure gravity or gravity coupled to matter. Examples of these[11] are the volume of the universe and the areas of maximal surfaces (when π^2 of the spatial manifold is non-trivial.) Of course, there must be an infinite number of diffeomorphism invariant observables, it is still an open problem to show that these techniques allow the construction of an infinite number of such operators.

3.14. Connection between finiteness and diffeomorphism invariance of operators

All of the diffeomorphism invariant operators which have so far been constructed are also finite. We may ask whether finiteness is a general property of diffeomorphism invariant operators constructed nonperturbatively. There is a general argument that this is the case; which I would like to sketch here.

The argument begins with the assumption that any diffeomorphism invariant operator that will exist in a nonperturbative quantum theory must be constructed through

a regularization procedure. All such procedures which are known require that one introduce both a background metric and a regulator scale. This is necessary, because the scale that the regularization parameter refers to must be described in terms of some metric and, since none other is available, it must be described in terms of a background metric or coordinate chart introduced in the construction of the regulated operator. Because of this, the dependence of the regulated operator on the cutoff parameter is related to its dependence on the background metric. This can be formalized into a kind of renormalization group equation [29]. When one takes the limit of the regulator parameter going to zero one isolates the nonvanishing terms. If these have any dependence on the regulator parameter (which would be the case if the term is blowing up) then it must also have a dependence on the background metric. Conversely, if the terms that are nonvanishing in the limit the regulator is removed have no dependence on the background metric, *they must be finite.*

This point has profound implications for the whole discussion of finiteness and renormalizability of quantum gravity theories. It means that any nonperturbative and diffeomorphism invariant construction of the observables of the theory must be finite. A particular approach could fail in that there could be no way to construct the diffeomorphism invariant observables as quantum operators. But if it can be done, without breaking diffeomorphism invariance, those operators will be finite.

3.15. Exact physical states of the quantum gravitational field

We come finally to the physical state space, $\mathcal{V}_{phys}$, which consists of those states which are solutions to all of the constraints of quantum general relativity, including the Hamiltonian constraint. Although it is logical to put the discussion of these last, this is not the order in which the subject actually developed. The discovery that the Hamiltonian constraint could be exactly solved was made first[37]; later the loop representation was invented to solve the diffeomorphism constraint [7]. At the time the first solutions to the Hamiltonain constraint were found, it was possible to imagine that the existence of an infinite dimensional space of exact solutions to the Hamiltonian constraint might be some kind of spurious result, having nothing to do with physics. Now, six years later, after the development of the loop representation, and after we have understood how the discreteness of the quantum representation acts, at the kinematical and diffeomorphism invariant level, to allow the existence of finite, background independent operators and discrete structure at the Planck scale, and after we have further understood the role of this discreteness in assuring the existence of the classical limit, that the Hamiltonian constraint can be solved in this way seems much more natural.

In order to define its action in any representation, the Hamiltonian constraint must be regulated. In the literature there are four different proposals for how to carry out this regularization, due to Rovelli and the author[7], Gambini [23], Blencowe [38] and Bruegmann and Pullin [39]. These are now understood to be equivalent, at least when acting on a certain class of states [39].

The result of one of these regularization procedures is a sequence of well defined operators $\mathcal{C}^\delta$ in $\mathcal{V}_{kin}$, for $\delta > 0$. A solution to the quantum Hamiltonian constraint is then taken to be one such that,

$$\lim_{\delta \to 0} \left(\mathcal{C}^\delta | \Psi > \right) = 0 \tag{59}$$

The physical states are then taken to be those states that are simultaneous solutions to this condition and the diffeomorphism constraint (44).

I would like to make several comments about this condition.

a) There is no necessity that the limit $\lim_{\delta\to 0} \mathcal{C}^\delta$ define an operator in $\mathcal{V}_{kin}$. One could construct such an operator by a renormalization procedure in which the operator was multiplied by the appropriate power of δ as the limit is taken. But, it is not clear what use this would be. The operator, in any case, must vanish on the space $\mathcal{V}_{phys}$ which we are interested in, and on $\mathcal{V}_{kin}$, for the reasons we discussed in section 3.7, it will not be diffeomorphism invariant.

b) The Hamiltonian constraint, even before regularization, is not diffeomorphism invariant, as it is the integral of an arbitrary density with the local function of the fields. Thus, it does not define an operator in $\mathcal{V}_{diffeo}$. It would be very interesting to have an expression for the projection of the diffeomorphism constraint into that space. One could do this, for example, by finding an infinite set of functions on the classical phase space that vanish on the same surface as $\mathcal{C}$, but which are diffeomorphism invariant, and then represent those as quantum operators.

c) An issue that is often raised is the question of whether the algebra of the constraints, quantum mechanically, has an anomaly which would prevent the existence of simultaneous solutions to all of them. In fact, as I am about to describe, we know of infinitely many simultaneous solutions, so there can be no anomalous terms in the algebra which is proportional to the identity operator in the state space. Furthermore, as we have just remarked, the Hamiltonian constraint does not define a good operator on $\mathcal{V}_{kin}$ unless we further break diffeomorphism invariance by the addition of a renormalization procedure, so it is not clear exactly what condition to ask from our quantum operator algebra. However, it would still be interesting to know what the algebra of the regulated operators is like; this problem is presently under study[40].

d) In the condition (59), the limit is taken in the pointwise topology, which means that the limit must vanish when taken over every point of the loop space. As the physically meaningful inner product is constructed on the space of solutions to the constraint, this is sufficient as long as that solution space is large enough. However, it would be surprising if the limit could also not be expressed in terms of a Hilbert structure on $\mathcal{V}_{kin}$. I believe that this can be done, but the details have not been worked out.

The result of the regularization procedures is that the Hamiltonian constraint can be given a kind of geometrical interpretation when acting in the loop representation. First of all, acting on states that have support only on loops without intersections, the limit (59) vanishes[37, 7, 23, 39]. Thus, the action of the operator is, in the limit, only sensitive to the behavior of the state at intersecting loops. At an intersection, the action of the operator consists of two parts: first, a rearrangement of the routings through the intersection and second, a loop derivative taken at the intersecting points[7, 23, 39].

At the present time, there are several different sectors of solutions to the physical state space which have been explicitly constructed. First, as I have just mentioned, any state that has support on only nonintersecting loops is a solution. This is an infinite dimensional space; among these is a state corresponding to every invariant of nonintersecting links[7]. Then, there are two different sectors of states which have been constructed which have support on intersecting loops. The first consists of characteristic states, which have support on only a finite number of diffeomorphism equivalence classes

of intersecting loops. These have been constructed for intersections at which two [37], three [41], four and five [42] lines meet. Then, very recently, a new sector of physical states has been discovered by Bruegmann, Gambini and Pullin, which are closely related to the Jones polynomial[43].

At present, the study of exact physical states is ongoing, and there are a number of open problems. It is clear that the full set of solutions is not known, and nothing is known about the relationship between the different sectors of the solution space. Most importantly, it has not been established the extent to which the known types of solutions characterize the general solution to the constraints.

Of course, the construction of the theory is not complete without further elements, particularly the physical observables and the physical inner product. I will discuss the open problems in the next section. For the remainder of this section, I would like to discuss a number of results concerning the application of these nonperturbative methods to models which are simpler than full $3+1$ dimensional quantum gravity.

3.16. Application of the loop representation to 2+1 gravity

As Witten first pointed out, quantum general relativity in $2+1$ dimensions is exactly solvable because for each spatial topology one has, after the solutions of the constraints, a finite dimensional phase space [44]. Thus, although the model has only a finite number of degrees of freedom, it provides a good test of many of the ideas and methods that were originally developed in the $3+1$ dimensional case. The theory can be completely solved in using both the connection representation[44, 9] and the loop representation[17, 9, 18]. The physical operators can be constructed, and they turn out to be closely related to the loop operators (2) and (4). The physical inner product is also easily constructed.

In addition, as Carlip has shown in a very elegant series of papers[45], the difficult problem of time can be resolved completely in this model, along the lines proposed by Rovelli [46].

A case which lies intermediate in difficulty between this case and the full $3+1$ case is that of $2+1$ gravity coupled to matter. In particular, with the addition of one scalar field one has a model that can also be interpreted as $3+1$ gravity with one killing field[47].

Recently Ashtekar and Varadarajan have studied this model, and have found a number of interesting results at the classical level[48]. The most important of these is that it is possible to define a notion of asymptotic flatness, such that the energy is bounded both from above and from below. Using very different methods, Bonacina, Gamba and Martellini have shown that this theory is also perturbatively renormalizable[49]. In addition, there have been two interesting papers in which such systems are treated in the loop representation; which treat the coupling of Maxwell fields to $2+1$ gravity[50], and $3+1$ gravity with one killing field[51].

3.17. Application of the loop representation and new variables to two killing field reductions

Another very interesting model is general relativity with two killing fields, as this is known to be an integrable system with an infinite number of degrees of freedom. This system has been studied using the new variables and the loop representation, and a

number of interesting results were obtained[52]. The key open problem in this area is to represent the generators of the Geroch group as canonical transformations, generated by an infinite dimensional algebra physical observables of the model. Some very interesting partial results in this direction have been obtained by Torre[53].

3.18. Nonperturbative quantization of the Bianchi models

A last type of model I would like to mention is the class of Bianchi cosmologies. These are finite dimensional model cosmologies, which offer good laboratories for ideas about quantum gravity and quantum cosmology. Most of these have not been solved, in spite of the fact that they have only a few degrees of freedom; these models thus serve as a reminder that in quantum gravity, as in ordinary quantum mechanics, finite dimensional does not imply solvable.

It would be very interesting to be able to solve these models, through approximation methods if not exactly. There are a number of interesting results concerning them which employ Ashtekar's variables[54, 55, 56]. Among these is the construction of a set of exact states for the Bianchi IX model by Kodama [55] but, as in the full theory, little is known about the physical observables or the inner product in this model.

4. What are the key open questions?

In quantum gravity, it is safe to assume that any important problem is difficult, until the occasion of some progress provides evidence to the contrary. Indeed, in the development of the work I have been describing here, almost every result came out in a surprising way. Actually, once understood correctly, most of the results are not very difficult; the key seems to be to ask the question in precisely the right way.

Given this, the key open questions are simply how to construct those elements of a quantum field theory that are, so far, missing. I give here a list of them; more detailed discussions about each of them may be found in the reviews [9, 10, 11, 5].

4.1. How can we construct the physical observables?

As I mentioned above, the problem of physical observables in quantum gravity is difficult partly because there is already a problem in the classical theory. The problem can be stated this way: any classical observable in general relativity, with cosmological boundary conditions, must be a constant of motion. This is because to be invariant under diffeomorphisms it must commute with the Hamiltonian constraint, but in the cosmological case the Hamiltonian is proportional to a linear combination of constraints. This must be; were there a meaningful nonvanishing Hamiltonian it would be meaningful to ask how fast the universe is evolving, so that evolutions that differed only by the rate at which time progressed would be physically distinct. As there can be no clock outside the universe, this cannot be meaningful.

Thus, the problem of physical observables is closely connected with the notion of time. As such, it is one of those great problems that are both conceptually and mathematically profound .

At the present time, a rather large number of ideas are being studied with an aim towards solving this problem. I list here the ones I am aware of, with references.

i) Coupling the theory to matter, and using this matter to provide a system of clocks with which to make observables meaningful. This is a very old idea[33], recently it has received a lot of attention[34, 32, 36].

ii) Imposing asymptotically flat boundary conditions, which provide an observer and a classical clock at infinity. The problem with such an approach is that spatial infinity is, in a certain sense, too far away, and only a limit number of observables, corresponding essentially to the globally conserved quantities, may be defined there. Still, this is undoubtably worth doing, and some recent results of Baez are very interesting in this regard[57]. The key open problem with this approach is to show that, with respect to the correct inner product, the quantum Hamiltonian is bounded from below.

iii) Imposing some other kind of boundary conditions, which may allow more observables to be introduced. One such idea, due to Crane, is to define observables on two dimensional surfaces, and use conformal field theory thereby as a kind of measurement theory for quantum cosmology[58]. Another related idea involves choosing boundary conditions so that a Chern-Simon theory is induced on the boundary [11].

iv) Studying certain limits of the theory, where observables can be constructed [32, 59].

v) Finding an approximation scheme, such as a strong coupling expansion or a new form of perturbation theory, that will allow observables to be constructed systematically.

vi) Modifying the interpretative rules of quantum cosmology, so as to make the problem easier to solve[60].

vii) Construct observables that are associated with global properties of the configuration of the gravitational field. This has led, during the last year, to the construction of the only explicit examples yet discovered of observables of the pure gravitational field[61]

ix) Construct a superposition of exact states that corresponds, in the semiclassical sense described above, to Minkowski spacetime. Small perturbations on this state should then correspond to gravitons traveling on Minkowski spacetime, at least for long wavelengths. To show this one can construct a map from a sector of the Fock space of linearized quantum gravity into a subspace the space of exact physical states. This map then can be used to construct an approximate interpretation of the exact states in that subspace. There is some preliminary evidence that this map exists[66, 67].

4.2. How can we construct the physical inner product?

The last structure that is necessary to do physics is the inner product. This problem is closely connected to the problem of the observables because, in the absence of a global Poincare covariance, the inner product must be picked by the requirement that a complete set of real classical observables are represented by self-adjoint operators. Further, since the observables are constants of motion, the problem of determining the inner product is a dynamical problem. As such, this is a problem that will probably have to be solved by some approximation scheme, following such a solution to the problem of the observables.

4.3. Completeness of the physical state space

Where do the exact physical states that have been found fit into the whole space of solutions? This is a problem that is clearly dependent on the inner product and physical

observable algebra; what we need in the end is to show that the physical states carry a representation of the physical observable algebra.

4.4. Coupling matter to gravity

The Ashtekar formalism allows coupling to all types of matter, including spin zero and one-half matter, Yang-Mills fields[62], supergravity[63] and antisymmetric tensor gauge theories[36]. It is easy to extend the loop representation to describe coupling to these matter fields at the kinematical and diffeomorphism invariant level. Nothing is known about solutions to the Hamiltonian constraint including matter.

5. Conclusions

The results that I have been describing constitute a collective work in progress, which has been undertaken by a number of people who share a common interest in the questions I outlined in the introduction. As with any result of a scientific endeavor, from Stonehenge down through the Macintosh computer on which I'm writing this, these results reflect both the knowledge and the aspirations of those who made them. While such a work remains unfinished, it is difficult to judge its ultimate worth. We certainly don't yet know whether there is a consistent quantum field theory that would go by the name of general relativity, although the steady progress we have been making keeps us confident that it will be possible to cleanly resolve this question. However, this was, and is, not the only goal of this program; it was equally hoped that this work would uncover some general features that would hold for any quantum theory of gravity that could be constructed nonperturbatively. I believe that it is fair to say that a number of such features have emerged, and that as a result of this work we are wiser about how the world will look when we have a satisfactory quantum theory of gravity then we were before. I would like to close by listing several morals that I believe we have learned from the work I've described here.

a) To solve the spatial diffeomorphism constraints it is necessary to take a different starting point already at the level of the quantum kinematics than is taken in conventional Minkowski space quantum field theories. To avoid introducing background structures, Fock space must be replaced by representations of the kinematical observable algebra that rely on no background metric and carry unbroken representations of the diffeomorphism group. That is, to get the diffeomorphism invariant physics right, we must make sure that our state space and regularization procedures are background independent already at the kinematical level.

b) At present the only representations known to have these properties are the discrete representations I discussed here. Whether or not there are others is presently an open question, however even without resolving this, these new representations have interesting structures that deserve more investigation. In essence, what they seem to do is to resolve the paradoxes that follow from the uncountable nature of the classical continuum, as each state in these representation has support only on a countable set of loops. It is exactly this structure that makes it possible to solve directly the diffeomorphism constraints, in a way in which the resulting space of diffeomorphism invariant states has a countable basis.

The price we pay for this is that at the kinematical level the state spaces are nonseperable. This would be a serious problem at the level of the physical state space; however it is only a technical inconvenience in our case.

c) It is a further property of this discrete representation that it allows us to construct finite and regularization independent operators to represent non-local functions of the gravitational field. This results in the quantization of the spectra of areas and volumes. This is, moreover, not a spurious result of the kinematics, for we can show by direct construction that the quantization of areas and volumes is maintained at the diffeomorphism invariant level, when they are measured by diffeomorphism invariant operators.

We believe that these results will survive further translation to the physical level. If this is the case they will be the first physical predictions made by a quantum theory of gravity. That is, we propose that any fine enough measurement will reveal that the area of any surface can only lie in a discrete spectrum consisting of integral multiples of $l^2_{Planck}/2$, and certain other values associated with intersections that are described in [11].

d) The necessary dependence of renormalization procedures for local operators on background metrics is, we believe, a general phenomena. As a result, I conjecture that in any quantum theory of gravity there will be no renormalized local operators. Further, all diffeomorphism invariant operators will be finite after an appropriate *regularization* procedure[11, 29].

e) I believe that another thing we have seen in our construction of a diffeomorphism invariant operator in section 3.13 is quite general. This is that all diffeomorphism invariant operators, and hence all physical operators, will measure topological properties of non-local structures.

f) This means that in the final quantum theory of gravity we will see the continuous geometry of the classical theory emerge from a quantum theory of purely topological structures at the Planck scale. This is a consequence of what we discovered in section 3.10 and 3.11, in which we saw that when the classical limit was formulated carefully, it follows that every state that behaves semiclassically at large scales must be far from the semiclassical limit when probed on Planck scales. So far, in fact, that what is revealed is the discrete structure required by the quantization of the area operator.

g) We believe that it is this behavior, which is apparent already at the kinematical level, and not a pathology of the dynamics of general relativity, that is responsible for the failure of perturbative quantizations of general relativity. That is, the perturbation theory is already wrong at the kinematical level because it is unitarily inequivalent to the correct kinematical state space. Moreover, while at large distances the correct physics can be well approximated by semiclassical states, this approximation becomes worse and worse at shorter and shorter distances.

h) Finally, while the possibility of solving the diffeomorphism constraint exactly is implied only by the existence of the loop representation, which implies only that it is possible to choose a connection as the canonical coordinate of the theory, that it is in exactly the same representation that the Hamiltonian constraint becomes exactly solvable seems the main miracle uncovered so far. (Here, by a miracle I mean something wonderful that happens for a reason we don't understand.) It seems to be the case that once the diffeomorphisms have been taken care of correctly, the information remaining

in the Hamiltonian constraint is very manageable.

i) Although I do not claim to understand completely what is behind this miracle, it is worth pointing out that it, is in fact, exactly the existence of the self-dual connection that makes it possible to write the Hamiltonian constraint as a single term, which in turn makes possible the exact solutions which have been discovered. It seems, as a result, very possible that self-duality is one of the keys to quantum gravity in the real $3+1$ dimensional world. Indeed, self-duality is the key to several very interesting results that have been recently uncovered about the classical theory [64, 65].

Note that only the last two of the nine morals in this list depended on the form of the Hamiltonian constraint, and hence on the conjecture that general relativity is the correct microscopic description of gravitation. The rest depend only on the existence of the loop representation, which needs only that the theory can be expressed in such a way that a connection is the canonical coordinate. Thus, the fact that so many of the key features are present at the kinematic and diffeomorphism invariant levels, before the dynamics has been imposed, makes it, in my opinion, quite likely that whatever dynamics turns out to be right, the description of Planck scale physics in the final theory of quantum gravity will look a great deal like the picture I have been sketching here.

Acknowledgements

This contribution was intended as a report of work done by a number of people and I am grateful to all of them for many discussions about nonperturbative quantum gravity. I am particularly indebted to my collaborators in this work: Abhay Ashtekar, Ted Jacobson and Carlo Rovelli; in many places here I am expressing ideas that I heard first from one of them. Conversations with Louis Crane over the years of the development of this program have been important in sharpening my understanding of what we are doing as well as for killing any number of bad ideas. I would also like to thank John Baez, Julian Barbour, Berndt Bruegmann, Riccardo Capovilla, John Dell, Rudolfo Gambini, Joshua Goldberg, Gary Horowitz, Viqar Husain, Karel Kuchar, Chris Isham, Lou Kauffman, Roger Penrose, Jorge Pullin, Paul Renteln, Rafael Sorkin, Madhavan Varadarajan and Edward Witten for critical discussions about this work. This work was supported, in part, by the National Science Foundation under grants from the division of Gravitational Physics and a US-Italy cooperative research program grant.

References

[1] J.B. Barbour, *Absolute or Relative Motion, vol 1: The Discovery of Dynamics* (Cambridge University Press,1989)

[2] P. K. Feyerabend, *Against Method* (Verso,London,1975).

[3] L. Smolin *Did the Universe evolve?* Classical and Quantum Gravity 9 (1992) 173-191.

[4] K. V. Kuchar, *Time and interpretations of quantum gravity* in the *Proceedings of the 4th Canadian Conference on General Relativity and Relativistic Astrophysics*, eds. G. Kunstatter, D. Vincent and J. Williams (World Scientific, Singapore,1992).

[5] L. Smolin, *Time, structure and evolution* Syracuse Preprint (1992); to appear (in Italian translation) in the Procedings of a conference on *Time in Science and Philosophy* held at the Istituto Suor Orsola Benincasa, in Naples (1992) ed. E. Agazzi.

[6] A. V. Ashtekar, Physical Review Letters 57, 2244–2247 (1986) ; Phys. Rev. D 36 (1987) 1587.

[7] C. Rovelli and L. Smolin, *Knots and quantum gravity* Phys. Rev. Lett. **61**, 1155 (1988); *Loop representation for quantum General Relativity*, Nucl. Phys. B133 (1990) 80.

[8] R. Gambini and A. Trias, Phys. Rev. D23 (1981) 553, Lett. al Nuovo Cimento 38 (1983) 497; Phys. Rev. Lett. 53 (1984) 2359; Nucl. Phys. B278 (1986) 436; R. Gambini, L. Leal and A. Trias, Phys. Rev. D39 (1989) 3127.

[9] A. Ashtekar , *Non-perturbative canonical gravity.* Lecture notes prepared in collaboration with Ranjeet S. Tate. (World Scientific Books, Singapore,1991).

[10] C. Rovelli, Classical and Quantum Gravity, 8 (1991) 1613-1676.

[11] L. Smolin, *Recent developments in nonperturbative quantum gravity* in the Proceedings of the 1991 GIFT International Seminar on Theoretical Physics: "Quantum Gravity and Cosmology, held in Saint Feliu de Guixols, Catalonia, Spain (World Scientific, Singapore,in press).

[12] K. Kuchar, in this volume.

[13] A. Ashtekar, C. Rovelli and L. Smolin, *Self duality and quantization*, Syracuse preprint (1991) to appear in J. Geometry and Physics, Penrose Festschrift issue.

[14] A. Ashtekar, C. Rovelli and L. Smolin *Gravitons and Loops*, Phys. Rev. D 44 (1991) 1740-1755.

[15] A. Ashtekar, C. Rovelli and L. Smolin, Physical Review Letters 69 (1992) 237-240.

[16] A. Ashtekar and C. Rovelli, *A loop representation for the quantum Maxwell theory*,Class. and Quant. Grav. 9 (1992) 1121-1150.

[17] A. Ashtekar, V. Husain, C. Rovelli, J. Samuel and L. Smolin *2+1 quantum gravity as a toy model for the 3+1 theory* Class. and Quantum Grav. L185-L193 (1989)

[18] L. Smolin, *Loop representation for quantum gravity in 2+1 dimensions*,in the proceedings of the John's Hopkins Conference on *Knots, Tolopoly and Quantum Field Theory* ed. L. Lusanna (World Scientific,Singapore,1989).

[19] R. S. Tate, *Constrained systems and quantization, Lectures at the Advanced Institute for Gravitation Theory*, December 1991, Cochin University, Syracuse Univesity Preprint SU-GP-92/1-4; *An algebraic approach to the quantization of constrained systems: finite dimensional examples* Ph.D. Dissertation (1992) Syracuse University preprint SU-GP-92/8-1.

[20] P. A. M. Dirac, *Lectures on Quantum Mechanics* Belfer Graduate School of Science Monographs, no. 2 (Yeshiva University Press, New York,1964).

[21] A. Ashtekar and C. J. Isham, *Inequivalent observer algebras: A new ambiguity in field quantization* and *Representations of the holonomy algegbra of gravity and non-abelian gauge theories*, Syracuse University and Imperial College preprints (1991).

[22] D. Rayner, Class. and Quant. Grav. 7 (1990) 651.

[23] R. Gambini, *Loop space representation of quantum general relativity and the group of loops*, preprint University of Montevideo 1990, Physics Letters B 255 (1991) 180. R. Gambini and L. Leal, *Loop space coordinates, linear representations of the diffeomorphism group and knot invariants* preprint, University of Montevideo, 1991.

[24] R. Gianvittorio, R. Gambini and A. Trias, Phys. Rev. D38 (1988) 702; C. Rovelli and L. Smolin. *Loop representation for lattice gauge theory*,1990 Pittsburgh and Syracuse preprint.

[25] B. Bruegmann, Physical Review D 43 (1991) 566.

[26] J.M.A. Farrerons, *Loop calculus for $SU(3)$ on the lattice* Phd. thesis, Universitat Autonoma de Barcelona (1990);

[27] H. Fort and R. Gambini *Lattice QED with light fermions in the P representation* IFFI preprint, 90-08.

[28] R. Loll *A new quantum representation for canonical gravity and $SU(2)$ Yang-Mills theory*, University of Bonn preprint, BONN-HE-90-02 (1990).

[29] C. Rovelli, unpublished.

[30] The theory of how to classify knots is an intriguing and active branch of mathematics; for a good introduction see L. Kauffman, *Knots and Physics* (World Scientific, Singapore, 1991).

[31] V. Husain and K. Kuchar, Phys. Rev. D 42 (1990) 4070.

[32] C. Rovelli, *A generally covariant quantum field theory* Trento and Pittsburgh preprint (1992).

[33] B. S. DeWitt, in *Gravitation, An Introduction to Current Research* ed. L. Witten (Wiley, New York,1960).

[34] C. Rovelli, Class. and Quant. Grav. 8 (1991) 297,317.

[35] K. Kuchar and C. Torre, Utah preprint (1991).

[36] L. Smolin, *Diffeomorphism invariant observables in quantum gravity from a dynamical theory of surfaces* Syracuse University preprint (1992).

[37] T. Jacobson and L. Smolin, Nucl. Phys. B 299 (1988) 295.

[38] M. Blencowe, Nuclear Physics B 341 (1990) 213.

[39] B. Bruegmann and J. Pullin, *On the Constraints of Quantum Gravity in the Loop Representation*, Syracuse University preprint, (1991).

[40] R. Gambini and J. Pullin, personal communication.

[41] V. Husain, Nucl. Phys. B313 (1989) 711- 724.

[42] B. Bruegmann and J. Pullin, Nucl. Phys. B 363 (1991) 221.

[43] B. Bruegmann, R. Gambini and J. Pullin, Phys. Rev. Lett. 68 (1992) 431-434; *Knot invariants as nondegenerate staes of four dimensional quantum gravity* Syracuse University Preprint (1991), to appear in the proceedings of the XXth International Conference on Differential Geometric Methods in Physics, ed. by S. Catto and A. Rocha (World Scientific,Singapore,in press).

[44] E. Witten, Nucl. Physics B311 (1988) 46.

[45] S. Carlip, Phys. Rev. D 42 (1990) 2647; D 45 (1992) 3584; UC Davis preprint UCD-92-23.

[46] C. Rovelli, Phys. Rev. D 42 (1991) 2638; 43 (1991) 442; in *Conceptual Problems of Quantum Gravity* ed. A. Ashtekar and J. Stachel, (Birkhauser,Boston,1991); Class. Quantum Grav. 8 (1991) 317-331.

[47] A. Ashtekar, personal communication.

[48] A. Ashtekar and M. Varadarajan, Syracuse preprint in preparation (1992).

[49] G. Bonacina, A. Gamba and M. Martellini, Physical Review D (1992)

[50] C. Nayak, Gen. Rel. and Grav. 23 (1991) 981-990.

[51] V. Husain and J. Pullin, Modern Phys. Lett. A5 (1990) 733-741.

[52] V. Husain and L. Smolin, Nucl. Phys. B 327 (1989) 205.

[53] C. Torre, Syracuse preprint (1991).

[54] H. Kodama, Prog. Theor. Phys. 80 (1988) 1024.

[55] H. Kodama, Phys. Rev. D 42 (1990) 2548-2565.

[56] A. Ashtekar and J. Pullin, Proc. Phys. Soc. Israel 9 (1990) 65-76.

[57] J. Baez, Wellsely preprints (1992).

[58] L. Crane, Kansas preprint (1992).

[59] L. Smolin, *The $G_{Newton} \to 0$ limit of Euclidean quantum gravity*, Class. and quant. grav. (1992), in press.

[60] J. B. Barbour,to appear in the *Procedings of the NATO Meeting on the Physical Origins of Time Asymmetry* eds. J. J. Halliwell, J. Perez-Mercader and W. H. Zurek (CAmbridge University Press, Cambridge, 1992); *On the origin of structure in the universe* in *Proc. of the Third Workshop on Physical and Philosophical Aspects of our Understanding of Space and Time* ed. I. O. Stamatescu (Klett Cotta); *Time and the interpretation of quantum gravity* Syracuse University Preprint, April 1992.

[61] J. Goldberg, P. Lewandowski and C. Stornaiolo, Commun. Math. Phys. 148 (1992) 377; T. Jacobson and J. Romano, U. or Maryland preprint, (1992).

[62] T. Jacobson, Class. and Quant. Grav. 5 (1988) L143-L148; A. Ashtekar, J. Romano and R. S. Tate, Physical Review D40 (1989) 2572-2587.

[63] T. Jacobson, Class and Quant. Grav. 5 (1988) 923-935.

[64] R. Capovilla, J. Dell and T. Jacoboson, Phys. Rev. Lett. 63 (1989) 2325; Class. and Quant. Grav. 8 (1991) 59; R. Capovilla, J. Dell, T. Jacoboson and L. Mason, Class. and Quant. Grav. 8 (1991) 41.

[65] E. T. Newman and C. Rovelli, Phys. Rev. Lett.69 (1992) 1300-1303

[66] J. Zegwaard, *Gravitons in Loop Quantum Gravity* Utrecht preprint, THU-91/13.

[67] J. Iwasaki and C. Rovelli, *Gravitons as embroidery on the weave*, Pittsburgh and Trento preprint (1992).

Paper presented at the 13th Int. Conf. on General Relativity and Gravitation
Cordoba, Argentina, 1992: Part 1, Plenary Lectures

Recent progress in the observations of compact objects and black holes

Luigi Stella[1]

Osservatorio Astronomico di Brera, Via Brera 28, 20121 Milano, Italy

Abstract. This paper reviews some of the recent observational results on compact objects and stellar mass black holes. Over the last decade, much of the progress in this field has been achieved through the study of transient X-ray binary sources, which shine only sporadically in the X-ray sky and undergo much larger luminosity variations than most persistent sources. The impact of the observations of the transient X-ray pulsar EXO 2030+375 on the study of accretion processes onto magnetic neutron stars is presented as one of the most significant examples. There are currently six stellar-mass black hole candidates for which the mass of the compact object is estimated to be well above the maximum neutron star mass. The three most recently identified black hole candidates of this kind belong to a class of transient X-ray sources with peculiar spectral properties. A number of transients of the same class which are currently being studied are likely to contain other black hole candidates. The increasing number of candidates and the large luminosity (and thus accretion rate) variations in transient sources will allow to study the chacteristics of accreting black holes in an unprecedented detail. The potential of iron K-shell lines around 6-7 keV as a new diagnostic for studying the innermost regions of accretion flows towards collapsed objects is also outlined and prospects for the future are discussed.

1. Introduction

A large number of collapsed objects has been discovered and made accessible to detailed study through observations in the X-ray band. Accretion of matter in the strong gravitational field of these objects is in most cases responsible for the efficient production of radiative energy. Up to $\sim 42\%$ of the rest-mass of the accreting flow can, in principle, be converted into radiation. In the case of accreting collapsed objects with stellar mass ($M \leq 20$ $M_\odot$) the bulk of the energy is radiated away in the X-ray band.

X-ray binaries consist of a neutron star or a stellar mass black hole accreting matter from a non-collapsed companion in a close binary orbit. Historically, these systems

[1] Affiliated to I.C.R.A.

have played a crucial role in the identification of the first black hole candidate and the measurement of basic physical parameters of compact objects (such as the mass and the magnetic field) (see e.g. Giacconi 1978). There is a growing body of evidence that the central region of active galactic nuclei (AGNs) hosts an accreting black hole with a mass of $\sim 10^5 - 10^8$ $M_\odot$. The observation and measurement of strong gravitational field effects are, in principle, possible in all these classes of sources, since most of the radiation is produced in the vicinity of the collapsed object. However, a number of largely unknown effects related e.g. to the dynamics, MHD and radiative-transfer of accretion complicates the interpretation of the observations and only relatively little information on the strong field regions has been obtained so far. Double neutron star binary systems containing a radio pulsar have provided so far a much *cleaner* laboratory for testing the predictions of gravitational theories. From the study of the propagation delays of the pulsar signal in these systems it has been possible to accurately test general relativity, but only in relatively weak fields ($R \leq 10^{-6} GM/c^2$) (Taylor 1993).

This review concetrates on recent insights regarding collapsed objects, obtained from the study of X-ray binaries. The progress in the understanding of accreting magnetised neutron stars achieved through the observations of the transient X-ray pulsar EXO 2030+375 is summarised. This example is to emphasise the role of X-ray transient sources in investigating accretion into collapsed objects over a wide range of source luminosities and, thus, mass accretion rates. A subclass of X-ray transient sources, in particular, seems to be associated with accreting black hole candidates. The potential of this class of sources for the identification of new candidates and for the study of accretion phenomena close to a black hole is outlined. The iron K-shell spectral lines around 6-7 keV, which are observed in virtually all classes of accreting collapsed objects, are likely to provide the first powerful probe of accretion flows at distances as short as tens of Schwarzschild radii.

2. X-ray binary basics

Luminous X-ray binaries ($L_x > 10^{35}$erg/s) are often classified as low mass and high mass systems depending on the mass of the donor star. While this classification leaves unspecified the nature of the accreting collapsed object (which indeed can be a neutron star or a black hole in either class), it allows to distinguish the phenomenology of the X-ray sources and their optical counterparts in a natural way (see Table 1).

2.1. High mass X-ray binaries

High mass X-ray binaries (HMXRBs) contain an early type (OB) star with a mass of > 5 $M_\odot$ and have a galactic disk distribution characteristic of young stars (population I). Mass transfer in most of these systems takes place because part of the intense stellar wind emitted by the OB star is captured by the gravitational field of the collapsed object. The energy production budget in HMXRBs is often dominated by the optical luminosity of the OB star, with the X-ray flux emitted in the vicinity of the collapsed object providing only a small perturbation. Correspondingly the optical spectra are stellar-like (Rappaport and Joss 1983, and references therein)

Periodic X-ray pulsations with periods ranging from ~ 0.1 to ~ 1000 s are present in a large number (~ 30) of HMXRBs. This signal originates from the beamed radiation

which is produced close to the magnetic poles of a young accreting neutron star with a surface field of $\sim 10^{12}$ Gauss. Due to the misalignement of the magnetic and rotational axes, the neutron star rotation modulates the X-ray intensity in a light-house fashion. Period (or phase) changes introduced by the binary motion allow to measure some of the orbital parameters of these systems. Together with the duration of the X-ray eclipse (which is present in several HMXRBs) and the Doppler velocity and photometric modulations of the optical star, these measurements provide the absolute orbital solution and the masses of the two components. Secular spin period changes arise because of the torque exerted on the neutron star magnetosphere by the accreting matter (Henrichs 1983 and references therein). X-ray pulsations from luminous X-ray binaries provide an incontrovertible signature of accretion onto a magnetised neutron star.

2.2. Low mass X-ray binaries

Low mass X-ray binaries (LMXRBs) typically contain a late type (K, M) low mass donor star (or, in the case of very short period binaries, a white dwarf, WD). About 10 LMXRBs are close to the core of globular clusters and many of the others are concentrated in the vicinity of the galactic bulge, therefore indicating a distribution characteristic of old stars (population II) (Lewin and Joss 1983 and references therein). The short orbital periods of LMXRBs are usually inferred from the orbital modulation of the optical and/or X-ray flux, rather than X-ray eclipses proper (which are rare) (Parmar and White 1988). In most cases mass transfer takes place because the low mass companion overfills a critical effective potential surface (the Roche lobe) and spills matter with high specific angular momentum towards the collapsed star, causing the formation of an accretion disk. The intrinsic optical luminosity of the low mass companion is orders of magnitude lower than the X-ray luminosity emitted by the accreting collapsed object. The spectral features of the late type companion are usually outshined by the reprocessing at optical wavelengths of the X-ray flux intercepted by the accretion disk and the star.

X-ray pulsations have been found only in a very small number of LMXRBs. Much more frequent, instead, is the phenomenon of (type I) X-ray bursts, sudden rises of X-ray luminosity which typically last for tens of seconds, show a characteristic cooling in the decay phase and recur on timescales of hours. These bursts account for only a small fraction of the time-averaged luminosity of LMXRBs. They originate from thermonuclear flashes in the freshly accreted matter on the surface of a neutron star. Therefore type I X-ray bursts provide another clear signature of accretion onto a neutron star.

In a few LMXRBs the bursting activity ceases when the persistent emission X-ray luminosity increases above a level of $\sim 10^{37}$erg/s (van Paradijs and Lewin 1988 and references therein). The X-ray properties of this state resemble those of a number of high-luminosity LMXRBs ($L_x \geq 10^{37}$erg/s), such as Sco X-1, Cyg X-2 and many others which populate the galactic bulge. It is thus inferred that also the latter systems contain accreting neutron stars.

2.3. X-ray pulsations versus type I X-ray bursts

X-ray pulsations and bursts are mutually exclusive properties of X-ray binaries: type I bursts have never been observed from an X-ray pulsar and no coherent pulsations have ever been detected from a type I burst source.

Table 1. Classification of X-ray binaries

Properties	HMXRBs	LMXRBs
donor star	O-B ($M > 5$ M$_\odot$)	K-M or WD ($M < 1$ M$_\odot$)
population	I	II
L_x/L_{opt}	0.001 – 10	100 – 1000
optical spectrum	stellar-like	reprocessing, no absorpt. lines
X-ray spectrum	usually hard	usually soft
orbital period	1 – 100 d	10 min – 10 d
X-ray eclipses	common	rare
X-ray pulsations	common ($P_s \simeq 0.1 - 1000$ s)	rare ($P_s \simeq 1 - 100$ s)
type I X-ray bursts	absent	common
X-ray QPOs	rare ($\nu_{QPO} \simeq 0.001 - 1$ Hz)	common ($\nu_{QPO} \simeq 1 - 100$ Hz)
collapsed object	young neutron star or black hole	neutron star or black hole

Pulsations are not expected from the old neutron stars in bursting sources if their magnetic field has decayed to $\sim 10^8$ Gauss, a value below which accretion is not significantly funneled close to the magnetic poles. On the other hand, LMXRBs are likely to be progenitors of the old, *recycled* millisecond pulsars which are found in increasing numbers especially in globular clusters (van den Heuvel 1991). In this case the neutron star magnetic field of $\sim 10^9 - 10^{10}$ Gauss inferred from the radio pulsar observations should also characterise the LMXRB stage, and low amplitude X-ray pulsations in the millisecond range would be expected if accretion occurs preferentially along the magnetic field lines. The search for fast pulsations in LMXRBs still continues.

Viceversa, type I bursts might not occur in X-ray pulsar binaries because the strong magnetic field ($\sim 10^{12}$ Gauss) of young neutron stars confines the infalling plasma to the polar caps, thereby increasing dramatically the accretion rate per unit area (compared to weakly magnetic neutron stars) and giving rise to steady (as opposed to flash-like) thermonuclear burning in the accreting material (Fujimoto *et al* 1981, Hanawa and Fuijmoto 1984).

While pulsations and type I X-ray bursts are clearly explained and imply the presence of an accreting neutron star, a variety of other phenomena characteristic of X-ray binaries still awaits for unambiguous interpretation. These include aperiodic variability (shot, red, very-low frequency, low-frequency, high-frequency noises), quasi-periodic oscillations (QPOs), some spectral and/or activity states, and continuum and discrete spectral components. At a phenomenological level, a number of regularities and correlations have emerged, which provide the basis for further classification and study. Some of these are discussed in some detail in the following sections.

3. Persistent versus transient X-ray binaries

Besides persistent sources the X-ray sky is populated by a number of transient sources which remain in their quiescent state for most of the time and sporadically undergo bright outbursts with peak luminosities of $10^{36} - 10^{38}$erg/s , durations ranging from weeks to months, and recurrence timescales of 1-10 years or longer. Throughout the years a very clear analogy of the X-ray characteristics of bright transient sources with those of persistent sources has emerged. In particular a number of X-ray transients display coherent X-ray pulsations or bursts, testifying to the presence of neutron stars, which undergo sporadic surges of accretion. Like in the case of persistent sources, X-ray bursting transients have low mass companions and relatively soft X-ray spectra, whereas X-ray pulsations are usually observed from transients in Be-star high mass binaries which are characterised by hard X-ray spectra extending up to a several tens of keV (White *et al* 1984).

The identification of the optical counterparts of bright transients, in crowded and often heavily absorbed regions of the galactic plane, is made easier by the optical flux increase associated with the outburst. In the case of low mass systems, in particular, the reprocessing of high energy radiation can induce an increase of more than a factor of 100 above the quiescent optical flux level (Lewin and Joss 1983). Contrary to the case of persistent low mass X-ray binaries, detailed photometric and spectroscopic studies of the companion star are often possible in low mass transient sources, due to the fact that in the quiescent state its optical spectrum is not dominated by the reprocessing of X-ray radiation or by the emission from the accretion disk around the collapsed object.

X-ray transients sources are also extremely useful in that they allow to investigate accretion onto collapsed stars over a much larger range of X-ray luminosities, and thus accretion rates, than persistent sources. To illustrate the importance of this point, we summarise in the next section the progress achieved in the understanding of the physics of accreting magnetic neutron star through the observation of an individual X-ray pulsar transient, EXO 2030+375.

4. Accretion onto magnetic neutron stars

The accretion flow towards magnetic neutron stars is characterised by a region inside which the dynamics of the infalling plasma is dominated by the magnetic field. Inside this *magnetosphere*, the majority of the material is channeled along the field lines and reaches the neutron star surface in the vicinity of the magnetic poles, therefore generating a pulsed signal at the spin period. Once the pulse arrival time delays introduced by the orbital motion are removed, secular changes in the spin period can be measured which provide information on the torques applied by the accreting matter. A secular spin up is observed in a number of X-ray pulsars, which implies the presence of an accretion disk outside the magnetosphere, transferring mass and angular momentum to the neutron star. Although this basic picture has been known and progressively confirmed by the observations of $\sim$ 30 X-ray pulsars over the last twenty years, a number of crucial predictions of the theory could not be addressed for a long time.

4.1. The X-ray pulsar transient EXO 2030+375

The study of the 42 s X-ray pulsar transient EXO 2030+375 has improved substantially our understanding of accretion onto magnetic neutron stars. The source was discovered and monitored for about six months in 1985 (Parmar *et al* 1988a). During that period the source underwent two outbursts and displayed X-ray luminosity variations of more than a factor of 1000. A heavily reddened high mass star was identified as the optical counterpart (Janot-Pacheco *et al* 1988). Delays in the arrival times of X-ray pulses were used to derive an orbital period of $P_{orb} \sim 45$ d, an eccentricity of $e \sim 0.4$, a projected semimajor axis of the neutron star orbit of $a_x \sin i \sim 260$ lt-s and an X-ray mass function of $f_x(M) = (M_{opt} \sin i)^3/(M_x + M_{opt})^2 \sim 5 - 10 M_\odot$ (here i is the inclination of the binary, M_x and M_{opt} the mass of the X-ray emitting compact object and the mass of the companion star, respectively). These values are similar to those measured in various other HMXRBs containing a high mass Be star and a transient X-ray pulsar.

4.2. Spin-up rate and disk-magnetosphere interaction

Two basic scenarios have been developed to describe the interaction of the accretion disk with the neutron star magnetosphere. In one model the disk is diamagnetic and developes surface currents which prevent the magnetic field lines from penetrating the disk (Aly 1980). In the Ghosh and Lamb (1979) model, instead, the magnetic field lines diffuse through the disk, giving rise to angular momentum transfer also by magnetic stresses. The predictions of both theories are in agreement when the mass accretion rate is high and the disk extends deeper in the neutron star magnetosphere. In this case the Keplerian frequency of the disk at the radius of the magnetosphere, $\nu_K(r_m)$, is substantially higher than the magnetospheric spin frequency, ν_s, and the angular momentum transfer towards the neutron star is dominated by the angular momentum carried by the accreting matter as it leaves the disk at r_m. The pulsar spin up rate in this case is expected to be $-\dot{P}_s \propto B^{2/7} P_s^2 L_x^{6/7}$, where P_s is the spin period, B the surface magnetic dipole field and L_x the X-ray luminosity of the neutron star. For lower accretion rates, the magnetosphere extends to larger radii and approaches the *fast rotator* regime ($\nu_K(r_m) \simeq \nu_s$) for which the consequences of magnetic field lines threading the disk become important and the predictions of the theory ambiguous. Before the discovery of EXO 2030+375, measurements of $\dot{P}_s$ were obtained over a limited range of luminosities for each of about 10 X-ray pulsars. Uncertainties related to the different distance, emission geometry and magnetic field strength of different sources allowed to confirm only tentatively the predictions of the theory (Henrichs 1983).

In the case of EXO 2030+375 the spin-up rate could be measured over a factor of ~ 100 variation in luminosity (from $\sim 10^{38}$ to $\sim 10^{36}$ erg/s). This provided a power law dependence of $-\dot{P}_s \propto L_x^{1.1-1.3}$ away from the *fast rotator regime*, which, in consideration of the uncertainties due to the varying emission geometry (see section 4.3), strongly supports the predictions of the theory in this regime ($-\dot{P}_s \propto L_x^{6/7}$). As the luminosity decayed further at the end of the outburst, EXO 2030+375 approached the *fast rotator* regime. An accurate measurement of the period change in this case was prevented by uncertainties in the orbital solution. Yet the application of the Ghosh and Lamb model allowed to derive approximate values of the surface magnetic dipole field of $B \sim 2 \times 10^{13}$Gauss, and the distance, $d \sim 5$ kpc. These results provide the best confirmation to date of accretion torque models for magnetic neutron stars (Parmar *et al* 1988a).

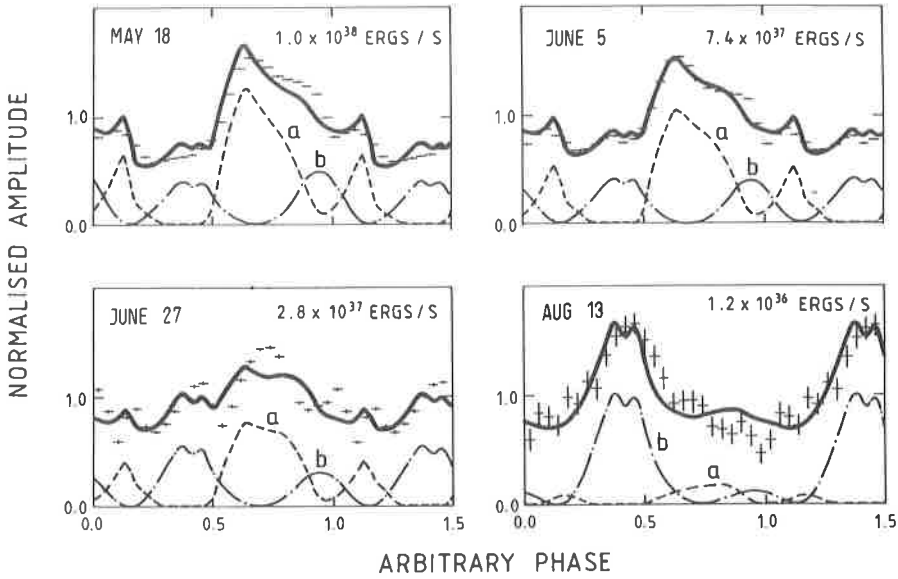

Figure 1. Evolution of the 1-10 keV pulse profile of EXO 2030+375 during the decay of the first 1985 outburst. The observed profiles are shown by the crosses; the solid lines represent the best-fit model profiles. These consist of the sum of fan and pencil-beam emissions from each of the two poles, plus unpulsed emission. The contributions from the fan and pencil beams (summed over both poles) are labeled as (a) and (b) (Parmar *et al* 1989b).

4.3. Pulse profiles and emission beaming from the neutron star magnetic poles

The pulse profiles of X-ray pulsars show great variety from source to source, ranging from sinusoidal-like profiles, to highly structured and energy-dependent modulations (Rappaport and Joss 1983, White *et al* 1983). The shape of the periodic signal contains information on the emission geometry from the regions close to the neutron star magnetic poles where accretion is concentrated. Detailed modelling of the emission pattern emerging from the accreting polar cap(s) of a neutron star has shown that beamed emission and complex pulse profiles can be produced primarily because of the effects of the magnetic field and the interaction of the radiation with the infalling matter. The preferred beaming direction depends on whether a stand-off shock and, therefore, a dense deceleration region are present above the polar cap. For high luminosities ($> 10^{37}$erg/s) a radiative shock is expected to form. In this case photons will escape preferentially from the sides of the high density post-shock accretion column, giving rise to a fan-beam. For lower luminosities ($< 10^{37}$erg/s) the infalling material might be decelerated in a collisionless shock above the polar cap possibly arising from plasma instabilities), or by Coulomb and nuclear collisions at the neutron star surface. If the latter occurs, the emission region will be located in a thin layer on the neutron star surface and radiative transfer effects in the strong magnetic field will favor photons escaping in the direction of the field lines, therefore giving rise to a pencil-beam. If a collisionless shock occurs a fan-beam is likely to form, as in the high-luminosity case (Basko and Sunyaev 1976, Meszaros 1984).

Comparative studies of the pulse profiles of the ~ 30 known X-ray pulsars, provided only marginal evidence in favor of fan-beams, producing the complicated pulse profiles in several high luminosity X-ray pulsars, and pencil beams, giving rise to the quasi-sinusoidal modulations of low luminosity X-ray pulsars (White *et al* 1983). However due to the inhomogeneities of the sample (different geometry, inclination, magnetic field strength etc.) the conclusion above was only tentative. EXO 2030+375 offered the first opportunity of observing and modelling the pulse profile of an individual X-ray pulsar over a factor of ~ 100 variation in luminosity. The points in Fig. 1 show the 1-10 keV pulse profiles of EXO 2030+375 for luminosities ranging from $\sim 10^{38}$ to $\sim 10^{36}$ erg/s. It is apparent that drastic changes in the shape of the pulses took place as the luminosity decreased. In particular the broad peak at a pulse phase of 0.6-0.9 became gradually less pronounced as the luminosity decreased and virtually disappeared for luminosities of $\sim 10^{36}$ erg/s. On the contrary, the importance of the feature centered around a phase of ~ 0.4 increased as the luminosity of EXO 2030+375 decreased during the outburst and became dominant around $L_x \sim 10^{36}$ erg/s (Parmar *et al* 1989b).

The solid lines in Fig. 1 show the results obtained from a semi-empirical modelling of the pulse profiles, consisting of the emission from two polar caps above the neutron star surface. The emission pattern from each of the poles was approximated by the sum of a fan-beam and a pencil beam. The luminosity of each beam was allowed to vary independently, as well as the longitude, latitude and luminosity of each magnetic pole. The effects of light bending in the strong field of the neutron star were approximately taken into account. It is apparent that the model can reproduce the basic characteristics, though not all the details, of the pulse profiles. The contribution due to fan beam emission from both poles is shown as a dashed curve at the bottom of each panel in Fig. 1, whereas the contribution due to pencil beam emission is shown by the dot-dashed curve. The modelling clearly indicates that, as the source luminosity decreased, the dominant emission pattern changed from a fan-beam to a pencil-beam, while the geometrical parameters remained roughly unchanged (as expected) (Parmar *et al* 1989b). These results provide the best evidence to date that for luminosities $\leq 10^{37}$ erg/s the material accreting on the neutron star polar caps is stopped at the neutron star surface, rather than being decelerated in a post-shock region.

4.4. QPOs and magnetospheric models

Periodic X-ray pulsations in the millisecond range are expected from LMXRBs if these systems are the sites where old neutron stars with magnetic fields of $\sim 10^9 - 10^{10}$ Gauss are spun up by accretion torques to very short rotation periods, therefore providing the progenitors of *recycled* radio pulsars. During the search for these fast pulsations (which have not been revealed yet), a different phenomenon was discovered in a number of bright LMXRBs: namely oscillations with frequencies of $1 - 50$ Hz and very poor coherence. These quasi periodic oscillations (QPOs) display large frequency variations on timescales as short as tens of seconds and, therefore, cannot represent the rotation of the neutron star. Their phenomenology is very complex, and different QPO modes have been identified which correspond to different spectral and activity states (Lewin *et al* 1988, Stella 1988). A number of sources show a mode in which the QPO frequency, ν_{QPO}, increases with the source luminosity. This property played a key role in the development of QPO models: the $\nu_{QPO} - L_x$ correlation is suggestive of the presence of a neutron star magnetosphere which is *compressed* for increasing accretion rates.

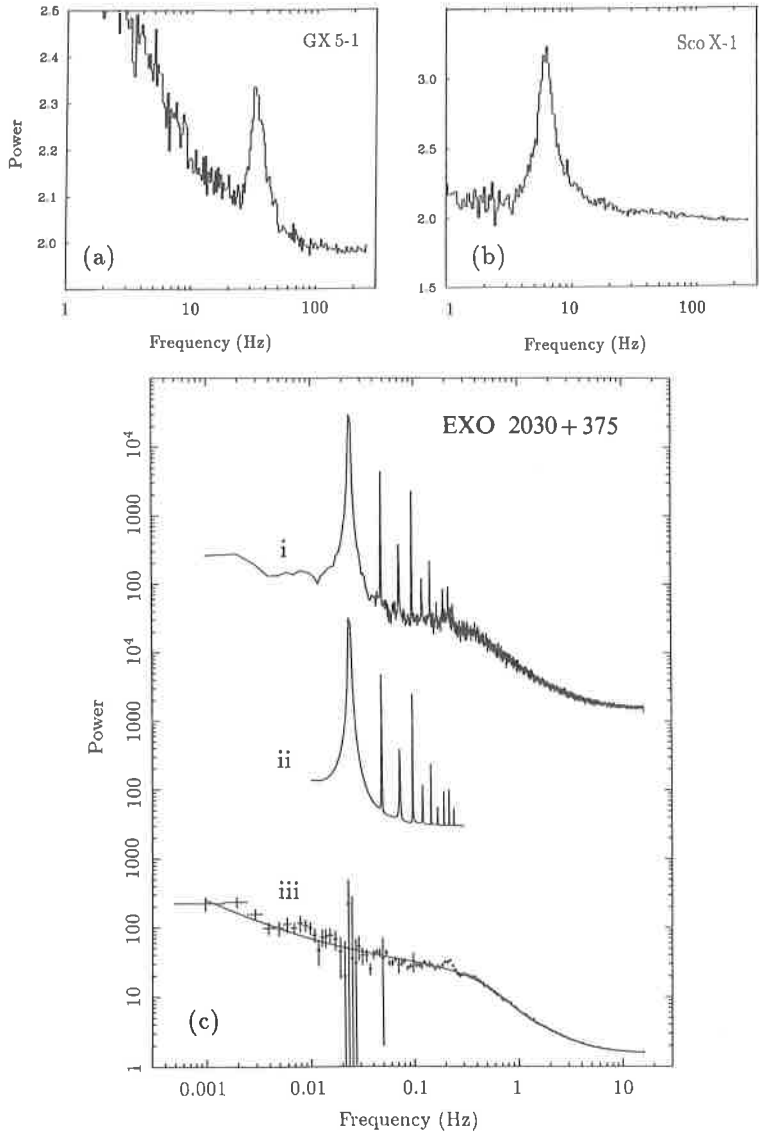

Figure 2. (a)(b) Power spectra of the intensity variations of two high luminosity LMXRBs, GX 5-1 and Sco X-1. The broad peaks around 30 and 6 Hz, respectively, testify to the presence of QPOs (van der Klis 1989). (c) (i) The power spectrum of the transient X-ray pulsar HMXRB EXO 2030+375 close to the maximum of the 1985 outburst; the very high peaks are due to the 42 s coherent pulsations. (ii) Best fit for a model including the first 10 harmonics of the coherent pulsar signal. (iii) The residual power spectrum after subtraction of the model for the coherent pulsations. The solid line shows the best fit for the extended continuum power spectrum components. The $\sim$ 20% broad peak at about 0.2 Hz reveals the presence of the QPOs (Angelini *et al* 1989).

In the best developed model for this kind of QPOs, the beat frequency model (BFM), the interaction between the disk and the magnetospshere causes the accretion flow to be modulated at the *beat frequency* between the disk Keplerian frequency at the magnetospheric boundary, $\nu_K(r_m)$, and the neutron star spin frequency, ν_s. In this case the QPO frequency is approximately given by

$$\nu_{QPO} \simeq \nu_K(r_m) - \nu_s \simeq B_{12}^{-6/7} L_{37}^{3/7} - \nu_s \qquad \text{Hz} ,$$

where B_{12} is the magnetic dipole field at the neutron star surface in units of 10^{12} Gauss, L_{37} the X-ray luminosity in units of 10^{37} erg/s. The equation above assumes also that the rest energy of the accreting matter is converted into X-rays with a *constant* efficiency of $\sim 10\%$. When used to fit the $\nu_{QPO} - L_x$ relation observed in bright LMXRBs like GX 5-1, Sco X-1 and Cyg X-2, this model predicts a neutron star spin frequency of $\nu_s \sim 100$ Hz, and a surface magnetic field of $\sim 10^{10}$ Gauss, in agreement with the idea that these systems contain a weakly magnetic neutron star which has been spun up by accretion (Lamb *et al* 1985). Periodic pulsations at the neutron star spin frequency are expected in the BFM, although their amplitude might be drastically reduced by the effects of electron scattering.

In the absence of measurements of the magnetic field strength and the neutron star spin frequency, it is difficult to use LMXRBs to verify the validity of the BFM. An alternative approach is to detect QPOs from an accreting X-ray pulsar for which an estimate of the magnetic field strength is available, either from the detection of cyclotron spectral features, or by the measurement of spin-period changes induced by accretion torques. This, together with the knowledge of the magnetospheric rotation period derived from the pulsed X-ray signal, removes the two largest uncertainties when investigating QPOs in LMXRBs.

QPOs are in most cases revealed and studied through the broad peak that they produce in the power spectrum of the intensity variations of the source (see Fig. 2a). In general the power spectrum of an X-ray pulsar contains the narrow peaks from the fundamental of the periodic signal from the neutron star rotation and some of its higher harmonics. These peaks must be removed in order to study the continuum power spectrum components, including the broad QPO peaks, which originate from noise-like variations of the source. The power spectra from EXO 2030+375 were investigated in this way. After the narrow peaks from the periodic modulation were modelled and subtracted, a broad peak was revealed around frequencies of ~ 0.2 Hz, which testifies to the presence of QPOs (see Fig. 2b) (Angelini *et al* 1990). By using the measured values of the spin frequency, the X-ray luminosity and the magnetic dipole field of the neutron star in EXO 2030+375 a predicted QPO frequency of $\nu_{QPO} \sim 0.2$ Hz is obtained from the BFM. This is in remarkable agreement with the observed value, especially in view of the fact that no free parameter has been adjusted to the data. Moreover, contrary to other models in which the QPO signal is generated directly by matter orbiting close to the magnetospheric boundary (for which $\nu_{QPO} \sim \nu_K(r_m)$), the BFM model can easily account for the relatively large amplitude of the QPO in EXO 2030+375 ($\sim 3.5\%$ *rms*).

These results provided the first quantitative confirmation of the BFM and showed that the model can work in the presence of an accretion disk interacting with a rotating neutron star magnetosphere (Angelini *et al* 1990). Whether QPO sources in nonpulsating LMXRBs possess a small rapidly rotating neutron star magnetosphere which can generate QPOs by the same physical mechanism remains an open question.

5. Black hole candidates in X-ray binaries

Despite the increasing evidence in favor of very massive black holes ($10^6 - 10^9$ $M_\odot$) in AGNs, black hole candidates in X-ray binaries provide still the strongest case for the existence of completely collapsed objects, characterised by an event horizon.

5.1. The mass criterion

The chain of arguments that leads to the identification of a black hole candidate in an X-ray binary through the *mass criterion* can be summarised as follows (Bahcall 1978, McClintock 1986):

(i) The luminosity of the X-ray source is high ($> 10^{36}$ erg/s) and characterised by fast variability ($\tau < 1$ s): therefore the binary must contain an accreting compact object.

(ii) Optical spectro-photometric observations allow to determine the orbital period and the Doppler velocity modulation of the donor star around the collapsed object. The *optical mass function* of the system, $f_{opt}(M) = (M_x \sin i)^3/(M_x + M_{opt})^2 = P_{orb}K^3/2\pi G$, is thus measured, where $K = v \sin i(1 - e^2)^{1/2}$ and $v \sin i$ is the semi-amplitude of the Doppler modulation.

(iii) The mass of the optical star and the inclination are inferred, or at least constrained, based on the distance, the optical spectrum and luminosity of the optical star, its size relative to the Roche-lobe and arguments related to the lack of X-ray eclipses and the presence of optical polarisation. By using $f_{opt}(M)$, the mass of the compact object, M_x, is thus measured or constrained.

(iv) If the mass of the compact object is determined to be larger than the canonical maximum mass of a neutron star predicted by the theory, $M_x > M_{max} \simeq 3.2\ M_\odot$ (Rhoades and Ruffini 1974, Hartle 1978), then the conclusion is reached that the accreting compact object is most likely completely collapsed and is therefore a *black hole candidate.*

The identification of the first black hole candidate in the early seventies, represents one of the most important achievements of X-ray astronomy.

5.2. Cyg X-1

Cyg X-1 is a highly variable and luminous HMXRB ($L_x > 10^{37}$erg/s), with an orbital period of 5.6 days an optical mass function of $f_{opt}(M) \simeq 0.25$ $M_\odot$. The most likely values of the companion mass, $M_{opt} \simeq 30$ $M_\odot$, and the inclination, $i \simeq 30°$, provide an estimate of the mass of the compact object of $M_x \sim 16$ $M_\odot$ (Liang and Nolan 1984).

It is important to realise that the value of the optical mass function provides an absolute *lower limit* on the mass of the compact object, corresponding to the unrealistic situation in which $M_{opt} = 0$ and $i = 90°$. In the case of Cyg X-1, however, $f_{opt}(M)$ is only ~ 0.25 $M_\odot$ and deriving a reliable lower limit on M_x is a difficult task. In a number of studies a devil's advocate approach was adopted in order to decrease the estimated value of M_x below 3.2 $M_\odot$. This proved very useful in verifying the validity of the assumptions underlying the mass estimate. Indeed the velocity curve of the companion star and, thus, the mass function might be affected by several sources of systematic uncertainties. These include tidal distortons, non-synchronous rotation and the effects of X-ray heating of the companion star, as well as contamination by emission lines from an accretion

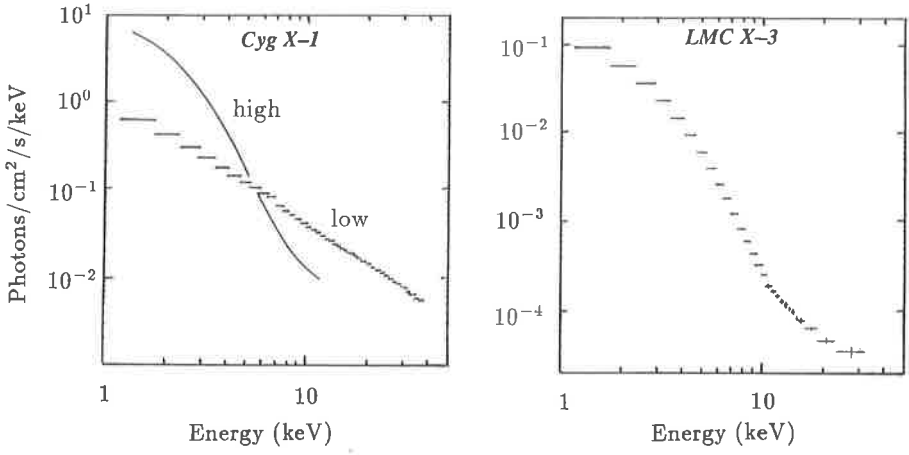

Figure 3. 1-40 keV X-ray spectra of the HMXRB black hole candidates Cyg X-1, LMC X-3. The high and low state spectra of Cyg X-1 are indicated (Tanaka 1989)

disk, gas streams or the companion's wind (Bahcall 1978). Moreover the value of M_{opt} determined from the companion's spectrum might be underestimated by a factor of two or more (in those X-ray pulsar HMXRBs for which a dynamical determination of the masses is available, the companion star is often undermassive for its spectral class). To circumvent this difficulty a different method was devised which sets an upper limit on the black hole mass as function of the distance, based only on the mass function, the lack of X-ray eclipses and the dereddened flux of the companion star (Paczynski 1974).

The most conservative assumptions lead to the conclusion that in Cyg X-1 $M_x > 3\ M_\odot$. However, M_x measures the total mass around which the supergiant star orbits. In a system like Cyg X-1, there is room to accomodate a neutron star and, e.g., a 8 $M_\odot$ B star in a tight orbit, which in turn both revolve around the O supergiant. The spectrum of such a B star could remain undetected, while M_x would be compatible with the measured value. This kind of triple star models, which clearly do not require the presence of a black hole, are still marginally viable. Despite this caveat, Cyg X-1 remained the most reliable black hole candidate for more than a decade.

The compact object in most HMXRBs has been determined to be a pulsating neutron star. Detailed studies of the often heavily reddened optical counterparts of X-ray binaries close to the galactic plane have proven difficult. For persistent LMXRBs the problem is even more pronounced, because of the intrinsically low optical luminosity and the absence (or near absence) of spectral features from the companion low mass star. Based on the X-ray characteristics of Cyg X-1, two *phenomenological signatures* of accreting black holes were suggested. These were aimed at selecting *tentative black hole candidates* among X-ray binaries, the optical counterparts of which could be studied in greater depth in order to apply the mass criterion.

5.3. Time variability criterion

The X-ray flux of Cyg X-1 shows a pronounced aperiodic flickering on timescales of the order of ~ 1 s or less. Significant structure has been detected down to a few milliseconds. These variations were modelled in terms of a simple shot noise model as early as 1972 and were suggested to arise by some form of instability in the accreting matter in the vicinity of the black hole (Liang and Nolan 1984).

In the early eighties, a transient source, V0332+53, was observed which, together with ~ 4.4 s periodic pulsations, displayed pronounced rapid aperiodic variations similar to those of Cyg X-1 (Stella *et al* 1984). This provided the first clear counterexample of the time variability criterion by showing that the accreting magnetic neutron star in an X-ray pulsar system could also produce fast aperiodic flickering. By now, virtually all classes of accreting compact objects in X-ray binaries are known to display rapid aperiodic variations somewhat similar to those observed from Cyg X-1 (Hasinger and van der Klis 1989, van der Klis 1989, Belloni and Hasinger 1990). More detailed investigations are required to characterise the time variability of black hole candidates and revise the criterion accordingly.

5.4. Spectral criterion

Cyg X-1 displays different spectral states. During most of the time the source is in the *low state*, characterised by a power-law like X-ray spectrum with a logarithmic slope of about -0.5, which steepens above ~ 100 keV and extends up to energies of several hundred keV. The X-ray luminosity is $3-4\times10^{37}$ erg/s. In the *high state* the luminosity increases by a factor of ~ 2 due to the presence of an additional spectral component for energies of < 10 keV, with a corresponding decrease of the high energy emission (Liang and Nolan 1984). The high state spectrum of Cyg X-1 is therefore much softer than the low state spectrum (Fig. 3).

The presence of spectral states similar to those of Cyg X-1 (for energies $< 20-30$ keV) and/or a high energy tail extending to hundreds of keV has also been used as a phenomenological signature of accretion onto black holes. As described in section 5.6, the application of this criterion has been very successfull over the last few years. However, out of the first two tentative black hole candidates that were suggested on the basis of the time variability and spectral criteria, only GX 339-4 might contain a black hole (Ilovaisky *et al* 1986; Dolan *et al* 1987). Type I X-ray bursts were instead observed from Cir X-1, which proved that the system contains an accreting neutron star (Tennant *et al* 1986). This fact emphasises that, despite its success, the spectral criterion is not fully reliable.

5.5. Black hole candidates in the LMC and X-ray colours

Two other black hole candidates were identified in the early eighties from detailed optical observations of two HMXRBs in the Large Magellanic Cloud. The likely mass of the compact object in LMC X-3 and LMC X-1 was determined to be ~ 9 $M_\odot$ and ~ 6 $M_\odot$, respectively (Cowley *et al* 1983, Hutchings *et al* 1987). The value of the optical mass function, though relatively large for LMC X-3 (~ 2.3 $M_\odot$), does not, by itself, exclude the presence of a neutron star in either of the two systems (see Table 2). Rather, systematic uncertainties in the luminosity and orbital velocity of the optical star in LMC X-3 (Mazeh

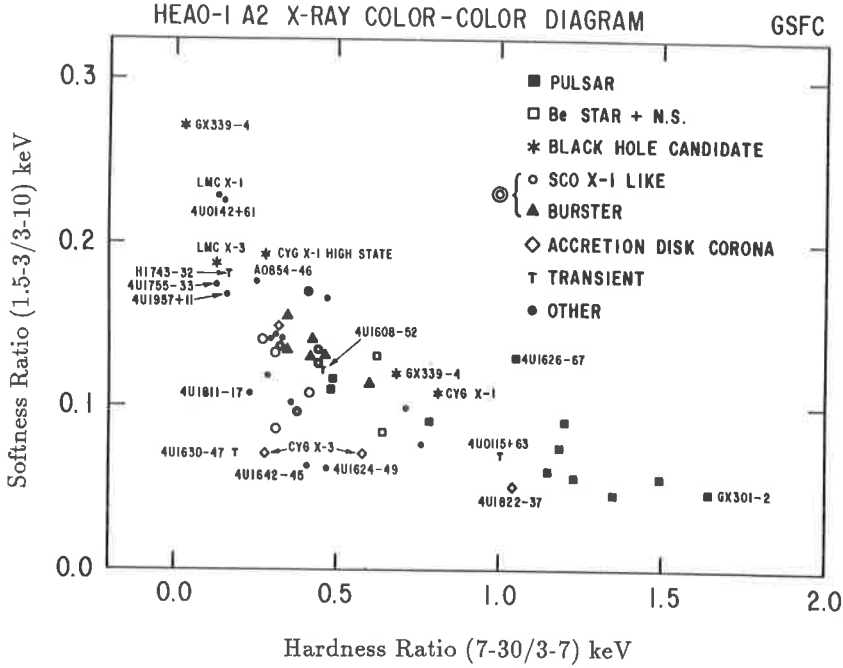

Figure 4. An X-ray colour-colour diagram. Different types of X-ray sources are indicated by different symbols (White and Marshall 1984)

et al 1986) and the relatively large error in the mass function of LMC X-1 do not allow ruling out the unlikely scenarios in which the mass of the compact object is $< 3\ M_\odot$. In this sense LMC X-3 and, especially, LMC X-1 provide less reliable black hole candidates than Cyg X-1. The X-ray spectra of LMC X-1 and LMC X-3 are extremely soft and resemble the high state spectrum of Cyg X-1 (Fig. 3).

When an X-ray colour-colour diagram was assembled from observations in the $1.5 - 30$ keV band (Fig. 4), it was noticed that the two LMC black hole candidates and Cyg X-1 and GX 339-4 in their high state lie in the left part of the diagram, characteristic of *ultrasoft* spectra (White and Marshall 1984). On the contrary, the low state colours of Cyg X-1 and GX 339-4 are harder and close to those of LMXRBs containing an old accreting neutron star. Should the colour-colour diagram be calculated for higher energies (from tens to hundreds of keV, where the observations are still sparse), then the high energy tail would probably make the hardness ratios of black hole candidates the highest (see Fig. 5a). In the $1.5 - 30$ keV energy range, however, X-ray pulsar binaries present with the hardest spectra and tend to occupy the right part of the diagram.

A number of *ultrasoft* sources, that do not display bursts or pulsations, have colours similar to those of black hole candidates in their high state. Shortly after the optical studies of LMC X-1 and LMC X-3, these sources were suggested as tentative black hole candidates. Among these there is a high incidence of *ultrasoft transient sources*, some of which have been identified as LMXRBs (White *et al* 1984). While in quiescence, these systems provide a rare opportunity to study the companion's optical spectrum and

thereby determine the distance and some orbital parameters.

5.6. Black hole candidates in X-ray transients

A0620-00 was the first ultrasoft X-ray transient to be studied in quiescence. For two months in 1975 it was the brightest X-ray source in the sky and achieved a peak luminosity of $L_x \sim 10^{38}$erg/s. Its optical counterpart was identified thanks to the ~ 6 mag brightening that accompanied the X-ray outburst. Detailed optical spectro-photometry in quiescence revealed an orbital period of ~ 7.7 hr and a low mass K-star companion, the spectral features of which are modulated with a velocity semiamplitude of $\simeq 450$ km/s (see Table 2) . This provides an optical mass function of ~ 2.9 $M_\odot$, almost sufficient in itself to identify A0620-00 as the first black hole candidate in a LMXRB (McClintock and Remillard 1986). The likely black hole mass is $M_x \sim 10$ $M_\odot$. A firm lower limit of $M_x > 3.2$ $M_\odot$ is obtained through considerations which are insensitive to the distance and the companion's mass. This is unlike the black hole candidates in HMXRBs. Moreover triple star models similar to those suggested for Cyg X-1 would not viable, because in a compact and optically faint binary like A0620-00 it would be impossible to hide a non-degenerate star of suffinciently high mass (see Fig. 6). A scenario in which two or more degenerate stars (neutron stars and/or white dwarfs) in a very close orbit revolve around the K-star companion can probably be constructed; such a system, however, would probably be very short lived due to the emission of gravitational radiation or the ejection of one (or more) degenerate star(s). As a black hole candidate, A0620-00 presents thus clear advantages with respect to Cyg X-1; some of these are related to the transient LMXRB character of the system.

The search for other black hole candidates in transient LMXRBs has been very successful over the last few years. New tentative candidates have been discovered through the application of the spectral criterion, owing to a more consistent monitoring of the X-ray sky with large field of view detectors. In particular observations in the hard X-ray band (≥ 30 keV) have recently allowed to identify several new transient sources, with spectra extending to several hundreds of keV. Discovered and monitored in 1989, GS2023+338 (V404 Cygni) is one such transient; its X-ray spectrum is similar to the low state spectrum of Cyg X-1 (see Fig. 5a). Observations of its optical counterpart have allowed to measure an orbital period of ~ 6.5 d and a velocity semiamplitude of ~ 210 km/s for the companion star. The very high value of the mass function, $f_{opt} \simeq 6.3$ $M_\odot$, establishes the compact object in GS2023+338 as the best black hole candidate available to date, independent of any consideration concerning the distance, the luminosity, the inclination or the mass of the companion star (Casares *et al* 1992). It is interesting to note that there is inderect evidence for a third star in a tight orbit around the collapased object in GS2023+338; the current constraint on the mass of such a star ($M < 0.5$ $M_\odot$) leaves unaffected the conclusion that GS2023+338 contains a black hole candidate. Another ultrasoft X-ray transient, GS1124-68 (Nova Muscae), has been recently determined to host a black hole candidate, based on a mass function of ~ 3.1 $M_\odot$ (Remillard *et al* 1992).

The number of black hole candidates in X-ray binaries has increased from one to six over the last decade; two of them were identified in 1992. Despite their faintness (often requiring observations with the largest telescopes), the optical counterparts of several other tentative black hole candidates in X-ray transient are currently being studied; some still remain to be identified. A0620-00 and GS2023+338 were observed as optical novae in

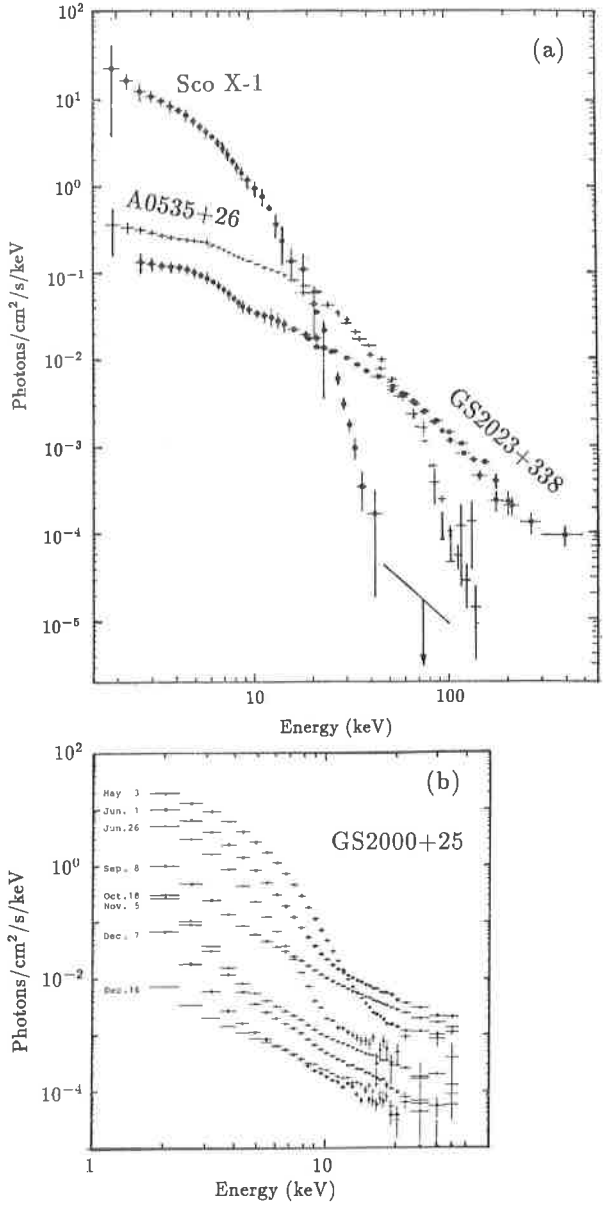

Figure 5. (a) Comparison of the wide energy range X-ray spectra from the bright LMXRB Sco X-1, the transient X-ray pulsar HMXRB A0535+26 and the black hole candidate GS2023+338 (Sunyaev *et al* 1991a). Note the spectrum of GS2023+338 extends to much higher energies than the spectra of the other two sources. (b) The 2-40 keV spectral variability of the potential black hole candidate GS2000+25. Note the independent variations of the ultrasoft and power law components (Tanaka 1989)

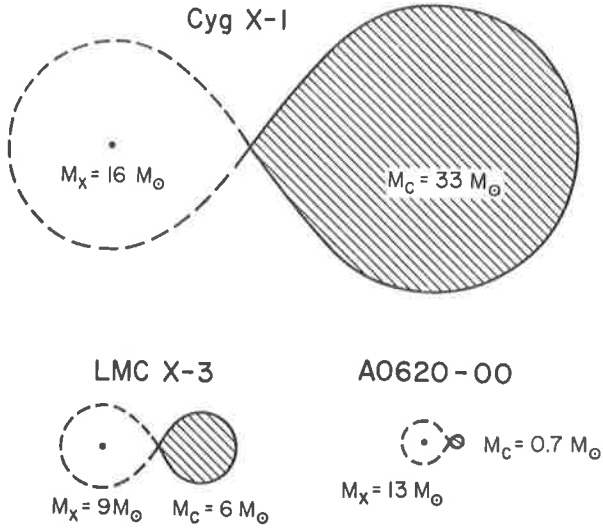

Figure 6. Schematic sketch, to scale, of likely models of the HMXRB black hole candidates Cyg X-1, LMC X-3 and the transient LMXRB black hole candidate A0620-00. The optical companions (shaded regions) are assumed to fill their Roche lobe (McClintock 1986).

1917 and 1938, respectively; intermediate outbursts may have been missed. Recurrence times are therefore in the range of tens of years. The total number of transient black hole candidate LMXRBs in the galaxy, although very uncertain, is estimated between ~ 100 and ~ 3000 (Tanaka 1991a).

5.7. Open issues in the physics of black hole accretion

In addition to the advantage which derives from the identification and study of their optical counterparts, ultrasoft and high energy tail X-ray transients offer the opportunity of investigating also the physics of accretion into black holes over a wide range of X-ray luminosities and, thus, accretion rates. The rapidly growing number of candidates, together with the coverage afforded by wide-field X-ray instruments and intensive follow-up programs, will allow to address in greater detail a number of open issues related to the physics of accreting black hole. Among these:

(a) What is the origin of the ultrasoft spectral component, characteristic of the *high state*, and the power law component, which is dominant in the *low state* ? Recent observations have revealed that the intensity of each of the two components can vary independently of the other; the tentative black hole candidate GS2000+25 provides the best example of this behaviour (Fig. 5b; Tanaka 1989, 1991b, Ebisawa 1991). Accretion disk models have been developed which comprise a relatively cold disk, emitting ultrasoft X-rays, and a hot phase, in the geometry of either a corona above the accretion disk or a central bulge, which produces the power law spectrum (Shapiro *et al* 1976, Galeev *et al* 1979, Haardt and Maraschi 1991). While the ultrasoft X-ray component is likely to originate

Table 2. Properties of black hole candidates

Properties	Cyg X-1	LMC X-1	LMC X-3	A0620-00	GS2023+33	GS1124-68
class[a]	pers. H	pers. H	pers. H	trans. L	trans. L	trans. L
L_x^{max} erg/s	10^{38}	2×10^{38}	3×10^{38}	1×10^{38}	4×10^{38}	10^{37}
donor star	O9.7I	late O	B3V	K5V	G/K	K2V
d kpc	2.5 (?)	55	55	1 (?)	2 (?)	2 (?)
$V - mag$	9	14	17	18	19	20
K km/s	75 ± 1	68 ± 8	235 ± 11	433 ± 4	211 ± 4	409 ± 18
P_{orb} d	5.6	4.2	1.7	0.32	6.5	0.43
$f_{opt}(M)$ $M_\odot$	$.25 \pm .01$	$.14 \pm .05$	$2.3 \pm .3$	$2.91 \pm .08$	$6.26 \pm .31$	$3.1 \pm .4$

[a] pers: persistent, trans: transient, H: HMXRB, L: LMXRB,

from the superposition of black body spectra (possibly modified by the effects of electron scattering, Treves *et al* 1987, White et al 1988), the power law component is probably produced by Comptonisation. The latter process may work by upscattering the soft photons from the cold disk in a hot plasma ($kT \geq 100$ keV) with electron-positron pairs providing a feedback mechanism for the temperature (Svensson 1990).

(b) Why is substantial emission above a few hundred keV characteristic of accretion into black holes candidates ? Only very preliminary explanations are available. Perhaps the absence of a star surface and of the associated emission of soft photons prevents cooling of the Comptonising region, and higher electron energies can be mantained than in the case of accreting neutron stars. Recent observations indicate, however, that also the persistent emission of X-ray burst source might possess a high energy tail extending to 100-200 keV (MXB1728-34 Cook *et al* 1991, Terzan 2 Barret *et al* 1991, KS1731-260 Barret *et al* 1992).

(c) What is the origin of the emission feature at energies of 400–600 keV or higher observed in some black hole candidates ? In the case of Cyg X-1 a variable broad bump at energies of 400–1500 keV has been observed, which could correspond to the Wien peak of a Comptonised spectrum or, alternatively, to a blueshifted $e^+ - e^-$ annihilation line (Ling *et al* 1987). A much narrower feature around $\sim$ 500 keV has been recently reported from the ultrasoft X-ray transient GS1124-68 which is strongly suggestive of an $e^+ - e^-$ annihilation line (Sunyaev *et al* 1992, Goldwurm *et al* 1992). Also the high energy source 1E1740.7-2942 near the galactic center displayed a variable bump at 300–600 keV, probably related to annihilation processes in an electron-positron plasma (Sunyaev *et al* 1991b, Bouchet *et al* 1991). Further observations with higher spectral resolution and a more extensive coverage of a number of sources are required to confirm this interpretation. Annihilation lines might provide an important diagnostic of pair-dominated plasmas.

(d) Is there cold optically thick matter in the vicinity of accreting black holes ? The low state spectrum of Cyg X-1 and, especially, the variable absorption spectra GS2023+33

display a braod excess at $\sim 10 - 40$ keV, above a simple power law spectrum (Tanaka 1989, 1991b, Ebisawa 1991). A similar excess is observed in the X-ray spectrum of a number of Seyfert I galaxies (Pounds *et al* 1990). This feature is best interpreted in terms of reflection by relatively cold optically thick matter in the vicinity of the collapsed object. Reflection is most effective above ~ 10 keV, where the photoelectric absorption by heavy elements is small, and below ~ 300 keV, where the effects of Compton recoil are not dominant (Lightman and White 1988). The study of this spectral component can provide insights into the physical state and geometry of matter in the vicinity of the black hole.

(e) Is there a type of short term X-ray variability which is characteristic of accretion into black hole candidates ? While virtually all classes of accreting compact stars are by now known to produce fast aperiodic variations ($\tau < 1$ s), more detailed studies based on power spectral and cross-spectral analyses suggest that there might exist a type of aperiodic variability which is specific to accretion into black hole candidates (Miyamoto *et al* 1991a). This would also allow to revise the time variability criterion for identifying tentative black hole candidates.

(f) Is there a connection between the QPOs in the X-ray flux of black hole candidates and those from accreting neutron star systems ? QPOs have already been revealed from several black hole candidates (and tentative black hole candidates). The QPO frequency is relatively low in the HMXRB black hole candidates LMC X-1 (~ 0.07 Hz, Kitamoto 1989) and Cyg X-1 (~ 0.04 Hz, Vikhlinin *et al* 1992, Angelini *et al* 1992). The ~ 6 Hz QPOs from GX339-4 display different modes, associated with different time variability and spectral states, which involve some interesting analogies with the QPOs observed from LMXRBs containing an old neutron star (Miyamoto *et al* 1991b). The frequency of the QPOs from the ultrasoft transient black hole candidate GS1124-68 is also quite high (~ 10 Hz). Models involving rotating magnetospheres are clearly not applicable to black hole candidate QPOs. Disk instability models, although still in their infancy, appear to be more promising.

6. Iron Kα lines in accreting collapsed objects

The iron emission line features observed at $\sim 6-7$ keV in virtually all classes of accreting collapsed stars provide probably the most promising diagnostic of the physical conditions and, possibly, the dynamics of the innermost regions of accretion flows towards collapsed objects. These lines have been studied in most cases only with modest spectral resolution detectors ($E/\Delta E \leq 10$), which could not resolve the profile of the line. Yet two very important conclusions have been reached. Firstly, the line centroid energy measured from Cyg X-1 (~ 6.2 keV, see Fig. 7a) and the ultrasoft transient 4U1543-47 (~ 5.9 keV) requires a substantial redshift, as the rest energy of Kα photons range from 6.4 keV for fluorescence from the lowest ionisations stages of iron, to 6.9 keV for recombination into hydrogenic iron ions (Barr *et al* 1985, van der Woerd *et al* 1989). Secondly, the iron Kα lines are broadened to widths of 1 keV or more for many accreting collapsed objects, such as old neutron stars in LMXRBs and a few black hole candidates (White *et al* 1986).

Blending of lines from different ionisation stages of iron can produce a maximum width of only ~ 0.5 keV. The observed line redshifts and widths can be caused by Compton scattering of line photons in an electron cloud, if the optical depth is ≥ 3

and the temperature is $\sim 10^7 - 10^8$ K (Kallman and White 1989). If the accretion flow is mediated by a disk, line emission from it is likely to be driven by photoionisation or Compton heating by radiation from closer to the central object. Compton line broadening is not expected to play an important role unless there exists a hot phase of the accretion disk with an optical depth > 1. This, in turn, is not likely to occur in a Compton heated disk corona, because photons from the central X-ray source, which are responsible for the heating, would not be able to penetrate to the base of the corona. The interpretation in which the broadening and shifting of the iron Kα lines arises because of the Doppler, transverse and gravitational shifts from high velocity plasma motions in the vicinity of the collapsed object appears to be more natural. The accretion disk modelling of the line profiles requires a general relativistic treatment, as the observed widths and redshifts suggest a line emitting region in the range of tens of Schwarzschild radii (Fabian *et al* 1989).

The profiles calculated for a Keplerian disk orbiting a non-rotating black hole display a characteristic double-horned profiles, similar to, although much broder than, those observed from spiral galaxies and cataclysmic variables. Fig. 7b shows an example of such profile. The inclination, the inner and outer radii of the line emitting region and the emissivity law were adjusted so as to approximate the parameters of the iron line of Cyg X-1. Relativistic effects (aberration, time dilation and redshift) combine to produce a blue horn brighter than the red horn, whereas, due to the relatively low inclination, the gravitational and transverse effects cause an overall redshift of the line. The solid state detectors (resolution of $E/\Delta E \sim 50$) to be flown on board the next generation of X-ray astronomy satellites should allow to resolve the profile of the iron Kα lines from a number of accreting collapsed stars. The observation of the characteristic double-horned profile would not only confirm the accretion disk scenario, but would also provide an unprecedentedly sensitive probe of regions as close as a few tens of Schwarzschild radii to the central object. In particular the dimensions (in units of Schwarzschild radii), inclination and emissivity distribution can be measured directly from the profile (Fabian *et al* 1989, Stella 1989, George and Fabian 1991, Matt *et al* 1992).

The strongest evidence to date that the iron Kα line is produced very close to the central object is given by the Seyfert galaxy NGC 6814, whose continuum X-ray variability (on timescales as short as ~ 50 s) is mimicked by intensity variations of the line, which take place with a delay of less than 250 s (Kunieda *et al* 1990). In the accretion disk scenario, changes of the line emissivity will arise in response to flux variations at the central source and will induce a time dependence of the line profile observed from infinity. After a sudden flux increase at the central source, two characteristic features are expected to form on opposite sides of the line which, owing to light travel time effects, gradually drift from the line wings, corresponding to the innermost disk regions, to the blue and red horns. The observation of this line-profile evolution can afford a measurement of the dimensions of the line emitting region in absolute units (*e.g.* cm). The combination of the relative and absolute measurements allows to infer the mass of the central object (Stella 1990).

This technique for estimating the mass of collapsed stars removes entirely the need of a companion star; moreover it relies upon dynamical measurements of the strong gravitational field regions a few tens of Schwarzschild radii away from the central object. The observability of the expected line profile changes depends on the number of X-ray photons at the Earth per Schwarzschild radius light crossing time. This is a factor of

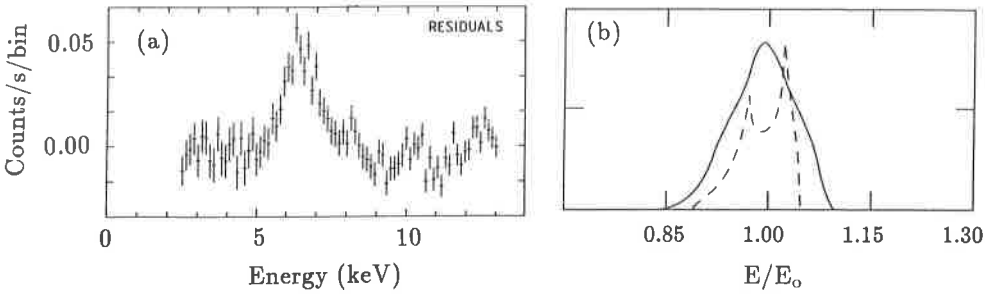

Figure 7. (a) The iron Kα line of Cyg X-1, after sutraction of the continuum spectral components (Barr *et al* 1985). The detector resolution is about 600 eV ($E/\Delta E \simeq 10$) (b) The dashed line shows the line profile expected from a relativistic accretion disk. The line emitting portion of the disk extends from 20 to 200 Schwarzschild radii, with an emissivity law $\propto r^{-2}$. The disk inclination is 30°. The solid line represents the line profile convolved with a rectangular window (with $E/\Delta E \simeq 10$) for comparison with the observed data. Note that despite the pronounced blue horn, the line centroid is redshifted

$\geq 10^4$ higher for bright AGNs than for X-ray binaries. Therefore, observational prospects for the application of the new technique are far more promising in AGNs.

7. Conclusions

Over the last few years, transient X-ray binaries have provided a number of insights into the physics of accreting collapsed objects. The case of accreting magnetised neutron stars is well illustrated by EXO 2030+375, an X-ray pulsar in which several physical processes could be investigated in detail. Crucial to these findings is the large range of luminosities and, thus, accretion rates over which properties can be studied in X-ray transient sources. Moreover a subclass of X-ray transients has been identified which is clearly associated with black hole candidates in low mass binaries. For three such systems a reliable upper limit on the mass of the collapsed object has been recently obtained from dynamical studies of the companion star in the quiescent state, when the reprocessing of X-ray radiation does not dominate the emission at optical wavelenghts. The optical counterparts of a number of tentative black hole candidates in X-ray transients remain to be identified and studied; moreover new transients are being discovered with large field of view X-ray instruments. Therefore the number of candidates is likely to increase rapidly in the near future.

Extensive X-ray monitoring of the outbursts, with improved spectral resolution and extended energy range, will allow to investigate in detail the properties of accreting black hole candidates over the large luminosity variations which are characteristic of X-ray transients. These studies will bring the level of our understanding of accreting

black holes closer to that of accreting magnetic neutron stars. The broad and sometimes redshifted iron Kα lines, which are observed in virtually all classes of accreting collapsed objects, are likely to provide a powerful diagnostic of strong gravitational field regions and a new, more accurate way of measuring the mass of the central object in AGNs.

The mass criterion is currently central to the identification of black hole candidates. While based on a small number of very likely hypothesis, the criterion provides only an indirect proof of the existence of black holes. Even within the framework of general relativity, compact stellar configurations substantially more massive than neutron stars could be made of, e.g., unknown light and stable fermions, or high-density nongravitationally bound states of baryon matter (Q-stars, Bahcall *et al* 1990). A new perspective for the observation and study of the properties of black hole candidates is clearly emerging, which will allow to probe regions closer and closer to the collapsed object. Identifying reliable *strong field signatures* of accreting black holes will require new synergetic efforts of theorists and observers. Ultimately, incontrovertible evidence should be found for the existence of an event horizon.

Acknowledgements

I am grateful to G. Ghisellini, S. Mereghetti and M. Tavani for useful discussions and to CGS for carefully reading the manuscript.

References

Aly J J 1980 *As. Ap.* **86** 192

Angelini L, Stella L and Parmar A N 1990 *Ap. J.* **346** 906

Angelini L, White N E and Stella L *et al* 1992, IAU Circ. no. 5580

Bahcall J N 1978 *Ann. Rev. As. Ap.* **16** 241

Bahcall S, Lynn B W and Sepilsky S B 1990 *Ap. J.* **362** 251

Barr P, White N E and Page C G 1985 *M.N.R.A.S.* **216** 65P

Barret D *et al* 1991 *Ap. J. (Letters)* **379** L21

Barret D *et al* 1992 *Ap. J.* **394** 615

Basko M M and Sunyaev R A 1976 *M. N. R. A. S.* **42** 311

Belloni T and Hasinger G 1990 *As. Ap.* **230** 103

Bouchet L. *et al* 1991 *Ap. J. (Letters)* **383** L45

Casares J, Charles P A and Naylor T 1992 *Nature* **355** 614

Cowley A P, Crampton D, Hutchings J B, Remillard R and Penfold J E 1983 *Ap. J.* **272** 118

Dolan J F, Crannell B R, Dennis B R and Orwig L E 1987 *Ap. J.* **322** 324

Ebisawa K 1991 PhD thesis (ISAS: Sagamihara, Japan)

Fabian A C, Rees M J, Stella L and White N E 1989 *M.N.R.A.S.* **238** 729

Fujimoto M Y, Hanawa T and Miyaji S 1981 *Ap. J.* **247** 267

Galeev A A, Rosner R and Vaiana G S 1979 *Ap. J.* **229** 318

George I M and Fabian A C 1991 *M.N.R.A.S.* **249** 352

Ghosh P and Lamb F K 1979 *Ap. J.* **234** 296

Giacconi R 1978 *Physics and Astrophysics of Neutron Stars and Black Holes* (North Holland: Amsterdam) Ed. Giacconi R and Ruffini R, p 17

Goldwurm A *et al* 1992 *Ap. J. (Letters)* **389** L79

Haardt F and Maraschi L 1991 *Ap. J.* **341** 955

Hanawa T and Fujimoto M Y 1984 *Publ. Astr. Soc. Japan* **36** 199

Hartle J 1978 *Phys. Rep.* **46** 201

Hasinger G and van der Klis M 1989 *As. Ap. 225* 79

Henrichs H F 1983 *Accretion Driven Stellar X-ray Sources* (Cambridge Univ. Press) Ed. W H G Lewin and E P J van den Heuvel, p 393

Hutchings J B, Crampton D, Cowley A P, Bianchi L and Thompson I B 1987 *As. J.* **94** 340

Ilovaisky S A, Chevalier C, Motch C and Chiappetti L 1986 *As. Ap.* **164** 67

Janot-Pacheco E, Motch C, Pakull M W 1988 *Astr. Ap* **202** 81

Kallman T and White N E 1989 *Ap. J.* **341** 955

Kitamoto S 1989 *Proc. 23rd ESLAB Symp.: Two Topics in X-ray Astronomy* (ESA SP-291) Ed. Hunt J and Battrick B, p 231

Kunieda H, Turner T J, Awaki H, Koyama K, Mushotzky R and Tsusaka Y 1990 *Nature* **345** 786

Lamb F K, Shibazaki N, Alpar M A and Shaham J 1985 *Nature* **317** 681

Lewin W H G and Joss P C 1983 *Accretion Driven Stellar X-ray Sources* (Cambridge Univ. Press) Ed. Lewin W H G and van den Heuvel E P J, p 41

Lewin W H G, van Paradijs J and van der Klis M 1988 *Space Sci. Rev.* **46** 273

Liang E P and Nolan P L 1984 *Space Sci. Rev.* **38** 353

Lightman A P and White T R 1988 *Ap. J.* **335** 57

Ling J C, Mahoney W A, Wheaton Wm A and Jacobsen A S 1987 *Ap. J. (Letters)* **321** L117

Matt G, Perola G C, Piro L and Stella L 1992 *As. Ap.* **257** 63

Mazeh T, van Paradijs J, van den Heuvel E P J and Savonije G J 1986 *As. Ap.* **157** 113

Meszaros P 1984 *Space Sci. Rev.* **38** 325

McClintock J E 1986 *The Physics of Accretion onto Compact Objects* (Springer: Berlin) Ed. Mason K O, Watson M G and White N E, p 211

McClintock J E and Remillard R A 1986 *Ap. J.* **308** 110

Miyamoto S, Kitamoto S, Iga S, Negoro H and Terata K 1991a preprint

Miyamoto S, Kimura S, Kitamoto K, Iga S, Dotani T and Ebisawa K 1991b *Ap. J.* **383** 784

Paczynski B 1974 *As. Ap.* **34** 161

Parmar A N and White N E 1988 *Mem. Soc. As. Ital.* **59** 147

Parmar A N, White N E, Stella L, Izzo C and Ferri P 1989a *Ap. J.* **338** 359

Parmar A N, White N E and Stella L 1989b *Ap. J.* **338** 373

Pounds K A, Nandra K, Stewart G C, George I M, and Fabian A C 1990 *Nature* **344** 132

Rappaport S A and Joss P C 1983 *Accretion Driven Stellar X-ray Sources* (Cambridge Univ. Press) Ed. Lewin W H G and van den Heuvel E P J, p 1

Remillard R A, McClintock J E and Bailyn C D 1992 *Ap. J. (Letters)* in press

Rhoades C E and Ruffini R 1974 *Phys. Rev. Lett.* **32** 324

Shapiro S L, Lightman A P ans Eardley D M 1976 *Ap. J.* **204** 187

Stella L 1988 *Mem. Soc. As. Ital.* **59** 185

Stella L 1989 *Proc. 23rd ESLAB Symp.: Two Topics in X-ray Astronomy* (ESA SP-291) Ed. Hunt J and Battrick B, p 19

Stella L 1990 *Nature* **344** 747

Stella L, White N E, Davelaar J, Parmar A N, Blisset R J and van der Klis M 1985 *Ap. J.* **288** L45

Sunyaev R *et al* 1991a *Sov. As. Letters* **17** 6

Sunyaev R *et al* 1991b *Ap. J. (Letters)* **383** L49

Sunyaev R *et al* 1992 *Ap. J. (Letters)* **389** L75

Svensson R 1990 *Physical Processes in Hot Cosmic Plasmas* (Kluwer Academic: The Netherlands) Ed. Brinkmann W *et al* p 357

Tanaka Y 1989 *Proc. 23rd ESLAB Symp.: Two Topics in X-ray Astronomy* (ESA SP-291) Ed. Hunt J and Battrick B, p 1

Tanaka Y 1991a ISAS research note 479 (ISAS: Sagamihara, Japan)

Tanaka Y 1991b *Iron Line Diagnostics in X-ray Sources* (Springer: berlin) Ed. Treves A, Perola G C and Stella L, p 98

Taylor J H 1993 this volume

Tennant A F, Fabian A C and Shafer R A 1986 *M.N.R.A.S.* **221** 27P

Treves A *et al* 1987 *Ap. J.* **325** 119

van den Heuvel E P J 1991 *Neutron Stars: Theory and Observation* (Kluwer Academic: The Netherlands) Ed. Ventura J and Pines D, P 171

van der Klis M 1989 *Ann. Rev. As. Ap.* **27** 517

van der Woerd H, White N E and Kahn S M 1989 *Ap. J.* **344** 320

van Paradijs J and Lewin W H G 1988 *Mem. Soc. As. Ital.* **59** 213

Vikhlinin A *et al* 1992, IAU Circ. no. 5576 .

White N E, Kaluzienski J L, and Swank J H 1984 *High Energy Transients in Astrophysics* (AIP: New York) Ed. Woosley S, p 31

White N E and Marshall F E 1984 *Ap. J.* **281** 354

White N E, Peacock A, Hasinger G, Mason K O, Taylor, B G and Branduardi-Raymont G 1986 *M.N.R.A.S.* **218** 129

White N E, Stella L and Parmar A N 1988 *Ap. J.* **324** 363

White N E, Swank J H and Holt S S 1983 *Ap. J.* **270** 711

Paper presented at the 13th Int. Conf. on General Relativity and Gravitation
Cordoba, Argentina, 1992: Part 1, Plenary Lectures

Testing relativistic gravity with binary and millisecond pulsars

J H Taylor

Joseph Henry Laboratories and Physics Department
Princeton University, Princeton, New Jersey 08544, USA

Abstract. Binary and millisecond pulsars offer unique opportunities for high precision experiments in relativistic gravity, probing well beyond the weak-field, slow-motion limit of all previous experimental tests. They also provide the means for accurate measurements of neutron star masses, placing rigorous constraints on the energy density of low-frequency gravitational radiation in the universe, and a number of other significant results. The first known binary pulsar, PSR B1913+16, has now been observed for more than 18 years. Its timing measurements have conclusively established the existence, quadrupolar nature, and propagation speed of gravitational waves; the results are presently in accord with general relativity at the 0.4% level. A more recently discovered binary pulsar, PSR B1534+12, has provided clean access to a test of gravity under strong-field conditions, independent of gravitational radiation effects. In this paper I summarize and update the status of experiments involving these two pulsars, and provide references to other related work.

1. Introduction

The discovery of pulsars 25 years ago introduced radio astronomers to a new type of natural clock, one that has proven remarkably useful for high-precision experiments in relativity. When a pulsar was found in a gravitationally bound orbiting system (Hulse & Taylor 1975), the way was opened for making explicit comparison of terrestrial atomic time with the time kept by a spinning macroscopic object located in a strong gravitational field and moving with mildly relativistic velocity. Measurements of this orbiting pulsar have been pursued intensively and have borne excellent fruit, as predicted. Meanwhile, a number of other pulsars have been found with characteristics and circumstances that make them interesting for applications in gravitational physics. Interesting results have now been obtained in areas as diverse as (1) determining neutron star masses, (2) placing limits on the cosmic gravitational wave background, (3) testing the constancy of the gravitational coupling parameter G, (4) establishing the existence and quadrupolar nature of gravitational waves, and (5) establishing constraints on possible departures of the

"correct" theory of gravity from general relativity, in the strong-field regime. A number of papers have been published on these topics in the last few years. I shall summarize and update a few of the results here, and provide references to more complete work published elsewhere.

2. Overview of experimental details

The first binary pulsar to be discovered, named PSR B1913+16 according to its position in the B1950 system of celestial coordinates, turned out to be the harbinger of a new subclass of pulsars whose evolution is substantially modified by the proximity of an orbiting companion star. More than 40 of these "recycled" pulsars have now been found, about half of them since 1990. They are believed to have passed through an evolutionary phase in which mass and angular momentum were accreted from an evolving companion after it left the hydrogen-burning main sequence and began shedding its outer layers. The recycled pulsars have much shorter rotation periods and weaker magnetic fields than other pulsars. Moreover, they exhibit even more remarkable long-term timing stabilities — approaching and possibly surpassing those of the best atomic clocks.

Pulsar timing experiments are conceptually straightforward, even though one of the "clocks" being compared may be located several kiloparsecs from Earth. The observations are usually signal-to-noise limited, so they are best carried out with the largest available radio telescopes, such as the 305 m spherical reflector of the Arecibo Observatory in Puerto Rico. Interstellar propagation effects and galactic background noise compromise the measurements below about 0.3 GHz, while steep radio frequency spectra make most pulsars hard to observe at frequencies much above 3 GHz. Therefore, timing observations are usually carried out somewhere within this one-decade frequency interval.

In the pulsar timing measurements that my colleagues and I make regularly at Arecibo, radio-frequency signals induced in the feed antennas are amplified, converted to intermediate frequency, and passed through a multi-channel spectrometer. Digital signal averagers accumulate the pulsar's periodic intensity waveform in each frequency channel, using circuitry under computer control and accurately synchronized with the observatory's time and frequency standard. A programmable synthesizer, whose output frequency is adjusted once a second in a phase-continuous manner, compensates for changing Doppler shifts caused by known accelerations of the pulsar and the observatory. Average pulse profiles are recorded every few minutes together with appropriate time tags. Phase offsets of individual profiles are measured relative to a high signal-to-noise standard profile, converted to time delays, and added to the recorded start times of the integrations, thereby yielding topocentric times of arrival (TOAs) according to the observatory's master clock. TOAs obtained for different spectral channels are combined after correcting for dispersive delays caused by the ionized interstellar medium, and clock offsets measured by means of Earth-orbiting GPS satellites (Lewandowski & Thomas 1991) are applied to correct all TOAs to the best available standard of terrestrial atomic time.

An extensive theoretical framework for analyzing pulsar timing data has been developed over the years (e.g., Manchester & Taylor 1977, Damour & Deruelle 1986, Taylor & Weisberg 1989, Ryba & Taylor 1991, Damour & Taylor 1992). Each observatory-centered TOA, say $t_{\rm obs}$, is first transformed to the reference frame of the solar system

barycenter, and then, for binary pulsars, to the pulsar co-moving frame (modulo an unknown constant velocity offset). In this frame the TOAs should be spaced so that pulse number $\mathcal{N}$ (an integer) is emitted at time $T_{\mathcal{N}}$, given implicitly by

$$\mathcal{N} = \nu T_{\mathcal{N}} + \frac{1}{2}\dot{\nu} T_{\mathcal{N}}^2 + \dots , \tag{1}$$

where $\nu = 1/P$ is the neutron star's rotation frequency and time derivatives higher than the first are generally found to be negligible. The transformation from terrestrial time to pulsar proper time is carried out in the weak-field, slow-motion limit of general relativity; to sufficient accuracy, other viable theories of gravity would yield identical results. The necessary equations include terms related to the positions, velocities, and masses of objects within the solar system and to frequency-dependent propagation effects in the interstellar medium. For binary pulsars there are also terms representing the consequences of orbital motion. The orbital effects have been worked out and parametrized in a general phenomenological way by Damour and Deruelle (1985, 1986).

To an accuracy consistent with the experimental state of the art, all significant terms appearing in the time transformation can be summarized in the single equation

$$\begin{aligned} T \;=\; & t_{\rm obs} - t_0 + \Delta_{\rm C} - D/f^2 + \Delta_{\rm R\odot}(\alpha, \delta, \mu_\alpha, \mu_\delta, \pi) + \Delta_{\rm E\odot} - \Delta_{\rm S\odot}(\alpha, \delta) \\ & - \Delta_{\rm R}(x, e, P_b, T_0, \omega, \dot{\omega}, \dot{P}_b, \dot{x}, \dot{e}, \delta_\theta) - \Delta_{\rm E}(\gamma) - \Delta_{\rm S}(r, s) - \Delta_{\rm A} \, . \end{aligned} \tag{2}$$

Here t_0 is a nominal equivalent TOA at the solar system barycenter; $\Delta_{\rm C}$ represents the measured offset between the observatory reference clock and the best terrestrial standard of time; D/f^2 is the dispersive delay for propagation at frequency f over the path from pulsar to Earth; $\Delta_{\rm R\odot}$, $\Delta_{\rm E\odot}$, and $\Delta_{\rm S\odot}$ are propagation delays and relativistic time adjustments within the solar system; and $\Delta_{\rm R}$, $\Delta_{\rm E}$, $\Delta_{\rm S}$, and $\Delta_{\rm A}$ are similar terms for a binary pulsar's orbit. Subscripts on the various Δ's indicate the nature of the delays, which include "Roemer," "Einstein," and "Shapiro" effects within the solar system, and these as well as "Aberration" effects in the pulsar orbit. Note that the Roemer terms have approximate amplitudes given by the orbital periods times v/c, where v is a speed characteristic of orbital motion and c the speed of light. The Einstein terms are proportional to (v^2/c^2), multiplied by the orbital eccentricities. The Shapiro delay (Shapiro 1964) in the solar system has a maximum value of $\approx 120\,\mu$s when the line of sight grazes the limb of the Sun, and depends logarithmically on the impact parameter. The corresponding delay within the binary orbit depends on the companion star's mass, the orbital phase, and the inclination i between the plane of the orbit and the plane of the sky. Full details on all of the terms in Eq. (2) can be found in the references quoted earlier.

Eqs. (1) and (2) have been written to show explicitly the nature of the most significant dependences of pulsar TOAs on the set of potentially measurable parameters. In addition to the pulsar rotation frequency ν and its first time derivative $\dot{\nu}$, these parameters include the reference arrival time t_0, dispersion constant D, celestial coordinates α and δ, proper motion terms μ_α and μ_δ, and annual parallax π. For binary pulsars, as many as 13 orbital parameters are also measurable, at least in principle. These include five that appear even in a purely Keplerian analysis of orbital motion: the projected semi-major axis $x \equiv a_1 \sin i/c$, eccentricity e, binary period P_b, longitude of periastron ω, and time of periastron T_0. If the experimental timing precision is high enough, relativistic effects give access to as many as eight "post-Keplerian" (PK) measurables: the

secular derivatives $\dot{\omega}$, $\dot{P}_b$, $\dot{x}$, and $\dot{e}$, the Einstein parameter γ, the "range" and "shape" of the orbital Shapiro delay, called r and s, and an orbital shape correction, δ_θ (see Damour & Taylor 1992, and references therein). Because seven quantities are required to fully specify the dynamics of a two-body orbiting system (up to theoretically uninteresting rotations about the line of sight), the measurement of any two PK parameters, in addition to the five readily measured Keplerian ones, provides a full description of the system, including predictions for the values of the remaining parameters. Thus, if a total of N post-Keplerian parameters are measurable one has immediate access to $N-2$ distinct tests of relativistic gravity.

In practice, parameter values are extracted from a set of TOAs by calculating the expected time of emission $T_{\mathcal{N}_i}$ for each observed pulse number $\mathcal{N}_i$ and then minimizing the weighted sum of squared residuals,

$$\chi^2 = \sum_i \left(\frac{T_i - T_{\mathcal{N}_i}}{\sigma_i} \right)^2 , \tag{3}$$

with respect to all phenomenological parameters to be determined. (Here T_i is the value of T corresponding to the i'th measured TOA, and σ_i is the measurement uncertainty. Note that with sufficiently frequent observations, the integers $\mathcal{N}_i$ are know exactly, so there is no ambiguity in determining $T_{\mathcal{N}_i}$.) In the analysis of a given set of data, some parameters will be more readily measurable than others. When TOAs are available for many observing dates distributed over a year or more, a pulsar's celestial coordinates, spin parameters, and Keplerian orbital elements are often measurable to accuracies of six or more significant digits. The PK parameters measure smaller effects, and are therefore more difficult to quantify. Extensive analyses of their measurabilities in practical circumstances have been carried out recently by Damour & Taylor (1992) and Taylor (1992).

As I have mentioned earlier, high-precision timing observations of pulsars have been put to many diverse uses. Even pulsars with no more than two measurable PK parameters can yield fundamentally important information. For example, the best available measurements of neutron star masses are derived from binary pulsar timing observations (Thorsett *et al* 1992, and references therein). Observations extending over many years have placed tight upper limits on the energy density of gravitational waves in the universe (Stinebring *et al* 1990) and on the constancy of the gravitational coupling parameter, G (Damour, Gibbons & Taylor 1988; Taylor 1992). Pulsar timing data are even proving useful as practical diagnostic tools, helping to establish time and frequency scales with the best possible long-term stabilities (Taylor 1991).

3. Binary pulsar PSR B1913+16

The first binary pulsar was found some 18 years ago, and its importance as a testbed for relativistic gravitation theories was recognized almost immediately (Hulse & Taylor 1975, Damour & Ruffini 1974, Brumberg *et al* 1975, Esposito & Harrison 1975, Wagoner 1975). In subsequent years, much effort has been put into making increasingly accurate measurements of its pulse arrival times and comparing the results with parametrized models such as the one summarized above. More than 4500 TOAs for PSR B1913+16 have now been recorded at the Arecibo Observatory (Taylor *et al* 1976; Taylor & Weisberg 1982,

1989, and unpublished work). These data determine the five Keplerian parameters and $\dot{\omega}$, the largest PK parameter, to a few parts per million or better (see Table 1). Two further PK parameters, γ and $\dot{P}_b$, have been determined with fractional accuracies better than 0.3%.

Table 1. Measured orbital parameters and derived masses for two binary pulsar systems. Figures in parentheses represent uncertainties in the last quoted digit; those in square brackets represent expected values of unmeasured quantities, according to general relativity.

	PSR B1534+12	PSR B1913+16
Keplerian phenomenological parameters:		
Orbital period, P_b (s)	36351.70270(3)	27906.9807804(6)
Eccentricity, e	0.2736779(6)	0.6171308(4)
Projected semi-major axis, x (s)	3.729468(9)	2.3417592(19)
Time of Periastron, T_0 (MJD)	48262.8434966(2)	46443.99588319(3)
Longitude of periastron, ω (°)	264.9721(16)	226.57528(6)
Post-Keplerian phenomenological parameters:		
Advance of periastron, $\dot{\omega}$ (° yr^{-1})	1.7560(3)	4.226621(11)
Time dilation, γ (ms)	2.05(11)	4.295(2)
Orbital period derivative, $\dot{P}_b$ (10^{-12})	−0.1(6)	−2.422(6)
Range of Shapiro delay, r (μs)	6.2(1.3)	[6.834]
Shape of Shapiro delay, $s \equiv \sin i$	0.986(7)	[0.734]
Derived masses:		
Pulsar mass ($M_\odot$)	1.34(7)	1.4410(7)
Companion mass ($M_\odot$)	1.34(7)	1.3874(7)

The PSR B1913+16 system thus has more measurable quantities than significant unknowns, and consequently provides a clean, accurate test of relativistic gravitation theories. It is this test that convincingly demonstrates the existence of quadrupolar gravitational radiation in general relativity. One uses the measured values of the five Keplerian parameters, $\dot{\omega}$, and γ to calculate the component masses and the expected rate of orbital period decay caused by gravitational radiation. This prediction is then compared with the observed value of $\dot{P}_b$. At present levels of accuracy it is necessary to include a small adjustment for the effect of galactic accelerations (Damour and Taylor 1991). An up-to-date error budget for the experiment is presented in Table 2, and Figure 1 illustrates the observed decay of the PSR B1913+16 orbit from September 1974 through April 1992. General relativity passes this "$\dot{\omega} - \gamma - \dot{P}_b$" test perfectly at the present level of accuracy, about 0.35% — an impressive confirmation of Einstein's theory in a regime where gravitation theories have not previously been testable.

Table 2. Error budget for the orbital period derivative of PSR B1913+16, and comparison of experimental result with general relativistic prediction.

Parameter	(10^{-12})
Observed value, $\dot{P}_b^{\rm obs}$	-2.4225 ± 0.0056
Galactic contribution, $\dot{P}_b^{\rm gal}$	-0.0124 ± 0.0064
Intrinsic orbital period decay, $\dot{P}_b^{\rm obs} - \dot{P}_b^{\rm gal}$	-2.4101 ± 0.0085
General relativistic prediction, $\dot{P}_b^{\rm GR}$	-2.4025 ± 0.0001
$\dot{P}_b^{\rm obs} - \dot{P}_b^{\rm gal}/\dot{P}_b^{\rm GR}$	1.0032 ± 0.0035

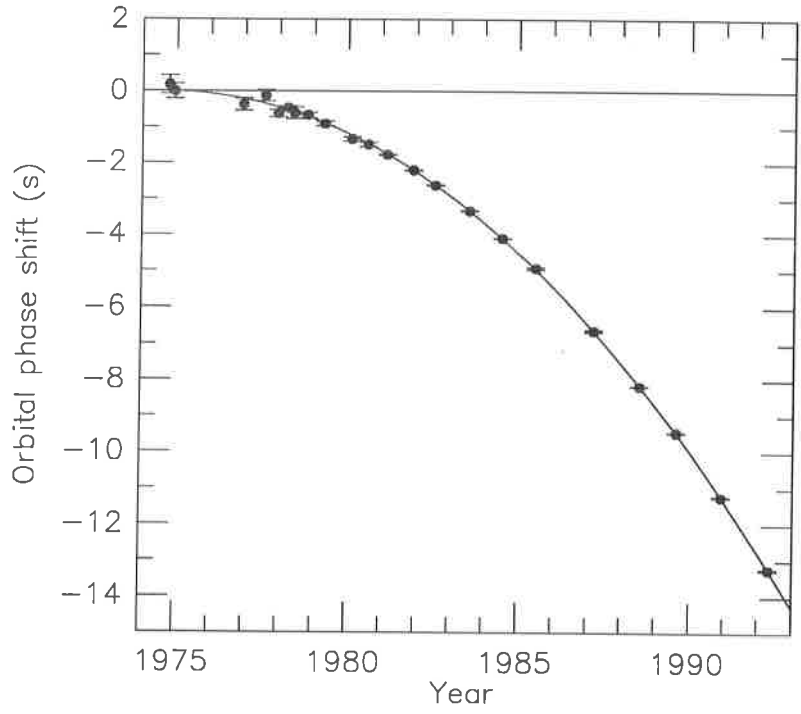

Figure 1. Filled circles represent the measured shifts of the times of periastron passage of PSR B1913+16, relative to a non-dissipative model in which the orbital period remains fixed at its 1974.78 value. The smooth curve illustrates the prediction of general relativity.

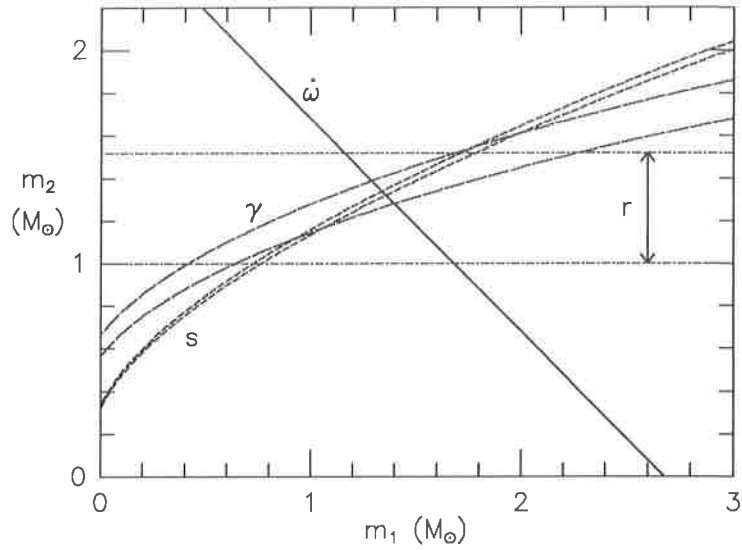

Figure 2. Parametric curves corresponding to experimental constraints on the post-Keplerian parameters $\dot{\omega}$, γ, r, and s for PSR B1534+12, according to general relativity. Pulsar and companion star masses $m_1 = m_2 = 1.34 \pm 0.07\,M_\odot$ are consistent with all of the measurements. (Observations carried out in collaboration with A. Wolszczan.)

4. Binary pulsar PSR B1534+12

About two years ago Wolszczan (1991) discovered a new binary pulsar system with relativistically interesting characteristics. Its orbit is less eccentric and slower than that of PSR B1913+16 (see Table 1), so most of its relativistic effects are smaller in magnitude. However, other circumstances contrive to make PSR B1534+12 an extremely attractive candidate for detailed study. The pulsar's radio signal is several times stronger than that of PSR B1913+16, and the pulse width is smaller. Consequently, in similar observations at Arecibo its TOAs can be measured about 5 times more accurately. In addition, the orbit of PSR B1534+12 is oriented more nearly edgewise to the line of sight ($\sin i > 0.97$, see Table 1), which greatly magnifies the Shapiro delay. PSR B1534+12 is also much closer to the Sun (approximately 0.7 kpc, compared with 7.1 kpc for PSR B1913+16; see Taylor & Cordes 1993), so contributions to measurable parameters from galactic kinematic effects are smaller and much easier to estimate.

The parameters of the PSR B1534+12 system measured by Taylor *et al* (1992) are listed in Table 1. For this pulsar, four PK parameters have already been measured with significance, and thus two distinct tests of relativistic gravity are available. The tests depend on the overall consistency of the parameter set, which is illustrated in Figure 2 by means of a plot of the constraints placed on the pulsar and companion masses, m_1 and m_2, by each of the measured PK parameters. The mass values $m_1 = 1.34 \pm 0.07$ and $m_2 = 1.34 \pm 0.07$ are consistent, within general relativity, with all of the directly measured parameters of PSR B1534+12 — and thus general relativity is found to be

fully consistent with the data. It is noteworthy that unlike the $\dot{\omega} - \gamma - \dot{P}_b$ test for PSR B1913+16, which is a mixed test of the strong-field and radiative aspects of gravity, the $\dot{\omega} - \gamma - r - s$ test for PSR B1534+12 is purely a strong-field test (Damour & Taylor 1992, Taylor *et al* 1992).

Experimental results of the sort described here will almost certainly be improved and extended in the near future. In a recent paper (Taylor 1992) I have shown that on a 10-year time scale two more binary pulsar systems, PSR B1534+12 and PSR B2127+11C, should yield $\dot{\omega} - \gamma - \dot{P}_b$ tests of gravitational radiation damping at the 1% level or better. By that time PSR B1534+12 will have at least 5 measurable PK parameters, providing cleanly separated probes of the coupling of matter to gravitational radiation and the behavior of gravity in very strong fields. Moreover, combining the results of timing measurements of several binary pulsars can provide even tighter theoretical constraints than available from each pulsar separately (Taylor *et al* 1992).

References

[1] Brumberg V A Zeldovitch Ya B Novikoff I D and Shakura N I 1975 *Soviet Astr. Lett.* **1** 2

[2] Damour T and Deruelle N 1985 *Ann. Inst. H. Poincaré (Physique Théorique)* **43** 107

[3] Damour T and Deruelle N 1986 *Ann. Inst. H. Poincaré (Physique Théorique)* **44** 263

[4] Damour T Gibbons G and Taylor J H 1988 *Phys Rev Letters* **61** 1151

[5] Damour T and Ruffini R 1974 *C. R. Acad. Sci. Paris* **A279** 971

[6] Damour T and Taylor J H 1991 *Astrophys J* **366** 501

[7] Damour T and Taylor J H 1992 *Phys. Rev.* D **45** 1840

[8] Esposito L W and Harrison E R 1975 *Astrophys. J.* **196** L1

[9] Hulse R A and Taylor J H 1975 *Astrophys J* **195** L51

[10] Lewandowski W and Thomas C 1991 *Proc. I. E. E. E.* **79** 991

[11] Manchester R N and Taylor J H 1977 *Pulsars* (San Francisco: Freeman)

[12] Ryba M F and Taylor J H 1991 *Astrophys J* **371** 739

[13] Shapiro I I 1964 *Phys Rev Letters* **13** 789

[14] Stinebring D R Ryba M F Taylor J H Romani R W 1990 *Phys Rev Letters* **65** 285

[15] Taylor J H 1991 *Proc. I. E. E.* **79** 1054

[16] Taylor J H 1992 *Phil. Trans Roy. Soc. London* **341** 117

[17] Taylor J H and Cordes J M 1993 *Astrophys J* in press

[18] Taylor J H Hulse R A Fowler L A Gullahorn G E and Rankin J M 1976 *Astrophys J* **206** L53

[19] Taylor J H and Weisberg J M 1982 *Astrophys J* **253** 908

[20] Taylor J H and Weisberg J M 1989 *Astrophys J* **345** 434

[21] Taylor J H Wolszczan A Damour T and Weisberg J M 1992 *Nature* **355** 132

[22] Thorsett S E Arzoumanian Z McKinnon M M and Taylor J H 1992 *Astrophys J* in press

[23] Wagoner R 1975 *Astrophys J* **196** L63

[24] Wolszczan A 1991 *Nature* **350** 688

Closed timelike curves

Kip S. Thorne

California Institute of Technology, Pasadena, California 91125, and
Institute for Theoretical Physics, UCSB, Santa Barbara, California 93106

Abstract. This lecture reviews recent research on closed timelike curves (CTCs), including these questions: Do the laws of physics prevent CTCs from ever forming in classical spacetime? If so, by what physical mechanism are CTCs prevented? Can the laws of physics be adapted in any reasonable way to a spacetime that contains CTCs, or do they necessarily give nonsense? What insights into quantum gravity can one gain by asking questions such as these?

1. Introduction

Much of the forefront of theoretical physics deals with situations so extreme that there is no hope to probe them experimentally. Such, largely, was the case nearly a century ago for Einstein's formulation of general relativity, and such is the case today for the attempt to quantize gravity. In these situations, thought experiments can be helpful. Of all thought experiments, perhaps the most helpful are those that push the laws of physics in the most extreme ways. A class of such thought experiments, which I and others have found useful in the last few years, asks [1] *What constraints do the laws of physics place on the activities of an arbitrarily advanced civilization?* In asking this question, we have in mind *all* the laws of physics that govern our universe, taken together—including those, such as quantum gravity, that are not yet well understood, and others, such as classical general relativity, that are, and with each set of laws holding sway only in its own domain of validity.

An especially fruitful question of this type is [1] *Do the laws of physics prevent arbitrarily advanced civilizations from constructing "time machines"* (machines for backward time travel), and if so, *by what physical mechanism are they prevented?* Hawking [2] has given the name *chronology protection* to the conjecture that there is such a mechanism, and that therefore *closed timelike curves* (CTCs) can never be created in the real Universe, no matter how hard advanced civilizations might try. In this lecture I shall review recent research on the chronology protection conjecture and related issues.

The laws of general relativity by themselves do not enforce chronology protection: it is easy to find solutions of the Einstein field equation that have closed timelike curves

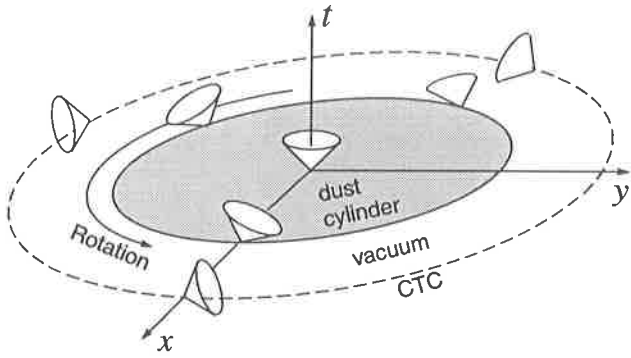

Figure 1. Van Stockum's spacetime.

(Section 2). However, the combination of general relativity's laws and the laws of quantum fields in curved spacetime may well provide a chronology protection mechanism—though we might not be sure of this until we understand the laws of quantum gravity much more deeply than today (Section 3).

Independently of whether chronology protection is correct, much insight into the laws of physics might be gained by studying how they behave in the presence of closed timelike curves (Section 4).

2. Spacetimes with closed timelike curves

A number of spacetimes with closed timelike curves have been exhibited in the literature, and much is now understood about the generic chronological structure of such spacetimes.

2.1. Spacetimes with eternal CTCs

The earliest example of a spacetime with CTCs is Van Stockum's 1937 solution of the Einstein field equation [4, 5], which represents an infinitely long cylinder made of rigidly and rapidly rotating dust. The dust particles are held out against their own gravity by centrifugal forces, and their rotation drags inertial frames so strongly that the light cones tilt over in the circumferential direction in the manner shown in Figure 1, causing the dashed circle in the figure to be a CTC. CTCs pass through every event in the spacetime, even an event on the rotation axis where the light cone is not tilted at all: one can begin there, travel out to the vicinity of the dashed circle (necessarily moving forward in t as one travels), then go around the cylinder a number of times traveling backward in t as one goes, and then return to the rotation axis, arriving at the same moment one departed. For the mathematical details of Van Stockum's solution see, e.g. Bonnor [5].

Physicists (but not science fiction writers) have generally dismissed Van Stockum's solution as "unphysical" because its source is infinitely long. Whether a finite-length, rotating body can also produce CTCs is not known; I shall return to this in Sec. 2.2.

A second old, famous example of a spacetime with CTCs is Gödel's solution of the Einstein equation [6], which describes a stationary, homogeneous cosmological model

with nonzero cosmological constant, filled with rotating dust. Again, the rotation tilts the light cones, creating CTCs. Because the spacetime is homogeneous and stationary, CTCs pass through every event. For the mathematical details of Gödel's spacetime, see, e.g. Hawking and Ellis [3], especially Figure 31.

Physicists have generally dismissed Gödel's solution as unphysical because it requires a nonzero cosmological constant and/or it doesn't resemble our own universe (whose rotation is small or zero).

2.2. Spacetimes with compactly generated chronology horizons

A spacetime whose CTCs are not eternal can be divided into *chronal regions* that are free of CTCs, and *nonchronal regions* that contain CTCs everywhere. The boundaries between the chronal and nonchronal regions are called *chronology horizons*; chronal regions end and CTCs are created at *future* chronology horizons; CTCs are destroyed and chronal regions begin at *past* chronology horizons.

A future chronology horizon is a special type of future Cauchy horizon, and as such it is subject to all the laws that govern any such horizon [3]; most importantly, it is generated by null geodesics that have no past endpoints but can leave the horizon when followed into the future. If the generators, when followed into the past, enter one or more compact regions $\mathcal{K}$ of spacetime and never thereafter leave them, the future chronology horizon is said to be *compactly generated* [2]; otherwise it is *non-compactly generated.* (If an arbitrarily advanced civilization were to create a time machine in a compact region of spacetime, then its chronology horizon obviously would be compactly generated.) A past chronology horizon is, in essence, the time reversal of a future one; and it therefore is generated by null geodesics that have no future endpoints but can leave the horizon when followed into the past.

When a future chronology horizon, at which CTCs arise, is compactly generated, its generators, followed into the past, can become confined into their compact region $\mathcal{K}$ in either of two ways: They can wander ergodically around $\mathcal{K}$ or some portion of it; or they can asymptote to one or more smoothly closed null geodesics in $\mathcal{K}$. Such smoothly closed null geodesics are called *fountains* because, when the generators are followed forward in time, they are seen to originate in the fountains and spew out of them, like streams of water, into the surrounding spacetime.

Hawking [2] has proved that in the generic case, $\mathcal{K}$ will contain such fountains, and it seems likely to me that generically all or almost all the horizon generators will emerge from them. Thus, fountains are generic, while ergodic wandering in $\mathcal{K}$ probably is not.

It is tempting to conjecture that, if a finite-sized, rapidly rotating body were to contract in some carefully designed, axially symmetric manner, it might create CTCs around itself in the manner of the Van Stockum solution without forming an event horizon. Figure 2 shows the chronological structure in the body's equatorial plane for such an evolution. At early times, when the body is large, there are no CTCs, so spacetime is chronal. At late times, when the body has settled down into its final, smaller state, the light-cone and chronology structures are similar to Van Stockum's solution, so spacetime is nonchronal. A future chronology horizon separates the chronal and nonchronal regions; it is shaped like a bowl with a small mountain in the center, i.e. like a "Mexican hat". The horizon's generators all originate on a single fountain. Two generators are shown. One, labeled A, spirals outward from the fountain and remains

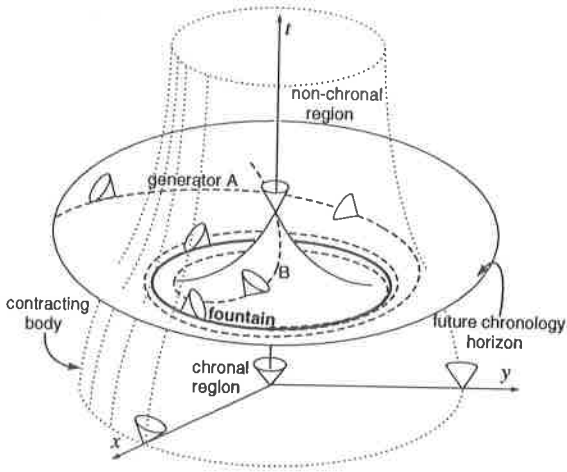

Figure 2. The chronological structure of a spacetime that might result from an axially symmetric contraction of finite sized, rotating body. This diagram is confined to the body's equatorial plane.

always on the horizon, eventually reaching future null infinity. The other, B, spirals inward, and then leaves the horizon at the tip of the Mexican hat. Hawking [2] has given mathematical details of a spacetime with this chronological structure.

In order for horizon generators to emerge from the fountain, there must be a net *defocusing* of any bundle of null geodesics that travels around the fountain. By the equations of geometric optics together with the Einstein field equation, this requires that

$$\oint_{\mathcal{F}} T_{\alpha\beta} l^\alpha l^\beta d\zeta < 0 \,, \tag{1}$$

where the integral is around the fountain $\mathcal{F}$, $T_{\alpha\beta}$ is the total stress-energy tensor for all matter and fields on $\mathcal{F}$, ζ is an affine parameter along $\mathcal{F}$, and $l^\alpha = dx^\alpha/d\zeta$ is the tangent to $\mathcal{F}$. Equation (1) says, in words, that *the averaged null energy condition (ANEC) must be violated around the fountain*; i.e., the integral in (1) must be negative. All ordinary, familiar forms of matter satisfy ANEC; therefore, no imploding body made of such matter can create CTCs in the manner of Figure 2. In Section 3 we shall return to the issue of whether ANEC can *ever* be violated, and shall learn that the answer is yes.

Lorentzian wormholes constitute a class of simple, explicit spacetimes that have generic-type, compactly generated chronology horizons [1, 7, 8]; as such, they have become a useful testbed for studies of chronology issues.

The simplest such wormholes are obtained by removing two balls from Euclidean space and identifying their surfaces in the manner of Figure 3a; the surfaces then become the wormhole's mouths. Such a wormhole necessarily violates ANEC: Any bundle of radially traveling null geodesics that passes through the wormhole is converging as it enters and diverging as it leaves, and therefore gets defocused by the wormhole, which means that $\oint T_{\alpha\beta} l^\alpha l^\beta d\zeta < 0$ along the bundle, with the negative contribution coming from a delta function $T_{\alpha\beta}$ at the junction between the two mouths. One can show more generally that every traversable wormhole, regardless of its shape or motion, violates ANEC [1].

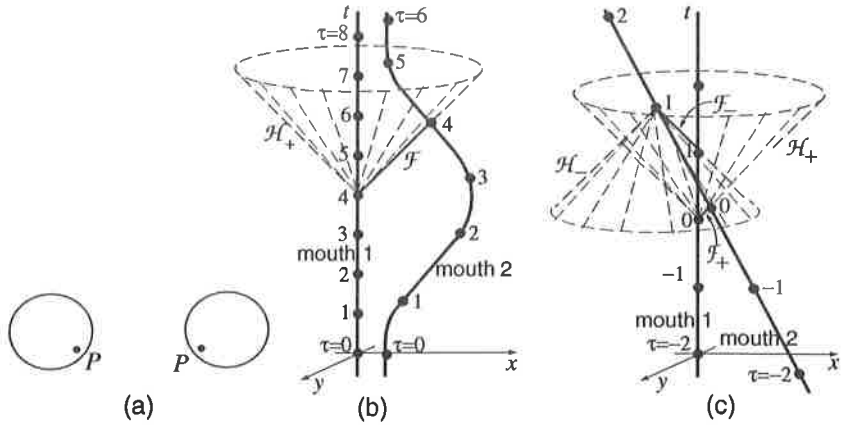

Figure 3. (a) A wormhole formed by removing two balls from Euclidean space and identifying their surfaces (the wormhole "mouths"); the identified points are reflections of each other in the midplane between the balls. (b) Chronology structure for a wormhole, one of whose mouths makes a "twins paradox trip." (c) Chronology structure for a wormhole whose mouths move past each other at uniform speed.

One can construct wormhole spacetimes, whose wormhole mouths travel along arbitrarily chosen world lines in Minkowski spacetime, by removing world tubes along those lines in a manner analogous to Figure 3a, and identifying their surfaces with each other [7]. Of course, the identification must be done in such a way that the intrinsic geometries of the two mouths' world tubes are the same. This may require a distortion of the spacetime geometry near the mouths if they are accelerated, but the distortion becomes vanishingly small in the limit that (acceleration)×(mouth radius)→ 0 [1, 7]. Since the mouths' intrinsic geometries are the same, the proper time interval $\Delta\tau$ between two identified neighboring events on the mouths must be same as seen through either mouth; it is this that dictates the form of the time markings in Figures 3b,c.

Figures 3b,c show the chronological structures of two wormhole spacetimes constructed in this way. In Figure 3b, one mouth remains at rest in a chosen Lorentz frame, while the other makes a "twins-paradox-type trip" into the external universe and returns [1]. As seen in the external universe there is a dilation of proper time on the moving mouth relative to the static one, but as seen through the wormhole there is no such time dilation. As a result, the relative motion of the mouths changes the manner in which time hooks up to itself through the wormhole. Initially the hookup is such that there are no CTCs; spacetime is chronal. After the trip, the hookup entails CTCs; spacetime is nonchronal. The future chronology horizon (denoted $\mathcal{H}_+$ in the figure) is the future light cone of the event $\tau = 4$ at the center of the right face of the left (static) mouth. The generators of this horizon (long-dashed lines) all originate in a single fountain (labeled $\mathcal{F}$): the smoothly closed null geodesic that travels from $\tau = 4$ on the left mouth to $\tau = 4$ on the right mouth, then through the infinitesimally short wormhole and back to where it started; cf. Figure 10 of Ref. [7].

In Figure 3c, one mouth moves past the other, creating CTCs that are confined to a

bounded nonchronal region of spacetime: the region that begins at the future chronology horizon $\mathcal{H}_+$ and ends at the past horizon $\mathcal{H}_-$. The generators of $\mathcal{H}_+$ (the future light cone of $\tau = 0$ on the static mouth) emerge from the fountain $\mathcal{F}_+$ and leave the horizon when they pass through $\mathcal{H}_-$. The generators of $\mathcal{H}_-$ (the past light cone of $\tau = 1$ on the moving mouth) enter the horizon at its intersection with $\mathcal{H}_+$ and ultimately asymptote to the fountain $\mathcal{F}_-$.

Figures 3b,c are prototypes for the consequences of generic relative motions of a wormhole's mouth: such motions will always produce CTCs [1], as will the gravitational redshifts that result from placing a wormhole in a generic external gravitational field [8]. Most physicists react to this by asserting that the laws of physics must prevent the existence of classical, traversable wormholes—perhaps by forbidding the existence of material that violates ANEC ("exotic material").

Not all compactly generated chronology horizons have the generic "generators-emerge-from-fountains" structure of Figures 2 and 3. An example that is different is Taub-NUT space [9]—a vacuum solution of the Einstein equation with a spatially compact chronal region, followed by a compact future chronology horizon, followed in turn by a non-compact nonchronal region with CTCs. Being a vacuum solution, Taub-NUT space satisfies ANEC and also satisfies the local null energy condition (NEC), $T_{\alpha\beta}l^\alpha l^\beta \geq 0$ everywhere. This means that the generators of the chronology horizon cannot peel off of fountains in the manner of Figures 2 and 3. Instead, every generator is itself a fountain (a smoothly closed null geodesic).

A simpler spacetime with this special type of chronology horizon is Misner space [10]. The relevant variant of Misner space can be obtained as follows: go into your bedroom in Minkowski spacetime, identify the back wall with the front wall (so when you walk into the back you find yourself emerging from the front), similarly identify the floor with the ceiling and the left wall with the right, and set the right wall moving toward the left. In other words, Misner space is Minkowski spacetime with identification under translations along y and z, and under a boost along x. Figure 4a shows the x-t portion of this spacetime. It initially is chronal, and then becomes nonchronal at a future chronology horizon whose generators are the closed null geodesics (fountains) y =const, z =const, $x = t - 2$.

Hawking [2] shows that *every* fountain $\mathcal{F}$ on a compactly generated chronology horizon must have a non-positive ANEC integral, $\int_{\mathcal{F}} T_{\alpha\beta}l^\alpha l^\beta \leq 0$. The generic case where generators peel off the fountain (Figs. 2 and 3) corresponds to "< 0" for this integral and thus to a violation of ANEC; the special Taub-NUT and Misner cases correspond to "$= 0$". Hawking points out that as soon as one allows energy of any sort to flow through the Taub-NUT or Misner horizon, or through any other horizon whose fountains have $T_{\alpha\beta}l^\alpha l^\beta = 0$ everywhere, that energy flow will carry a nonzero local value of $T_{\alpha\beta}l^\alpha l^\beta$, and therefore in order to keep the ANEC integral nonpositive, the local null energy condition (NEC) must be violated somewhere along each perturbed fountain. This means that on any physically realistic, compactly generated chronology horizon, NEC must be violated, even if ANEC is not. (This conclusion strengthens an earlier result due to Tipler [11].)

2.3. Spacetimes with non-compactly generated chronology horizons

The simplest example of a spacetime with a non-compactly generated chronology horizon is Grant space [12], which is a slight generalization of Misner space. Go into your bedroom

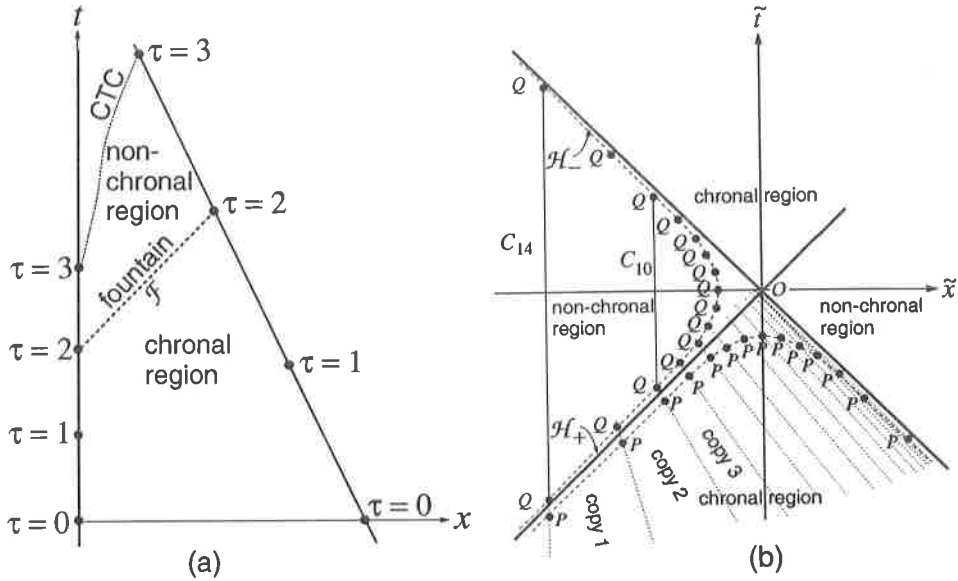

Figure 4. (a) The t-x portion of Misner space. (b) The covering space for Misner space (with the points P identified and Q identified) and for Grant space (with each successive P or Q identified only after a displacement by distance a into the paper).

in Minskowski spacetime, translate the left wall relative to the right by a distance a, and then identify them and set the right wall moving toward the left with speed β; the result is Grant space. In other words Grant space is Minkowski spacetime identified along the x direction then boosted in x and translated in y. Misner space (without identification in y and z) is the same as Grant, but with vanishing y translation ($a = 0$).

Figure 4b is the t-x portion of the covering space for Grant space. In this covering space, a sequence of copies of Grant space (labeled "copy 1", "copy 2", etc.) are lined up side by side, each one boosted by β with respect to the previous one. The (fictitious) wall at which the boosts occur is shown dotted. The events P and Q lie on the wall, with each successive copy of P or Q displaced into the paper by a distance a relative to the preceding one.

With the aid of Figure 4b, one can show that the translation along y does not change the location of the chronology horizon; it is the dark line labeled $\mathcal{H}_+$ for $a = 0$ (Misner space) and also for a finite (Grant space). To show this for finite a, we need only demonstrate that through any event Q which is in the claimed nonchronal region but arbitrarily close to the claimed $\mathcal{H}_+$, there passes a CTC. One can connect Q to itself by many geodesics C_n, with each one circling around Grant space a different number of times, n. The figure depicts projections of the geodesics C_{10} and C_{14} on the t-x plane; they also extend distances $10a$ and $14a$ down the y axis (into the paper). Since C_{14} has twice as long a temporal duration $\Delta\tilde{t}$ in the covering space as C_{10}, but only makes 40% more trips around Grant space and thus has only a 40% longer extent in the y direction, $dy/d\tilde{t}$ is $1.4/2.0 = 0.7$ as large on C_{14} as on C_{10}. As one goes to an ever larger number n

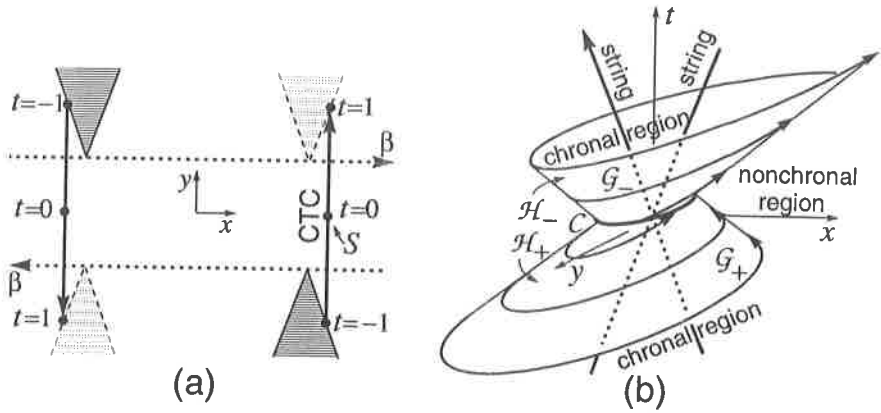

Figure 5. (a) Spatial diagram of Gott space. (b) Spacetime diagram showing the chronological structure of Gott space. For ease of visualization, the wedges are not removed around the strings, and to compensate for this, the chronology horizons and generators are shown as curved rather than flat.

of traversals, $dy/d\tilde{t}$ gets ever smaller (as one can readily show by a detailed calculation), so that eventually, for a sufficiently large number n of traversals, $dy/d\tilde{t}$ is small enough for the geodesic C_n to be timelike.

Although the translation $a \neq 0$ leaves the location of the horizon unchanged, it alters radically the character of the horizon generators. For $a = 0$ (Misner space), the generators are all smoothly closed null geodesics (fountains) and the horizon is compactly generated. For $a \neq 0$, the generators originate at past null infinity in the covering space and, never intersecting themselves or each other, they travel to the covering space's spacetime origin $\mathcal{O}$. Correspondingly, after the identification that produces Grant space, the generators travel around and around Grant space an infinite number of times without ever intersecting themselves or each other. When followed to the past, through an infinite affine parameter, they never leave the future chronology horizon $\mathcal{H}_+$. When followed to the future, after a finite affine parameter and an infinite number of circuits, they reach the end of $\mathcal{H}_+$ and leave it.

A famous example of a spacetime with a noncompactly generated chronology horizon is Gott space [13], which is a solution of the Einstein field equation representing two infinitely long, parallel, straight cosmic strings that move past each other at high speed. Figure 5a depicts the strings in a spatial diagram; they are at the vertices of the wedges and extend into and out of the paper (z direction) infinitely far. The figure is drawn in the strings' mean rest frame; the upper string moves rightward at speed β and the lower, leftward at speed β. Each string is surrounded by a flat but conical spatial geometry, which can be obtained by removing the indicated wedge from Euclidean space and identifying its edges. The identification is synchronous in the rest frame of the string, which means that for the upper, rightward moving string the event labeled $t = 1$, at Lorentz time $t = 1$, is identified with that labeled $t = -1$, at Lorentz time $t = -1$, and similarly for the lower string.

For a suitable choice of parameters, the dark vertical line in the diagram is a CTC.

It begins at the "starting point" labeled S, at time $t = 0$ in the mean rest frame, when the two strings are just passing each other. It moves upward to meet the right edge of the upper string's wedge at $t = 1$. It passes through the wedge, emerging at $t = -1$ when the right string was near the left edge of the diagram. It then travels downward to meet the lower, left-moving string's wedge at $t = 1$, passes through the wedge emerging at $t = -1$ when the wedge was near the right edge of the diagram, and then travels upward to its starting event S.

Cutler [14] has deduced the chronological structure of Gott space. It is shown in Figure 5b, topologically correctly but not geometrically correctly. (The geometrically correct depiction, shown in Fig. 3 of Cutler's paper, takes some work to decipher because of the string wedges that are removed; their removal permits the chronology horizons and their generators to be flat planes and straight null lines instead of curved surfaces and curved lines as here.) The future chronology horizon $\mathcal{H}_+$ has null geodesic generators $\mathcal{H}_+$ that originate at spatial infinity and, spiraling around and around the moving strings, work their way in to the closed spacelike geodesic $\mathcal{C} = \mathcal{H}_+ \cap \mathcal{H}_-$, where they leave $\mathcal{H}_+$. The past chronology horizon $\mathcal{H}_-$ is generated by null geodesics $\mathcal{G}_-$ that enter $\mathcal{H}_-$ at $\mathcal{C}$, and then spiral their way around and around the moving strings until they reach spatial infinity. The CTCs are confined to the nonchronal region outside $\mathcal{H}_+ \cup \mathcal{C} \cup \mathcal{H}_-$, which means that they are bounded away from the strings. In the mean rest frame—indeed, in any Lorentz frame—the horizons extend to temporal infinity; thus, at all times there are CTCs, but at arbitrarily early or late times they are confined to arbitrarily large radii.

If one parallel transports a set of vectors around the strings and back to their starting event, the local Lorentz transformation that relates the returning vectors to the starting vectors is called the *holonomy* of Gott space. One can similarly transport vectors around the closed x-dimension of Grant space and compute the resulting holonomy. Grant [12] has shown that for suitable choices of parameters, the two holonomies, that of Gott space and that of Grant space, are identical; and he has argued that this, plus the fact that both spaces are flat (except at the string locations) implies that Grant space must actually be the same as a portion of Gott space.

Because of its translation and boost invariance in the z direction, Gott's two-string space can be regarded as a solution of the $2 + 1$ dimensional vacuum Einstein equation for two point masses moving past each other at high speed. This has enabled Deser et. al. [15] and Carroll et. al. [16] to infer, using ideas from the 2+1 dimensional theory, that, despite the fact that each of the strings moves at less than the speed of light, taken together they have a *tachyonic* total momentum. Stated more precisely, the strings' holonomy (in a suitable Lorentz frame) is a pure boost and not a rotation, and this implies in the 2+1 theory that their total momentum is spacelike. Deser et. al. and Carroll et. al. argue that this means Gott space is *unphysical* within the framework of 2+1 dimensional theory, and it suggests to them that Gott space might also be unphysical (not creatable by realizable initial conditions) in our real 3+1 dimensional universe. (For further interesting results on CTCs in 2+1 dimensional, point-particle spacetimes, both spatially closed and spatially open, see the references in Carroll et. al. [16].)

Two other arguments have been used to cast doubt on cosmic strings as generators of CTCs: Gott [13], by order-of-magnitude estimates, suggests that, if one tries to make CTCs by the relative motion of two *curved* strings, the strings' energies in their center of mass frame will become so great that they *might* form a black hole around themselves before the CTCs can arise. More firmly and convincingly, Hawking [2] points

out that finite loops of cosmic string, by themselves, cannot create CTCs because their stress-energy tensor satisfies NEC, and any physically realizable, compactly generated chronology horizon must violate NEC (see the end of Sec. 2.2 above).

Another famous vacuum solution of the Einstein equation that has a non-compactly generated chronology horizon is Kerr spacetime. The exterior of Kerr's outer horizon ($r > r_+$ in the usual notation) and the region between the outer and inner horizons ($r_+ > r > r_-$) are chronal; the inner horizon ($r = r_-$) is a chronology horizon; and the region inside there ($r < r_-$) is nonchronal. It is conventionally argued that, although the chronal region is likely to occur in our real universe as the exterior and interior of an old rotating black hole, the spacetime near and inside the chronology horizon will be altered by an instability due to infalling, blueshifted perturbations; and this alteration (hopefully) will prevent CTCs from arising [17].

3. Is there a Chronology Protection Mechanism?

The examples in Section 2 show that, according to classical general relativity, a wide variety of circumstances can give rise to CTCs. What attitude should a physicist take to this? The most common attitude is to assert that all such circumstances are *unphysical*: Infinitely long, rotating cylinders are unphysical and (presumably) finite ones will not produce CTCs; our universe does not rotate as fast as the Gödel universe, so Gödel is unphysical; traversable wormholes are unphysical; infinitely long, straight cosmic strings are unphysical; ....

I do not find such assertions at all satisfying. Physicists' past records in labeling various things as unphysical are not good. For example, Oppenheimer, Wheeler, and others in the 1930s through the 1950s claimed on physical grounds that the trace of the stress-energy tensor cannot be negative, and therefore superdense matter cannot have a pressure that exceeds 1/3 its energy density. They were wrong, as Zel'dovich showed in the early 1960s by a simple quantum-field-theory model, and nowadays several plausible equations of state for nuclear matter entail $T_\alpha{}^\alpha < 0$. As another example, it was widely asserted several decades ago that negative energy densities are unphysical, but we now know they are not: quantum field theory predicts negative renormalized energy densities under a variety of circumstances—e.g. in the Casimir vacuum between two electrically conducting plates and in squeezed vacuum states of light, both of which are realized in the laboratory.

This poor record cautions us to keep an open mind about CTCs until we have found a concrete chronology protection mechanism (or mechanisms): a mechanism that will prevent CTCs from arising under *all* conceivable circumstances—e.g. when a hypothetical arbitrarily advanced civilization is using all means at its disposal to produce CTCs. It seems likely to me that the search for such a firm chronology protection mechanism may teach us much about the laws of physics.

It would be rather surprising to me if Nature uses one protection mechanism in one situation (e.g., collapsing, spinning bodies), a different one in another situation (e.g., moving cosmic strings), and a third mechanism in a third situation (e.g., the interior of a spinning black hole). More likely there is one universal mechanism, that always does the job if other mechanisms fail. (Visser [18] has argued for a number of universal

mechanisms, i.e. a "defense in depth" against CTCs. On this I am agnostic; I would be happy to find just one firm, universal mechanism).

In the following subsections I shall discuss the three mechanisms that have seemed most promising in recent years: An enforcement of NEC or ANEC by quantum field theory, a classical instability of future chronology horizons, and a quantum-field instability.

3.1. Enforcement of NEC or ANEC

Because compactly generated chronology horizons across which energy flows must always violate NEC, if we knew that NEC is always enforced by the laws of physics, then we could rule out CTCs ever being generated in compact regions of spacetime. This might not be a fully universal chronology protection mechanism, but it would come close.

Unfortunately, quantum field theory—the ultimate arbiter of obedience to energy conditions—insists that NEC *can* be violated; for example, it is violated in the Casimir vacuum and in squeezed states of light.

It may well be that in the real universe, for reasons that we do not yet understand firmly, compactly generated chronology horizons must be of the generic sort illustrated in Figure 2 (contracting, rotating cylinder) and Figure 3 (wormholes): the horizon generators emerge from fountains that necessarily violate ANEC. This makes enforcement of ANEC an attractive possible chronology protection mechanism.

With this motivation, there has been considerable effort in the last several years to determine quantum field theory's attitude toward ANEC. It has been shown that ANEC is enforced for noninteracting quantized scalar and electromagnetic fields in Minkowski spacetime [19, 20], and in generic, curved 1+1 dimensional spacetimes [21]. On the other hand, in 3+1 dimensions (the real universe), both nontrivial topology [19] and spacetime curvature [21] can induce ANEC violations. Indeed, as Wald and Yurtsever have shown, there are generic classes of curved spacetimes in which quantum fields violate ANEC.

It could still turn out that ANEC is enforced under all circumstances where CTCs try to form, thereby protecting chronology; for example, it might be impossible for quantum fields ever to produce the specific ANEC-violating stress-energy tensors that are required to hold a wormhole open, and therefore wormhole-based CTCs might be forbidden. However, the fact that ANEC *can* be violated under a wide class of generic situations suggests to me that ANEC enforcement is *not* a very promising, universal chronology protection mechanism.

3.2. Classical instability of future chronology horizons

The future chronology horizons in the Kerr, Taub-NUT, and Misner spaces are infamously classically unstable. Particles or fields falling into a Kerr black hole, or traveling around the spatially closed Taub or Misner space, become infinitely blue shifted as they near the horizon; and it seems reasonable to hope that the resulting divergent energy density will always act back on the spacetime, via the Einstein equation, to prevent the CTCs from forming.

The example of Misner space is depicted in Figure 6a: A high-frequency electromagnetic wave packet moving along the solid world line gets blue shifted by a factor $\xi \equiv \sqrt{(1+\beta)/(1-\beta)}$ with each passage around the "universe", and it traverses the universe an infinite number of times as it nears the chronology horizon's fountain, thereby

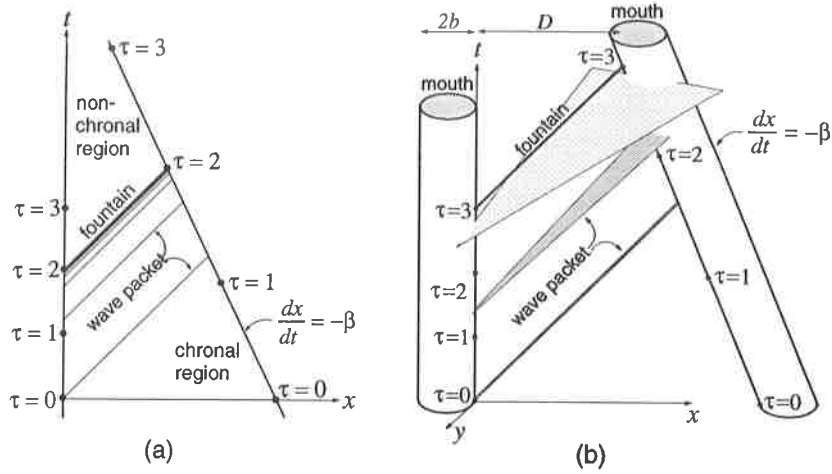

Figure 6. (a) The motion of a high-frequency electromagnetic wave packet in Misner space. (b) The same wave packet in a wormhole spacetime where CTCs are forming.

piling up on itself in spacetime and producing a divergent energy density just before the chronology horizon forms.

Until a few years ago, it was widely thought that such instabilities *always* occur at future chronology horizons, thereby protecting chronology. However, wormhole spacetimes provide a counterexample [1], and generalizing this result, Hawking [2] has shown that a generic subset of all compactly generated chronology horizons are counterexamples: they are all classically stable.

A wormhole counterexample is depicted in Figure 6b. This wormhole spacetime is identical to Misner space (Fig. 6a), except that Misner's identified flat walls are converted into spherical "walls" (the wormhole's mouths). The high-frequency wave packet still gets blue shifted by a factor $\xi \equiv \sqrt{(1+\beta)/(1-\beta)}$ with each circuit through the wormhole, and still tries to pile up on itself at the fountain. However, the wormhole's ANEC-induced diverging-lens action causes the wave packet to spread laterally, driving its amplitude down by a factor $b/2D$ with each circuit (where b is the wormhole radius and D the distance between the mouths, as seen in the left mouth's reference frame, when the horizon forms). If $(b/2D)\xi < 1$, i.e. if the distance between the mouths is large enough, then the packet's energy density decreases with each circuit, and the total energy density at the horizon remains finite, despite the pileup. Chronology is not protected—at least not by this mechanism.

3.3. Quantum-field instability of future chronology horizons

Our greatest hope—indeed, it seems, a very realistic hope—for universal chronology protection lies in a quantum field instability of all future chronology horizons. This instability was first discovered in 1982, in the context of Misner space, by Hiscock and Konkowski [22]. After Morris, Yurtsever, and I discovered that Misner space's *classical* instability is removed by curving its walls (i.e. by going to a wormhole spacetime) [1], we presumed the same would be true of the quantum instability. We were wrong, as

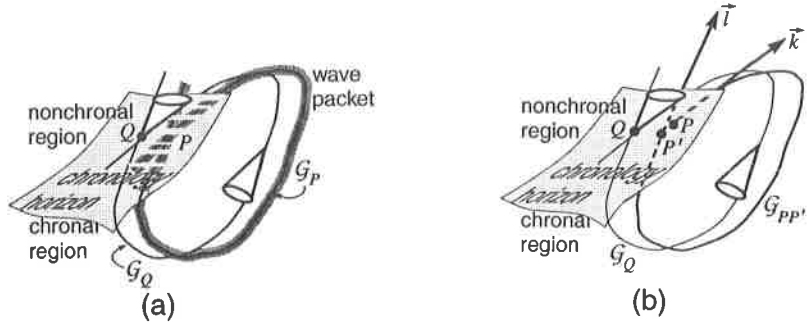

Figure 7. (a) Heuristic explanation of the quantum-field instability of the chronology horizon. (b) Geometric construction for point-splitting computation of the renormalized stress-energy tensor, which induces the instability.

Kim and I [23], Frolov [24], and Gnedin and Kompaneets [26] all independently realized in 1989. The quantum instability is universal; it must arise at every location on every future chronology horizon in any spacetime.

This instability can be described heuristically as due to a piling up of the vacuum fluctuations of any quantum field in the vicinity of any chronology horizon. This pile up causes the fluctuations to have a nonzero renormalized energy density (nonzero *vacuum polarization*) that diverges as one approaches the horizon. The diverging energy density in turn, via the semiclassical Einstein equation, *might* distort the spacetime geometry in such a way as to protect chronology. I shall return to the "might" in the next subsection.

To understand this heuristic explanation in greater detail, consider an arbitrary location on any future chronology horizon $\mathcal{H}_+$. Since the horizon is the dividing line between a region with CTCs and one with none, arbitrarily close to $\mathcal{H}_+$, on its nonchronal side, there is an event Q through which passes a CTC. As one pushes Q closer to $\mathcal{H}_+$, the CTC through Q becomes more nearly null, then null, and then a null geodesic that I shall call $\mathcal{G}_Q$; see Figure 7a. This $\mathcal{G}_Q$ travels from Q around a closed loop and back to Q, but in general does *not* return pointing in the same direction as it started; for this reason it is sometimes called a *self-intersecting null geodesic.*

Now, let P be an event very close to Q, but on the chronal side of $\mathcal{H}_+$. There will be an almost closed null geodesic $\mathcal{G}_P$ that starts out at P, travels along nearly the same route as $\mathcal{G}_Q$, and returns very near P, but cannot quite close itself up at P because P is in the chronal region. High-frequency wave-packet modes of any massless quantum field can travel along this $\mathcal{G}_P$; the world tube of such a mode is shown as a dark strip in Figure 7a. The closer P is to the chronology horizon $\mathcal{H}_+$, the closer will $\mathcal{G}_P$ come to closing up on itself, and if it comes close enough, then the wave packet, with its finite size, will pile up on itself in spacetime near P, and its piled-up vacuum fluctuations will interfere with themselves in such a way as to produce a nonzero energy density after renormalization.

As P is pushed closer and closer to $\mathcal{H}_+$ (and thence to Q), $\mathcal{G}_P$ comes closer and closer to closing up on itself, and correspondingly modes of higher and higher frequency manage to pile up on themselves, with each contributing a nonzero amount to the vacuum polarization. With more and more modes of higher and higher frequency contributing, the renormalized energy density grows larger and larger in magnitude, as P is pushed

up to $\mathcal{H}_+$.

This heuristic picture has been justified by a point-splitting calculation of the renormalized stress-energy tensor $T^{\alpha\beta}$ for a quantized, noninteracting, massless scalar field $\hat{\phi}$ [23, 24]. The single point P is split into two points P and P' (Fig. 7b), and the field's regularized Hadamard function $G^{(1)}_{\rm reg}(P,P') = \langle\hat{\phi}(P)\hat{\phi}(P')+\hat{\phi}(P')\hat{\phi}(P)\rangle$ is evaluated. The dominant contribution—one due to vacuum fluctuations and therefore independent of the state of the field—comes from scalar-wave propagation of $\hat{\phi}(P)$ and thence $G^{(1)}$ around routes close to the null geodesic $\mathcal{G}_P$ (see Refs. [23] and [24] for careful justifications); it has the usual Hadamard normal form

$$G^{(1)}_{\rm reg} = \frac{\Delta^{1/2}}{4\pi^2}\left(\frac{1}{\sigma_{PP'}}+\frac{1}{\sigma_{P'P}}\right) . \tag{2}$$

Here Δ is the Van Vleck-Morette determinant, which measures the amount of focusing ($\Delta > 1$) or defocusing ($\Delta < 1$) that occurs around $\mathcal{G}_P$ or equally well around $\mathcal{G}_Q$; $\sigma_{PP'}$ is the geodetic interval along the spacelike (but nearly null) geodesic $\mathcal{G}_{PP'}$ (Fig. 7b) that leads from P to P' by a route that is very close to the null geodesic $\mathcal{G}_P$; and $\sigma_{P'P}$ is the interval along the similar route that begins at P' and ends at P. More specifically, $\sigma_{PP'} = \int_{\mathcal{G}_{PP'}} g_{\alpha\beta}(dx^\alpha/d\zeta)(dx^\beta/d\zeta)d\zeta$, with ζ an affine parameter that goes from 0 at P to 1 at P', and similarly for $\sigma_{PP'}$. The closer is P (and thus also P') to Q (which was arbitrarily close to the horizon), the closer will the σ's be to zero, and thus the larger will be the Hadamard function.

The renormalized stress-energy tensor is computed from the Hadamard function by the standard point-splitting relation

$$T_{\mu\nu} = \ell^2_{\rm Pl}\lim_{P'\to P}\left(\frac{2}{3}\nabla_\mu\nabla_{\nu'} - \frac{1}{3}\nabla_\mu\nabla_\nu - \frac{1}{6}g_{\mu\nu}\nabla_\alpha\nabla^{\alpha'}\right)G^{(1)}_{\rm reg} , \tag{3}$$

where $\ell_{\rm Pl}$ is the Planck length and $G = c = 1$ so $\ell_{\rm Pl} = \sqrt{\hbar}$. The dominant contribution to $T_{\mu\nu}$ comes from differentiating twice the nearly-zero and sharply varying σ's; the result is

$$T_{\mu\nu} = -\frac{\Delta^{1/2}}{6\pi^2}\frac{\ell^2_{\rm Pl}}{\sigma^3}(2k_\mu l_\nu + 2l_\mu k_\nu + k_\mu k_\nu + l_\mu l_\nu + g_{\mu\nu}l_\alpha k^\alpha) . \tag{4}$$

Here $k_\alpha = -\lim_{P'\to P}\nabla_\alpha\sigma_{PP'}$ is the outgoing tangent to $\mathcal{G}_{PP}$ or equally well to the self-intersecting null geodesic $\mathcal{G}_Q$, and $l_\alpha = +\lim_{P'\to P}\nabla_\alpha\sigma_{PP'}$ is the returning tangent; see Figure 7b. For Q arbitrarily close to the horizon (as we have assumed), σ becomes arbitrarily small as P approaches the horizon, while Δ, l_μ, and k_ν remain finite; and therefore the renormalized stress-energy tensor becomes arbitrarily large.

The divergence is actually a little more complicated than this, because there is an infinite sequence of events Q_1, Q_2, Q_3, ... in the nonchronal region that asymptote to any chosen event on $\mathcal{H}_+$, and each of which is connected to itself by a self-intersecting null geodesic $\mathcal{G}_{Q_n}$. Each of these events Q_n with its own $\mathcal{G}_{Q_n}$ gives rise to a term of the form (4) in the renormalized stress-energy tensor, and the total stress-energy tensor is a sum over all these contributions

$$T^{\mu\nu} = -\sum_{n=1}^{\infty}\frac{\Delta_n^{1/2}}{6\pi^2}\frac{\ell^2_{\rm Pl}}{\sigma_n^3}(2k_n^\mu l_n^\nu + 2l_n^\mu k_n^\nu + k_n^\mu k_n^\nu + l_n^\mu l_n^\nu + g^{\mu\nu}l_{n\alpha}k_n^\alpha) . \tag{5}$$

One can see that there is such an infinite sequence of Q's by the following argument [23]: Take the original Q, and construct a new causal curve that connects this Q to itself

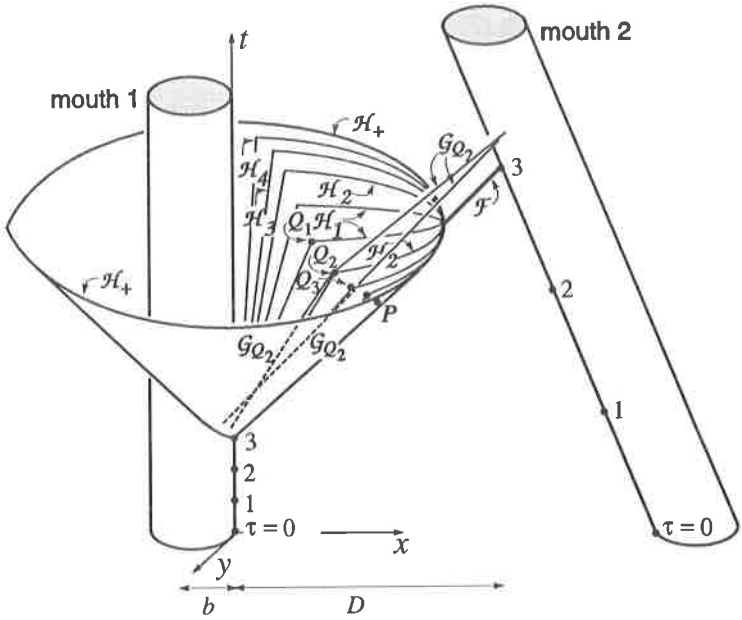

Figure 8. The future chronology horizon $\mathcal{H}_+$ and polarized hypersurfaces $\mathcal{H}_n$ for the simple wormhole spacetime of Figure 6b. The kink in the left edge of $\mathcal{H}_+$ is due to a caustic there.

by traveling around the original $\mathcal{G}_Q$ twice. That curve (call it $\mathcal{G}_{Q_2}$) has a kink in its middle (a discontinuous jump from the null direction l_α to k_α). By moving Q nearer the horizon (and giving it the new name Q_2), while keeping $\mathcal{G}_{Q_2}$ null, one forces the kink to smooth out and converts $\mathcal{G}_{Q_2}$ into a null geodesic. Repeating the process indefinitely, one obtains a subsequence of the infinite sequence of Q's alluded to above.

Each Q_n lies on a distinct hypersurface $\mathcal{H}_n$ made of events that are connected to themselves by self-intersecting null geodesics; and these $\mathcal{H}_n$, which are called *polarized hypersurfaces*, asymptote to the future chronology horizon $\mathcal{H}_+$ in the limit $n \to \infty$. We can regard the order-n term in the vacuum-polarization stress-energy tensor (5) as produced by the presence nearby of the n'th polarized hypersurface $\mathcal{H}_n$, with its specific event Q_n and associated self-intersecting null geodesic $\mathcal{G}_{Q_n}$.

The wormhole spacetime of Figure 8 provides an example [23]. The polarized hypersurfaces $\mathcal{H}_1$, $\mathcal{H}_2$, $\mathcal{H}_3$, $\mathcal{H}_4$, ... are nested, one inside the next, within the chronology horizon. The event $\mathcal{Q}_n$ on $\mathcal{H}_n$ is connected to itself by a null geodesic $\mathcal{G}_{Q_n}$ that traverses the wormhole n times, and that gives rise to the order-n term in the stress-energy tensor (5) at the event P.

Explicit evaluations of the vacuum fluctuational stress-energy tensor (5) have been carried out near the fountain of the wormhole spacetime of Figure 8 by Kim and Thorne [23] (and for arbitrarily slow wormhole motions, $\beta \to 0$, by Visser [18]), near the fountain of other wormhole spacetimes by Frolov [24], near the fountain of a generic, compactly generated horizon by Klinkhammer [25], and near the non-compactly-generated horizon

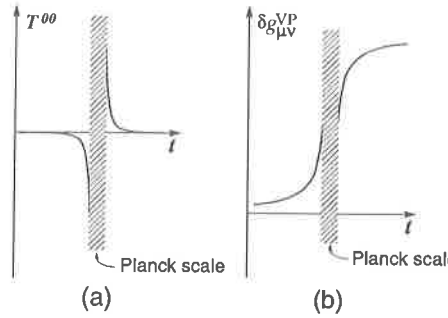

Figure 9. The vacuum-polarization-induced stress-energy tensor, and the metric perturbations it produces as one passes through $\mathcal{H}_+$ or $\mathcal{H}_n$.

of the Grant/Gott spacetime by Grant [12]. A central issue in these and all other cases is this:

3.4. Does the back-action of the vacuum-polarization energy protect chronology?

By inserting the $T_{\mu\nu}$ of Eq. (3) into the semiclassical Einstein equation and performing a rough order-of-magnitude integration, one obtains an estimate for the metric perturbations created by the vacuum-polarization energy of the quantized scalar field (and also of any other nongravitational field):

$$\delta g^{\rm VP}_{\mu\nu} \sim G^{(1)}_{\rm reg} \sim \sum_n \Delta_n^{1/2} \frac{\ell^2_{\rm Pl}}{\sigma_n} \,. \tag{6}$$

Although these metric perturbations diverge at the chronology horizon and at each subsequent polarized hypersurface, the divergences can be remarkably slow—so slow that it is conceivable, under some circumstances, that quantum gravity will invalidate the above analysis before the spacetime has been altered substantially [23]. Note that, as one passes through the n'th polarized hypersurface, σ_n passes through zero and reverses sign. Correspondingly, if quantum gravity were simply to smooth out the divergences in Eqs. (5) and (6), one would see the vacuum polarization produce, on the nonchronal side of $\mathcal{H}_+$ and $\mathcal{H}_n$, a $T_{\mu\nu}$ equal in magnitude but opposite in sign to that on the chronal side. Therefore, as observers approach and then pass through $\mathcal{H}_+$ or $\mathcal{H}_n$, they might see this $T_{\mu\nu}$ first distort the spacetime geometry, and then undo the distortions it had produced, as illustrated in Figure 9.

Such a scenario is highly speculative, but seems to me plausible *if* the divergence is sufficiently weak.

Just how strong is the divergence? The most important place to ask this question is at the chronology horizon $\mathcal{H}_+$ rather than at a polarized hypersurface $\mathcal{H}_n$, because that is where CTCs first arise. For a compactly generated $\mathcal{H}_+$, the divergence is much stronger at the horizon's fountains than away from the fountains. This is because the polarized hypersurfaces $\mathcal{H}_n$ (at which the $\sigma_n = 0$) are all tangent to $\mathcal{H}_+$ at any fountain $\mathcal{F}$, but each $\mathcal{H}_n$ is finitely separated from $\mathcal{H}_+$ away from the fountains [23]; see Figure 8 for an example. If, as I have conjectured in Section 2.2, almost all the horizon generators originate on fountains, then a divergence that is strong enough at the fountains to distort

the spacetime geometry significantly there will have its influence propagate over the entire horizon and perhaps thereby protect chronology everywhere. Conversely, if the divergence is too weak to protect chronology at the fountains, it probably is too weak to protect chronology elsewhere.

As an example, consider the wormhole spacetime of Figure 8, as examined by an observer who sits on the left mouth and on the x-axis, where the fountain $\mathcal{F}$ arises. Assuming the mouth speed β is not too close to zero or one, at a time Δt before crossing the horizon this observer sees a value $\sigma_n \sim D\Delta t$ for the geodetic interval along the closed, spacelike geodesic that traverses the wormhole n times, and he sees $\Delta_n^{1/2} \sim (b/2D)^{n-1}$ for the amount of defocusing around that geodesic [23]. (There is no net defocusing on the geodesic's first trip because it first passes through the wormhole—the "diverging lens"—only at the end of that trip; however, the first wormhole passage produces a net defocusing of $b/2D$ during the second trip, the second wormhole passage produces another defocusing $b/2D$ during the third trip, etc.) Correspondingly, the spacetime distortion measured on the wormhole throat is

$$\delta g_{\alpha\beta}^{\rm VP} \sim \frac{\ell_{\rm Pl}^2}{D\Delta t} \sum_n \left(\frac{b}{2D}\right)^{n-1} . \tag{7}$$

When Kim and I first computed this back-action of the vacuum polarization, it seemed to me to be extremely weak. "Surely," I said to myself, "the analysis will break down during a time interval $\Delta t \sim \ell_{\rm Pl}$ around the passage through the horizon (the 'Planck region' of Fig. 9), since 'time' does not make classical sense on such short scales." If this were true, and quantum gravity were to smooth out the divergence, then the metric distortion just before smooth-out would be much too small, $\delta g_{\alpha\beta}^{\rm VP} \sim \ell_{\rm Pl}/D \sim 10^{-35}$ for $D \sim 1$ meter, to protect chronology.

Hawking [2] has convinced me that this assessment is wrong [23, 27]. The distance D between the mouths and the time Δt until the Chronology horizon depend on the observer's reference frame, he points out, but the product $D\Delta t$ does not (as one can see from the fact that $\sigma_n \sim D\Delta t$ is an invariant). Therefore, he conjectures, it may well be that the spacetime remains classical, near the chronology horizon, and the computed $\delta g_{\mu\nu}^{\rm VP}$ of Eq. (7) remains correct, until the product $D\Delta t$ gets as small as $\ell_{\rm Pl}^2$, and correspondingly $\delta g_{\mu\nu}^{\rm VP}$ reaches unity. The resulting distortion of the classical spacetime geometry might then be sufficient to protect chronology, Hawking speculates.

If Hawking were right, and the relevant Planck region were $D\Delta t \sim \ell_{\rm Pl}^2$, then there is a strategy that an arbitrarily advanced civilization could use to circumvent chronology protection. The civilization need only make sure that the fountain encounters two or more widely separated regions of defocusing, instead of only one as in Figure 8. This could be done, for example, by using two wormholes to make CTCs, with time through each wormhole synchronously identified in the wormhole's rest frame, and the two wormholes moving in the manner of Figure 10. (Such a spacetime was suggested to Mike Morris and me several years ago by Tom Roman.) The fountain would have the indicated form, the defocusing as measured on a wormhole mouth would be $\Delta_n^{1/2} \sim (b/2D)^{2n-1}$ rather than the $(b/2D)^{n-1}$ of Figure 8, and thus for a large wormhole separation, $D \gg b$, the resulting metric perturbation on the mouth (which is dominated by $n = 1$) would be $\delta g_{\mu\nu}^{\rm VP} \sim (b/2D)(\ell_{\rm Pl}^2/D\Delta t)$ [28]. By making $b/2D$ arbitrarily small, the advanced civilization could force $\delta g_{\mu\nu}^{\rm VP}$ to be arbitrarily small at the beginning of Hawking's conjectured Planck

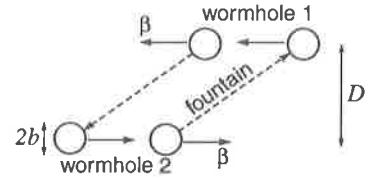

Figure 10. Spatial diagram of Roman's wormhole spacetime.

region, $D\Delta t \sim \ell_{\rm Pl}^2$. If quantum gravity were then to provide a cutoff and smooth-out, chronology would not be protected.

Even weaker is the vacuum-polarization divergence in Gott/Grant spacetime, as computed by Grant [12]. Since the chronology horizon is non-compactly generated, it has no fountains, and every polarized hypersurface is everywhere separated from the horizon by a finite distance. (More specifically, in the covering space of Figure 4b, the n'th polarized hypersurface $\mathcal{H}_n$ consists of events connected to themselves by self-intersecting null geodesics such as $\mathcal{C}_{10}$ of the figure, that circle around the $\tilde{x}$-direction of the universe n times; this $\mathcal{H}_n$ is the hyperbola $\tilde{x}^2 - \tilde{t}^2 = a^2 n^2 \xi^{-n}(1-\xi^{-n})^{-2}$, where $\xi = \sqrt{(1+\beta)/(1-\beta)}$ is the blue shift produced by each propagation around the universe and a is the y-translation that makes the horizon non-compactly generated.) When Grant sums over the vacuum polarization contributions from all the polarized hypersurfaces, each one finitely displaced from the horizon but approaching the horizon in the limit $n \to \infty$, and when he then computes the resulting back action on the spacetime metric, he obtains

$$\delta g_{\alpha\beta}^{\rm VP} \sim \frac{\ell_{\rm Pl}^2}{a^2} \ln\left(\frac{\tilde{t}^2 - \tilde{x}^2}{a^2}\right). \tag{8}$$

Although this metric perturbation diverges as one approaches the horizon, $\tilde{t} \to \tilde{x}$, the divergence, being logarithmic in time with a coefficient $\ell_{\rm Pl}^2/a^2$ that can be made arbitrarily small, is extremely weak. It is even harder here than in the Roman spacetime to see how such a divergence can protect chronology.

Nevertheless, I suspect that it may do so. It may well be that quantum gravity invalidates the semiclassical analysis only when metric fluctuations, treated as a spin-two field on the classical background, develop mean-square fluctuations of order unity *as a result of the same pileup process as induces the vacuum polarization of nongravitational fields.* If so, then the semiclassical analysis, just before *every* chronology horizon, might remain valid up to the location where $\delta g_{\mu\nu}^{\rm VP} \sim 1$, and only then fail.

To determine whether this is so, and to determine the nature of the subsequent evolution of spacetime, will require an understanding of quantum gravity. Indeed, it may be that efforts to decipher these issues will teach us useful things about quantum gravity.

4. Physics in the presence of closed timelike curves

It may turn out that on macroscopic lengthscales chronology is *not* always protected, and even if chronology *is* protected macroscopically, quantum gravity may well give

finite probability amplitudes for microscopic spacetime histories with CTCs [29]. For these reasons, some effort has been devoted recently to exploring whether and how the laws of physics might adapt themselves to CTCs [7]. In this concluding section, I shall summarize very sketchily what has been learned.

The cleanest of such explorations are carried out in spacetimes, such as Figure 3c, that have a chronal "IN" region, followed by a compact nonchronal region, followed by a chronal "OUT" region. Initial data are posed in the IN region for some physical system, and the system is then evolved from the IN region through the nonchronal region and into the OUT region. The evolutionary laws are generally chosen to be the most conservative possible—the same laws, at least locally, in the nonchronal region as one is accustomed to in everyday, chronal physics—and one asks whether the evolution problem is well posed, i.e. whether standard initial data in the IN region produce a unique evolution through the chronal region and into the OUT region.

For *noninteracting, classical* systems (particles [7] and fields [30]) the answer appears to be yes; there does exist a unique evolution. However, just as interactions produce evolutionary problems in science fiction (e.g., one can go back in time and kill one's younger self), so also interactions produce trouble for classical particles [31] and presumably also for classical fields: One finds that a large number of classical evolutions can follow from a single, standard set of initial data. It was thought, at first, that for some initial data there might be *no* self-consistent evolutions; but thus far no clean examples of such a thing have been exhibited in classical, continuum physics [31, 32]. (On the other hand, there *are* examples in simple, highly idealized, discrete models [33].)

Of course, physics is quantum mechanical at heart, not classical, and it is in the quantum domain that these studies become especially fruitful. Just as in quantum cosmology, where there is *no a priori* notion of "time", so also in nonchronal spacetimes, where CTCs alter the nature of time, the only viable approach to quantum mechanics seems to be Feynman's sum over histories. Indeed, spacetimes with CTCs have become a useful testbed for the sum-over-histories formulations of quantum theory that are being developed for use in quantum cosmology [34].

It turns out that for nonrelativistic particles [35] and also for relativistic fields [37], the sum-over-histories formalism enables one to compute unique probabilities for the outcomes of all measurements that one might reasonably try to make, even in the nonchronal region of spacetime. However, when the particles or fields are self-interacting, their interactions produce peculiar phenomena: (i) the propagators from the IN region to the OUT region are not unitary—but nevertheless, there is no loss of probability [35, 36, 37]; and (ii) although one recovers standard Hamiltonian quantum mechanics in the chronal OUT region, one does not recover it in the chronal IN region, and the fact that CTCs exist to the future of the IN region influences probabilities in the IN region itself [37]. The strength of this influence, and how it grows as one approaches the future chronology horizon, are not as yet understood.

In summary, these studies are giving us glimpses of how CTCs influence physics; but whether those glimpses are teaching us something deep and important, or we are just playing fun mental games, is far from clear.

Acknowledgments

The author's research on this topic was supported by the National Science Foundation Grant AST-8817792 until February 1992. Since then, both the Astronomy and the Physics Divisions of NSF have declined to support this research, so it has been supported solely by Caltech's Feynman Research Fund. This manuscript was written at the Institute for Theoretical Physics in Santa Barbara, California, with support from National Science Foundation Grant PHY-8904035.

References

[1] Morris M S, Thorne K S, and Yurtsever, U 1988 *Phys. Rev. Lett.* **61** 1446–1449

[2] Hawking S W 1992 *Phys. Rev. D* **46** 603–611

[3] Hawking S W and Ellis G F R 1973 *The large scale structure of space-time* (Cambridge: Cambridge University Press)

[4] Van Stockum W J 1937 *Proc. Roy. Soc. Edin.* **57** 135–154

[5] Bonnor W B 1980 *J. Phys. A: Math. Gen.* **13** 2121–2132

[6] Gödel K 1949 *Rev. Mod. Phys.* **21** 447

[7] Friedman J, Morris M S, Novikov I D, Echeverria F, Klinkhammer G, Thorne K S, and Yurtsever U *Phys. Rev. D* **42** 1915–1930

[8] Frolov V P and Novikov I D 1990 *Phys. Rev. D* **42** 1057-1065

[9] Taub A H 1951 *Ann. Math.* **53** 472
Newman E T, Tamburino L, and Unti T J 1963; *J. Math. Phys.* **4** 915–923.

[10] Misner C W 1967 in *Relativity Theory and Astrophysics I. Relativity and Cosmology* ed J Ehlers (Providence RI: American Mathematical Society) pp. 160–169
Thorne K S 1993 in *Directions in General Relativity* eds BL Hu et al (Cambridge: Cambridge University Press)

[11] Tipler F J 1977 *Ann. Phys.* **108**, 1–36.

[12] Grant J D E 1993 *Phys. Rev. D* in press

[13] Gott J R 1991 *Phys. Rev. D* **66** 1126–1129

[14] Cutler C 1992 *Phys. Rev. D* **45** 487–494

[15] Deser S, Jackiw R., and /t Hooft, G 1992 *Phys. Rev. Lett.* **68** 267–269

[16] Carroll S M, Farhi E, and Guth, A H 1993 *Phys. Rev. D* in press

[17] Ori A 1992 *Phys. Rev. Lett.* **68** 2117 and references therein

[18] Visser M 1993 *Phys. Rev. D* **47** 554–565

[19] Klinkhammer G 1992 *Phys. Rev. D* **43** 2542–2548

[20] Folacci A 1993 *Phys. Rev. D* **47** 2726–2729

[21] Wald R M and Yurtsever U 1991 *Phys. Rev. D* **44** 403–416

[22] Hiscock W A and Konkowski D A 1982 *Phys. Rev. D* **26** 1225–1230

[23] Kim S-W and Thorne K S 1991 *Phys. Rev. D* **43** 3929–3947

[24] Frolov V P 1991 *Phys. Rev. D* **43** 3878–3894

[25] Klinkhammer G 1992 *Phys. Rev. D* **46** 3388–3394

[26] Gnedin N N 1991 private communication
Kompaneets D A 1991 private communication.

[27] Thorne K S 1991 *Ann. New York Acad. Sci.* **631** 182–193.

[28] Lyutikov M 1993 research in progress.

[29] Friedman J 1992 in *Proceedings of the 4th Canadian Conf. on General Relativity and Relativistic Astrophysics* eds G Kunstatter et al (Singapore: World Scientific) pp. 183–199

[30] Friedman J L and Morris M S 1991 *Phys. Rev. Lett.* **66** 401–404

[31] Echeverria F, Klinkhammer G, and Thorne, K S 1991 *Phys. Rev. D* **44** 1077–1099

[32] Novikov I D 1992 *Phys. Rev. D* **45** 1989–1994; Lossev A and Novikov I D 1992 *Class. Quant. Grav.* **9** 2309–2321

[33] Deutsch D 1991 *Phys. Rev. D* **44** 3197–3217

[34] Hartle J B 1993 "Unitarity and causality in generalized quantum mechanics for acausal spacetimes" in preparation

[35] Politzer H D 1992 *Phys. Rev. D* **46** 4470–4476

[36] Boulware D G 1992 *Phys. Rev. D* **46** 4421–4441

[37] Friedman J L, Papastamatiou N J, and Simon J Z 1992 *Phys. Rev. D* **46** 4442-4455, and **46** 4456–4469

Paper presented at the 13th Int. Conf. on General Relativity and Gravitation Cordoba, Argentina, 1992: Part 1, Plenary Lectures

The present status of general relativity

Robert M. Wald

Enrico Fermi Institute and Department of Physics, University of Chicago, 5640 S. Ellis Ave., Chicago, IL 60637, USA

Abstract. The present status of research in classical general relativity, cosmology, and quantum gravity is discussed, and some prospects for future developments in these fields are indicated.

1. Introductory Remarks

It is customary at the end of a meeting of this sort to have a "Conference Summary" talk, to aid the participants in distilling the key new ideas that have been presented here. To give a comprehensive, balanced Conference Summary is an extremely difficult task. Fortunately, I have the advantage at this meeting that my presentation does not even pretend to be a Conference Summary. Rather, this contribution represents my best attempt to summarize the current status and trends of research in the broad subject area covered by this meeting. This will be done from my own personal perspective. I state this obvious fact to emphasize that, although none of my comments are intended to be frivilous, there is no reason for anyone else to take them seriously. In particular, if I felt it likely that any of my remarks would be used to give a stamp of approval (or disapproval) to any given line of research, I probably would have refused to give this talk.

At least one other disclaimer should be made before I begin. I cannot pretend to keep up with all developments even in the areas in which I have done substantial research, no less the very broad area I am attempting to review here. My intention is to review only the basic trends, with an eye toward developments in the field which may be expected in the not too distant future. Names of individual researchers will be kept to an absolute minimum, and references will be limited to other plenary talks at this meeting. The reader who wishes to obtain a comprehensive survey of recent research results in general relativity would be much better served by systematically browsing through all the other contributions to this volume than by reading this contribution.

For the purposes of this talk, I define the term "general relativity" to mean the topics that the people who come to a GR meeting do research on. This probably is best reflected by the subjects covered in the workshops at this meeting. I shall organize this review by dividing it into three main categories: (1) Classical general relativity, (2) Cosmology, and (3) Quantum gravitational physics.

2. Classical General Relativity

General relativity has had a rather strange history. It was formulated more than 75 years ago, and was immediately recognized as being both "beautiful" and "deep". Within a very short time after its formulation, some of its key predictions were confirmed; specifically, the 1919 eclipse expedition confirmed the "light-bending" prediction, and the calculation of the perihelion precession of Mercury accounted for the previously observed "residual" precession of that planet. Furthermore, some key exact solutions (particularly, the Schwarzschild solution) were discovered almost immediately, which could have enabled researchers to investigate some of the new "strong field" phenomena predicted by the theory. With all of these factors working in its favor, it seems remarkable, in retrospect, that so little attention was paid to the theory over the next forty years or so. There were, of course, several notable investigations concerning gravitational collapse and the nature of (homogeneous, isotropic) cosmology, but there appears to have been very little attempt to really take the theory seriously by working out its predictions and consequences in a systematic way. Indeed, general relativity appears to have had more the status of a mathematical curiosity than of a theory of the physical world during this period, and vestiges of this attitude persist even today.

One of the reasons contributing to the unusual status of general relativity in physics is the unusual relationship general relativity has had throughout most of its history between theory and experiment/observations. Most of the dramatic and exciting advances in physics occur when the experimentalists are one step ahead of the theorists, finding new phenomena that challenge the theorists either to find an explanation within existing theory (thereby probing its structure more deeply) or to modify the existing theory. Such a relationship between theory and experiment is particularly healthy when the existing theory is only a partial one, since experiments and observations then serve to define the limits of the theory and suggest appropriate generalizations. The decades of research in particle physics preceding the formulation of the present-day "standard model" of electroweak and strong interactions provides an excellent example of this kind of vigorous interplay.

However, general relativity was "born whole" and – unless it is wrong – its limits, presumably, are defined by the Planck scale, which is entirely inaccessible to direct observation. Throughout most of its history, the contact of general relativity with experiment and observation (apart from cosmology, to be discussed in the next section) has been limited to "solar system" tests: light bending, Mercury's precession, the gravitational redshift, and (within the past 20 years) the gravitational time delay. These tests have been extremely important for validating the theory. The ever increasing precision of these observations – such as the confirmation of the light bending prediction of general relativity to within .3% achieved with long baseline interferometry – should be appreciated and applauded by all general relativists. New experiments planned for the future – such as the measurement of geodetic precession – will provide further tests. However, these "solar system tests" have not, as yet, posed any significant challenges to the theorists: The approximation schemes needed to derive the predictions are quite straightforward, and the data has been in beautiful accord with the theory. Unless they eventually demonstrate that general relativity is wrong, I do not believe that these tests are likely to have much impact upon the direction and progress of the field in the foreseeable future.

However, in other arenas, there is a good chance that theorists will be challenged by experiments and observations in the not to distant future, even assuming that all future experiments and observations will yield results consistent with the predictions of general relativity. To some extent, this is happening already with the high precision observations of binary pulsar systems. Observations of the original Hulse-Taylor binary pulsar system dramatically confirmed the existence of gravitational radiation as predicted by general relativity. Perhaps even more significantly for the interplay of theory and observation, the effects being measured for these systems are of a sufficiently "strong field" nature that the approximation schemes needed to derive them are not straightforward. Indeed, even the approximation which leads to the standard "quadrupole formula" for gravitational radiation (and corresponding back-reaction) is nontrivial to justify rigorously in the case of self-gravitating systems. The binary pulsar observations should continue to provide a stimulus to theorists to develop better and more rigorously justified approximation schemes.

Observations of millisecond pulsars also may result in significant interplay with general relativity theory in the foreseeable future. Already, the rotation rates of the fastest pulsars – when interpreted within the framework of general relativity – are not far from providing significant restrictions on the equation of state of matter composing neutron stars. It is quite possible that with better statistics afforded by observation of more millisecond pulsars, an upper limit on rotation rates (as expected from the instability of rotating stars in general relativity) will be deduced. When combined with additional information about neutron stars obtained from other observations, we stand a good chance of learning new things about the strong field behavior of general relativity, as well as about the properties of matter at nuclear densities.

As I shall comment further upon in the next section, observational astronomy has undergone a revolution in the past decade, and new and higher quality information about astrophysical systems is likely to continue to be obtained at a rapid rate. In particular, it seems safe to predict that in the foreseeable future, important new observations will be made of binary X-ray sources within our galaxy, of the central regions of the nucleus of our galaxy and nearby galaxies, and/or of quasars and other active galactic nuclei. Such observations may yield some stringent tests to the models for these systems which involve black holes. The discovery of some strikingly new phenomenon in these systems would probably afford us the best opportunity we have of learning more about phenomena occurring in the strong field regime of general relativity. Indeed, it was the original discovery of quasars in the early 1960's that provided the first real stimulus to systematically study the strong field predictions of general relativity. This stimulus probably was largely responsible for what I consider to be the "golden era" of classical general relativity – the period from the mid-1960's to the early 1970's when "global methods" were formulated, the singularity theorems were proven, and the theory of black holes was developed.

Perhaps the best opportunity of all for vigorous interaction of general relativity theory with experiment will be attained if the new generation of gravitational wave detectors – presently under construction – succeed in achieving the sensitivity for which they are ultimately designed. The goal of obtaining an unambiguous detection of gravitational radiation would then easily be met. However, for the same reason as the solar sytem tests, by itself this probably would not have a much impact upon the field – unless, of course, the observed characteristics of the radiation differ measureably from the

predictions of general relativity. Indeed, the binary pulsar observations have already confirmed the existence of gravitational radiation as predicted by general relativity, with a quantitative precision much greater than any direct detection is likely to attain. Rather, the true potential impact upon the field arises from the possibility of doing "gravitational wave astronomy". Observations of the wave forms of gravitational radiation signals – possibly in conjunction with observations of electromagnetic signals emitted by the same sources – would provide the kind of challenges and stimuli to theorists that could lead to major advances in our understanding of phenomena involving strong gravitational fields.

However, despite the above remarks, it would be quite optimistic to believe that any significant interplay between general relativity theory and experiments and observations will occur within this decade. Thus, for the foreseeable future, the development of classical general relativity is likely to continue to be driven mainly by the study of "old problems", as well as by new developments in cosmology and quantum gravity.

Of the "old problems", there is one which, in my view, stands out as dominant both on account of its fundamental importance and because of the possibility that some significant progress can be made within the coming decade: the nature of singularities. During the "golden era" of classical general relativity alluded to above, mathematical techniques of differential geometry were used to establish the existence of singularities in solutions to Einstein's equations in a wide variety of circumstances relevant to gravitational collapse phenomena and to cosmology. In these theorems, Einstein's equation (together with energy conditions on matter) is used only to obtain inequalities on the Ricci curvature. The fact that the detailed properties of Einstein's equation do not play much role in the singularity theorems is largely responsible for their great power and generality: The occurrence of singularities cannot be evaded by modifying the matter content (provided that the energy conditions are still satisfied) or even by a wide class of modifications of Einstein's equation itself. However, this generality is probably the root cause of the one significant deficiency of the singularity theorems: For the most part, they say nothing about the nature of the singularities they predict apart from the fact that some inextendible causal geodesic must be incomplete.

The study of the nature of singularities in classical general relativity is of fundamental importance for at least two reasons. First, it is crucial for understanding strong field behavior. The physical relevance of black holes is entirely premised upon the hypothesis that the singularities resulting from gravitational collapse are confined to black holes, i.e., that "naked singularities" do not occur. If this "cosmic censor hypothesis" should turn out to be false, our present beliefs concerning strong field phenomena in general relativity would undergo drastic modifications; indeed, even my statement above that Planck scale phenomena are inaccessible to direct observation probably would be wrong. It should be recalled that at the present time, support for belief in the validity of the cosmic censor hypothesis comes entirely from some linear perturbation analyses and from the beauty and internal consistency of the theory of black holes, rather than from analysis of the general behavior of solutions to Einstein's equation in situations corresponding to gravitational collapse. Similarly, in cosmology one would like to understand whether initial singularities generically have a "spacelike character" (so that horizons are present in the early universe) and the manner in which the Ricci and/or Weyl curvatures diverge near an initial cosmological singularity.

Secondly, an analysis of the nature of singularities would provide a major step

toward our understanding of the manner in which classical general relativity breaks down. It could provide some important clues as to the kinds of new phenomena that might occur in a quantum theory of gravitation.

Singularities are present and can be studied in detail in some of the simplest and most basic solutions in general relativity, such as the Schwarzschild solution and the Robertson-Walker models. However, it is quite possible that these simple solutions may give a very misleading picture of the general properties of singularities. Clearly, what is required is a very general analysis of properties of solutions to Einstein's equation. As already indicated above, it does not appear that the "global methods" used to prove the singularity theorems can be pushed much further to enable a detailed description of their properties. However, I believe it likely that progress can be made on this issue on two broad fronts.

First, as described in the contribution of Evans, it appears that "numerical relativity" is finally coming of age. It now is feasible to reliably study via numerical simulation the dynamical evolution of nonspherical spacetimes in which gravitational collapse to a singularity occurs. Some interesting examples already have been obtained by Shapiro and Teukolsky, as reported in Teukolsky's contribution. The study of the evolution of spacetimes without any symmetries imposed (i.e., with "4-dimensional codes") may be possible in the foreseeable future. Of course, numerical codes tend to break down near singularities at a much earlier stage than the breakdown of classical general relativity itself, so it is not likely that we will be able to learn about the detailed structure of singularities from numerical experiments. However, at the very least, numerical experiments should be able to provide strong hints concerning strong field behavior near singularities. They also should be able to stringently probe the validity of the cosmic censor hypothesis.

Secondly, the global properties of solutions to Einstein's equation are now being studied with modern methods of partial differential equations. As discussed further in the contribution of Klainerman, the main results obtained thus far are "nonsingularity theorems" – in particular, a proof that globally nonsingular solutions to Einstein's equation exist for all initial data sufficiently close to flat spacetime (i.e., that singularities cannot be created starting with weak gravitational waves). These methods will have to be developed considerably further in order to have a chance at obtaining general results on the properties of singularities in general relativity. However, the advances in this area which have been made in the past decade are quite encouraging, and they hold out hope that analysis of such issues as the validity of cosmic censorship may be possible in the not too distant future.

My extended discussion of the issue of the nature of singularities should not be interpreted as indicating a belief that there are no other issues in classical general relativity worthy of intensive study. However, I do not have space here to discuss these issues, and I fear that any short list of problems which I might attempt to compile would end up being most notable for its inadvertant omissions. Thus, I will simply remark that a substantial portion of my own recent research has been on other issues in classical general relativity, and I have every expectation that this will continue in the future.

The most promising source of new ideas and issues in classical general relativity is the research efforts in the "border areas" of cosmology and quantum gravity. I now turn to a discussion of the first of these topics.

3. Cosmology

The field of cosmology has undergone very significant development in the past two decades. In my view, the most remarkable – and certainly the most solid – of its achievements has been the (in my opinion, convincing) demonstration of the success of the "standard cosmological model" – i.e., the Friedman-Robertson-Walker (FRW) solution with matter in thermal equilibrium in the early universe – in accounting for the basic features of our universe from the era of nucleosynthesis onwards. Thirty years ago, the main reasons for believing in the validity of the "standard model" were the following: (1) Its assumed homogeneity and isotropy appeared to be in good (rough) agreement with the observed distribution of the galaxies. (2) It accounted nicely for the Hubble expansion. (3) The relationships between the observed values of Hubble's constant, the "deceleration parameter", the age of the universe, and the mass density of the universe were in (very rough) accord with the predictions of the model – at least, after serious errors in the determination of Hubble's constant were corrected.

The discovery of the cosmic microwave background in 1965 provided dramatic further evidence in favor of the "standard model." Such "relic" radiation is naturally predicted by the model, and it is very difficult to find other plausible explanations for its existence, Planckian spectrum, and isotropy.

Another strong piece of evidence in favor of the standard model became evident by the late 1960's. The standard model predicts that a substantial amount of He^4 should have been synthesized in nuclear reactions beginning several seconds after the "big bang" singularity of the model, and ending about 15 minutes later. The predicted abundance of He^4 produced in this manner is in excellent agreement with observations. Since far more He^4 is produced in this manner than plausibly could have been produced in stars, the agreement of the predictions with observed helium abundance is not something that could be easily accounted for in other ways.

Today, we probably have, if anything, less grounds than thirty years ago for advancing reason (1) above in support of the FRW models. As discussed further below, redshift surveys during the past decade have enabled the determination of the "three dimensional" distribution of galaxies, and very significant departures from homogeneity have been observed on much larger scales (at least $\sim$50 megaparsecs) than anticipated thirty years ago. The status of the observational support for reasons (2) and (3) above has not changed significantly in the past thirty years.

Nevertheless as already indicated, evidence in support of the standard model has been enormously strengthened in the past two decades. One reason is the greatly improved precision of the measurements of the microwave background radiation. As far as can be determined by COBE, the spectrum of the microwave background radiation is exactly Planckian, and only very recently has COBE finally has detected some tiny departures from exact isotropy. This has provided strong support for believing that this radiation is, indeed, the "relic radiation" predicted by the standard model.

However, perhaps the strongest new evidence for the standard model has come from new observations (and experiments!) related to nucleosynthesis occurring in the early universe. In addition to He^4, trace amounts of H^2, He^3, and Li^7 also are predicted to be synthesized. The predicted abundance of these elements depends sensitively on the baryon density, and the first measurements of deuterium abundance twenty years ago were used primarily to estimate the baryon density – yielding the result

that baryons provide about 5% of the mass density that a flat FRW model would have. However, the measurement of additional light element abundances provides a test of the model itself. The fact that the primordial abundances of these elements inferred from observations during the past decade also agree with the predictions of nucleosynthetic calculations is strong support for the standard model (as well as for baryon density $\sim$5% of the "closure density", i.e., $\Omega_B \sim .05$) . Second, the percentage of He^4 which is synthesized is not very sensitive to the baryon density but is sensitive to the expansion rate of the universe during the era of nucleosynthesis. This expansion rate, in turn, depends upon the matter content, and calculations within the standard model showed that the presence of more than 3 light neutrino species in thermal equilibrium in the early universe would yield too high a helium abundance. The existence of only 3 light neutrino species has now been confirmed by the experiments on the decay of the Z^0 conducted at CERN.

Of course, if some discrepancies between calculations and observations had been found, the theorists surely would have found plausible ways of modifying the standard model so as to reproduce the observations (or would have found plausible reasons for rejecting or re-interpreting the observations). However, the fact that the model now has stood up to quite a number of nontrivial, quantitatively precise tests with little or no "fudging" should be taken very seriously. Any present or future difficulties associated with supplementary hypotheses to the standard model – particularly with regard to the origin of the departures from homogeneity – should not be confused with the remarkable success of the essential features of the model. To repeat the advice of a particle experimentalist who had just paid off his wager with a cosmologist on the number of neutrino species: "Don't bet against the big bang!"

It is worth mentioning that the success of the standard cosmological model has had an unfortunate side effect for general relativists. It does not require much knowledge of general relativity to write down the Robertson-Walker line element. Had difficulties with the standard model arisen, the possibility of curing them by going to anisotropic or inhomogeneous models would have been explored, and general relativists undoubtedly would have played a leading role in these efforts. Since such efforts have not been necessary, general relativists have stayed largely on the sidelines, and not much stimulation to the field of general relativity has resulted from these developments in cosmology.

Given the success of the standard model, it is only natural that attempts would be made to extrapolate its description of the universe back to earlier epochs, when the energy and density of matter was much higher than accessible or testable in laboratory conditions. Such research necessarily is of a speculative nature, and most of the research in cosmology during the past decade has been of this speculative kind, concerned primarily with phenomena which may have occurred at times corresponding to the grand unification scale (i.e., $t \sim 10^{-35}$ sec.) or even the Planck scale ($t \sim 10^{-43}$ sec.). Some innovative ideas have been introduced, and some interesting developments have taken place, most notably in theory of inflation and ideas concerning the formation of cosmic strings and other "topological defects."

It would take me too far afield to discuss these ideas here. However, I do wish to briefly comment upon what I view as a shortcoming of the nature of some of the theoretical research activity in this area. In my opinion, insufficient distinction often is drawn between "physical issues" and "metaphysical issues." (Here, by "metaphys-

ical issue" I mean simply an issue lying outside the nature and scope of present-day theories of physics; I do not intend the negative connotation usually implicit in the use of that term by physicists.) To illustrate this point, consider the following two widely discussed problems which often are discussed as though they were on the same footing: the "monopole problem" and the "flatness problem." The "monopole problem" refers to the prediction of grossly overabundant monopole production in the early universe, assuming that matter is described by a grand unified field theory. It is very much a "physical problem", i.e., from well defined initial assumptions and well defined physical laws, one obtains a prediction inconsistent with observation. Its solution must be sought in abandonment or modification of the grand unified model, or in a mechanism to dilute the monopole density (such as inflation), or in the modification of other cosmological assumptions. On the other hand, the "flatness problem" refers to the fact that in standard FRW models, in the early universe the spatial curvature must have been enormously smaller than the energy density of matter; equivalently, the lifetime of our universe is enormously larger than the Planck time or any of the fundamental timescales of elementary particle theory. This problem has much more of a metaphysical character. It is not at all clear that there is any "problem" at all, except possibly "naturalness" (and the creator of the universe might well have a rather different concept of "naturalness" than we do!). A "solution" to this "problem" presumably consists of a model where the conditions of the early universe arise in a manner which seems less artificial to us. Many of the other widely discussed problems of cosmology also have a similar metaphysical component. Our civilization has made enormous progress in the development of physical theories, but I am not at all confident that we are better equipped than the ancient Greek philosophers or medieval scholars to deal with metaphysical issues. It is very important to the progress and direction of science to raise metaphysical issues; important breakthroughs can be made by attempting to address issues which lie outside the scope of present physical laws. Furthermore, the distinction between physical and metaphysical issues is not always entirely clear cut. However, in my view, a serious effort to draw these distinctions as sharply as possible would be very helpful for clarifying the goals and aims of research in cosmology.

Without question, the area of research in cosmology which presently is undergoing the most active development and where many further exciting developments can be anticipated in the near future concerns the "origin of structure", i.e., the processes which led to the formation of galaxies and clusters of galaxies. These developments have been fueled by what can best be termed a revolution in observational astronomy. Without question, the most important single technological advance involved in this revolution was the replacement of the photographic plate by charge-coupled devices (CCD's), which came into wide use beginning about a decade ago. Present day CCD's have a photon detection efficiency of order 75% (as compared with $\sim$1% for photographic plates), and their digital character makes it possible to do accurate and efficient subtractions of the night sky background. As a direct consequence of the advent of CCD's, several orders of magnitude less observing time is required to obtain galactic redshifts than was possible using photographic plates. This has made it possible to take redshift surveys which are much more complete, much deeper, and not nearly as subject to selection effects as prior surveys. As already mentioned above, this enables one to obtain reliable "3-dimensional" maps of the distribution of galaxies, and the surveys taken thus far have shown significant departures from homogeneity and the

presence of coherent structures in galactic clustering on much larger scales than had been anticipated. A large number of very ambitious new redshift surveys are currently in progress or planned for the near future, so in the coming decade, we will obtain a great deal of new information concerning the large scale clustering of galaxies.

Other major advances also are occurring in our ability to make observations relevant to cosmology. A new generation of telescopes is being built with adaptive optics, which will greatly improve resolution and could enable phenomena like supernovae in progenitor galaxies to be studied. A large number of satellites have been launched or will soon be launched, giving us a capacity to probe the universe much more fully and in more detail in the microwave, infra-red, visible, ultraviolet, X-ray, and gamma ray regimes. It would be rather surprising if some dramatic new discoveries do not arise from the wealth of information we will obtain from these sources in the coming decade.

This new observational data undoubtedly will provide stringent tests for theories of the origin of structure. The two major new theoretical ideas of the past decade – inflation and cosmic strings – presently provide competing explanations for the origin of structure. Inflationary models provide a simple mechanism for amplifying quantum fluctuations to a magnitude and power and fluctuation spectrum suitable for formation of the observed large scale structure. Strings (or other topological defects) could provide seeds for structure formation either directly through gravitational attraction or via "wakes" resulting from their motion. Models based upon these ideas have provided us with valuable theoretical lampposts under which to search for the keys to the origin of structure. It is certain that much will be learned from the confrontation between these models and new observations, such as the microwave temperature anisotropy observations recently reported by COBE.

The wealth of new observational data also is likely to finally provide convincing evidence as to whether our universe is closed (as had traditionally been favored by theorists prior to inflationary models), very nearly flat (as predicted by inflationary models), or open (as favored by some present observational evidence and as indicated by the observed light element abundances arising from "big bang nucleosynthesis" – unless the present mass density of the universe is dominated by matter in non-baryonic form).

In summary, it seems likely that we are in the midst of a "golden age" in cosmology.

4. Quantum Gravity

Without question, the key issue in quantum gravity is the determination of the fundamental character of the formulation of the theory itself. In all non-gravitational theories, the causal and metrical structure of spacetime are fixed in advance. The notion of an "instant of time" is represented by a spacelike hypersurface in a given, background spacetime. In an ordinary quantum field theory, the fundamental observables of the theory are the values of the field and its correlation functions – modulo gauge if the field is a gauge field – at any instant of time. Methods exist for calculating these field correlation functions in perturbation theory. Although many of the formal expressions for terms occurring in the perturbative expansion are infinite, for

renormalizable theories well defined rules exist for extracting finite results without introducing any new parameters into the theory. Furthermore, it should be noted that in most applications, one is interested only in the behavior of the field at asymptotically early and late times, when it can be treated as "free." A particle interpretation of the states is available in these asymptotic regimes, and, in most applications, the relevant information is encoded in the S-matrix, for which (in renormalizable theories) a well defined perturbative expansion exists.

The key phrase in the preceding paragraph is "the field . . . modulo gauge . . . at any instant of time." In general relativity or other gravitational theories based upon a spacetime metric, the gauge group is (or includes) the diffeomorphism group of the spacetime manifold. Unlike other gauge theories, this gauge group includes all "time translations," so in order for a quantity defined at an "instant of time" to be "gauge invariant", it is necessary for it to be "time independent". However, quantities of this sort – referred to as "perennials" in Kuchař's contribution – would appear to be essentially trivial, and do not encompass the usual dynamical variables of general relativity, such as the induced metric and extrinsic curvature of a spacelike hypersurface. Thus, we have what appears to be an essential conflict between general relativity and quantum theory: General relativity demands that only "histories" are well defined (i.e., "gauge invariant"), whereas the fundamental structure of quantum theory requires that only observables defined at an instant of time be well defined, (i.e., "histories" are ill defined in quantum theory – except in an idealized limit of perfect decoherence).

Parametrized theories of particles or fields have formal properties very similar to general relativity. A well defined quantum theory of these systems can be obtained by "de-parametrization", i.e., by explicitly identifying the variable which (secretly, in the initial formulation) plays the role of time in these theories and interpreting the constraint equation as a time evolution equation in this variable. A sensible Hilbert space structure on states also can be defined by making use of this distinction between the "time variable" and the "true dynamical degrees of freedom". However, although the issues involved have been discussed for more than two decades, very little, if any, progress has been made in the direction of "de-parametrizing" general relativity, or, for that matter, in precisely defining the states and observables of the theory by any other means. This is the "problem of time".

There are additional difficulties which any proposed quantum theory of gravity must overcome. One difficulty (undoubtedly closely related to the "problem of time") has to do with the lack of a background causal structure of spacetime. The fact that quantum fields commute (or anticommute) at spacelike separated events is a fundamental property of all non-gravitational quantum field theories, but it is far from clear whether a similar idea can even be expressed in quantum gravity, since the notion of whether two points on the spacetime manifold are "spacelike related" depends upon the (quantum) metric and is not sharply defined (and also is state-dependent). An additional serious difficulty (whose exact relationship, if any, to the "problem of time" is unclear) is that a simple dimensional argument indicates that even if a quantum theory of gravity could be written down, it would be nonrenormalizable. Hence, either each term in its perturbative expansion would have to be finite or (presumably, infinitely many) new parameters would have to be introduced to define the quantum theory.

Two decades ago, the (rather small number of) researchers in quantum gravity

divided into two main groups. The first group – comprised largely of theorists with substantial background in particle theory – favored the "covariant approach" to formulating a quantum theory of general relativity. In this approach, one writes the spacetime metric, g_{ab}, as a flat metric, η_{ab}, plus a remainder, h_{ab}, and treats h_{ab} as a quantum "nonlinear spin-2 field" in a flat spacetime. One makes free use of the causal structure of η_{ab} in the formulation of the theory, and treats the theory as though the fundamental observables were the correlation functions of h_{ab} (or the S-matrix elements for graviton-graviton scattering). Eventually, in this approach, one should have to confront the fact that the causal structure of η_{ab} should play no role in the theory, and that η_{ab} and h_{ab} should not have any physical significance as separate entities, since the theory should depend only upon g_{ab}. In other words, initially one does not attempt to impose the condition that the theory one obtains should be independent of the choice of the (artificially introduced, physically unmeasurable) flat metric, η_{ab}, and that h_{ab} is not an observable. In this manner, one can postpone dealing directly with some of the difficult formulational issues of quantum gravity, and focus attention upon issues involving the nonrenormalizability of the theory. One then immediately confronts the difficulty that, although S-matrix elements are finite at one-loop order in perturbation theory, they are infinite at two-loop order (and, presumably, all higher orders); S-matrix elements for typical theories of gravity coupled to matter are infinite at one loop order.

The second group of quantum gravity researchers of twenty years ago – comprised largely by theorists with a substantial background in classical general relativity – favored the "canonical" approach to formulating a quantum theory of general relativity. As discussed further in the contribution of Kuchař, the idea here is to express classical general relativity in Hamiltonian form (with constraints), write down the fundamental canonical commutation relations for the configuration and momentum observables, and then impose the constraints as conditions on the state vector. The quantum version of the classical Hamiltonian constraint equation yields the Wheeler-DeWitt equation. Since this equation enforces the gauge invariance of the state vector under time translation diffeomorphisms, the problem of time is confronted head-on in this approach, and has caused severe difficulties in the definition of the Hilbert space of states and the observables of the theory. Even if these difficulties can be overcome, the difficulties associated with the nonrenormalizability of the theory presumably would remain to be confronted.

Today, there still are two rather distinct groups of researchers investigating the formulation of quantum gravity: one – comprised mainly by particle physicists – taking an approach in much the same spirit as the covariant approach, and the other – comprised mainly by general relativists – taking the canonical approach. Many of the fundamental issues in both approaches remain unresolved. Nevertheless, some interesting new ideas have been introduced, and some notable progress has been made.

The covariant approach has evolved through higher derivative gravity theories (which cure nonrenormalizability but introduce new serious problems), supergravity theories (which yield finite scattering amplitudes to higher loop order in perturbation theory than ordinary general relativity, but apparently yield no fundamental advantages), and on to superstring theory. Superstring theory has the substantial achievement of providing us with a finite theory of gravity in the following sense: It is "finite" in that there is every expectation that the scattering amplitudes for ("first quantized")

strings in a background flat (10-dimensional) spacetime are finite at each order in perturbation theory. It is a "theory of gravity" in the sense that one of the modes of oscillation of the string corresponds to a massless, spin-2 field, and there are arguments indicating that the theory can be reinterpreted as a theory involving a metric field which satisfies Einstein's equation in the "low energy limit." Nevertheless, many significant (and, undoubtedly, fundamental) difficulties remain regarding the interpretation of the theory and how to extract predictions from it for quantities other than string scattering amplitudes. In particular, it is not clear what local observables (as opposed to S-matrix quantities) are defined in the theory or how they are to be calculated. Some attempts have been made to formulate a "string field theory" (which, however, would appear to involve "local observables" on an abstract loop space, not on spacetime), but I am not aware of much progress in this direction. The initial surge of optimism in the mid-1980's that superstrings could provide the ultimate "theory of everything" appears to have been replaced in the past several years by a compensating pessimism – based primarily on the non-uniqueness of string theory and the difficulty of doing calculations rather than the uncovering of any inconsistency or wrong predictions of the theory. I have not followed developments in superstring theory closely and, thus, am not in a position even to speculate upon the future directions of research in this area. However, it is clear that the wealth of new ideas and mathematical tools introduced by superstring theory will have a lasting impact upon research in quantum gravity.

Research in canonical quantum gravity has gotten a significant boost in the past five years from Ashtekar's introduction of new Hamiltonian variables for general relativity. Instead of using the induced metric and extrinsic curvature of a hypersurface as the fundamental, canonically conjugate variables on phase space, Ashtekar takes a (complex) SU(2) connection and a "soldering form" as the fundamental variables. The constraint equations of general relativity simplify considerably in terms of these variables, and take a form closely analogous to that of Yang-Mills theory. Probably the most promising new idea toward formulating a quantum theory of gravity to arise from this approach is the "loop quantization" program, reviewed in the contribution of Smolin. This approach has not, as yet, provided a solution to the "problem of time" and issues related to regularization and renormalization of the theory remain to be confronted. However, this approach is still very much in its developing phase, and considerable further progress can be anticipated. At the very least, the "Ashtekar variables" have provided some new life to an approach to quantum gravity that for nearly 20 years had contributed very little in the way of new ideas toward solving the fundamental problems of the quantum gravity.

The direct attacks upon the problem of formulating a quantum theory of gravitation discussed above are by no means the only research activity related to quantum gravity presently being pursued. The theory of linear quantum fields in curved spacetime has been developed to a mathematically complete and precise theory, though some issues with regard to treating "back-reaction" remain to be resolved. Some additional insights into quantum phenomena occurring in strong gravitational fields – in particular, in the early universe – can be expected to result from further research in this area. Research continues in "quantum cosmology", i.e., quantum gravity with all but finitely many degrees of freedom eliminated by symmetry restrictions. Quantum cosmology provides an excellent testing ground for ideas related to the interpretation of quantum

gravity, since the usual infinities of quantum field theory are absent, but the "problem of time" remains present. Proposals for selecting a preferred "wavefunction of the universe" also can be formulated and tested in the context of quantum cosmology. Such proposals make a serious attempt – for the first time in the era of modern physics – to provide us with a theory of initial conditions.

Some of the most penetrating insights into the nature of quantum gravity have come from the analysis of particle creation near black holes and its implications for black hole thermodynamics and the loss of quantum coherence. In the past year, research in this area has been revitalized by the study of a two-dimensional ("string-inspired") field theory which appears to have the necessary properties to model a situation corresponding to the gravitational collapse of a body which subsequently emits Hawking radiation. The model is sufficiently tractable that, in the semiclassical approximation, it should be possible to study in detail the nature of singularities produced by the collapse and quantum back-reaction, as well as issues related to the loss of quantum coherence. Unless the model turns out to be seriously flawed, it is likely that some new insights into the nature of the black hole formation and evaporation process will be achieved.

In summary, it undoubtedly is much too early to assess how far down the road we have come toward obtaining a quantum theory of gravitation. However, at least it is encouraging that some progress down that road presently seems to be taking place.

Acknowledgements

I wish to thank Richard Kron for a discussion on observational astronomy. This research was supported, in part, by NSF grant PHY 89-18388 to the University of Chicago.

Paper presented at the 13th Int. Conf. on General Relativity and Gravitation
Cordoba, Argentina, 1992: Part 1, Plenary Lectures

Large-scale structure

Simon D.M. White

Institute of Astronomy
Madingley Road
Cambridge CB3 0HA, U.K.

Abstract.
Recent observational surveys have made substantial progress in quantifying the structure of the Universe on large scales. Galaxy density and galaxy velocity fields show deviations from the predictions of a homogeneous and isotropic world model on scales approaching one percent of the current horizon scale. A comparison of the amplitudes in density and in velocity provides the first direct dynamical evidence in favour of a high mean density similar to that required for closure. The fluctuations observed on these scales have the amplitude predicted by the standard Cold Dark Matter (CDM) model when this model is normalised to agree with the microwave background fluctuations measured on much larger scales by the COBE satellite. However, a CDM model with this amplitude appears *inconsistent* with observational data on smaller scales. In addition it predicts a scale dependence of fluctuation amplitude which disagrees with that observed for galaxies in the APM survey of two million faint galaxies. The COBE measurement also strongly excludes the standard neutrino-dominated Hot Dark Matter model. Finally, the baryon fraction in rich clusters of galaxies appears much larger than the baryon fraction allowed in an Einstein-de Sitter universe by the theory of Big Bang nucleosynthesis, and so conflicts with both models. Several modifications of these standard models have been proposed in order to avoid some of these difficulties, but none avoids all of them.

1. Introduction

Current theories for the formation of structure in the Universe embrace aspects of quantum gravity, of high energy particle physics, and of nonlinear gravitational collapse, together with a lot of rather messy astrophysics. The global geometry of space-time presumably finds its origin at the Planck time, either as an initial condition, or as a result of some consistency requirement in quantum gravity. The very large (possibly infinite) length-scale characterising the curvature of the universe may reflect an initial phase of

chaotic inflation, or may arise from a later inflationary phase occurring, for example, at the symmetry breaking epoch of a Grand Unified Theory. The latter phase-transition could also produce the baryon asymmetry of the Universe, as could processes at later times. Quantum fluctuations present in the gravitationally dominant field during either inflationary phase could create small-scale structure on the world model which might ultimately develop into present-day galaxies and larger structures. As an alternative, breaking of the symmetry of some non-dominant field to a state with non-trivial local orientation properties might lead to topological defects, regions of space where topological constraints force the field to remain in its unbroken state. The energy density associated with these defects could then induce gravitational perturbations in the dominant component of the Universe and so lead to galaxy formation. Thus well before the end of the first second, the structure and the particle and radiation content of the Universe may all be determined [1].

A minute or so later the temperature drops to the point where atomic nuclei can bind. The abundances of the light elements (^{2}H, ^{3}He, ^{4}He and ^{7}Li) produced at this time can compared with observation. Good agreement is found for a simple model where the early universe is effectively homogeneous and its baryon content is a few percent of that required for closure. This has long been taken as one of the main pieces of evidence in favour of the Hot Big Bang [1, 5]. The techniques needed for calculating the linear evolution of fluctuations on an FRW background are now well developed and can be used to predict the statistical properties of angular fluctuations in the microwave background [2]. The detection of such fluctuations by the COBE satellite has opened up what should prove to be a very rich source of information about the contents and structure of the early universe [3]. However, comparison with observations of structure in the present universe requires some treatment of the nonlinear evolution of structure, and in particular of galaxy formation (since it is galaxies that we are able to observe). This has been done most effectively through large-scale computer simulations, although such work has so far treated only a subset of the relevant physical processes [4, 2].

2. The Standard Model

The complex of ideas briefly sketched in the last section has led to a "standard" model for the content and structure of the Universe. This model, in variety of forms, is currently the subject of intensive exploration and testing. Its major elements may be enumerated as follows:

1. The inflationary model indicates that the present Universe should be flat to a very good approximation. Thus in the absence of a cosmological constant, the present density parameter, $\Omega_0 \approx 1$ [1].

2. Comparison of light element abundances with cosmic nucleosynthesis calculations leads to the conclusion that Ω_b, the mean baryon density in units of the critical density, and h, Hubble's constant in units of 100 km/s/Mpc, must satisfy, $\Omega_b h^2 = 0.0125 \pm 0.0025$ [5].

3. The bulk of the present matter content of the Universe must therefore be in some nonbaryonic form, for example neutrinos with a mass of ~ 30 eV (Hot Dark Matter

or HDM), or more exotic particles with smaller thermal motions, such as axions or the lightest supersymmetric partner of known particles (generically such particles are termed Cold Dark Matter, CDM).

4. Structure is imposed either: (a) by quantum fluctuations in the field which dominates during the inflationary period, giving rise to a gaussian fluctuation field obeying the Harrison-Zel'dovich scaling of amplitude with spatial scale, or (b) by gravitational effects due to topological defects (*e.g.* cosmic strings) or to the realignment of orientable random fields (*e.g.* cosmic texture) [1, 6].

5. Observed structure grows primarily through gravitational instability, and the galaxy distribution reflects the underlying mass distribution in the simple way suggested by schematic models for galaxy formation [7]

In the absence of a cosmological constant, a flat universe can only approach consistency with the inferred ages of globular star clusters if Hubble's constant is small, $h \sim 0.5$. The evolution of structure has so far been studied most thoroughly in the gaussian case, 4a, for which a HDM model seems unable to produce the observed structure [4].

Recent progress in assessing the viability of this picture has come from a variety of directions. Surveys of galaxy motions are now providing the first direct dynamical evidence in favour of $\Omega_0 \sim 1$, but other observational developments sit less well with these ideas. Although the COBE fluctuation measurement appears to confirm that structure grew through gravitational instability, the actual fluctuation amplitude measured is not in agreement with prior expectations for the simplest models. For standard CDM the result is a factor of two larger than predicted based on the strength of galaxy clustering, while for standard HDM it gives such a low amplitude that the model develops almost no structure at all and can therefore be ruled out. (Neutrino dark matter may still be viable if structure originates through cosmic strings or textures.) Another discrepancy comes from measurements of galaxy clustering which imply a scaling of fluctuation amplitude with spatial scale which is inconsistent with a standard CDM model. Finally, new estimates of the baryon fraction in rich clusters of galaxies appear much too large to be compatible with the baryon fraction allowed by nucleosynthesis constraints in an $\Omega_0 = 1$ universe. The rest of this contribution amplifies these points.

3. Ω_0 estimates from streaming motions

For the growing mode of linear density fluctuations in a dust-filled FRW universe, the present *peculiar velocity* of matter relative to the local fundamental (unperturbed) frame is related to a quasi-Newtonian gravitational potential through:

$$\mathbf{v}(\mathbf{x}) = \frac{2}{3} H_o^{-1} g(\Omega_0) \nabla \Phi$$

where $\mathbf{x}$ is a comoving spatial coordinate, H_0 and Ω_0 are the present values of Hubble's constant and the density parameter, g is a dimensionless function well approximated by $\Omega_0^{-0.4}$, and Φ is related to the overdensity, $\delta(\mathbf{x}) = \rho(\mathbf{x})/\overline{\rho} - 1$, by

$$\nabla^2 \Phi = \frac{3}{2} H_0^2 \Omega_0 \delta.$$

Recent improvements in observational techniques have made it possible to measure distances to large numbers of galaxies. Subtracting H_0 times the distance from the observed recession velocity of an object gives the projection along the line-of-sight of the difference beween the values of $\mathbf{v}$ at its position and at our own. Since $\mathbf{v}(\mathbf{x})$ is curl-free, estimates of this quantity for a dense enough sample of galaxies suffice to reconstruct the whole field up to a constant value. This constant is the peculiar motion of our own Galaxy which can be measured directly through the dipole asymmetry it induces in the apparent temperature of the microwave background. Hence the observational data permit the reconstruction of the full peculiar velocity field. Its divergence is then $\Omega_0 g(\Omega_0)\delta(\mathbf{x}) \approx \Omega_0^{0.6}\delta$.

A galaxy overdensity field, $\delta_g(\mathbf{x})$, can be estimated quite independently by measuring the spatial density of objects directly in a suitable (different) sample of galaxies. Clearly, if the structures seen in the field, δ_g, correspond well to those seen in the field, $\Omega_0^{0.6}\delta$, this suggests that the observed peculiar motions are indeed induced gravitationally, that the measurements are not dominated by observational error, and that the observable galaxy density field is closely related to the invisible mass density field. This programme was first suggested and applied by [8]. There really does seem to be quite a good correspondance between the density field measured by counting galaxies, and that inferred from peculiar velocities. Furthermore, if complete samples of galaxies from catalogues constructed using the Infrared Astronomical Satellite (IRAS) are used to define the galaxy density field, the amplitudes also agree approximately *i.e.* $\delta_g(\mathbf{x}) \approx \Omega_0^{0.6}\delta(\mathbf{x})$. Thus if IRAS galaxies trace the mass distribution (so that $\delta_g \approx \delta$) we infer that $\Omega_0 \approx 1$.

In this subject it is often assumed that the galaxies are *biased* relative to the mass in the sense that the contrast of their density field is enhanced; in the simplest model $\delta_g = b\delta$, where the bias constant is taken to be $b > 1$ in order to account for the small mass to luminosity ratios measured for galaxy clusters. In this model the comparison of peculiar velocity and galaxy density fields leads to the estimate, $b \approx \Omega_0^{0.6}$, for the IRAS galaxy samples. Hence mass would need to be substantially *more* clustered than the observed IRAS galaxies in order to produce the observed peculiar velocities in a low density universe (*e.g.* $\Omega_0 \sim 0.2$). This is difficult to reconcile with our present, admittedly poor understanding of how galaxies form, and so the current situation is usually taken as a positive indication in favour of $\Omega_0 = 1$. At present a number of long-term projects are acquiring larger and more accurate datasets to carry out this test, so the conclusions should become much more solid over the next few years.

4. The COBE amplitude measurement

The Differential Microwave Radiometer on board COBE has detected fluctuations on all angular scales larger than the instrument's resolution of 7°. Smoothed to 10° the *rms* temperature fluctuation on the sky is $30 \pm 5\mu K$ in the first year's data, or one part in 10^5 [9]. The spatial scale corresponding to 10° is considerably larger than any scale for which we have an estimate of fluctuation amplitude from studies of clustering in the present universe. Comparison of the COBE result with such data therefore requires not only an assumption about the evolution of fluctuation amplitudes (so that the present amplitude on the COBE scale can be inferred from the observed amplitude on the surface where the radiation was last scattered, at a redshift between 30 and 10^3) but also an

assumption about the scaling of fluctuation amplitude with spatial scale. The time evolution depends only on the global cosmological parameters, Ω_0 and the cosmological constant, Λ, but the spatial scaling depends in addition on the details of the fluctuation generation mechanism, on the nature of the dark matter, and on the baryon density, Ω_b. A further complication is that the COBE measurements could contain a significant contribution from gravitational wave modes which would have no effect on structure formation and hence would not be reflected in measurements of galaxy clustering.

In one of the announcement papers [9] the COBE team discussed the consequences of their fluctuation measurement for the models of §2. For standard CDM models in which inflation-generated gaussian fluctuations have the Harrison-Zel'dovich scaling of amplitude with spatial scale, the COBE measurement implies that the *rms* amplitude of mass fluctuations on small scales is about equal to the observed amplitude of galaxy fluctuations, *i.e.* that the bias parameter, $b \sim 1$. This amplitude is almost exactly that required in a CDM universe to produce the observed peculiar motions discussed in the last section. However, it has generally been argued that substantially larger values of b, (and hence smaller fluctuation amplitudes) are required for a flat universe to be consistent with the observed dynamics of galaxy clustering on smaller scales and with the observed, relatively low abundance of massive objects such as galaxy clusters [7, 10]. A dissenting opinion that $b \sim 1$ is actually required to explain the observed properties of galaxy clustering was expressed in [11] and the situation is still somewhat controversial. If it is accepted that CDM with $b \sim 1$ is unacceptable, then a variety of possibilities have been suggested which might reconcile the COBE observations with observed clustering within a CDM-like model (see, for example [12]).

For the standard HDM model with $H_0 = 50$ km/s/Mpc, with $\Omega_0 = 1$ contributed predominantly by a single species of massive neutrino, and with the Harrison-Zel'dovich spectrum of inflationary perturbations, the COBE amplitude implies an *rms* neutrino density fluctuation from point to point in the present universe of only $\langle \delta^2 \rangle^{1/2} \approx 0.7$ [13]. For such a small amplitude only a few percent of all the matter in the universe is predicted to be in nonlinear collapsed objects by the present day, and virtually no nonlinear objects should exist at redshifts of one or greater. This is clearly inconsistent with the substantial structures seen in the present universe and with the existence of quasars at redshifts approaching five.

In universes where fluctuations result from the presence of topological defects or the realignment of orientable random fields (*e.g.* cosmic strings or textures) the present nonlinear structure of the mass distribution is considerably more difficult to calculate than in the gaussian fluctuation models just discussed. As a result it has not yet been possible to compare their predictions for galaxy clustering and for the COBE data to observation at the same level of precision as for the other models. At present it still appears that these models may be viable [6].

5. The shape of the galaxy correlation function

The amplitude of fluctuations as a function of spatial scale can be estimated for the galaxy distribution by measuring its spatial autocorrelation function. So far the most sensitive estimate of this two-point statistic has come from the Automated Plate-measuring Machine (APM) survey which catalogued the positions of more than 2×10^6 galaxies over a

large area of the southern sky [14]. Individual distances to these galaxies are not known. However, the very large number of objects means that estimates of galaxy clustering from their projected positions are still more precise than those obtained from the much smaller samples for which complete three-dimensional information is available. The scaling of fluctuation amplitude with spatial scale found from this survey is not consistent with that predicted for the *mass* in a standard CDM model [14]. If the model is matched to the data on scales where the amplitude is of order unity ($\sim 5h^{-1}$Mpc), it predicts fluctuations which are too weak on somewhat larger scales ($\sim 20h^{-1}$Mpc). This is often stated as showing that standard CDM has too little power on large scales. In fact, however, if the overall level of mass fluctuations is taken to be fixed by COBE, then the problem seems rather to be that standard CDM predicts too much small-scale power.

The APM result has survived a number of challenges and seems to be supported by similar results for a number of three-dimensional catalogues. At present the latter are less statistically significant than the original, and in addition their interpretation is complicated by the distortion of the distribution which results from the peculiar velocities of galaxies. Extensions of the CDM model which reconcile the COBE amplitude with the dynamics of galaxy clustering generally change the scaling of fluctuation amplitude in a way which makes it more compatible with the APM data [12]. An alternative resolution of the difficulty may lie in the physics of galaxy formation. The assumption that $\delta_g = b\delta$, and thus that the shape of the galaxy correlation function should parallel that of the mass, is based on a rather simple and schematic model for galaxy formation [7]. This assumption does not hold in models where the formation of galaxies is modulated by some long-range nongravitational effect (for example, is inhibited or is stimulated by radiation from nearby quasars). Such models can reproduce the shape of the APM *galaxy* correlations within a standard CDM universe [15].

6. The baryon fraction in rich clusters of galaxies

If the first two assumptions of the standard model of §2 are correct, then the fraction of the matter in the universe which is baryonic is,

$$F_b = \frac{\Omega_b}{\Omega_0} = 0.0125 \pm 0.0025 \quad h^{-2}.$$

The largest objects for which it is possible to get reliable estimates both of total mass and of baryon content are rich clusters of galaxies. These are the most massive quasi-equilibrium systems known, and baryons are observed within them both in the form of stars within the individual galaxies, and in the form of a pervasive intergalactic medium which is sufficiently hot that it emits X-rays. The total mass of a cluster can be obtained using the virial theorem, or by applying the equations of hydrostatic equilibrium either to the gas or to the galaxy population. The baryonic mass in stars can be estimated from the optical luminosity of the cluster, and the baryonic mass in hot gas from X-ray imaging and spectroscopy. For the best observed rich cluster, the nearby Coma cluster, the results found by applying standard techniques are,

$$M_{tot} = 7.4 \times 10^{14} h^{-1} \mathrm{M}_\odot, \quad M_* = 3.2 \times 10^{13} h^{-1} \mathrm{M}_\odot, \quad M_{gas} = 5.6 \times 10^{13} h^{-2.5} \mathrm{M}_\odot,$$

where all three masses are estimated within a radius of $1.5h^{-1}$ Mpc and have errors of 20 to 30% [16]. Since some of the unseen dark matter might be made of baryons, we then get a lower limit to the baryon fraction in the Coma cluster:

$$F_{b,Coma} \geq \frac{M_* + M_{gas}}{M_{tot}} = 0.043 + 0.076h^{-1.5}.$$

Although the observational uncertainties in this limit are substantial, it is clearly much too high to be compatible with the baryon fraction in the standard model. Retaining the standard model requires the total mass of the Coma cluster to be much larger than is usually believed and the gas and star masses to be much smaller. Since Coma does not appear to be in any way atypical, similar systematic errors would have to apply to mass estimates in other clusters. One might hope to resolve this paradox by arguing that baryons are preferentially concentrated into the centres of rich clusters during cluster formation. Numerical simulations of cluster formation can put upper limits on the enhancement attainable, and, for a region as large as that used above, the maximum possible enhancement is a few tens of percent; indeed, most simulations of cluster formation conclude that the gas should end up slightly *less* concentrated than the galaxies, thus making the discrepancy worse [17].

7. Conclusions

While the observational data reported in §3 tend to support the standard picture outlined in §2, those discussed in §§4 – 6 disagree with various parts of it. The fluctuation amplitude measured by COBE rules out a standard HDM model with a Harrison-Zel'dovich spectrum of gaussian initial fluctuations. *A priori* this is, of course, the most attractive of all the models dominated by nonbaryonic dark matter. The corresponding standard CDM model is also in trouble because COBE requires a higher normalisation of the fluctuation amplitude than seems consistent with the abundance of rich galaxy clusters and with the dynamics of galaxy clustering on scales of a few Mpc. The shape of the APM galaxy correlation function also suggests the need for initial density fluctuations with less small-scale power than the standard CDM model, and several modifications of the model have been suggested which accomplish this (for example, replacing some of the CDM by HDM, or adding an additional relativistic component such as nonthermalised neutrinos [12]).

Unfortunately the difficulty highlighted in §6 applies in any Einstein-de Sitter universe and so to these modifications of the standard CDM model. Unless some flaw can be found in the interpretation of the observational data, it forces the abandonment of:

1. the Einstein-de Sitter model, or
2. the standard theory of cosmic nucleosynthesis, or
3. the growth of structure through hierarchical clustering driven by gravity.

An example of the first way out is the introduction of a cosmological constant. In an otherwise standard flat CDM model this allows a large value of $F_b = \Omega_b/\Omega_0$, a correlation function shape consistent with the APM survey, and an age for the Universe consistent with those of globular star clusters despite a high value of Hubble's constant.

However, such a model does not produce large enough peculiar velocities to be consistent with the data described in §3. An example of the third way out is explosive galaxy formation. A first generation of objects (early quasars?) might produce hydrodynamic flows which pile the baryonic gas into large-scale structures. These could then act as accretion nuclei for the dark matter. Such a model is unattractive because it decouples the properties of present large-scale structure from primordial density fluctuations. In addition, the energetic shocks it requires are almost certainly inconsistent with the very precise black-body spectrum and the relative weak angular fluctuations measured by COBE. No current model appears consistent with a straightforward interpretation of all the observational data.

References

[1] Kolb E W, Turner M S 1990 *The Early Universe* (New York: Addison-Wesley)

[2] Efstathiou G 1990 *Physics of the Early Universe* (ed. Peacock J A, Heavens A F, Davies A T) (Edinburgh: Ed.Univ.Press)

[3] Mather J 1993, this volume.

[4] Frenk C S 1991 *Physica Scripta* **T36**, 70.

[5] Walker T P, Steigman G, Schramm D N, Olive K A, Kang H-S 1991 *Astrophys.J.* **376**, 51.

[6] Kibble T 1993, this volume.

[7] Davis M, Efstathiou E, Frenk C S, White S D M 1985 *Astrophys.J.* **292**, 371; Bardeen J, Bond J R, Kaiser N, Szalay A S 1986 *Astrophys.J.* **304**, 15.

[8] Dekel A, Bertschinger E, Faber S M 1990 *Astrophys.J.* **364**, 349; Bertschinger E, Dekel A, Faber S M, Dressler A, Burstein D 1990 *Astrophys.J.* **364**, 370; Kaiser N, *et al.* 1991 *Mon.Not.R.astr.Soc.* **252**, 1.

[9] Smoot G F, *et al.* 1992 *Astrophys.J.* **396**, L1; Wright E L, *et al.* 1992 *Astrophys.J.* **396**, L13.

[10] Henry J P, Arnaud K A 1991 *Astrophys.J.* **372**, 410; White S D M, Efstathiou G, Frenk C S 1993 *Mon.Not.R.astr.Soc.*, in press.

[11] Couchman H M P, Carlberg R G 1992 *Astrophys.J.* **389**, 453.

[12] Efstathiou G, Bond J R, White S D M 1992 *Mon.Not.R.astr.Soc.* **258**, 1P.

[13] Holzmann J A 1989 *Astrophys.J.Supp.* **71**, 1.

[14] Maddox S J, Efstathiou G, Sutherland W J, Loveday J 1990 *Mon.Not.R.astr.Soc.* **242**, 43P.

[15] Babul A, White S D M 1991 *Mon.Not.R.astr.Soc.* **253**, 31P; Bower, R G, Coles P, Frenk C S, White S D M 1993 *Astrophys.J.*, in press.

[16] The L S, White S D M 1988 *Astron.J.* **95**, 15; Hughes J P 1989 *Astrophys.J.* **337**, 21; Briel U G, Henry J P, Böhringer H 1992 *Astron.Astrophys.* **259**, L31.

[17] Evrard A 1990 *Astrophys.J.* **363**, 349.

PART 2

WORKSHOP SUMMARIES

Exact solutions and their interpretation

G. Neugebauer

Max-Planck-Gesellschaft, Arbeitsgruppe "Gravitationstheorie" an der Friedrich-Schiller-Universität, Max-Wien-Platz 1, D-6900 Jena, Germany

Sometimes the search for exact solutions resembles the attempt to set up a record in sports: The more free parameters the better the solution (the more famous the author). This review article is not meant to report on new records born or published at GR13. The chairperson rather wants to emphasize the contributions to new methods, interpretations and interesting mathematical structures of Lorentzian manifolds.

61 authors contributed to the Workshop. 16 of them were asked to give an oral presentation (of 10 or 20 minutes), the other ones had the opportunity to display posters.

1. New techniques and solutions

The classical Kerr-Schild ansatz

$$g_{\alpha\beta}(\lambda) = \eta_{\alpha\beta} + \lambda h_{\alpha\beta} \ ,$$

where $\eta_{\alpha\beta}$ is flat and the perturbation $h_{\alpha\beta}$ is the square of a null vector, is an example of a pencil of metrics that is a solution to the field equations for all values of the parameter λ. R.P. KERR (speaker), Z. PERJÉS and C. HOENSELAERS (1992) considered a generalisation of Kerr-Schild by investigating all $g_{\alpha\beta}(\lambda)$, where the base metric $\eta_{\alpha\beta}$ must be itself an Einstein space, but need not be flat, and obtained the following theorem:
A necessary and sufficient condition for the determinant of a pencil to be independent of λ,

$$| \ \eta_{\alpha\beta} \ - \ \lambda h_{\alpha\beta} \ | = | \ \eta_{\alpha\beta} \ |$$

is that the matrix $H = (h_{\alpha\sigma}\eta^{\sigma\beta})$ *should satisfy* $H^3 = 0$. *When* $H^2 = 0$ *they get the usual KS Ansatz. Otherwise, the metric and its inverse can be expressed as*

$$\begin{aligned} g_{\alpha\beta} &= \eta_{\alpha\beta} - \lambda(k_\alpha n_\beta + n_\alpha k_\beta) \\ g^{\alpha\beta} &= \eta^{\alpha\beta} + \lambda(k^\alpha n^\beta + n^\alpha k^\beta) + \lambda^2 k^\alpha k^\beta \end{aligned}$$

where $\boldsymbol{n.n} = \mathbf{1}$, $\boldsymbol{k.k} = \boldsymbol{k.n} = \mathbf{0}$.
A typical example of a metric close to this type is the Goldberg-Kerr type III space (Kerr and Goldberg 1961), which can be thought of as such a perturbation of a flat space, but only if *quadratic perturbations* are allowed,

$$ds^2 = 2d\zeta d\bar{\zeta} - 2dudv + 2\lambda \boldsymbol{kn} + \lambda^2 H_2 \boldsymbol{k}^2.$$

The authors think it unlikely that there are any significant examples of linear perturbations that are not Kerr-Schild. There are two simple Ansätze where the perturbations are functions of two independent orthogonal vectors. For the first there exists a pair p_α and q_α such that

$$\begin{aligned} g_{\alpha\beta}(\lambda) &= \eta_{\alpha\beta} + \lambda(p_\alpha q_\beta + q_\alpha p_\beta) + \lambda^2(q^\sigma q_\sigma)p_\alpha p_\beta \ , \\ g^{\alpha\beta}(\lambda) &= \eta^{\alpha\beta} - \lambda(p^\alpha q^\beta + q^\alpha p^\beta) + \lambda^2(p^\sigma p_\sigma)q_\alpha q_\beta \ . \end{aligned}$$

This has been studied by Bonanos (1991), but there do not appear to be any nontrivial Einstein spaces of that type known. For the second,

$$\begin{aligned} g_{\alpha\beta}(\lambda) &= \eta_{\alpha\beta} + \lambda(k_\alpha n_\beta + n_\alpha k_\beta) + \lambda^2 A(n^\sigma n_\sigma)k_\alpha k_\beta \ , \\ g^{\alpha\beta}(\lambda) &= \eta^{\alpha\beta} - \lambda(k^\alpha n^\beta + n^\alpha k^\beta) + \lambda^2(1-A)(n^\sigma n_\sigma)k_\alpha k_\beta \ , \end{aligned}$$

where $\boldsymbol{k.k} = \boldsymbol{n.k} = \boldsymbol{0}$ and A is arbitrary. Using Reduce on a Sparcstation they established the following theorem:
Given any quadratic pencil of the type in the last equation, where the base space $\eta_{\alpha\beta}$ is flat, one of the following is true: (i) $A = 0$, (ii) $A = 1$, or (iii) k is a geodesic principal null vector of the curvature for all values of the pencil parameter. The authors emphasized that they did not know whether the cases (i) and (ii) allow k to be non geodesic or whether there are any interesting solutions of these types. Any metric in the third class must either be one of those considered by Newman and Tamburino, or algebraically special, in which case it must be of the Kundt class.

A modification of an integration formalism was presentes by S.B. EDGAR (1992), who analysed the different stages of Held's integration procedure (Held 1974) in the G.H.P. formalism and effected some improvements.

In order to describe compact objects in astrophysics more effort must be made to analyse the structure of the Einstein equations with (more or less) realistic source terms. For nearly 25 years the well-known Wahlquist solution (Wahlquist 1968) for a rotating perfect fluid waits for interpretation, i.e. for a matching to an exterior vacuum field. M. TINTO and H.D. WAHLQUIST (speaker) did not solve this problem (Tinto and Wahlquist 1992), but they presented, in computergenerated imbedding diagrams, configurations of rotating rigid bodies encompassed by the Wahlquist solution. The configurations include single bodies with simple, approximately spheroidal, exterior surfaces ($p = 0$) having either oblate or prolate intrinsic geometry. For certain values of the solution parameters, the body becomes stretched along the axis of rotation into more bizarre, dumbbell-like shapes with regions of negative Gaussian curvature. As the parameters are varied further, these solutions develop into configurations with an arbitrary number of disjoint bodies symmetrically distributed on the rotation axis.

There was only one contribution to wormholes in our workshop. E.E. DONETS and D.V. GAL'TSOV (1992) showed that for some value of the cosmological constant the negative semi-definite action wormhole solutions of the self-consistent Einstein-Yang-Mills SU(2) system form a continuous family, i.e. the separation constant is allowed to take an arbitrary value on the interval [0,1]. Their considerations complement the discrete wormhole family found by Hosoya-Ogura, Verbin-Davidson and S. Rey. For each member of the new family the period of the square of the scale factor coincides exactly with the period of the Yang-Mills function.

F.J. CHINEA (1992) presented a new formalism (developped by himself and L.M. GONZÁLEZ-ROMERO) (Chinea and Gonzáles-Romero 1992) for treating the gravitational fields of stationary, axially symmetric differentially rotating perfect fluids. Among the new solutions there is an interior metric of a rotating fluid in irrotational motion (Chinea and Gonzáles-Romero 1990).

A radiating body with heat flow was considered by D. KRAMER (1992). He matched a spherically symmetric time-dependent interior metric to the (exterior) Vaidya solution. The regular bounded source consists of a shearfree fluid with radial heat flow. In the remote past the solution coincides with the interior Schwarzschild solution, the mass parameter of which is now taken as a function of time. The evolution as determined by the junction conditions (Santos 1985) leads to a collapsing model; the area of the boundary surface decreases monotonically. Finally, the model runs into a physical singularity.

The gravitational fields of rigidly rotating perfect fluids with nonlinear Born-Infeld electromagnetism were investigated by H. SALAZAR (speaker) and R. CORDERO (1992). They found two type D solutions for the equation of state $e + 3p =$ constant (Salazar et al. 1987).

The collapse of a charged body with vanishing radial pressure was studied by C.H.A. BECERRA and C.A. LOPEZ (speaker) (1992).

J. GRIFFITHS (speaker) and P.C. ASHBY (1992) reported on colliding plane wave solutions. In the colinear case the main equation is linear but the subsidiary equations are not. GRIFFITHS, HOENSELAERS and ASHBY adopted an unconventional coordinate system in which an explicit integral of the subsidiary equations can be obtained for a general solution of the main equation. Further exact solutions for colliding non-colinear gravitational waves coupled with fluid motions have been obtained by GRIFFITHS and ASHBY by relating the potential of Chandrasekhar and Xanthopoulos to the solution-generating potential of Wainwright, Ince and Marshman (1979).

Soliton methods are still used to generate asymptotically flat gravitational fields and to interprete the source afterwards. Using Alekseev's Inverse Scattering Method A.D. DAGOTTO, R.J. GLEISER (speaker) and C.O. NICASIO (1992) obtained a two soliton solution to the Einstein-Maxwell equations representing an isolated cosmic string in the presence of a pair of solitonic pulses of electro-gravitational radiation. The gravitational field of a rotating mass endowed with a magnetic dipole moment was presented by V.S. MANKO (speaker) and N.R. SIBGATULLIN (1992).

2. Mathematical structures

The study of symmetries in General Relativity is often done on an ad hoc basis. The idea of G. HALL'S (1992) lecture was to show how the study of symmetry in General Relativity can be accomplished in quite general terms and, whilst on a more systematic mathematical basis, can be very useful in establishing direct results for use in the study of exact solutions of Einstein's equations. Isometries, homotheties, conformal and affine symmetries as well as curvature collineations were all covered and attention was paid to the orbit structure, the isotropy structure and the algebraic consequences for the energy-momentum and Weyl tensors. In particular, the precise nature of the metrics that admit certain symmetries can be established in many cases (Hall and da Costa 1991). The

maximum dimension of the algebra of each of the above symmetry vector fields can also be found by these techniques. A more precise statement of the Bilyalov-Defrise-Carter theorem regarding the conformal reduction of conformal vector fields to Killing vector fields also results.

A wide variety of exact solutions (e.g. all the solutions admitting a group G_3 of motions acting on 2-dimensional orbits, Robertson-Walker, de Sitter etc.) belong to the Warped space-time family[1]. J. CAROT (speaker) and J. DA COSTA (1992) provided invariant characterisation of Warped space-times in terms of the existence of vector fields with special properties, and put forward a classification scheme based on such properties. Thus, they defined three classes; Class A_1, formed by all space-times which admit a global, nowhere zero, unit vector field that is geodesic, shearfree, hypersurface orthogonal and such that the gradient of its expansion is orthogonal to it. Class A_2 consists of all those space-times which also admit a global, nowhere zero, unit vector field that is shearfree, non-expanding, hypersurface orthogonal and such that its acceleration is a gradient. Finally, Class B contains all those space-times admitting two global null, nowhere zero, linearly independent vector fields whose covariant derivatives take on a precise and well defined form. The curvature structure was looked into showing that all Petrov and Segre types may occur in general, but that they are highly restricted and easily charaterizable in most cases. They also studied the problem of the isometry group that each class of warped space-time may admit, giving expressions for the Killing vector fields (KV's) in terms of the KV's of (M_1, h_1) and the conformal Killing vectors of (M_2, h_2). The maximal dimension of the Lie algebra of motions was studied, showing that it is 10 for class A_1 (de Sitter models), 7 for class A_2 (either the static Einstein Universe, or a pure radiation field (Petrov)), and 6 for class B.

A group of projective collineations is a continuous group of mappings that map geodesic curves to geodesic curves without necessarily preserving affine parameterisation. A. BARNES (1992) showed that the only four-dimensional proper Einstein spaces admitting a proper projective collineation are Petrov Type N or have constant curvature. Furthermore the vector field ξ^i generating the collineation is (up to the addition of a Killing vector field) the principal null vector of the Weyl tensor and is a gradient vector field. For the vacuum case a projective collineation is obviously both a curvature and a Ricci collineation (Katzin 1969). In this case the space-time admits a constant vector field and so is a *pp* wave or is flat.

The scalar invariants of the Riemann tensor are important in general relativity since they allow a manifestly coordinate invariant characterization of many geometrical properties of space-times in particular the existence of curvature singularities and the question whether two space-time metrics are equivalent. Most attention has been directed to the fourteen second-order invariants. Two contributions were devoted to this set of problems. J. CARMINATI, R.G. MCLENAGHAN (speaker) and P.E. WIEBE (1992) reported on one of their recent papers (Carminati and McLenaghan 1991), in which they constructed a set of invariants satisfying the following properties:

(i) The set consists of invariants of lowest possible degree and
(ii) it contains a minimal independent set for any Petrov type and for any specific choice

[1] Given two manifolds (one Lorentzian and one Riemannian) (M_1, h_1) and (M_2, h_2) and given a smooth function $\theta : M_1 \to \Re$; one can build a new Lorentz manifold (M, g) by setting $M = M_1 \times M_2$ and $g = h_1 \oplus e^{2\theta} h_2$. If dim $M = 4$, (M, g) is called a "Warped space-time"

of Ricci tensor type.

This was possible by adding a new complex mixed invariant of degree five to the set of Géhéniau and Debever including a real invariant of degree four that they had excluded. The resulting set contains the equivalent of sixteen real invariants. The lengthy algebra required to obtain the new invariant was performed using the npspinor package (Czapor, McLenaghan and Carminati 1993) in the Maple computer algebra system. The procedure used leads naturally to invariants of lowest possible degree. The authors establish that the set of invariants contains complete minimal sets for the Einstein-Maxwell and perfect fluid cases.

C.B.G. McIntosh's (1992) main purpose was to discuss the geometrical significance of the three mixed invariants of the Riemann tensor. Furthermore, be was able to show that the fourteen invariants can all be written in the form *trace(A.transpose(B))* where A and B are 3×3 complex matrices formed from the spinor components of the Riemann tensor; thus the invariants are easily calculated using a computer algebra program. Further comments were made about (a) the situation in degenerate cases, (b) the invariants when the metric is complex and (c) the general problem of classifying metrics. Finally, some examples were given.

The author of this review article gave a brief report on a global exact solution describing the gravitational field of a rotating disc of dust. R. Meinel and he found this solution by applying the inverse scattering method. It contains the Bardeen-Wagoner approximative solution.

References

Barnes A 1992 *13th Int. Conf. on General Relativity and Gravitation, Cordoba, Argentina (June 28 - July 4, 1992) Abstracts of Contributed Papers* p 9

Becerra C H A and Lopez C A 1992 *13th Int. Conf. on General Relativity and Gravitation, Cordoba, Argentina (June 28 - July 4, 1992) Abstracts of Contributed Papers* p 11

Bonanos S 1991 *Class. Quantum Grav.* **9** 697

Carminati J and McLenaghan R G 1991 *J. Math. Phys.* **32** 3135

Carminati J McLenaghan R G and Wiebe P E 1992 *13th Int. Conf. on General Relativity and Gravitation, Cordoba, Argentina (June 28 - July 4, 1992) Abstracts of Contributed Papers* p 15

Carot J and da Costa J 1992 *13th Int. Conf. on General Relativity and Gravitation, Cordoba, Argentina (June 28 - July 4, 1992) Abstracts of Contributed Papers* p 17

Chinea L M and González-Romero 1990 *Class. Quantum Grav.* **7** L 99

Chinea F J 1992 *13th Int. Conf. on General Relativity and Gravitation, Cordoba, Argentina (June 28 - July 4, 1992) Abstracts of Contributed Papers* p 28

Chinea F J and González-Romero 1992 *Class. Quantum Grav.* , to appear

Czapor S R McLenaghan R G and Carminati J 1993 *Gen. Rel. Grav.*, to appear

Dagotto A D Gleiser R J and Nicasio C O 1992 *13th Int. Conf. on General Relativity and Gravitation, Cordoba, Argentina (June 28 - July 4, 1992) Abstracts of Contributed Papers* p 31

Donets E E and Gal'tsov D V 1992 *13th Int. Conf. on General Relativity and Gravitation, Cordoba, Argentina (June 28 - July 4, 1992) Abstracts of Contributed Papers* p 32

Edgar S B 1992 *13th Int. Conf. on General Relativity and Gravitation, Cordoba, Argentina (June 28 - July 4, 1992) Abstracts of Contributed Papers* p 35

Griffiths J B and Ashby P C 1992 *13th Int. Conf. on General Relativity and Gravitation, Cordoba, Argentina (June 28 - July 4, 1992) Abstracts of Contributed Papers* p 40

Hall G and da Costa J 1991 *J. Math. Phys.* **32** 2848 and 2854

Hall G 1992 *13th Int. Conf. on General Relativity and Gravitation, Cordoba, Argentina (June 28 - July 4, 1992) Abstracts of Contributed Papers* p 45

Held A 1974 *Commun. Math. Phys.* **37** 311

Katzin G H Levine J and Davis W R 1969 *J. Math. Phys.* **10** 617

Kerr R P and Goldberg J N 1961 *J. Math. Phys.*

Kerr R P Perjés Z and Hoenselaers C 1992 *13th Int. Conf. on General Relativity and Gravitation, Cordoba, Argentina (June 28 - July 4, 1992) Abstracts of Contributed Papers* p 50

Kramer D 1992 *J. Math. Phys.* **33** 1458

Manko VS and Sibgatullin N R 1992 *13th Int. Conf. on General Relativity and Gravitation, Cordoba, Argentina (June 28 - July 4, 1992) Abstracts of Contributed Papers* p 58

McIntosh C B G 1992 *13th Int. Conf. on General Relativity and Gravitation, Cordoba, Argentina (June 28 - July 4, 1992) Abstracts of Contributed Papers* p 59

Salazar H et al. 1987 *J. Math. Phys.* **28** 2171

Salazar H and Cordero R 1992 *13th Int. Conf. on General Relativity and Gravitation, Cordoba, Argentina (June 28 - July 4, 1992) Abstracts of Contributed Papers* p 67

Santos N O 1985 *Mon. Not. R. Astron. Soc.* **216** 403

Tinto M and Wahlquist H D 1992 *13th Int. Conf. on General Relativity and Gravitation, Cordoba, Argentina (June 28 - July 4, 1992) Abstracts of Contributed Papers* p 70

Wahlquist H D 1968 *Phys. Rev.* **172** 1291

Wainwright J Ince W C W and Marshman B J 1979 *Gen. Rel. Grav.* **10** 259

Paper presented at the 13th Int. Conf. on General Relativity and Gravitation
Cordoba, Argentina, 1992: Part 2, Workshop Summaries

Complex methods/twistors/new Hamiltonian variables

Carlos N. Kozameh

FaMAF, Laprida 854
Universidad Nacional de Córdoba
5000 Córdoba, Argentina

Abstract.
Papers presented at this session were divided in two areas; one covering complex methods and Twistor theory and the other devoted to new hamiltonian variables in general relativity. The topics discussed ranged from new solutions in classical relativity to degeneracy states in the loop variables for quantum gravity.

The use of complex methods, and in particular the concept of self-duality, plays a very important role in different areas of general relativity. Using fundamentally complex theories one obtains the general solution of the self-dual (or anti self-dual) vacuum equations[1, 2]. Self-dual SL(2,C) connections are the basic variables in the Ashtekar formulation of canonical gravity[3]. The list of applications of these methods is extensive and this review does not aim to cover the subject exhaustively. Rather, the focus will be on the topics presented at the A2 workshop.

The talks given at this session reflected a growth in the range of applications of complex methods in general relativity, both at the classical and quantum level. Since papers on new hamiltonian variables did not have a direct connection to other papers presented, the workshop was split in two parts. This review reflects this fact.

However, it was very suggestive to see that talks given on different formalisms used similar methods and, to some extent, were solving similar problems. As an example one could mention that the goal of obtaining the solution space of the Einstein equations is not only important for non-local formalisms in classical relativity but also is a needed result in canonical quantization of gravity. Also, the concept of holonomies for SL(2,C) connections is used in several programs not necessarily connected with each other. It thus appears that the results here presented as well as future developments in this area should be useful to most of the formalisms that use these methods.

1. Complex Methods and Twistors

R. Penrose presented a new twistorial approach to the Einstein vacuum equations[4]. One of the main goals of twistor theory has been to obtain the general solution of the full vacuum equations. Although it was shown in 1976 that twistors describe the self-dual solutions of these equations[5], the general solution has remained elusive.

As an introduction, a brief review of the main results of twistor theory, a complex, non-local, and mathematically sophisticated theory, was given. In particular, a description of null geodesics, momentum, angular momentum, and charges for helicity $\frac{3}{2}$ massless fields was presented.

The Rarita-Schwinger equation, the field equation for helicity $\frac{3}{2}$ massless fields, written in terms of a two spinor valued connection was used as the starting point for the new ideas relating twistors to the Einstein vacuum equations. Using the following results:

1. Twistors describe the charges for helicity $\frac{3}{2}$ massless fields in flat space-times.
2. The Einstein vacuum equations are the consistency conditions for the Rarita-Schwinger equation in curved space-times.

R. Penrose suggested the construction of the twistor space for vacuum space-times as the space of of spin $\frac{3}{2}$ charges. The second step would be the reconstruction of the underlying space-time in terms of the twistor geometry.

Although details of this construction were not presented, the main idea is to generalize the spin $\frac{3}{2}$ fields to self-dual connections on SL(3,C) bundles and then to define the twistors as source charges for those connections.

L. Mason presented a broad review of twistor theory and its main goal - to reformulate the laws of basic physics in terms of twistor geometry.

Part of this review covered an area where twistor theory has been very successful; integrable systems. It is well known that the general solutions of self-dual Yang-Mills and Einstein equations can be described in terms of twistors[6]. By imposing symmetry reductions on the self-dual Yang-Mills equations one obtains almost all integrable equations in two dimensions[7]. Thus, twistors play an active role in the study of integrable systems providing a geometric generalization of the inverse scattering transform for these equations.

The review also included some of the various approaches to the study of the full Einstein vacuum equations. The main idea here is to obtain the general solution for General relativity in terms of these non-local, complex variables. Although none of the several approaches presented (Hypersurface Twistors, Space of Complex Null Geodesics and Light Cone Cuts) have yet solved the problem, there is work in progress.

T. Newman presented an outline of the basic ideas an results of the Theory of Light Cone Cuts of Null Infinity[8]. The goal of this formalism is to reformulate the vacuum Einstein equations in terms of these new variables. The main motivation is to find the general solution of these equations which represent regular, radiative space-times. It is worth mentioning that a solution of this class is yet to be obtained.

After a brief review of the kinematical properties of the light cone cuts (LCC), T. Newman showed that the reformulation of the Einstein vacuum equations in terms of

the holonomy operator does not yield meaningful results unless the LCC are treated as basic variables in the theory[9]. Thus, one is forced to find field equations for the LCC which should reflect the conformal invariance of these variables.

An explicit relationship between the LCC and the holonomy operator was presented. This relation is then used in the field equations for the holonomy operator to yield, after involved calculations, the sought for equations for the LCC[10]. Although at present the equations are only known as a perturbative series, up to second order they are equivalent to the Bach equations. It is worth mentioning that the vanishing of the Bach tensor is essentially the conformal Einstein equations[11]. Thus, the formalism seems to provide a natural splitting of the vacuum equations into a conformal invariant equation for the LCC and an equation for the conformal factor.

Another complex object which has played a relevant role in general relativity is the self-dual spin connection. First introduced by J. Plevanski to be used as a primary field variable[12], it was later rediscovered by others[13] and applied in different formalisms[3, 14, 15]. In this session R. Capovilla presented, in collaboration with J. Plevanski, a very simple ansatz on the self-dual spin connection to construct a space-time of the Kasner type. This result is a first step to produce explicit solutions of the field equations using this connection as the basic variable[16].

Other results on complex methods were reported by F. Estabrook, E.Wilson and C. Sobczyk.

F. Estabrook, in collaboration with H. Wahlquist, showed that the immersion of four dimensional Riemannian geometries in a 10-dim euclidean space produce twistor like bundles. Cartan analyses for Ricci flat spaces, Einstein-Maxwell theory, as well as other immersed geometries were presented[17].

E. Wilson showed[18] that, given a CR geometry with an integrable distribution of two-spaces, one can construct an associated family of space times with a null, shear-free congruence. E. Wilson and I. Robinson considered space-times arising from Cauchy Riemann geometries of maximal symmetry. The class of solutions so obtained contains the Taub-NUT geometry and the Hauser twisting type N solutions as a limiting case.

Finally, G. Sobczyk showed that using Clifford's geometric algebra one can have an unambiguous geometric interpretation of quantum field theory[19].

2. New Hamiltonian variables

In the Ashtekar formalism for canonical relativity the usual energy and diffeomorphic constraints are supplemented by three SU(2) gauge conditions, thus yielding seven constraint equations. C. Rovelli reported work with T. Newman on a new set of canonical variables that solve six of the seven constraints[20]. This result arises as an application of a powerful method developed to obtain *"the true degrees of freedom"* for Hamiltonian systems with first class constraints[21].

The essential idea is, for these systems, to solve the constraint equations via a Hamilton - Jacobi technique. The formulation provides both the algebraic solution to the constraint equations and the complete set of canonical variables that have vanishing Poisson brackets with the constraints.

When the formalism is applied to Maxwell theory, the gauge invariant observables obtained are two scalar functions that define a congruence of curves. These lines have a

straightforward physical meaning: they are the Faraday electric lines of force. For SU(2) Yang-Mills theory one obtains three non-abelian generalizations of Faraday's lines. They are three congruences of pairs of lines which capture the degrees of freedom of the theory.

Since the constraint equations in the Ashtekar formulation are the SU(2) Yang-Mills plus the Hamiltonian constraint, the same three congruences automatically solve the diffeomorphic and gauge constraints for general relativity.

One should point out that if the energy constraint could also be solved, one could obtain the general solution of the Einstein equations.

J. Frauendiener reported on a characteristic initial value formulation of general relativity using a formalism developed by R. Penrose[22]. It is known that the *"heart"* of the Einstein vacuum equations in the N-P formalism are the spin-2, zero rest mass equation $\nabla^A_{A'}\Psi_{ABCD} = 0$. In 1966, R. Penrose showed how to obtain a formal solution of this equation from data given on an initial null cone. At the workshop, J. Frauendiener presented a new formulation of this problem together with an algorithm to obtain the solution which can be implemented on an algebraic manipulation program. Apart from its own relevance, these new results could also be useful to recent work on canonical formalism on null surfaces using the new variables[23].

Properties of the holonomy group for SL(2,C) Yang - Mills connection were addressed by J. Golberg and T. Jacobson at their respective presentations.

The parallel propagator of SL(2,C) (or SU(2)) Yang-Mills connections associated with closed paths plays an important role in the hamiltonian formulation of general relativity using the Ashtekar variables. The collection of these propagators for arbitrary loops form the holonomy group. Denoting each element by $h_\gamma(x)$ with x the starting point of the loop γ, one can show that h_γ transforms covariantly under a gauge transformation. Moreover, one can also show that the trace of h_γ is a gauge invariant operator. Thus, these variables automatically solve the gauge constraints of the Ashtekar formalism.

C. Rovelli and L. Smolin[24] proposed to use the pair $tr(h_\gamma)$ and $tr(\sigma^a h)$ with σ^a the spinorial triad, as new coordinates on the phase space instead of the usual triad and SL(2,C) connection. A natural question then arises. Is there a one to one correspondence between the two set of variables?

J. Goldberg reported on work together with J. Lewandowski and Stornaiolo addressing this question[25]. They showed that this correspondence is unique for generic connections. Degeneracies can only occur if the holonomy group is in the subgroup of null rotations of SL(2,C). By degeneracies the authors mean either 1) a lack of uniqueness in determining a gauge inequivalent connection and triad, or 2) when at a point in phase space there is a direction orthogonal to the gradients of the $tr(h_\gamma)$ and $tr(\sigma^a h)$ functionals and this direction is transverse to the orbits of the gauge group.

In addition, the authors showed that the holonomy group based at a given point x_o is time-independent in vacuum general relativity. They also pointed out that the holonomy group is invariant under spatial diffeomorphisms and SL(2,C) gauge transformations that are the identity at x_o. Thus, the holonomy group qualifies as an observable in general relativity.

T. Jacobson presented work done in collaboration with J. Romano providing further analyses of observables in GR. They gave an alternate proof of the results of J. Goldberg *et al* using the 4-dim self-dual connection.

It is known that if this connection corresponds to a vacuum space-time then the vanishing of the trace free part of the Ricci tensor implies that its curvature two form is self-dual. Writing these equations in a suitable gauge one shows that if the connection belongs to a given subalgebra at some initial time, then the subalgebra will be preserved under time evolution. Since only the vanishing of the trace free part of the Ricci tensor is used in the proof, as a bonus, these results extend those given in [25] to include space-times with cosmological constant.

(The same analysis also shows that in the presence of matter the given subalgebra will not be preserved under time evolution.)

However, since the self-dual connection is not gauge invariant, it does not qualify as an observable. Following J. Goldberg *et al* and using the *reduction theorem* for a connection on a principal fiber bundle they also showed that the holonomy associated with this connection is preserved under time evolution.

In addition, T. Jacobson presented new observables that arise when the structure of the holonomy group is not equal to all of SL(2,C). The local forms of the solutions for those reduced holonomy groups were also given[28].

Further results on the loop representation were given by J. Pullin reporting on work together with B. Brügmann[26]. The authors presented a simple and novel version of the constraint equations in the loop representation. To construct this constraint operator they introduced the *area derivative* in loop space. The action of this operator is obtained by first building wavefunctions based on analytic knot invariants and then applying the area derivative on each knot. The method presented provides another tool to search for physical states of the quantum theory of gravity[27].

The last talk, given by G. Immirzi, addressed the issue of reality conditions for degenerate metrics. The author argued that when degeneracy does occur, one cannot in general impose consistently those conditions. Thus, the theory describes complex general relativity[29].

References

[1] Ko M, Ludvigsen M, Newman E T, and Tod K P 1981 *Phys. Rep.* **71**, 53.

[2] Penrose R and McCallum M A H 1973 *Phys. Rep.* **6C**, 242.

[3] Ashtekar A 1991 *Lectures on Non-perturbative Canonical Gravity*, (Singapore: World Scientific).

[4] Penrose R 1991 *Gravitation and Modern Cosmology* (London: Plenum)

[5] Penrose R 1976 *Gen. Rel. Grav.* **7**, 31-52.

[6] Ward R S 1981 *Comm. Math. Phys.* **80**, 563-74; 1983 *Gen. Rel. Grav.* **15**, 105-9; 1984 *Nucl. Phys.* **B 236**, 381-396; 1985 *Phil. Trans. R. Soc. A* **315**, 451-7.
Woodhouse N M J 1987, Class. Quantum Grav. , **4**, 799-814.

[7] Mason L and Sparling G 1992 *J. Geom. Phys.* **8**, 243-271; 1989 *Phys. Lett.* **A 137** (1,2), 29-33.
Woodhouse N and Mason L 1988 *Nonlinearity* **1**, 73-114.

[8] Kozameh C N and Newman E T 1986 *Topological Properties and Global Structure of Space-Time*, Plenum, 121; 1983 *J. Math. Phys.* **24**, 2481.

[9] Kozameh C N, Lamberti P W, and Newman E T 1991 *Ann. Phys.* **206**, 193.

[10] Iyer S, Kozameh C N, and Newman E T 1992, *J.Geom. Phys.* **8**,195-209.

[11] Kozameh C N, Newman E T, and Tod K P 1985 *Gen. Rel. Grav.* **17**, 343.

[12] Plebański J F 1977 *J. Math. Phys* **18**, 2511; *J. Math. Phys* **16**, 2395.

[13] Samuel J 1987 *Pramana* **28**, L429.
Jacobson T and Smolin L 1987 *Phys. Lett. B* **196**, 39; 1988 Class. Quantum Grav. **5**, 583

[14] Capovilla R, Dell J, and Jacobson T 1991 Class. Quantum Grav. **8**, 59; 1989 *Phys. Rev. Lett.* **63**, 2325.

[15] Kozameh C N and Newman E T 1991, *Gen.Rel. Grav.* **23**, 87.

[16] Capovilla R and Plebański J F 1993; *J. Math. Phys* **34**, 130.

[17] Estabrook F and Wahlquist H 1991, Class. Quantum Grav. **8**, L151; 1989, *Class.Q. Grav* **6**, 263.
Yang, K. 1992, *Exterior Differential Systems and Equivalence Problems*, Kluwer.

[18] Wilson E and Robinson I 1992, University of Dallas preprint.

[19] Sobczyk G 1990 *Clifford ALgebras and their Applications in Mathematical Physics*, eds. J Chisholm and A Common (Reidel Pub Co.) pp 227-244.

[20] Newman E T and Rovelli C 1992 *Hamilton Jacobi theory for constrained systems and gauge invariant degrees of freedom in general relativity and Yang-Mills theory*, University of Pittsburgh preprint; 1992 to appear in *Phys. Rev. Lett.*.

[21] Goldberg J N, Newman E T, and Rovelli C 1991 *J. Math. Phys.* **32**, 2739.
Komar A 1978 *Phys. Rev.* D **18**, 1881.
Kuchar K 1972 *J. Math. Phys.* **13**, 758.

[22] Penrose R 1966 *Perspectives in Geometry and Relativity*, ed. B. Hoffman (Indiana Univ. Press).

[23] Goldberg J N, Robinson D C, and Soterou C 1992 Class. Quantum Grav. , to appear.

[24] Rovelli C and Smolin L 1990 *Nucl. Phys. B* **331**, 80.

[25] Goldberg J N, Lewandowski J, and Stornaiolo C 1991 *Degeneracy in Loop Variables*, Syracuse University preprint.

[26] Brügmann B and Pullin J 1991 *On the constraints of quantum gravity in the loop representation*, Syracuse University preprint.

[27] Brügmann B, Gambini R, and Pullin J 1992 *Phys. Rev. Lett.* **68**, 431.

[28] Jacobson T and Romano J 1992 *The spin holonomy group in general relativity*, University of Maryland preprint.

[29] Immirzi G 1992 *The reality conditions for the new canonical variables of general relativity*, Universitá di Perugia preprint.

Paper presented at the 13th Int. Conf. on General Relativity and Gravitation
Cordoba, Argentina, 1992: Part 2, Workshop Summaries

Mathematical studies of Einstein's and other relativistic equations/alternative gravity theories

Rafael D. Sorkin

Department of Physics,
Syracuse University,
Syracuse NY 13244-1130

Abstract.
The topics represented in this workshop spanned a wide range, reflecting the diversity of the very large number of abstracts submitted (nearly 100). In view of this breadth, it would be impossible to organize the present report according to subject. Instead I will just summarize the talks of the three sessions individually, pointing out now and then some connections which exist between them.

The first session began with a presentation by Daniel Sudarsky (co-author Robert Wald) devoted mainly to the understanding of some old exact solutions of pure gravity and some new ones of gravity plus gauge fields. The basic ideas involve the extremal properties of the total energy-functional E on the space of field configurations. It is known in general that, when the field equations derive from an Action-principle, a solution extremizes E iff it is stationary. Now for a black hole, stationarity represents a kind of thermal equilibrium, and this strongly suggests the following analogous assertion, special cases of which have long been known: the region outside a horizon of area A is stationary $\Longleftrightarrow$ one has $dE = TdA + \Omega dJ$ for appropriate constants T and Ω and for arbitrary variations of the exterior field configuration. Under the assumption of a "bifurcate Killing horizon", Sudarsky quoted a theorem which is essentially the assertion's "$\Rightarrow$" part, a strong form of the so-called first law of black hole thermodynamics. Under the additional assumption of a maximal slicing, he quoted a theorem which is a strengthened form of a special case of the assertion's "$\Leftarrow$" part, namely that initial data which extremize E at fixed A are actually static. Together these results show that stationary implies static for non-rotating holes, thereby filling a gap in the black-hole uniqueness theorems. Moreover the theorems still apply when gauge-fields are present, and for non-abelian gauge groups, there exist disconnected minima of the energy differing from each other by "large" gauge transformations (ones disconnected from the identity). A differential topology argument then implies the existence between these minima of saddle points of the energy, and therefore of unstable static solutions. In this way several types of recently found "colored" black hole solutions can be understood, and analogous "colored excitations" of the Kerr metric can also be predicted.

The next pair of presentations, by David Meyer and Alan Daughton, reported on work inspired by the "causal set" approach to quantum gravity. In his brief summary of this approach, which is based on the hypothesis that the Lorentzian manifold of General Relativity is only an approximation to a discrete substratum whose basic structure is that of a causal set (= locally finite partial order), Meyer pointed out that, although significant progress has been made in understanding the kinematics of causal sets, there is as yet no convincing candidate for the quantum amplitude [the discrete analog of $\exp(iS)$] on which the theory of causal set dynamics would be based. Invoking recent string-theoretic work on discretized "Euclidean" two-dimensional gravity coupled to statistical mechanical models, Meyer introduced the analogous study of Ising models on causal sets. Remarkably, one is able to solve these models exactly in several different cases. In the most interesting of these, both the Ising "spins" and the causal set itself are summed over (the latter sum being restricted to causal sets generated by random sprinkling into 2-dimensional Minkowski space) and the "partition function" turns out, for certain critical values of the parameters, to be a modified Bessel function. In this model (because 2-dimensional metrics are conformally flat) one is effectively summing over *all* (1+1)-dimensional causal sets, and in that sense is dealing with a full quantum gravity coupled to a particular kind of "matter". However the amplitude being used is not very realistic, in part because of its likely non-locality, a general difficulty highlighted in Daughton's talk. In this connection, Meyer suggested that putting the Ising model at a critical point might lead to a more satisfactory form of "induced gravity", though perhaps one still limited to the very special subclass of two-dimensional causal sets.

Daughton's presentation (co-authors Jorge Pullin, Rafael Sorkin and Eric Woolgar) also was concerned with a kind of "matter" living on "flat" causal sets, specifically with

a scalar field on a *background* causal set generated by sprinkling N points randomly into a so-called interval-subset of Minkowski space. As contrasted with Meyer's work, the additional motivation here – beyond the same ones of learning how to couple matter to a causal set, and the possibility of thereby producing a theory of "induced gravity" – was to investigate whether it is possible to recover effective locality in theories based on causal sets. (That this is not easy, is due to the inherent contradiction among discreteness, locality, and Lorentz invariance, these being three properties which any successful discrete theory must combine. For *regular* spacetime "lattices" locality proves easy to implement but Lorentz invariance difficult; for *random* "lattices" such as a sprinkled causal set, the situation is reversed. The promise of a successful combination would be great however, as locality in the presence of local Lorentz invariance is very restrictive, leading almost inevitably to an effective Lagrangian of the Einstein-Hilbert form.) Daughton presented several approaches to defining a scalar field Action which would reduce to the the usual one in the (naive) continuum limit, but he concentrated on a scheme which is based on thinking of the d'Alembertian as the inverse of a Green's function rather than vice versa. This scheme is most natural in two and four dimensions, and Daughton discussed mainly the lower-dimensional case, offering both analytic and numerical evidence for its feasibility there. Specifically, he showed that the inverse of the retarded Green's function on the (perfectly regular) "trellis-causal-set" is precisely a discretized d'Alembertian; and he described preliminary computer simulations which indicate a similar relationship – on average – for the half-retarded half-advanced Green's function on a randomly sprinkled causal set.

The presentation by John Friedman (co-authors Nicolas Papastamatiou and Jonathan Simon) concerned a possibility in a sense opposite to that animating the two previous talks, namely the possibility that spacetime might contain causality violations in the form of closed timelike curves. Such a possibility (reviewed in Kip Thorne's plenary lecture) is of interest, not only for its own intrinsic fascination, but because it provides one way for the spacetime topology to change without forsaking the globally regular Lorentzian metric mandated by the "equivalence principle". However there is grave doubt whether closed timelike curves are physically consistent, and in particular whether quantum fields can consistently propagate in spacetimes containing them. The difficulty raised by Friedman (citing also similar conclusions by David Boulware) was that of unitarity, in a situation where the causality violation is confined to a compact spacetime region. He explained that, even though free fields experience no problem, it is quite otherwise when interactions are present. Then perturbative unitarity reduces to a set of spacetime "cutting identities" whose validity relies on the Feynman propagator taking the form of a "time-ordered two-point function", a form which loses its sense in the absence of a consistent causal ordering of events. Given this destruction of unitarity by interactions, Friedman considered whether a more general sum-over-histories interpretation could salvage quantum field theory. He argued that such an interpretation does allow a consistent assignment of probabilities to histories, but at a price: the results of present experiments would depend on whether closed timelike curves form in the future.

If it is only recently that the possibility of causality violations has begun to be taken seriously, then the desire to localize gravitational energy is as old as General Relativity itself. Having pretty much abandoned the hope of defining a true energy-momentum current, people would now settle for a "quasi-local" expression giving the energy, angular momentum, etcetera, contained within some 2-surface. A particularly

natural approach to finding such an expression is to seek it as the variation of some corresponding quasilocal Action, and this is the approach adopted by David Brown, whose presentation inaugurated the second session. Defining an Action appropriate for a spatially bounded region, and varying the lapse-part of the metric on the timelike boundary led him to an energy given in terms of the extrinsic curvature of the 2-surface (with respect to the hypersurface in which it lies). This energy behaves favorably at spatial infinity and in the Newtonian limit, but has the drawback (in my opinion) that it appears difficult to generalize to a full 10-parameter set of conserved quantities.

The contributions by Brian Edgar and Graham Hall (co-author B.M. Haddow) were in the realm of pure differential geometry, but have obvious implications for the possibility – in either standard Relativity or one of its generalizations like Weyl's theory – of taking a connection (or possibly a curvature tensor) to be the fundamental physical field, rather than, for example, a metric. A series of related results was presented, but the focus was on the case of so-called Weyl connections, i.e. ones which preserve a compatible metric up to a conformal factor in the sense that $\nabla_c g_{ab} = g_{ab}\lambda_c$. The main result of Edgar asserted that a given symmetric connection ∇_a is Weyl iff there exists a putative curvature tensor and a putative metric tensor such that the former has the correct algebraic symmetries and fulfills the Bianchi identity, and the latter yields a symmetric index-pair when used to lower the raised index of the former. (Here a genericity condition has been assumed; in its absence conditions on the derivatives of the curvature enter as well.)

The main result of Hall was somewhat more global in character, being couched in terms of the holonomy group of the connection rather than its curvature tensor. It asserts that a connection in an n-manifold is Weyl iff its holonomy group is a Lie subgroup of some conformal group (rotations plus scalings) in n-dimensions. A companion result characterizes the ambiguity in the pairs (g_{ab}, λ_c) which "fit" a given connection. In addition to the conformal ambiguity from which "gauge transformations" in the original sense got their name, there can be further ambiguities only in the non-generic case where the holonomy group at a point (by definition a group of linear transformations of the tangent space) is reducible in the sense of admitting an invariant subspace; moreover there *will* be such ambiguities if the subspace is not null. However, even in these special cases, Weyl's "electromagnetic field", curl λ, remains unique.

Closely related to the Weyl "unified field theory" are the scalar-tensor gravity theories, and in particular the so-called Brans-Dicke alternative gravity theory, which is distinguished by a special choice for the coupling between ordinary matter and the scalar field. Although this theory allows in principle for an additional "scalar charge", in addition to the mass, the electric charge, etcetera, it is known that, as with Einstein gravity, the only spherically symmetric "vacuum" solution with a horizon is again the Schwarzschild metric, this being especially evident when the scalar field is constant, in which case the field equations reduce to the vacuum Einstein equations. In his contribution to the second session, however, Carlos Lousto (co-author Manuela Campanelli) presented a family of non-Ricci-flat spherically-symmetric solutions of the Brans-Dicke field equations, which remain non-Ricci-flat even in a certain $\omega \to \infty$ limit in which the scalar field becomes constant. In a certain formal sense these solutions do have a horizon, but that "surface" has infinite area, and is perhaps better interpreted as a kind of internal infinity of the spacetime. For certain values of their parameters the solutions presented are supposed to be astrophysically viable, and what is also interesting, it is

claimed that their "surface-gravity" vanishes, apparently lending them a zero horizon temperature. What seems needed now is a better understanding of the global structure of these solutions.

Another set of unified field theories in which extra scalar fields play a prominent role are those named for Kaluza and Klein. This approach to unification was represented in the workshop by the talk of Alfredo Macías (co-author H. Dehnen), who considered an 8-dimensional spacetime with "internal" dimensions in the form of a $SU(2) \times U(1)$ group-manifold. In order to provide explicit fermionic degrees of freedom, a Dirac field is introduced to complement the 8-dimensional metric. With the Higgs-like modes (the dilaton, etc.) frozen out by hand, and with a prescribed dependence of the spinor field on the internal coordinates, one obtains a dimensionally reduced Lagrangian resembling the electro-weak sector of the standard model. However, the fermions, including the neutrino, all have large masses, and of course, no Higgs fields are present. In addition the neutrino acquires an anomalous electromagnetic moment.

The third and last session of the workshop commenced with a talk by James Isenberg (co-authors Vincent Moncrief and Yvonne Choquet-Bruhat), who reported on a new theorem guaranteeing under certain conditions the existence of vacuum initial data with non-constant mean curvature. He began, however, by reviewing the rather satisfactory understanding which we have of those solutions of the initial-value constraints which do possess constant mean curvature. In that case, one specifies certain free hypersurface data (the conformal metric, λ_{ab}, the "conformal extrinsic curvature", σ^{ab} (a symmetric tensor with zero trace and divergence), and a number, τ, giving the mean curvature) and solves an elliptic equation for the conformal factor relating this data to the true metric and extrinsic curvature. The question of when this equation has a solution is fully understood, the answer depending on which of 12 cases the free data falls into, as determined by the Yamabe class of the conformal metric and by the vanishing or not of τ and σ^{ab}. Unfortunately, not all solutions of the Einstein equations admit constant mean curvature slices, and so one is forced to consider a more complicated scheme in which τ is non-constant and one has a pair of coupled equations to solve, rather than a single one. As a first step in sorting out necessary and sufficient conditions for this set to be soluble, Isenberg presented a sufficient condition which requires roughly that λ_{ab} be in the negative Yamabe class and that τ be sufficiently close to a non-zero constant. The method of proof is one which he expects to yield existence for a much more general class of data, as well.

The final two presentations to be discussed were connected more or less closely with string theory. That by Robert Mann, dealt more generally with 2-dimensional gravity, or rather with what he called "dilaton gravity", in which a scalar field multiplies the scalar-curvature term in the Lagrangian, and may couple also to the "matter fields" if such are present (there is of course no "Einstein-Hilbert gravity" in 2-dimensions, at least classically, since the unadorned scalar curvature is a total divergence there). Working with a rather general class of such theories, Mann introduced a certain vector field ξ^μ for which he could show, in many cases, that $J^\mu := T^{\mu\nu}\xi_\nu$ is conserved, even though ξ^μ itself is not in general a Killing vector. Interpreting the potential for this current as a "mass function", he illustrated the relations found on some "black-hole" solutions, ending with the provocative conclusion that the same black-hole metric can be associated with two very different entropies, depending on which theory one interprets it within.

Finally Tevian Dray (reporting on behalf of Corinne Manogue) described a group-

theoretic result of interest for octonionic string theory, as well as more generally. The relevant mathematical coincidences here are that $SO(9,1)$ is isomorphic to $SL(2,\mathbf{O})$ ($\mathbf{O}$ being the octonions), and that correspondingly vectors in 10-dimensional Minkowski space can be represented in terms of octonions analogously to how 4-vectors can be represented as 2×2 complex matrices. Now, however, there is a puzzle. If X is the 2×2 octonionic-valued matrix representing a 10-vector, and M a matrix in $SL(2,\mathbf{O})$, then $X \rightarrow MXM^{\dagger}$ does not yield all possible Lorentz transforms on X, despite the isomorphism we began with. The resolution is that one obtains the missing transformations by *iterating* the simple ones, the former not collapsing to the latter precisely because octonions are non-associative. Dray showed explicitly how this works in a simple example.

Asymptotia, singularities and global structure

C J S Clarke
Faculty of Mathematical Studies, University of Southampton, Southampton, UK, SO9 5NH

1. Asymptotia

A dominant theme in the study of asymptotic properties of space-time has been that of $\mathscr{I}$: the null boundary attached 'at infinity' to certain asymtoticaly flat space-times by embedding them in a larger space-time with a conformally related metric called the unphysical metric (Geroch, 1976). When $\mathscr{I}$ was first introduced, there was a spate of activity in exploring its geometrical structure, leading to the establishment of concepts such as Bondi mass, peeling, conserved quantities and news functions. Subsequently activity turned to the complex extensions of $\mathscr{I}$ and many attempts to incorporate angular momentum along with the Bondi mass (e.g. Bramson, 1975). All this early work depended on assuming that $\mathscr{I}$ was "decent": the unphysical metric being required to be smooth and to satisfy causal properties such as strong causality. More recently this has been called into question as it was realised that such apparently innocuous properties represented subtle and highly non-obvious global properties of the physical space-time. Consequently Newman (1989) was able to show, by relaxing the causality requirements in the unphysical space-time (but not in the physical) that the global topology of $\mathscr{I}$ could be very far from the simple $S^3 \times \mathbb{R}^1$ usually assumed.

An increasing worry became the question of the existence of the traditional $\mathscr{I}$, in generic cases (or, more specifically, radiative cases) other than the well-known stationary examples and minor modifications of these: the work of Christodoulou and O'Murchadha (1981), which investigated existence in a neighbourhood of i^0 which did not extend as far as $\mathscr{I}$, indicated that the sort of asymptotic flatness conditions that seemed appropriate to an initial hypersurface did not lead to a smooth $\mathscr{I}$. The first existence theorems for $\mathscr{I}$ were due to Friedrich (1986,1988), but while establishing the existence of some genuinely radiative space-times with a traditional $\mathscr{I}$, they reinforced the view that strong restrictions were required for this.

The one paper on asymptotia in the workshop was important, therefore, for announcing the development of techniques for showing the existence of $\mathscr{I}$ more generally, and for advancing the question of differentiability. Anderson [1] showed that if data is prescribed on a spacelike hypersurface intersecting $\mathscr{I}$, then in general the solution obtained will be C^2 but not C^3 in the unphysical space. This suggests that many of the properties traditionally associated with asymptotia (peeling, the Penrose conserved quantities and so on) may well not obtain, in general.

2. Singularities

2.1. Strong cosmic censorship

It has become abundantly clear that some singularities (ideal points where a smooth solution to Einstein's equations breaks down) are censored – hidden behind event

horizons – while others are not. The key question is how to characterize which is which. Strong cosmic censorship holds that, under conditions yet to be determined, singularities are not only censored in being invisible from infinity, but also are censored "locally"; roughly speaking, they cannot be timelike.

The main area of exploration recently has been that of a distribution of collisionless particles, either having a distribution of velocities at each point (the Einstein-Vlasov equations) or comoving with a single velocity at each point (pressureless fluid or 'dust'). Much of the work has concerned spherically symmetric situations, although exciting numerical work by Teukolsky (this volume) has simulated a naked singularity for the axisymmetric Einstein-Vlasov case. Naked singularities also appear in the Vaidya metrics, which can be regarded as a limiting case of the dust solutions in which the flow vector becomes null. A useful survey unifying these various metrics was provided by P Lemos [7].

So far it has been possible to argue that these naked singularities are either "special", or are not really singularities. Their specialness lies partly in their exceptional equation of state, considered in the next paragraph, and partly in their high degree of symmetry. The Teukolsky example just quoted certainly weakens spherical symmetry, which is highly significant, but it is still non-generic even within the class of axially symmetric Einstein-Vlasov spaces because the initial distribution of velocities is degenerate (the particles start from rest). The idea that some of the examples (those called shell-crossing singularities) are not singularities lies in the fact that in is possible to extend through the apparent singularity into a larger space-time in which Einstein's equations are satisfied distributionally (Clarke, 1992).

The obvious question then becomes, whether these singularities are still naked if pressure is included. P Szekeres [12] gave a very interesting presentation in which null coordinates were used to examine the singularity asymptotically, from which he concluded that generically the pressure had to be negative in order to achieve a naked singularity. More work is needed to understand the relation of this to the work of Ori and Piran (1988) which demonstrated naked singularities with positive pressure for self-similar solutions; though this does not contradict Szekeres whose genericness condition rules out self similarity.

2.2. Weak cosmic censorship

This is the idea that singularities, though perhaps locally naked, are at least shielded from infinity by a horizon. Since horizons are implied by trapped surfaces (though this is only part of the story) the question then becomes, when does a trapped surfce form? There is a large gap between the unproved, and even unformulated, "hoop conjecture" (Teukolsky, this volume) and the only proved criterion (Schoen and Yau, 1983) which is so strong as to be practically unattainable. In spherical symmetry, however, considerable progress has been made by Bizon et al (1989); work which was extended to the cosmological context in a presentation of a paper by Brauer and Malec [2]. The conditions in this latter paper were unfortunately not so clean as in the earlier one, but they still provide useful criteria as to whether or not there is a trapped surface in the spherically symmetric case.

2.3. Stability arguments

These enter both forms of cosmic censorship. The original arguments for strong cosmic censorship was that the singularity structure of the Reissner-Nordström solution was unstable, linearised perturbation theory suggesting that the horizon became singular, thus turning the "timelike" singularity into one compatible with strong cosmic censorship. Królak presented a non-perturbative proof [6] that the horizon was indeed unstable: specifically, he and Rudnicki had showed the instability of the compactness of the intersection of the past of a point on the horizon with the partial Cauchy surface. This is in itself interesting: this compactness property is one that would naturally be expected if the singularity arose from data on a finite region of the initial surface, and so the result may call into question the picture one sometimes has of singularity formation as an event arising from a finite region of collapse.

Stability also enters one program for proving weak cosmic censorship. The program involves showing that (i) generically singularities are in some sense "strong"; and then (ii) that strong singularities have to be censored from infinity. The first step is supported by results such as that of Brauer [2] on Newtonian fluids, described below, but in general genericness seems to be crucial: one has to show that the weak singularities (such as those where all scalars constructed from the Riemann tensor remain bounded) are in some way unstable, either disappearing or becoming strong under perturbations. A series of results by Konkowski and Helliwell have supported this view, and the latest, giving evidence that the singularity in Type V cosmological solutions was unstable against becoming a scalar curvature singularity, was presented to the workshop [5].

2.4. Global Solutions

The only cast-iron results on naked singularities, one suspects, come from global existence theory of the sort discussed by Klainerman (this volume). Here there was a contribution from Rein and Rendall [10], showing that the spherically symmetric Einstein Vlasov system, with bounded momenta for the particles, remained singularity free for all time provided the initial data were sufficiently small.

Further information about global solutions came from Newtonian work presented by Brauer [2], showing that for the Euler-Poisson system (gravitating compressible fluid) a breakdown in the global solution has to be related to unboundedness of physical quantities. While this is no surprise (unlike the case of general relativity where many singularities are known in which physical quantities are bounded) it is striking that a proof in the Newtonian case has only just emerged. The main technical feature was the introduction of a new variable to cope with the loss of differentiability at the boundary of the compact region occupied by the matter (the edge of the "star").

2.5. Boundaries

At one time a popular approach to singularities was through the provision of a boundary to space-time whose points could be characterised either as inessential (through which extension was possible) or singularities. This fell into disfavour because of the wide variety of possible boundaries and the difficulty of actually constructing any of them. Recently Scott [11] has revived the idea, in collaboration with Szekeres, in a radically new form by defining an abstract boundary based purely on the topology of the manifold, in a way that is both very general and reasonably computable. This

then provides a framework within which a variety of specific properties of singularities can be discussed. The idea is certainly elegant and free from many of the objections raised against previous uses of the idea; but it remains to be seen whether any really new results can be obtained by its use.

3. Global Structure

3.1. Causality Violation

This has always been a major topic of global analysis, and has recently become fashionable (Thorne, this volume). The workshop received one contribution to this topic from S M Carroll [3], who showed that the presence of Gott-style string time-machines in an open (2+1)-dimensional universe required the total momentum to be spacelike, which he argued was unphysical. For me the attraction of the paper was not so much its physical implications, which remain uncertain, but the elegance of handling the group theory associated with the Lorentz transformations generated by the strings. This was done by regarding the Lorentz group as itself a space-time, and then using techniques of global relativity on it.

3.2. Horizons

Two papers discussed the properties of horizons. Wald and Racz [9] considered Killing horizons, and showed that the only physically relevant ones were either bifurcate or had the surface gravity κ equal to zero (thus making rigorous an argument suggested by de Felice and Clarke, 1990). This allows one to use the uniqueness theorems for stationary space-times, which require the existence of a bifurcate horizon.

The second paper was by P Chrusciel, who examined extensions through the Cauchy horizon of Robinson-Trautman space-times. By definition, predictability breaks down and not surprisingly, therefore, an infinite number of differentiable solutions are possible which extend the space-time through the horizon. An amusing curiosity of the result is that the extension can be chosen to have differentiability C^{117} but not more than C^{122}!

3.3. Caustics

Null surfaces propagating in curved space-time generically develop caustics, a feature that imposes a major complication of some approaches to numerical relativity. The general form of these is known in the generic case from the work of Arnol'd (1970), and has been taken into relativity by Friedrich and Stewart (1983). The one poster displayed in the workshop, by V Perlick [8], extended this work by considering null cones specifically and giving examples of explicit metrics that produced the various singularity types for caustics.

4. Conclusions

The workshop exhibited well the need for bringing together a broad range of techniques in order to make progress on the outstanding problems of global structure. The

cleanness and elegance of some of the early work may be fading as we get to grips more closely with real astrophysical problems, and realise that the classical $\mathscr{I}$ may be an idealisation; that very sophisticated boundary constructions have nothing to tell us; and that topological methods fail to make contact with real physics. All these approaches are none the less essential as tools to be used in conjunction with each other.

By the next GR meeting I would expect that work on asymptotia will have fixed on a broad class of smoothness-conditions for $\mathscr{I}$ which are much weaker than those used in the past, thereby cutting out a lot of the classical work in the subject. But at the same time there could well be a move away from an increasing refinement of the conditions for asymptotic flatness, and a start being made on asking what such concepts mean in terms of real observations at finite distances from the source, within a cosmological context.

Continuing this futuristic speculation, on singularities we might expect to be obtaining a more realistic grasp of which singularities might be observable, from the point of view of the classical theory, and what this might mean in reality. There is still astonishingly little real geometrical/physical understanding of what is going on in singularity formation, even in simple cases. It is not sufficient to dismiss examples like that of Ori and Piran (1988) as non-generic: we need to understand precisely what aspect of the non-genericness is responsible for singularity formation in what is apparently a very realistic model. The way forward here will certainly involve much more numerical simulation, to supply more hypotheses and physical hunches in an area where new ideas are urgently needed.

References

Titles and authors of abstracts, appearing in the GR13 volume of abstracts, for workshop talks and posters.

1 L Anderson. On the Obstructions to Smoothness of Scri
2 U Brauer and E Malec. Trapped Surfaces due to Spherical Inhomogeneites in Expending Open Universes
3 S M Caroll, E Farhi and A H Guth. Obstacles to Time Machine construction in 3+1 dimensions
4 P T Chrusciel. Global structure of the Robinson-Trautman Black Holes
5 D A Konkowski and T M Helliwell. Stability analysis of a Nonscalar Curvature Singularity
6 A Królak and W Rudnicki. On Instability of the Kerr-like Cauchy Horizons
7 J P S Lemos. Naked singularities: A Unified Approach for Gravitationally Collapsing Spheres of Dust or Radiation
8 V Perlick. Classification of Caustics
9 I Racz and R M Wald. Local Extensions of Space-times with Killing Horizons
10 G Rein and A D Rendall. Global Dynamics of Solutions of thee Spherically Symmetric Vlasov-Einstein System
11 S M Scott. An Introduction to the Abstract Boundary
12 P Szekeres. Singularities of Sherically Symmetric Space-times and Strong Cosmic Censorship

Other references.

Arnol'd, V I, Gusein-Zade, S M and Varchenko, A N (1970) *Singularities of Differentiable Maps* I

Bizon, P, Malec, E & O'Murchadha N (1989) Trapped surfaces due to concentrations of matter in spherically symmetric geometries, *Class. Quant. Grav.* **6** 961—976

Bramson, B (1975) Relativistic angular momentum for asymptotically flat Einstein-Maxwell manifolds, *Proc. Roy. So. Lond. A* **341** 463—490

Christodoulou, D & O'Murchadha, N (1981) The boost problem in general relativity, *Comm. Math. Phys.* **80** 271—300

Clarke, C J S (1992) in *The Renaissance of Relativity and Cosmology*, ed. Lanza, A, Cambridge University Press.

de Felice, F & Clarke, C J S (1990) *Relativity on curved manifolds*, Cambridge University Press

Friedrich, H (1986) On the existence of n-geodesically complete or future complete solutions of Einstein's field equations with smooth asymptotic structure, *Comm. Math. Phys.* **107** 587—609

Friedrich, H (1988) On static and radiative space-times *Comm. Math. Phys.* **119** 51—73

Geroch, R (1976) in *The asymptotic structure of space-time* ed. Esposito P & Witten L, p. 1, Plenum

Friedrich, H & Stewart, J M (1983) Characteristic initial data and wave-front singularities in general relativity *Proc. Roy. Soc. Lond. A* **385** 345

Newman, R P A C (1989) The global structure of simple space-times *Comm. Math. Phys.* **123** 17—52

Ori A & Piran T (1988) Self-Similar Spherical Gravitational Collapse and the Cosmic Censorship Hypothesis *Gen. Rel. Grav.* **20** 7—13

Schoen, R C & Yau, S T (1983) The existence of a black hole due to condensation of matter *Comm. Math. Phys.* **90** 575—579

Paper presented at the 13th Int. Conf. on General Relativity and Gravitation
Cordoba, Argentina, 1992: Part 2, Workshop Summaries

Approximation and perturbation methods

B.R.Iyer [1]

Raman Research Institute, Bangalore, 560 080, INDIA

Few problems in nature are amenable to an exact solution and hence when one proceeds from elegant problems of theory to messy complicated problems of practice one is forced to recourse to methods of approximation and perturbation. The development of such techniques has been natural in attempts to extract physically verifiable consequences from either exact solutions of general relativity or from specific astrophysical systems for which an exact solution is impossible to find. However, this should not be taken to imply giving up of mathematical rigour and an appeal to only physical intuition.

1. Approximation Methods

Though the topic of approximation methods have been with us since the inception of the theory of general relativity it received a fresh impetus with the observation of the secular acceleration in the mean orbital motion of the Binary Pulsar. Suddenly, the observations were getting to be accurate enough to make measurable higher order effects coming from general relativity and the theorist had also to update his tools and make more precise the conceptual foundations that formed the paradigm for matching the increasingly accurate observations to a theoretical model. The main class of approximation schemes that have been developed so far are the Post Newtonian Approximation(PNA) and the Post Minkowskian Approximation(PMA). Originally developed in the context of the solar system these old approximation schemes used a global coordinate system, a global weak field assumption and a single asymptotic expansion. The need to treat binary systems containing two neutron stars or black holes requires looking at regimes where strong field effects come into play. A more detailed description incorporating the clumpiness of the universe is also called for in cosmology. These new problems require new approximation methods characterised by the use of several coordinate systems and several asymptotic expansions[1].

What is the relation between the approximation methods and the exact theory? How do we go about investigating this? This was the theme of Alan Rendall's talk on 'Approximation methods in theory and practice'. It was concerned with the passage from practical use of approximations which are heuristically defined to rigorous theorems on how well and in what sense, the results of calculations of this kind, approximate solutions of the exact equations. A possible programme for such an undertaking could be : First, find a definition which on the one hand looks likely to provide a basis for rigorous theorems and on the other hand is relevant to practical calculations. Next, use this

[1] E-mail: bri@rri.ernet.in

definition to prove that the approximations are qualitatively good in some sense. Finally, obtain quantitative estimates of the difference between exact and approximate solutions. For PMA a definition was given by Blanchet and Damour [2] which was tailored to their practical calculations and proved useful for questions of rigorous justification. Damour and Schmidt[3] showed on the basis of these definitions that in general circumstances PMA are asymptotic to solutions of Einstein's equation. That this is essentially optimal follows from the fact that the series obtained does not converge[4]. Quantitative estimates seem out of reach at present and is a challenge to the theory of partial differential equations which rarely yield quantitative information. On the other hand though a good definition of the newtonian limit of general relativity has existed for some time[5] its relation to PNA was unclear(however, see[6]). Recently, a definition of PNA has been given[7] which appears to have a good chance of linking theory and practice. Work is in progress towards justification of the PNA. The first step is the proof that the spherically symmetric Vlasov-Einstein system has a regular newtonian limit[8]. A comparison of the new definition of PNA with the traditional approach e.g. Chandrasekhar[9] shows that the most obvious difference is that in the former case all the variables including the matter variables are expanded in powers of $c\lambda-1$ whereas in the latter case only the variables describing the gravitational field are expanded. This leads to problems in regard to secular effects and quantitative information seems absolutely necessary to pin down good properties of PNA.

The new approximation methods are used mainly to investigate problems of motion and the generation problem in gravitational radiation theory. They are also used in problems of relativistic celestial mechanics as explained by Chongming Xu. He summarized the new formalism of Damour, Soffel and Xu[10] for studying general relativistic celestial mechanics of systems of N arbitrary, weakly self- gravitating, rotating bodies using a sophisticated version of the first PNA. It is characterised by use of a multi-reference system; a global one for describing overall dynamics of N bodies and N local systems to describe the internal structure of each body. The special features of the scheme are that the field equation and transformation laws are linear; the structure of the energy-momentum tensor is left open; each body is characterised by the Blanchet-Damour PN multipole moment[11] and external PN tidal moments. This allows one to obtain complete and explicit results for laws of translational motion at 1PN level for bodies with arbitrary composition and shapes. One can obtain an expression for the tidal moment in terms of PN multipole moments of other bodies. The discussion of PN spin motion requires in addition the Damour-Iyer[12] spin moments.

Moreschi talked about approximation methods around stationary systems. He argued that since the asymptotic symmetry group is the infinite dimensional BMS group and not the ten dimensional Poincaré group it led to an arbitrariness both in the definition of physical concepts associated with the system and also the best flat background to expand around. Consequently, approximation methods around a fixed stationary background metric can give a consistent description of the system at a fixed time at most[13].

Cutler commented briefly on their recent work[14] on calculation of inspiral waveforms using the Regge-Wheeler-Teukolsky perturbation formalism. The relevant equations were solved numerically to very high accuracy and a post newtonian expansion was fit to the numerical results. This lead to the surprising result that higher order terms were not getting smaller causing the template waveform to go out of phase with the

general relativity waveform very quickly.

Will presented the recent results of Iyer and Will[15] on gravitation radiation reaction in equation of motion of binary systems at PN order beyond the quadrupole approximation. The method is based on PN expression for energy and angular momentum flux to infinity and an assumption of energy and angular momentum balance. The arbitrariness in the formula is related to the coordinate system dependence of the radiation reaction formula. As mentioned earlier it is important to know this secular damping very accurately so that the theoretical template not lose phase with the observed signal.

2. Perturbations

The subject of perturbations in general relativity has developed into a specialized discipline on its own. Bulk of the work in this area has been in one of the following two topics: Cosmological Pertubations and Black Hole Perturbations.

2.1. Cosmological Perturbations

The basic questions that started these investigations was the attempt to understand the formation of structure in Cosmology and the growth of inhomogeneities in the expanding universe. Recently, a new covariant approach has been given to study the perturbations and this was summarized by George Ellis in his presentation.The gauge problem of perturbations in cosmology is the arbitrariness in the perturbed quantities arising from the arbitrariness in the choice of the map between the background spacetime and the real spacetime. The gauge problem in perturbed Robertson-Walker cosmologies has not been resolved in a satisfactory way. Bardeen's[16] introduction of gauge invariant variables was a major triumph. However, the formalism and method are not geometrically transparent, the split into scalar, vector and tensors is nonlocal/nonunique, it is not easily related to observations and cannot be easily extended beyond linear order because it is linearized ab initio. Moreover, the analysis is not invariant under general gauge transformations but only under a restricted set that respects the harmonic splitting. A more transparent gauge invariant formalism has been set up using fully covariant methods in terms of variables that are both gauge invariant and covariantly defined, leading to covariant evolution equations. The basic variables are spatial gradients of density, pressure and expansion of the cosmological fluid taken orthogonal to the fluid flow vector. Since they vanish in a Robertson-Walker universe they are gauge invariant and characterise inhomogeneities in the universe. The basic formalism is set up and applied to pressure free matter[17], perfect fluid [18], scalar field[19], multi-fluids and imperfect fluids[20]. Subtle effects due to rotation[21], relation to the Bardeen's approach[22], density waves in cosmology [23] and applications to newtonian cosmology[24] have also been investigated. The main advantages of this formalism is that the geometrical definition is clear, it is defined in an arbitrary spacetime(no background is needed), nonlinear equations can be obtained, the variables are observable in principle and finally there exist newtonian analogues of all equations. The effect of this programme has so far been to rederive standard results in a more transparent gauge invariant way as also more generally valid equations before linearizing about Robertson-Walker. The idea of density waves in cosmology is new. The formalism is also being used to study the Sachs-Wolfe effect[25] and clarify the gauge invariance of these calculations.

2.2. Black Hole Perturbations

The issues that led to the development of this subject were the following: Are black holes stable against small changes? Can one compute what happens when a test particle or radiation scatters off the black hole? What are the frequencies in which the black hole rings and how do the notes die off? The studies were made with all the analytically known black hole solutions: Schwarzschild, Reissner-Nordstrom, Kerr and Kerr-Newman e.g[26]. A variety of methods have been used in these investigations: scalar, vector and tensor harmonics, Newman-Penrose formalism, Debye-Hertz potentials, gauge invariant approaches e.g[27].

2.2.1. Separability of Wave Equations in Curved Backgrounds. One of the most remarkable results was the separability of the perturbation equations of Hamilton Jacobi, scalar, electromagnetic, gravitational, neutrino and electron fields in the background of the Kerr-Newman black hole. The miracle continues and some time back the separability of a Nambu-Goto string configuration in the Kerr background was also established[28]. All this led on to a more critical investigation of the question of separability, the operators commuting with the wave operator and operators whose eigenvalues the separation constants are. The status of issues related to the separability of wave equations on curved backgrounds was the subject of Ray McLenaghan's presentation. A symmetry operator of the equations satisfied by the variables of a physical system is a linear differential operator that maps the space of solutions into itself. The most familiar examples of symmetry operators are operators which commute with the differential operator appearing in the field equations. They are called constants of the motion and their eigenvalues are interpretable as quantum numbers of the system[29]. A remarkable example of such an operator is given by Carter and McLenaghan's[30] discovery of a first order commuting operator for the Dirac operator on Kerr spacetime by an analysis of the separation of variables procedure devised by Chandrasekhar[31]. They showed that these operators admit the separable solutions as eigenfunctions with the corresponding eigenvalues as separation constants and characterised one of them in terms of a valence two Killing spinor satisfying a skew-hermiticity condition. McLenaghan and Spindel[32] gave a tensorial expression for the most general first order commuting operator with the charged Dirac operator on a general curved background in terms of Killing-Yano tensors of valence one, two and three. A different situation arises when dealing with the conformally invariant Klein-Gordon, Dirac and Maxwell equations for zero rest mass particles where symmetry operators which are not necessarily commuting operators must be considered. In the case of the conformally invariant Klein-Gordon equation and the Dirac equation for the neutrino, the symmetry operators appear in the form of R-commuting operators that is operators whose commutators with the wave operator are proportional to it. All such operators up to the second order for the Klein-Gordon equation and up to first order for the Dirac equation have been characterised by Kamran and McLenaghan[33] in terms of conformal Killing vectors and conformal Killing tensors of valence two and conformal Killing-Yano tensors of valence one, two and three respectively. A corresponding analysis for Maxwell's equations has proved more elusive. However, Kalnins, Miller and Williams[34] recently found a second order symmetry operator for Maxwell's equations in the Kerr solution which characterises the separable solutions found previously by Teukolsky. The most general second order symmetry operator for Maxwell's equations on a general curved spacetime has been constructed[35]. This uses a conformal Killing

vector, a conformal Killing tensor of valence two and a new valence four tensor with the same algebraic symmetries as the Weyl tensor which satisfies a first order differential equation which has similarities to both the conformal Killing equation and the conformal Killing-Yano equation. This tensor corresponds to a valence four Killing spinor; its properties and integrability conditions for its existence have been studied.

2.2.2. Quasi-Normal Modes. At late times, all perturbations of the black hole are radiated away like the last pure dying tones of a ringing bell. To describe this 'ringing' the notion of quasinormal modes (QNM) was introduced around 1970. QNM's are the sourceless perturbations of spacetime and in the case of the time-evolution of small perturbations of a Schwarzschild black hole are governed by a one-dimensional wave equation. The QNM frequencies are characteristic of the black hole, and (in the Schwarzschild case) depend only on its mass. QNM's excited during for example a gravitational collapse may be eventually detected. Thus the determination of QNM's of black holes is an important problem on which considerable effort and progress has been made in the last three years. Nils Andersson summarized the current status of these calculations as follows. Recently, Nollert and Schmidt [36] proved that the QNM's should be properly defined as poles of the Green's function to the Laplace transformed wave equation. In a simplified picture, the desired solutions correspond to boundary conditions of purely outgoing waves arriving at spatial infinity, and purely ingoing waves crossing the event horizon. The desired solutions to the radial problem increase exponentially towards spatial infinity and the event horizon. To identify a QNM solution an exponentially decreasing solution must be singled out from the exponentially increasing one in the asymptotic region. Hence, the determination of QNM frequencies is a delicate problem. During the last ten years several attempts to determine the QNM frequencies have been made. Leaver[37] determined accurate values for them using a continued fraction approach. Recently Leaver's results have been confirmed as reliable using numerical integration in the complex coordinate plane [38]. For gravitational perturbations, recent double-precision calculations by Leaver (unpublished) agree to nine decimal places with the numerical integration results. Nollert and Schmidt have also verified Leaver's results. These three independent investigations of the problem yield results that agree perfectly for the first ten modes. A semi-analytical phase-integral method has also been applied to the problem [39]. A powerful phase-integral formula determining QNM frequencies for a Schwarzschild black hole has been derived by Andersson and Linnæus[40] . The results obtained from this formula are in good agreement with the numerical results mentioned above. This method may also prove to be powerful in situations more general than the Schwarzschild case, where the methods mentioned above are difficult to apply.

Omar Ortiz talked about recent work[41] on hyberbolizing the heat equation. Systems described by parabolic or hyperbolic-parabolic systems have arbitrarily high propagation velocities incompatible with relativity. One therefore looks for a one parameter family of hyberbolic systems that goes over to the parabolic system in an appropriate limit. For the heat diffusion equation, such an analysis can be done, using as parameter the reciprocal of the maximum speed of propagation. The solution to this family equals the solution to the heat equation plus terms that vanish when the maximum speed of propagation becomes infinite.

3. Open Questions

We conclude with a list of important open questions that remain in these areas[42, 43]. (i) Characterisation of the Teukolsky parameter for gravitational perturbations in Kerr background by a symmetry operator. (ii) Solve the perturbation problem for Kerr-Newman *i.e* exhibit the relevant symmetry operator and its solutions. (iii) Extend the results of separability to other backgrounds like Cosmological and String backgrounds. Do generalized Hertz potentials exist? And if they do, how do we find them? (iv) What can we say about the completeness of QNM's in the neighbourhood of the black hole? (v) Investigate the relation between Einstein's Theory and PNA. (vi) Prove existence theorems for various sources since validity of approximation methods cannot be judged otherwise. (vii) Put procedures used to relate near zone approximations with far fields on a sound basis since they are always used in most applications to astrophysical systems. Normally one relates a higher order PN description of sources to PMA of higher order. (viii) Are different PNA schemes like Chandrasekhar's, Damour- Soffel- Xu and Rendall's equivalent? (ix) Do PNA methods have a fundamental limitation? Can their convergence be improved using better numerical techniques? Are some variables better than others? Or do we need a very different approximation scheme?

Acknowledgements

It is a pleasure to thank Cliff Will and Bernard Schutz for hospitality at the Washington University, St. Louis and University of Wales, Cardiff respectively, where parts of this article were written.

References

[1] Damour T 1987 in *300 years of Gravitation* eds. S. Hawking and W. Israel (Cambridge: Cambridge University Press)

[2] Blanchet L and Damour T 1986 *Phil. Trans. R. Soc.* A **350** 379

[3] Damour T and Schmidt B 1990 *J. Math. Phys.* **31** 2441

[4] Rendall A D 1990 *Class. Quantum Grav.* **7** 803

[5] Ehlers J 1991 in *Classical Mechanics and Relativity: Relationship and Consistency* ed. G. Ferrarese (Naples: Bibliopolis)

[6] Schutz B 1985 in *Relativity, SUSY and Cosmology* eds. O. Bressan, M. Castagnino and V. Hamity (Singapore: World Scientific)

[7] Rendall A D 1992 *Proc. R. Soc.* A To Appear

[8] Rein G and Rendall A D 1992 *Comm. Math. Phys* To Appear

[9] Chandrasekhar S 1965 *Astrophys. J* **142** 1488
Chandrasekhar S and Nutku Y 1969 *Astrophys. J* **158** 55

[10] Damour T, Soffel M and Xu C 1991 *Phys. Rev.* D **43** 3273

[11] Blanchet L and Damour T 1989 *Ann. Inst. Henri Poincaré, Phys. Théor.* **50** 377

[12] Damour T and Iyer B R 1991 *Ann. Inst. Henri Poincaré, Phys. Théor.* **54** 115

[13] Moreschi O M 1988 *Class. Quantum Grav.* **5** L53

[14] Cutler C *et al* 1992 The last three minutes: issues in the measurement of coalescing compact binaries by LIGO *Phys. Rev. Lett.* Submitted for Publication

[15] Iyer B R and Will C M 1992 Post-Newtonian gravitational radiation reaction for two-body systems *Phys. Rev. Lett.* Submitted for Publication

[16] Bardeen J 1980 *Phys. Rev.* D **22** 1882
Stewart J 1990 *Class. Quantum Grav.* **7** 1169

[17] Ellis G F R and Bruni M 1989 *Phys. Rev.* D **40** 1804

[18] Ellis G F R, Hwang J and Bruni M 1989 *Phys. Rev.* D **40** 1819

[19] Bruni M, Ellis G F R and Dunsby P K S 1992 *Class. Quantum Grav.* To Appear

[20] Dunsby P K S, Bruni M and Ellis G F R 1992 *Astrophys J* To Appear

[21] Ellis G F R, Bruni M and Hwang J C 1990 *Phys. Rev.* D **42** 1035

[22] Bruni M, Dunsby P K S and Ellis G F R 1992 *Astrophys. J* To Appear

[23] Ellis G F R, Hellaby C and Matravers D R 1990 *Astrophys. J***364** 400

[24] Ellis G F R 1990 *Mon. Not. R. Ast. Soc* **243** 509

[25] Stoeger W R, Ellis G F R and Schmidt B G 1991 *Gen. Rel. Grav* **23** 1169

[26] Chandrasekhar S 1983 *The Mathematical Theory Of Black Holes* (New York: Oxford University Press)

[27] Vishveshwara C V 1988 in *Highlights in Gravitation and Cosmology* eds. B. R. Iyer, A. K. Kembhavi, J. V. Narlikar and C. V. Vishveshwara (Cambridge: Cambridge University Press)

[28] Carter B and Frolov V 1989 *Class. Quantum Grav.* **6** 569

[29] Carter B 1977 *Phys. Rev.* D **16** 3395

[30] Carter B and McLenaghan R G 1979 *Phys. Rev.* D **19** 1093

[31] Chandrasekhar S 1976 *Proc. R. Soc.* A **349** 571

[32] McLenaghan R G and Spindel P 1979 *Phys. Rev.* D **20** 409

[33] Kamram N and McLenaghan R G 1984 *J. Math. Phys.* **25** 1019
——1984 *Phys. Rev.* D **30** 357
——1985 *Lett. Math. Phys.* **9** 65

[34] Kalnins E G, Miller W and Williams G C 1989 *J. Math. Phys.* **25** 2360

[35] Kalnins E G, McLenaghan R G and Williams G C 1992 in *Proc. of 4th Canadian conf. on Gen. Rel. and Rel. Astrophys.* (Singapore: World Scientific)

[36] Nollert H and Schmidt B G 1992 *Phys. Rev.* D **45** 2617

[37] Leaver E 1985 *Proc. R. Soc.* A **402** 285

[38] Andersson N 1992 *Proc. R. Soc.* A To Appear

[39] Fröman N, Fröman P O, Andersson N and Hökback A 1992 *Phys. Rev.* D **45** 2609

[40] Andersson N and Linnæus S 1992 *Phys. Rev.* D Submitted for Publication

[41] Nagy G B, Ortiz O E and Reula O A 1992 *Hyperbolising the Heat equation* In Preparation

[42] Kalnins E G 1992 Private Communication

[43] Schmidt B G 1992 in *Advances in Gravitation and Cosmology* eds. B. R. Iyer, A. R. Prasanna, R. K. Varma and C. V. Vishveshwara (Delhi: Wiley Eastern) In Press

Paper presented at the 13th Int. Conf. on General Relativity and Gravitation Cordoba, Argentina, 1992: Part 2, Workshop Summaries

Numerical relativity

Takashi Nakamura

Yukawa Institute for Theoretical Physics, Kyoto University, Kyoto 606 , JAPAN

In GR13 we heard many reports on recent progress as well as future plans of detection of gravitational waves. According to these reports (see the report of the workshop on the detection of gravitational waves by Paik in this volume), it is highly probable that the sensitivity of detectors such as laser interferometers and ultra low temperature resonant bars will reach the level of $h \sim 10^{-21}$ by 1998. In this level we may expect the detection of the gravitational waves from astrophysical sources such as coalescing binary neutron stars once a year or so. Therefore the progress in numerical relativity is urgently required to predict the wave pattern and amplitude of the gravitational waves from realistic astrophysical sources. The time left for numerical relativists is only six years or so although there are so many difficulties in principle as well as in practice.

Apart from detection of gravitational waves, numerical relativity itself has a final goal:

Solve the Einstein equations numerically for *any* initial data as accurately as possible and clarify physics in strong gravity.

In GR13 there were six oral presentations and 11 poster papers on recent progress in numerical relativity. I will make a brief review of six oral presentations. The Regge calculus is one of methods to investigate spacetimes numerically. Brewin from Monash University Australia presented a paper *Particle Paths in a Schwarzshild Spacetime via the Regge Calculus*. One of the merits in the Regge Calculus is that the metric interior to each pair of adjacent blocks is Minkowskian so that the computations of particle or photon orbits can be performed using only the rules of special relativity. Williams and Ellis[1] formulated how to compute

particle paths in the Regge spacetime. Unfortunately the value they obtained for the precession of the perihelia of Mercury was at best 1800 times larger than the correct one. Brewin argued how to recover the correct value by developing the approach in the context of the Regge spacetime. The first point is that he uses the continuous time 3+1 approach, that is, he discritizes space while retaining a continuous time. Next he argues the convergence properties for several Schwarzschild geodesics in his method. The question here is : if the number of blocks (N) is increased (i.e. $N \to \infty$), will we obtain the correct geodesics? The answer is NO. If the error in each block is of order $O(h)$ where h is a linear size of a block the global error is of order unity because the total number of blocks is proportional to $1/h$. It became clear from this that the only way in which accurate paths could be obtained from the Regge calculus is to modify the Regge equations. He used a linear interpolation of the original equations so that the error in each block becomes $O(h^2)$. This guarantees that a global error is $O(h)$ which implies the convergence. As for advance of the periherion of Mercury he obtained the correct value.

Clarke from Southampton UK presented a paper entitled *Numerical Relativity in a Transputer Array* by A C W Garret, R A d'Inverno and C J S Clarke. They are now using a parallel processor in characteristic initial value problem with compactified equations in which $r = \infty$ corresponds to the finite value of a new coordinate z ,where r is the luminosity distance. For an axisymmetric problem all the quantity q is a function of $q(u, z, y = \cos\theta)$ where u is a time coordinate. Each processor corresponds to each value of y and they are linked by fast communication link. At present they are using 32 commercial (Parsy) array. The performance for vacuum problem with 500 time levels is 52 sec and in near future they will have 10 times faster array. They are also testing a code with matter. In this case they are working with Bishop[2] from University of South Africa who explained the method. The characteristic initial value problem is appropriate in vacuum but in the presence of matter it loses the advantage because the characteristics of matter do not coincide with those of the gravitational fields. So he uses usual 3+1 Cauchy problem within some distance R_+ which is larger than R_m where matter exists

only for $R \leq R_m$. For the development of the gravitational fields outside R_+ the characteristic method will be adopted. So one must (1) construct the coordinate and metric for $R \geq R_+$ from Cauchy data and (2) obtain the boundary condition at R_+ for the Cauchy evolution from the results of the characteristic initial value problem. The scheme has difficulty at the start up stage. To calculate the null geodesics from $R = 0$ to $R = R_+$, Cauchy data at t=0 is not enough. The numerical implementation of the scheme is now developing.

Suen from Washington St.Louis reported the paper *Horizon Boundary Conditions in Numerical Relativity* with Seidel . One of the major problems in numerical relativity has been how to avoid the space-time singularities which exist from the beginning like in collisions of black holes or which are formed from the non- singular initial data. Many different types of singularity avoiding slicings have been proposed since in general relativity we have a freedom to choose time coordinate. However the question to the capability of long time integration of the space-time has not been answered. Suen and Seidel proposed a horizon locking coordinate to answer this problem. They choose a spatial coordinate in spherical symmetric space time such that the location of the apparent horizon has the constant coordinate value in time. After the horizon is locked all grid points are tied to it by requiring the radial metric function to be constant in time. In this case the problem is that the shift vector becomes so large that x^i=constant line may not be time like in an extreme case, which causes numerical instabilities. To avoid this instability they propose causal finite difference such that 1) Return to the zero shift coordinate and make the finite difference. 2) Transform the finite difference version of equations to coordinates with large shift vector. The final results are similar to the up-wind finite difference method in hydrodynamics which is numerically stable. They demonstrated the ability of long time integration of their method by evolving the Schwarzschild geometry with initial data of Einstein-Rosen bridge.

Numerical simulations of dynamical black hole systems (oscillating black holes and head-on two black hole collisions) were presented by D. Hobill and E. Seidel[3]

Both simulations are axi-symmetric and utilize maximal slicing conditions to

calculate the lapse function. A two component shift vector is introduced to control axis instabilities. The conditions imposed on the shift vector are: i) the three-metric is diagonal and ii) the shift vector components are determined from partial derivatives of a scalar function that is obtained from solving a linear second-order elliptic equation. The evolution is calculated using a leapfrog (with half time step extrapolation) finite difference technique and the elliptic equations for the lapse and shift are solved with a multigrid method. Essentially the codes are the same modulo initial data and boundary conditions, although a number of different gauge and coordinate choices have been tried for the 2 black hole collision. For the oscillating black hole, a Brill wave is superposed with a black hole and for the two black hole collision, Misner initial data is used.

The codes have been run on Cray-YMP, Cray-2 and NEC SX-3 supercomputers. The standard resolution involves 200 radial zones and 56 angular zones to cover one quadrant (equatorial symmetry is maintained in addition to axial symmetry). Various methods for analyzing these spacetimes have been developed. The Zerilli function (for the $\ell = 2$ and $\ell = 4$ modes) can be extracted and its propagation compared to analytic perturbation theory for quasi-normal mode generation. For low amplitude Brill waves the agreement is good to a percent or better. All of the Newman-Penrose spin coefficients and Weyl tensor components can be constructed as can various Bel-Robinson quantities. The mass loss rates can be calculated from the above quantities and they all agree to within numerical errors. The (quasi-local) ADM mass is measured at the outer boundary of the grid and the black hole mass can be measured by calculating the area of the apparent horizon or alternatively from measurements of the wavelength and damping factor of the quasi-normal mode wave functions. These black hole mass measurements are consistent with each other to within a few per cent. Furthermore, the difference between the ADM mass and the black hole mass can be accounted for in the energy loss associated with the emission of gravitational waves.

Computer graphics animations were presented for the simulations. For both large and small amplitude distortions of a single black hole the values of Ψ_0 and

Ψ_4 were tracked until late times (of order $100M$). The standard quadrupole waves were seen to dominate the perturbation case where a pure $\ell = 2$ Brill wave was introduced in the initial data. Nonlinear interactions between the $\ell = 2$ and $\ell = 4$ modes were evident in the high amplitude wave case. The two metric induced on the apparent horizon was embedded into a 3D flat space and a color map showing the local Gaussian curvature of the surface was used to track the incoming waves. The geometry of the horizon surface oscillates with the quasinormal frequency of the black hole. In the two black hole collision, it was shown where and when a second outer most trapped surface forms as the holes collide. This surface surrounds the original separated trapped surfaces associated with each hole, and then oscillates as in the case of a single, distorted black hole. It was also shown that for the parameters studied in the two black hole collision, that the nonlinearities were not as strong as those generated from highly distorted black holes. In fact all simulations to date seem to have gravitational waveforms that are clearly dominated by quasi-normal mode waveforms.

Schutz from Cardiff UK reported papers *An ADI Scheme for a Black Hole Problem* with Allen and *Time-Symmetric ADI and Causal Reconnection* with Alcubierre. He first pointed out various difficulties in calculating coalescing binary black hole in 3D. 1)One needs a quasi-rectangular 3D grid which can remain fixed at infinity and through which the holes move. This makes stringent requirement on the gauge and slicing conditions. 2) If black holes move through the grid, then grid points go down into a hole and then pop back out the other side. This popping out requires that the grid move faster than light, which causes numerical instability. 3)Coalescing black hole may begin with the holes relatively far apart so that it will take much longer time-steps compared with dynamical time. Since conventional explicit integration schemes are restricted by the Courant condition, one may like to use implicit method, which is less well known.

As for 2) they propose causal reconnection of grids. When a grid is moving faster than light, the grid point at the desired time-steps is outside the light cones of grid points at previous time steps. For a simple wave equation, the standard

numerical schemes go unstable in such a situation. Their remedy for this is that they simply reformulate the computational grid so that its members are within each others's light cones. As for 3) they argue ADI(Alternating Direction Implicit) method on moving grids. They have found that all the standard ADI schemes for the simple wave equation go unstable on grids that move. They also noticed that none of the standard methods preserve the fundamental time-symmetry of the original equation. Requiring that the time-symmetry be maintained, they found a stable scheme for all grid speeds up to the speed of light. For grids moving faster than light, one can perform causal reconnection of the grids.

Nakamura reported the present status of the construction of a 3D code in Kyoto University, Japan. To construct fully GR 3D code is highly difficult. So his group divided the problem into easier problems. 1)In 1988 they constructed a fully 3D GR code in which only metric part of the Einstein equations is included (see T. Nakamura and K Oohara in Proceedings of GR12 p 61.). They used the Cartesian grids of 80^3 and showed that they could trace propagation of the l=m=2 quadrupole linear localized gravitational wave within a few percent error. 2) In 1989 Oohara and Nakamura[4] succeeded in determining initial data for coalescing binary neutron stars using the Cartesian Coordinate and ICCG (Incomplete Cholesky decomposition and Conjugate Gradient) method to solve coupled Poisson equations. 3) From 1989, Oohara, Shibata and Nakamura started the Post-Newtonian 3D simulations of coalescing binary neutron stars including radiation reaction by gravitational waves for various initial data.[5] If one can combine all these three numerical codes, a fully GR code will be completed. For this purpose we need the good gauge condition and the time slice. For the time being, they are using the *quasi -minimal* shear conditions and the conformal time slicing. In the minimal shear condition all the equations are coupled so that one need too much computing time. In the *quasi-minimal* shear conditions one uses the shift vector one time step before or changes the basic equations so that this coupling is resolved. In the conformal time slicing[6], the lapse function is determined as a function of conformal factor of the 3-space metric. They are now using 47^3 grids

and developing their code for coalescing binary neutron stars running on FACOM VP2600.

There were so many good papers in the workshop but the space is limited. So I just show the list of papers that I could not mention.

Bishop, N. T. *The Numerical Calculation of Gravitational Radiation on a Bondi Sphere*

Dubal M.R., Oliveira S. R. and Matzner R. A., *Three Dimensional Initial Data for the Two-Black-Hole collision problem*

Gorgoulhon E., Bonazzola S. and Marck J. A. *A High Precision Numerical code for Spherically Symmetrically Gravitational Collapse Based on a Chebyshev Spectral Method*

Harleston, L.H. *Numerical Solution of the Einstein-Boltzmann Equations in Spherical Symmetry: Results and Perspectives*

Shinkai H. and Maeda K. *Gravitational Waves in a Planar Universe with Cosmological Constant*

REFERENCES

1. Williams R M and Ellis G F R 1981 Gen.Rel.Grav.**13** 361,
   1984 Gen.Rel.Grav.**16** 1003.
2. Bishop N T ,Preprint of Univ. South Africa, 116/92(2).
3. A. Abrahams, D. Bernstein, D. Hobill, E. Seidel and L. Smarr, 1992 *Phys. Rev.*, **D45**, 3544.
4. Oohara K and Nakamura T 1989 Prog. Theor. Phys. **81** 360.
5. Oohara K and Nakamura T 1992 Prog. Theor.Phys. **88** 307. and references therein.
6. Shibata M and Nakamura T 1992 Prog. Theor.Phys. **88** 317.

Computer methods in general relativity: algebraic computing

Marcelo E Araujo, Departamento de Matemática, Universidade de Brasília, 70919 Brasília DF, BRAZIL, E-mail: meda@v1.astro.cf.ac.uk.

Tevian Dray, Department of Mathematics, Oregon State University, Corvallis, OR 97331, USA, E-mail: tevian@math.orst.edu.

James E F Skea, School of Mathematical Sciences, Queen Mary and Westfield College, London, E1 4NS, UK, E-mail: jimsk@cbpfsu1.cat.cbpf.br.

Andreas Koutras, School of Mathematical Sciences, Queen Mary and Westfield College, LONDON, E1 4NS, UK, E-mail: andreas@maths.qmw.ac.uk.

Andrzej Krasinski, Polish Academy of Sciences, Bartycka 18, 00 716 Warszawa, Poland, E-mail: akr@camk.edu.pl.

David Hobill, Department of Physics and Astronomy, University of Calgary, Calgary, Alberta T2N 1N4, Canada, E-mail: hobill@acs.ucalgary.edu.

R G McLenaghan, University of Waterloo, Ontario N2L 3G1, Canada, E-mail: rgmclena@daisy.waterloo.edu.

Steven M Christensen, MathSolutions, Inc., P.O. Box 16175, Chapel Hill, NC 27516 USA, Email: steve@physics.unc.edu.

• The first presentation in this session was **Finding isometry groups in theory and practice** by Araujo, Dray, and Skea. They summarized their work as follows:

Karlhede & MacCallum [1] gave a procedure for determining the Lie algebra of the isometry group of an arbitrary pseudo-Riemannian manifold, which they intended to implement using the symbolic manipulation package SHEEP but never did. We have recently finished making this procedure explicit by giving an algorithm suitable for implementation on a computer [2]. Specifically, we have written an algorithm for determining the isometry group of a spacetime (in four dimensions), and partially implemented this algorithm using the symbolic manipulation package CLASSI, which is an extension of SHEEP.

Our procedure is the following: Apply the classification algorithm built into CLASSI to determine the isotropy group and the set $\{J^\mu\}$ consisting of the functionally independent quantities in the frame components of the curvature tensor of M and its covariant derivatives. By construction, the J^μ are constant on each orbit of the isometry group. Let $\hat{\omega}^i$ be the frame in standard form produced by CLASSI as being appropriate to the isotropy, and let ω^i be the corresponding frame on the isotropy bundle $I(M)$ (the frame bundle restricted to the isotropy group). Calculate the connection 1-forms $\omega^i{}_j$, and note that $\{\omega^i, \omega^j{}_k\}$ is a basis for the cotangent space of $I(M)$. From this point onwards, there are at least three possible approaches which will produce a basis for the orbits of the isometry group in the frame bundle, namely

1) Find the dual vectors, take the subbasis orthogonal to the dJ^μ (which are therefore a basis in the orbit), and take their commutators (see MacCallum & Skea [3]). Though conceptually the most intuitive method, it has the disadvantage, for a computer

Summary compiled by Steven M. Christensen, Session Chair

algebra system, of involving the calculation of a determinant to obtain the vectors dual to the basis one-forms, and this is not necessary as we shall see.

2) The "functional independence" of $\{J^\mu\}$ means that the 1-forms dJ^μ are independent on M (and hence also on $I(M)$); extend them to a basis on M by adding as many ω^i as possible and then extend this to a basis on $I(M)$ by adding the independent $\omega^i{}_j$. Call this basis $\{\sigma^\alpha\}$, and take its exterior derivative. We can now further restrict ourselves to a particular orbit of the isometry group in M by solving the equations $dJ^\mu = 0$, then substituting appropriately on both sides of the remaining equations obtained by exterior differentiation. We thus obtain a basis of the cotangent space of the restriction of $I(M)$ to an isometry group orbit in M. This method has the disadvantage that the calculation of the structure constants is done from a basis which spans a space larger than that of the orbits of the isometry group, and so requires more computation than necessary.

3) use the relations $dJ^\mu = 0$ to rewrite the remaining 1-forms in terms of a reduced set of variables and calculate their exterior derivatives. This method has the disadvantage that dependencies on the reduced set of variables need to be calculated.

The method adopted in the examples presented in [2] is method (3), though which of the methods will eventually be adopted in CLASSI remains to be seen. In any case, since the $\{\sigma^\alpha\}$ form a basis, their exterior algebra closes. But from the Cartan structure equations, the structure constants involve only ± 1, the Riemann tensor of (M, g), and its covariant derivatives, whereas by the construction of the isotropy group given by Karlhede & MacCallum [1], the Riemann tensor and all of its covariant derivatives are constant on each orbit. Thus, the structure constants are constant on each orbit. As pointed out by Karlhede & MacCallum, they are therefore precisely the structure constants of the Lie algebra of the isometry group.

Most of our algorithm has been implemented using CLASSI although the restriction of the 1-forms to a basis in the orbit has not yet been fully automated and thus needs to be done by hand. Automating this part of the algorithm is a reasonably straightforward programming task which when completed will mean that the structure constants for the isometry Lie algebra can be automatically determined from any metric for which CLASSI can compute the isotropy group.

• The second paper was **An algorithm for determining whether a space-time admits a homothety** given by Andreas Koutras and James E. F. Skea who said:

As part of the equivalence problem, and as a problem in its own right, it is useful to have an algorithm which decides whether or not a spacetime admits a homothety. For such a spacetime there exists a vector field x, such that any geometrical quantity q obeys the relation

$$\underset{x}{\pounds} q = cq$$

for some constant c (the weight of q).

To do so, we use a result of Defrise-Carter (later refined by Hall) that a spacetime which admits a homothety group H_n is conformal to a spacetime which admits an isometry group G_n, together with the scaling properties of geometrical quantities in a homothetic spacetime.

¿From CLASSI we can determine a canonical set, $\{R^\mu\}$, of frame components of (the spin decomposition of) the Riemann tensor and its covariant derivatives to an order k which completely determines the local character of the spacetime. As a byproduct, CLASSI gives us the dimension of the isometry group.

We construct a set $\{S^\mu\}$ of the independent ratios of members of $\{R^\mu\}$ such that each member of $\{S^\mu\}$ has zero weight. For a homothetic spacetime there exists a (pseudocanonical) basis related to the canonical basis for $\{R^\mu\}$ by a pure boost in which the number of functionally independent functions of the coordinates in $\{S^\mu\}$ is one less than that in $\{R^\mu\}$.

The problem is now reduced to the determination of the pseudocanonical basis of a homothetic spacetime. This problem parallels the determination of the canonical basis, requiring an analysis of all the possible isotropy groups which can arise at each stage of derivation of the Riemann tensor, and has been solved for all Weyl spinors and physical Segre types.

• Andrzej Krasinski presented **The program Orthocartan**:

The program Ortocartan is written in Lisp and designed for automatic calculation of curvature tensors (Riemann, Ricci, Einstein and Weyl) from a given metric tensor. The input data are the components of an orthonormal tetrad of exterior forms representing the metric, the output are the tetrad components (and, on request, also the coordinate components) of all the quantities calculated at intermediate stages. An "abacus" program Calculate is provided together with Ortocartan for performing simple algebraic operations on expressions defined by the user. The program is presently available on the Atari Mega STE computers, and is being implemented on an iPSC-SYM1 parallel computer by M. Perkowski and his student. Unlike most other programs, Ortocartan was never extended into an elaborate system allowing many kinds of specialized procedures to be carried out automatically (like e.g. determining the Petrov type of a metric or the Segre type of its energy-momentum tensor, or calculation in other tetrads). However, much effort was invested into making its algorithms efficient (in the sense of speed and core-economy), general and fail-safe. By this we mean that the user will not be forced to use the facility of substitutions to correct obvious but nasty little failures. This is the list of advantages that the creators of the program think the users may find attractive:

1. Rational powers of rational numbers are handled automatically, producing the result in the standardized form: either an integer number or an integer base with the exponent having the numerator equal to one.

2. Substitutions can automatically replace sub-sums and sub-products of sums and products, respectively.

3. The chain rule for differentiating composite functions is automatically applied to the maximal depth necessary in a given case.

4. The algebraic simplification is automatically called on the result of each single user-defined substitution.

5. Substitutions by pattern-matching are directly available to the user.

6. The printing procedure will correctly handle nested exponentials to an arbitrary height.

The program, together with the examples and the user manual, is available on a single diskette. A general user-oriented description by A. Krasinski is now in press in the Gen. Rel. Grav. journal, and it contains references to more literature about Ortocartan.

• David Hobill presented **Symbolic computing for numerical relativity**:

In its attempts to model realistic physical systems whose behavior is governed by the Einstein equations, numerical relativity must employ very general expressions for the spacetime metric. This makes the derivation of the Einstein equations and their

subsequent translation into high level computer languages difficult due mainly to the large number of terms that must be manipulated. Computer algebra software packages have long been used to provide relief from the tedium associated with keeping track of the hundreds to thousands of terms associated with analytic calculations in general relativity theory and now it is being used to pre- and post-process numerical codes that model black hole dynamics, stellar collapse and gravitational wave generation.

For the most part numerical relativity uses variables that are quite different from the variables used in most symbolic tensor packages for general relativity. Therefore a new relativity package has been written that produces the ADM or 3+1 form of the Einstein equations in order that a numerical evolution of Cauchy data may be performed. The package presently sits on top of MACSYMA and accepts as input the lapse function, shift vector, three-metric, and extrinsic curvature components to be used in a numerical code. In addition the dependency of these objects upon the spacetime coordinates needs to be specified, but only implicitly. The package then calculates the connection coefficients, Ricci tensor components and the scalar curvature associated with the t=const. spacelike slices. The evolution equations for the three-metric and extrinsic curvature components as well as the scalar and vector constraint equations are then calculated as are other expressions that are important for 3+1 numerical relativity (e.g. the trace of the extrinsic curvature, the trace-free part of the extrinsic curvature, etc.).

As numerical simulations progress to two and three spatial dimensions and as one experiments with different coordinate and gauge conditions, the need to quickly derive new forms of the 3+1 equations becomes evident. In addition these equations must be transcribed into a high level computer language used by the numerical code. In order to avoid typographic and transcription errors another package has been written that receives the output from the 3+1 equation generation package, allows the modeller to replace the partial derivatives appearing in the Einstein equations with Fortran expressions, and then breaks a complicated expression into simple expressions that can be summed over in a Fortran DO loop. The code produced is readable to both humans and machines in order to facilitate the eventual debugging process.

While these examples demonstrated the use of symbolic manipulation for the pre-processing of numerical codes, examples of post-processing were provided. In particular it was shown how a Newman-Penrose formalism packages written specifically for 3+1 variables was used to determine the spin coefficients and Weyl curvatures in terms of hypersurface information. This package also was written to run on top of MACSYMA and interfaces with the 3+1 equation generator.

Examples of the use of Mathematica for producing graphical images from numerical data and for fitting the hole quasi-normal mode parameters to the extracted wave forms from numerical simulations were also demonstrated.

• R.G. McLenaghan gave the fifth talk, **General relativity calculations in Maple**:

A collection of packages and procedures for performing calculations in general relativity and differential geometry that have been written in the Maple computer algebra system was described. Maple is an interactive system for symbolic mathematical computation currently supported by the following operating systems: UNIX and various UNIX-like systems, 386 DOS, Macintosh Finder, DEC VMS, IBM VM/CMS, NeXT and Amiga DOS.

Packages and/or procedures have been written to perform the following tasks: (i)

computation of the connection and curvature, in the natural basis, in a general moving frame and in a complex null frame; (ii) determination of the Petrov type of the Weyl tensor and the Segré type of the Ricci tensor; (iii) computer aided integration of the field equations both in the Newman-Penrose (NP) formalism and in the Ellis-MacCallum formalism. The NP package has been extended [4] to include the transformation of spinor equations into NP form. (iv) calculations with differential forms; (v) determination of the Bianchi-type of a three-dimensional Lie algebra for a given set of structure constants. Most of these packages and procedures are contained in the library of the currently distributed version of Maple (Maple V). Descriptions of these packages and procedures are given in [5] and [6]. A Maple version of the muTENSOR system of Harper and Dyer [7] is also available.

The power of the NPspinor package was illustrated by a description of its use to obtain the following new results in the theory of second order differential invariants of the metric tensor [8]. The spinor equivalent of the Géhéniau-Debever (GD) [9] invariants $D_{(5)}$ and $\dot{D}_{(5)}$ contains the spinor expression $m_6 := \Psi_A{}^B{}_{KL}\Psi_B{}^C{}_{MN}\Phi^{KL}{}_{\dot{A}\dot{B}}\Phi^{MN}{}_{\dot{B}}{}^{\dot{C}}\Phi_C{}^A{}_{\dot{C}}{}^{\dot{A}}$. By means of the NPspinor functions *contract* and *dyad* and the Maple commands *expand* and *factor* it may be shown that $m_6 = (1/6)Ir_2$, where $I := \Psi_{ABCD}\Psi^{ABCD}$, is the complex quadratic Weyl invariant and $r_2 := \Phi_{AB\dot{A}\dot{B}}\Phi^B{}_C{}^{\dot{B}}{}_{\dot{C}}\Phi^{CA\dot{C}\dot{A}}$, is the cubic Ricci invariant. It follows that the GD invariants $C^{(1)}_{(2)}$, $E_{(3)}$, $D_{(5)}$ and $\dot{D}_{(5)}$ are related by the algebraic equation $12(D_{(5)} - \dot{D}_{(5)}) + C^{(1)}_{(2)}E_{(3)} = 0$. It may thus be concluded that their set of fourteen invariants contains at most thirteen independent invariants. The package was subsequently used to obtain a new complex invariant of fifth degree which when adjoined to the GH set yields a set that contains complete minimal sets in the Einstein-Maxwell and perfect fluid cases. The package has also been used to obtain some new results on Huygens' principle [10].

• The final talk was **A practical application of MathTensor** by Steven M. Christensen:

MathTensor [11] is a *Mathematica*-based [12] system with over 250 functions and objects added into *Mathematica* for the purpose of doing both abstract and concrete tensor analysis. One of the earliest reasons for producing *MathTensor* was to perform coincidence limit computations in quantum field theory in curved spacetimes. These calculations involve tensor equations with thousands and ultimately millions of terms or more. The answers are in terms of complicated products of Riemann tensors and their derivatives or contractions.

The basic object we study first is the bi-scalar of geodetic interval, $\sigma(x, x')$, which measures the square of the distance along a geodesic between to spacetime points. The defining equations for σ are:

$$\sigma - \frac{1}{2}\sigma_{;\rho}\sigma_{;}{}^{\rho} = 0, \lim_{x'\to x}\sigma_{;\alpha} = 0, \lim_{x'\to x}\sigma_{;\alpha\beta} = g_{\alpha\beta}.$$

In some applications, we need to determine the coincidence limit of as many as twelve derivatives of σ. Some test computations show that this might involve equations with more than a billion terms! Clearly, a computer program will be needed to generate the terms and then simplify and combine them into a more usable form.

With *MathTensor* it is very easy to set up the equations above and automate taking the derivatives and then substituting in lower covariant derivative coincidence

limits. The only problem that remains is that the products of Riemann tensors that appear may not be independent of each other. Using a program now under development called Schur, Lie group analysis tools like Young Tableau can be used to determine the number and structure of a linearly independent set of Riemann tensor invariants so that the coincidence limits can be put into a canonical form.

Once the computation of the σ coincidence limits is done, then these results can be put into the Schwinger-DeWitt recursion relation formula so that the coincidence limits of the famous Hamidew coefficients can be computed. These are then used in things like the computation of stress tensors, anomalies, point-splitting algorithms, divergences, and so forth. All this can be done with *MathTensor*.

A brief discussion of the functionality of *MathTensor* in general relativity, mathematics, and engineering was also given.

References

[1] Karlhede A and MacCallum M A H 1982 *Gen. Rel. Grav.* **14** 673

[2] Araujo M E, Dray T and Skea J E F 1992 *Gen. Rel. Grav.* **24** 477

[3] MacCallum M A H and Skea J E F 1992 *SHEEP: A Computer Algebra System for General Relativity*, in *Algebraic Computing in General Relativity, Lecture Notes from the 1st Brazilian School on Computer Algebra* Roque W L and Rebouças M J Eds. Oxford University Press to appear

[4] Czapor S R, McLenaghan R G and Carminati J 1992 *Gen. Rel. Grav.* in press

[5] McLenaghan R G 1992 *Maple applications to general relativity* in *Algebraic computing in general relativity, Lecture notes from the 1st Brazilian School on Computer Algebra* **2** Roque W L and Rebouças M J Eds. Oxford University Press to appear

[6] Char B W, Geddes K O, Gonnet G H, Leong B L, Monagan M B and Watt S W 1991 *Maple V Library Reference Manual* (Springer-Verlag, New York)

[7] Harper J F and Dyer C C 1985 *The muTENSOR User Manual*, Department of Astronomy and David Dunlap Observatory, University of Toronto, Toronto, Ontario, Canada

[8] Carminati J and McLenaghan R G 1991 *J. Math. Phys.* **32** 3135; Report on Workshop A1 in these proceedings

[9] Debever R 1964 *Cah. Phys.* **168-169** 303

[10] Carminati J, Czapor S R, McLenaghan R G and Williams G C 1991 *Ann. Inst. Henri Poincaré* - Phys. théor. **54** 9

[11] Parker L and Christensen S M 1991 *MathTensor: A System for Doing Tensor Analysis by Computer*, MathSolutions, Inc., P.O. Box 16175, Chapel Hill, NC 27516, USA

[12] Wolfram S 1991 *Mathematica: A System for Doing Mathematics by Computer*, Wolfram Research, Inc., Champaign, IL 61820, USA

Paper presented at the 13th Int. Conf. on General Relativity and Gravitation Cordoba, Argentina, 1992: Part 2, Workshop Summaries

Relativistic astrophysics

Richard H. Price

Department of Physics
University of Utah
Salt Lake City, UT 84112
USA

Abstract. Work reported in the workshop on relativistic astrophysics spanned a wide variey of topics. Two specific areas seemed of particular interest. Much attention was focussed on gravitational wave sources, especially on the waveforms they produce, and progress was reported in theoretical and observational aspects of accretion disks.

The title "relativistic astrophysics" is rather broad, and has ill-defined boundaries separating it from the concerns of other branches of relativity, and of other workshops at the conference. As I interpret it, the subject should be driven (although sometimes indirectly) by observations.

On this basis the following topics would seem to be some of the interesting developments in the past year or so: (i) The discovery of the isotropy of γ-ray burst sources suggests a cosmological origin. (ii) The identification of black-hole candidates in compact systems, systems with signatures rather different from Cygnus X-1, is a sign of maturation of the field. (iii) There is increasing evidence that our own galaxy has a mass concentration near its center, suggesting a central black hole, and/or other strong-field phenomena. (iv) Data from the Hubble telscope confirm the longstanding belief that there is a mass cusp at the center of M87. (v) Recent work, especially the numerical simulations by Shapiro and Teukolsky [1], reported at this conference, presage the downfall of cosmic censorhip. What are the astrophysical consequences of Shapiro-Teukolsky spindle singularities?

Along with these new, or reactivated, problems there remain some venerable issues, problems on which progress continues, and on which further work is needed. We are still far from having an adequate understanding the physics of accretion disks around black holes. Perhaps related to this is one of the classic question of relativistic astrophysics: the central structure of active galactic nuclei.

The range of subjects covered was broad and somewhat scattered. I shall review here two subjects, which were well represented: sources of gravitational radiation, and accretion disks/toroids. Both, are questions the "long-standing problem" variety.

The first of these, gravitational wave sources, was a subject of marked concentration in the workshop, and was the focus of five of the thirteen oral presentations. The focus,

in fact, was rather narrow. The great interest in details of source waveforms leaves no doubt that the anticipated advent of laser interferometer gravitational wave detectors is having a considerable impact.

One presentation holds out an interesting hope of tying together observations of gravitational waves and the recent discoveries about γ-ray bursts. Perhaps the reason these bursts have not created the same flurry among relativists as among astrophysicists is that in the 20 or so years since they were first discovered the bursts have been assumed to be dramatic, but not relativistic, events involving strongly magnetized neutron stars. All this has changed with the findings of BATSE (Burst and Transient Source Experiment) on the Compton GRO (Gamma Ray Observatory) [2]. There are two important features of bursts that it has discovered: (i) the bursts are distributed isotropically, and (ii) plots of number of bursts vs. burst strength show a fall off at low strength. These results almost certainly point to sources at $z \sim 1$. At this point the field is open for much speculation, and the data leave room for some differences of interpretation. For one thing, the PVO (Pioneer Venus Orbiter) which has higher threshold (and different energy range) than BATSE has been monitoring strong γ-ray bursts since 1978. A comparison of the strong burst data of PVO with the extensive statistics of weak bursts from BATSE suggests a deficiency of weak bursts that points to evolution of the source population and rules out early universe sources [3].

The way the data are to be interpreted is an open question. There might, for example, be two different classes of phenomena contributing to the observations. The nature of the source is a question that is open even wider. Exotic possibilities, such as stars made of strange matter, have been suggested [4], as has the tidal disruption of an ordinary star by a massive ($M \sim 10^8 M_\odot$) black hole [5]. Perhaps the favored model–at least the most conservative model–is that of the coalescence of two compact objects, either a neutron star binary as suggested by Piran [6] and others, or a neutron star-black hole binary, as suggested by Paczynski [7]. The expected frequency of neutron-star coalescences is on the order of 10^{-6} year^{-1} per galaxy) which would be compatible with the rate at which bursts are being detected. The energy requirement for a $z \sim 1$ burst is around 10^{50} ergs, and would be compatible with the expected 10^{53} ergs for a binary coalescence.

This all should, of course, means that γ-ray bursts are of great potential interest to relativists, since coalescence of compact binaries is also the presently favored source of detectable gravitational waves. At the workshop David Nicholson reported on work with Schutz that points out how this can be exploited for gravitational wave (GW) observations. In particular, the γ-ray burst and the GW burst for the coalescence should coincide within 1 sec The pattern matching necessary for the search for GW bursts would be simplified by searches limited to the data within 1 sec of an observed γ-ray burst. This allows, for example, the use of a finer mesh of filters and improved sensitivity. The detection of a burst by BATSE would provide directional information on the burst source, and would therefore constrain the form of the signal arriving in GW detectors. Monte Carlo simulations were used by Nicholson and Schutz to explore the magnitude of the improvement to be gained from a search for coincident bursts. They find (for an optimally configured detectors) that 2 GW detectors would be able to detect a binary neutron star coalescence out to $z = 0.4$.

Other work reported at the workshop was also motivated by the problem of the pattern matching necessary to find GW signals from coalescing binaries. Very recent

results by the CalTech group [8] show that much improved detailed understanding of the final inspiral of compact binaries will be needed if astrophysical information is to be efficiently extracted from GW waveforms. Larry Kidder reported on work with Will and Wiseman [9] on one facet of the progress that will be needed. They study the existence of innermost stable orbits for compact binary systems. The approach is a hybrid of $(\text{post})^2$-Newtonian analysis [so that $(\text{post})^{5/2}$ radiation damping is omitted] and test body effects from the Schwarzschild geometry. The results show that, for a given mass of the binary pair, the radius of the innermost stable orbit is 20% larger for an equal mass binary than for the test body case.

The need for more precise predictions of source waveforms was part of the motivation for the presentation, by Alan Wiseman, of a study with Will of "heredity effects" in GW phenomena. These are nonlinear effects that in general require integration over the past history of the system. Another motivation was the relation of this work to the "Christodoulou nonlinear memory" [10]. A gravitational wave train is said to have "memory" if the strain $h_{ij}(t)$ is different after the the passage of the waves than it was before the arrival of the waves. In principle, this would leave the free masses of a GW detector in positions different from their positions before the waves arrived, and the shifted positions would constitue a souvenir of the burst, the "memory." In practice this DC offset of the burst would be swamped by noise, but it can be inferred by measurement of the low frequency content of the burst, the components of the burst at frequencies below any characteristic frequency of the source.

A major reason for the interest in GW memory is the recent correction of a widespread and longstanding conceptual error concerning its calculation. Only $1/r$ contributions to h_{ij} are radiative; for the memory only the initial and final values of h_{ij} are needed; before and after the strong accelerations producing the fat middle of the burst, the stress energy is simply that of "particles" moving at uniform velocity; the $1/r$ fields of these particles is just the Coulomb fields due to their masses, and those contributions are completely within the scope of linear theory. It was therefore the common wisdom that the calculation of the memory produced by an astrophysical event required only linearized gravity theory, even for a strong field events like black hole collisions. As long as the waves were weak at the detectors, it was thought, linearized theory should suffice for calculations of the memory.

Christodoulou [10], with a rigorous mathematical analysis, showed that this was wrong, that there are in fact nonlinear contributions to the memory that are of the same order as the linear contributions. It was quickly shown to be equivalent, for waves from astrophysical sources, to including the emerging GWs themselves as part of the source [11][12]. Though this nonlinear part of the memory is present in principle, it is a separate question whether it could be detected. Preliminary model calculations [12] suggest that detection with laser interferometers will be very difficult, but may be possible.

A presentation by Viqar Husain dealt with an aspect of GW waveforms not directly related to binary coalescences. Model calculations of GWs from strong-field sources, black holes or relativistic stars, have shown that GW wavetrains tend to be dominated by "quasinormal" (QN) ringing, oscillations at complex (i.e., damped) frequencies which are characteristics of the spacetime of the black hole or relativistic star. This QN dominance has some apparent parallels with normal mode phenomena. In Newtonian theory, for example, a stellar model with no fluid dissipation would have hydrodynamical motions consisting of a superpostion of the normal modes of the star. Given the initial conditions

of the system one can calculate the amplitude of excitation of each mode, and subsequent motions of the system are always the same superposition of these normal modes. The dynamics of the system is in this way relatively easily extracted from the initial data. The damping of the modes could then be calculated to find the imaginary part of the damped normal mode frequency. These simplifications of calculation and understanding follow from the propery of "completeness" normal modes. Intuition suggests that the same sort of picture, in *some* sense, must apply for QN dominated sources, at least for not-very-relativistic stars described with general relativity. There is, however, a crucial mathematical difference between normal modes and QN modes.

Completeness and other useful properties of normal modes follow from the fact that they are solutions of a self-adjoint problem. For QN modes the causal boundary condition of outgoing waves gives a problem which is not self-adjoint (the complex frequency eigenvalues are a sure sign of this) and for which none of the familiar features of normal mode systems are guaranteed to be valid. There is then no guarantee of completeness, but neither is there a prohibition against it, and consideration of not-very-relativistic stars argues for some sort of completeness [13]. But what sort?

Husain[14] uses a model problem of mechanical system with radiative damping. Special features of the system make a complete analysis straightforward and allows some insights to be gained about the sense in which QN modes can be complete. The results show that, for the model, the QN modes are complete for initial data that is "purely outgoing," that is, initial data that generate no ingoing waves at large distances. For such data the excitation of the QN modes, and the subsequent motions of the system can be extracteda from the initial data. It is likely that this result can be extended to an approximate description of a relativistic stellar model.

One presentation at the workshop dealt with aspects of GWs much broader than that of waveforms. An analysis of the exterior Schwarzschild geometry by Kundu had led him to claim [15] that when propagating in the neighborhood of a body of mass M, the amplitude of GWs of frequency ω is smaller by a factor $\sqrt{1+\omega^2G^2M^2/4c^6}$ than what would have been expected on the basis of the quadrupole formula. This would mean, for example, that waves from a binary coalescence with $\omega = 10^3$ Hz at the center of a galaxy with a mass of $10^{11}M_\odot$ will be suppressed by nine orders of magnitude.

Kundu's analysis assumes that the quadrupole moment of the radiating source is described by ψ_0^0, the coefficient of the r^{-5} part of the Newman-Penrose quantity Ψ_0. This is known to be the case for stationary spacetimes, and for linearized gravity, but on the basis of a scalar model calculation Kozameh *et al* [16] have argued that ψ_0^0 does not describe the source quadrupole; rather, in the scalar model the analog of ψ_0^0 is "enhanced" with respect to the quadrupole, and this enhancement cancels Kundu's calculated suppression.

A gravitational source calculation [17] has been done to show that for GWs from a source at the center of a galaxy the only effects due to the mildly curved background are those small effects that are expected. On the other hand, Kundu has presented the basic formalism for a source calculation of ψ_0^0, and finds no enhancement [18]. At the workshop Jorge Pullin reported that all results are now reconciled. For quadrupole-dominated sources there is indeed an enhancement of ψ_0^0 which cancels Kundu's suppression. Thus, long wavelength sources deep inside a massive galaxy will be detectable on the outside with negligible suppression. Sources far (many wavelengths) from the center of the galaxy will suffer a strong suppression of the radiation in the "quadrupole mode," if by

"quadrupole" is meant the formal $\ell = 2$ mode with respect to the center of of the galaxy. But sources far from the center, even if they are slow motion sources, will have their radiation dominated by high order ℓ-pole moments. The mathematics of the suppression is therefore correct, but if does not affect any sources of astrophysical interest.

The second topic of concentrated interest at the workshop was accretion disks. Studies of disks around holes has, so far, always featured some approximations. In particular, Abramowicz *et al* [19] studied self-gravitating disks around pseudo-Newtonian black holes and found evidence that massive disks would be unstable; Wilson [20] studied non-self-gravitating disks around Kerr holes and found no evidence on instability.

At the workshop Shogo Nishida reported work, with Eriguchi and Lanza, on a numerical code that is able to handle a rotating black hole and toroid in equilibrium. The numerical approach uses the full equations of general relativity and is not limited in the range of masses or angular velocities. The particular results reported assume that the toroid is made of a perfect fluid obeying a polytropic equation of state and having a uniform distribution of angular momentum. Studies were reported on several interesting strong-field phenomena such as the appearance of ergotoroids, and the absence of evidence for prolate black holes. But the results which were probably of most immediate astrophysical interest were those on stability. By considering sequences of models with fixed total angular momentum Nishida *et al* conclude that self-gravitating disks are in fact unstable if (for their model assumptions) the toroid mass is greater than 10% of the hole mass.

It is generally thought that massive black holes play a crucial role in the energetics of active galactic nuclei, though the nature of the mechanism is not clear. A conservative model assumes that the angular momentum of infalling matter leads to the formation of an accretion disk, and uses "standard" accretion disk theory to produce the AGN energy and the high energy part of its observed spectrum. The predicted spectra depend on details of the disk (thick vs. thin, source of viscosity) and of the dominant radiative processes. A careful and extensive study was made by Sun and Malkan [21] of what kind of agreement with observation could be achieved by such a conservative approach. They assumed physically thin and optically thick disks around both Schwarzschild and rapidly rotating ($a = 0.998$) Kerr holes, and included the general relativistic effects on the emerging radiation (gravitational and Doppler boosting, focussing) and inclination effects. The results were compared with 60 quasars and AGNs selected for well determined spectra, and the spectra were corrected for reddening, intergalactic absorption and other observational effects, to arrive at an estimate of the inherent spectrum generated directly by the energy source. For each object the black hole mass, the accretion rate and the inclination angle were varied to achieve a best fit. Sandip Chakrabarti, reporting at the workshop on work with Wiita, noted that the overall results were quite good but that in several cases (he concentrates on 1202+281 and 2130+099) the models predict too little emission in the far UV. (It should be noted, however, that the Sun-Malkan models even more clearly predict *too much* UV emission in other cases, in particular 1421+330.) Chakrabarti has argued that the conditions necessary for the formation of standing shocks will often be present in accretion disks, and that these shocks can have significant observational effects [22]. At the workshop he pointed out that such shocks will result in a hotter inner disk region with the potential to produce greater UV than the Sun-Malkan models. He presented results for geometrically thin, optically thick disks around Schwarzschild holes. General relativistic effects on boosting and focussing emis-

sion were omitted (and were found to be important by Sun and Malkan only for rapidly rotating Kerr models). The results [23] do indeed show better agreement in the UV than the Sun-Malkan models, and suggest that standing shocks may indeed play some role in some AGNs and quasars.

References

[1] Shapiro S L and Teukolsky S A 1991 *Phys. Rev. Lett.* **66** 994–7

[2] Meegan C A *et al* 1992 *Nature* **355** 143-6

[3] Fennimore E E, Epstein R I, Ho C, Klebesadel R W and Laros J 1992 *Nature* **357** 140–141

[4] Haensel P, Paczyński B and Amsterdamski P 1991 *Astrophys J.* **375** 209-215

[5] Carter B 1992 *Astrophys J.* **391** L67–L70

[6] Piran T 1992 *Astrophys J.* **389** L45–L48

[7] Paczyński B 1991 *Acta Astron.* **41** 257
Lindley D 1991 *Nature* **354** 20–21

[8] Cutler C, Bildstein L, Flanagan E, Markovic D, Sussman G and Thorne K 1992 reported at workshop C1 of GR13

[9] Kidder L E, Will C M and Wiseman A G 1992 submitted to *Class. Quantum Grav. Letters*

[10] Christodoulou D 1991 *Phys. Rev. Lett.* **67** 1486–9

[11] Thorne K S 1992 *Phys. Rev.* D45 520–524

[12] Wiseman A G and Will C M 1991 *Phys. Rev. (Rapid Communications)* R2945–9

[13] Friedman J 1991 in *Recent Advances in General Relativity* eds. A. Janis and J. Porter (Boston: Birkhauser)

[14] Price R H and Husain V 1992 *Phys. Rev. Lett.* **68** 1973–6

[15] Kundu K 1990 *Proc. R. Soc. London* **A431**, 337
Kundu K Ohio University report (unpublished)

[16] Kozameh C, Newman E T and Rovelli C 1991 *Phys. Rev.* **D44** 551–554

[17] Price R H and Pullin 1992 *Phys. Rev.* in press

[18] Kundu K 1992 submitted to *Proc. R. Soc.*

[19] Abramowicz M A, Calvani M and Nobili L 1983 *Nature* **302** 597

[20] Wilson D B 1984 *Nature* **312** 620–621

[21] Sun W-H and Malkan M A 1989 *Astrophysical J.* **346** 68–100

[22] Chakrabarti S K 1990 *Theory of Transonic Astrophysical Flows* (Singapore: World Scientific)

[23] Chakrabarti S K and Wiita P J 1990 to appear in *Astrophysical J. Letters*

Paper presented at the 13th Int. Conf. on General Relativity and Gravitation
Cordoba, Argentina, 1992: Part 2, Workshop Summaries

Mathematical cosmology

G. F. R. Ellis

Department of Applied Mathematics
University of Cape Town
South Africa

Abstract. Many topics were covered in the submitted papers, showing much life in this subject at present. They ranged from conventional calculations in specific cosmological models to provocatively speculative work. Space and time restrictions required selecting from them, for summarisation here; the book of Abstracts should be consulted for a full overview.

There is continuing interest in various forms of imperfect fluid solution; for example M. L. Bedran and M. O. Calvao [Feederal Univeristy of Rio de Janeiro] discuss universe models where imperfect fluids evolve reversibly owing to the presence of a conformal Killing vector field, and L. P. Chimento and A. S. Jakubi [University of Buenos Aires] consider stability of solutions with causal viscous fluids, concluding that qualitative asymptotic behaviour in the future is not altered by relaxation processes but that in the past it is significantly changed.

The ongoing study of Mixmaster universe dynamics was represented by two papers based on numerical simulations of its dynamical behaviour. The problem is that the standard indicators of chaotic behaviour, such as Lyaponov exponents, give different results when applied on the one hand to the one-dimensional Return Map, characterising the evolution as a change of parameters in a series of Kasner epochs, and on the other hand to the exact field equations, represented in terms of evolution of parameters in a two-dimensional anisotropy plane. A. Burd and R. Tavakol [Queen Mary College, London] argue that the gauge freedom in general relativity makes all such standard indicators of chaotic behaviour problematical, and that indeed chaos is an inherently gauge-dependent phenomenon. B. Berger [Oakland University, Michigan] however argues that use of Minisuperspace proper time gives a definitive answer, showing that there is chaotic behaviour in the full solutions, in agreement with analyses based on the Return Map.

More general dynamics of homogeneous models is studied in papers by K. Rosquist [Stockholm University], discussing the nature of the symplectic structure needed in order to represent Bianchi Class B dynamics in Hamiltonian form, and by C. Uggla [Syracuse University] and R. Jantzen [Villanova University], indicating a hierarchical structure emerging from the study of invariant manifolds in the space of solutions. Thus "simpler models constitute building blocks for the construction of the dynamical structure of more

complicated ones". A similar theme emerges in the paper by C. Hewitt [University of Waterloo], showing how self-similar solutions appear to be asymptotic states at late time for more general (diagonal G2) inhomogeneous cosmologies. These approaches seem to be very helpful in obtaining an overview of the kinds of dynamics possible in cosmological models.

The use of piecewise Friedmann-Tolman models for the expanding universe (generalised "Swiss-Cheese" models) is discussed by A. Chamorro [Bilbao], in a paper representative of studies by a number of authors. Overdense or underdense regions can be imbedded in external expanding universes, provided they are surrounded by compensating intervening Tolman zones. One can thus construct a model of an intermediate scale inhomogeneous universe made up of Friedmann underdense and overdense spherical regions surrounded by compensating thick Tolman shells imbedded in a Friedmann expanding background, in line with current ideas about the cell structure of the universe.

The study of inhomogeneous models will be considerably helped by a survey project reported on by A. Krasinski [Copernicam Astronomical Centre, Warsaw], who emphasizes that while in the old days there was a view that solutions of the Einstein Field Equations are so difficult to come by that any new solution was worth having, now the situation is different. There are so many published solutions (most discovered many times) that the first thing to do when looking for exact solutions is to see if what you are planning has already been done, for there is a good chance it will already be in the literature; and the need is to understand the solutions obtained and their relations to each other, rather than just to find new solutions.

The project assembles and classifies exact inhomogeneous solutions of the Einstein equations that contain the FLRW (Friedmann-Lemaitre-Robertson-Walker) universes as limiting cases, and so can be understood as inhomogeneous cosmological models; results of 247 papers have been included in this compilation so far. The relationships between the models (in particular, specialisations that lead from one to another) have been examined, leading to a broad classification into five main types, and characterising which models are subcases of others; in many cases, multiple discoveries of the same model have been catalogued.

Krasinski points out that many interesting inhomogeneous models were already studied in the 1930's and 1940's, particularly papers by R. C. Tolman [1] and by N. R. Sen [2] contain proposals that are still attractive today. Sen showed that the Lemaitre [3] solution predicts a behaviour of density distribution that today would be called formation of voids, also implying that the Einstein-Strauss "swiss-cheese" model is unstable to velocity perturbations (i.e. to perturbations that allow non-comoving walls). Krasinski also comments that despite all the work done to the present day, no rotating generalisation of the expanding FLRW models are explicitly known; we also lack explicit shearing and accelerating FLRW generalisations.

Given the special nature of exact solutions, perturbation solutions are inevitable; two important issues arise.

One is their linearisation stability, that is, how well the linearised solution represents the behaviour of the exact solutions. An interesting study by J. Frauendiener and B. G. Schmidt [Max Planck Institute for Astrophysics, Garching] looks at this issue in the case of spherically symmetric spacetimes, comparing the linearised and exact solutions with each other. Not surprisingly, the linear and exact solutions deviate from each other

more and more as the density contrast grows. Such studies are invaluable in analysing the reliability of perturbation theory.

The second issue is the averaging problem in cosmology, raised in a survey talk on cosmology at the GR10 meeting and now gradually becoming a focus of activity. The point is that practical cosmological models, being patently unable to represent all the structure in the universe down to the finest details, represent the universe averaged over some suitable scale; and different models represent it at different averaging scales (for example some may contain perturbations representing clusters of galaxies but others only a smoothed out cosmological substratum). There are two implications. Firstly, it is clear that cosmological models should state explicitly state the averaging scale envisioned in their application, for this is crucial to their interpretation. The issue then is that averaging does not commute with process of working out the Einstein field equations [4]. Consequently the field equations in cosmology at smoothed out scales should include an effective polarisation term resulting from the averaging process (as in the well-known Isaacson term in the case of gravitational radiation).

Various authors have examined this issue in an interesting manner, for example Bildhauer and Futamase [5] claim that the effect could be large enough to seriously change the relation between the Hubble constant and the age of the universe. Now in a series of papers summarised in a poster presented to the meeting (being unable to give an oral presentation for financial reasons), R. Zalaletdinov [Uzbek Academy of Sciences, Tashkent] gives a systematic way of tackling the problem by use of bitensors that enable averaging of tensor fields over a finite volume. He works out the consequences for averaging covariant derivatives, and so the effect on the field equations, in terms of structural functions and a series of correlation tensors, the latter determining splitting rules for averages. The result is a scheme for averaging out a Riemann space resulting in the appearance of an averaged space with a metric and two equi-affine symmetric connections. He obtains the averaged Einstein equations and contracted Bianchi identities. The result is a very promising scheme for tackling this fundamental problem from the foundations; its implications, and its relation to other proposals such as those of Futamase and Kasai, have still to be determined.

To broaden the scope of the discussion , some studies considered more general issues in cosmology. D. H. King [Vancounver] and C. Klein and H. Pfister [Tubingen] consider Machian properties of rotating universe from different viewpoints, both claiming (in different contexts) that a FLRW universe cannot rotate with respect to its inertial frame, in agreement with previous work by D. J. Raine. In a different spirit, R. Tavakol [Queen Mary College, London] asks the questions "Is general relativity fragile?", examining its stability under various possible changes both in terms of imposing symmetries on models (which are never truly satisfied) and in terms of various ways of generalising General Relativity. Various examples show that in general structural stability will not hold. In the long term this kind of issue will become important in determining the questions we ask and the models we use. This kind of issue underlies some of the other papers presented, for instance that by Chimento and Jakubi mentioned above.

Finally, the most speculative paper of the session, by L. Smolin [Syracuse University], considered the possibility of natural selection in cosmology. The issue here is that, as emphasized by Dawkins and others, Darwinian selection is a powerful mechanism for creating apparently purposeful structure and order where none existed before, and indeed is the only mechanism known that can do so. The issue then is whether this

might be introduced as an explanatory principle in cosmology, explaining some of the coincidences that are otherwise inexplicable (except perhaps on an anthropic basis, that many people reject). This becomes a possibility if one conceives of situations where collapse of a black hole gives rise to new expansion phase (a "daughter universe"), with the possibility of the constants of physics being different in the new expansion phase than in the old universe. Thus there is a source of variation of conditions, one of the necessities for evolution. One also needs some mechanism for selection: here the proposal is that it is simply numbers of progeny universes that is the mechanism acting, leading eventually to an overwhelming likelihood of universe models existing with a maximal creation of daughter universes. The issue is to show that the constants realised around us are indeed such as to maximise black hole production and consequent creation of daughter universes.

This is highly speculative, but certainly in the spirit of much modern theoretical cosmology. Smolin presents a detailed argument for his proposal [6]. It can be criticised in detail, as was shown by the workshop discussion, and there is room for development and testing of the proposal; however it provides an exciting prospect of uniting two of the major paradigms of scientific understanding (evolution through natural selection, and the expanding and evolving universe) into a new way of understanding cosmology. Smolin proposes to explain in this way why many of the dimensionless numbers which characterise particle physics and cosmology take unnatural values. This will be regarded with skepticism by many, but is certainly an interesting idea.

Overall there is much interesting activity in this area. It was a pleasure to have Charles Misner, one of the pioneers in much innovative work in theoretical cosmology, join us for some of the discussions; we wish him well on the occasion of his 60th birthday.

References

[1] R. C. Tolman 1934 *Proc Nat Acad Sci* **20** 169

[2] N. R. Sen 1934 *Z Astrophys* **9** 215, 1935 *Z Astrophys* **10** 291

[3] Lemaitre 1933 *Ann Soc Sci Bruxelles* **A53** 51

[4] G. Ellis 1983 *GR10 conference proceedings, Ed. B. Bertotti et al, Plenum Press*

[5] Bildhauer and Futamase 1991 *GRG* **23** 1251

[6] Smolin 1992 *Class. Qu. Grav.* **9** 173

Paper presented at the 13th Int. Conf. on General Relativity and Gravitation
Cordoba, Argentina, 1992: Part 2, Workshop Summaries

Early Universe phenomena

Alexander Vilenkin

Tufts Institute of Cosmology,
Physics Department,
Tufts University,
Medford, MA 02155, USA

Abstract.
In this review I will be able to give only a brief summary of the talks presented at the Workshop. The topics discussed can be divided into three categories: (i) inflation, (ii) cosmic strings, and (iii) misellaneous.

1. Inflation

The opening talk was given by Leonid Grishchuk who disussed the generation of gravitational waves during inflation. He pointed out that the waves are not generated with totally random phases, as it was usually assumed. The gravitational field after inflation is in a "squeezed" quantum state with strong correlations between waves having the same frequency but travelling in opposite directions [1]. The resulting gravitational wave background can be thought of as a stochastic ensemble of standing waves. The same conclusions apply to scalar field perturbations which can evolve into cosmological density fluctuations and later lead to structure formation in the Universe. It is not clear, at present, what will be the effect of "squeezing" on structure formation scenarios and on the pattern of the microwave background fluctuations observed by COBE [2]. But this is an interesting question which deserves further investigation.

The implications of COBE for various versions of inflationary scenario were discussed by Andrew Liddle [3]. He focussed in particular on extended inflation, which is based on a Brans-Dicke-type theory of gravity and in which inflation ends through bubble nucleation. The spectrum of density fluctuations produced in this scenario has a power-law form, $\delta\rho/\rho \propto k^n$, where k is the wave number and the spectral index n depends on the Brans-Dicke parameter. The index n cannot exceed 0.75, since otherwise the large bubbles formed early during inflation would be seen on the microwave sky. On the other hand, Liddle argued that the spectrum has insufficient power on small scales unless $n > 0.82$. The reason is that, for values of n in the allowed range ($n < 0.75$), the

microwave background anisotropies are dominated not by the scalar density fluctuations, but by the gravitational waves generated during inflation. The amplitude of density fluctuations should therefore be smaller than a direct comparison with COBE would suggest. The conclusion is that the structure formation scenario based on extended inflation is inconsistent with observations. This version of inflation may still provide a viable model explaining the homogeneity and flatness of the universe, but the origin of cosmological density fluctuations will have to be explained by another mechanism.

As observations are beginning to catch up with models of structure formation, the theorists come up with more complicated models having a larger number of adjustable parameters. In the case of extended inflation, a possible modification is to allow the Brans-Dicke parameter to be a function of the "inflaton" scalar field ϕ. The corresponding model is called "hyperextended inflation". At the Workshop, Franco Occhionero [4] discussed an inflationary model with ordinary (not Brans-Dicke) gravity in which the scalar field ϕ is coupled to the curvature through the interaction term $L_{int} \propto \phi^2 R^2$. It can be shown that the phase space of this model has an attractor trajectory with two consecutive periods of inflation. With a judicious choice of parameters, the transition between the two periods will fall into the observable range of scales and will introduce a "break" into the otherwise featureless spectrum of density fluctuations.

The calculation of bubble nucleation rate in a false vacuum is a problem of considerable importance for inflationary scenarios. William Hiscock [5] presented a detailed numerical study of this problem, extending the original work of Coleman and de Luccia [6]. He showed in particular that the thin-wall approximation, which gives a simple analytic estimate for the decay rate, has a very limited range of validity. As the wall thickness is increased, the Coleman-De Luccia instanton approaches the homogeneous Hawking-Moss instanton. The instanton method has also been used to calculate the rate of bubble nucleation in the presence of compact objects, such as black holes.

2. Cosmic Strings

A general review of string evolution was given by Tom Kibble. Strings can be produced as linear defects at a phase trnsition in the early universe. They form a stochastic network with most of the string length, about 80%, in the form of infinite strings which have the shape of random walks, and the remaining 20% in closed loops. As the strings start moving under the action of their tension, they intersect, the wiggles on long strings get chopped off in the form of closed loops, the loops fragment into smaller loops, which oscillate and eventually decay into gravitational waves. This is a very complicated system, and despite a tremendous amount of work, both analytic and numerical, the string evolution is not yet fully understood.

It is generally expected that, after some transition period, the strings will get into a scaling regime of evolution, in which all the characteristic length scales of the network grow proportionally to the horizon, t. This expectation is supported by numerical simulations which show a few long strings in each horizon-size volume at any time, indicating that the persistence length of strings and the typical string separation are always comparable to t. However, the simulations also revealed that long strings have a substantial small-scale structure in the form of kinks and wiggles propagating along the strings at the speed of light. The scale of the smallest wiggles and the size of the loops produced

by the network were not observed to grow proportionally to t, thus casting doubt on the scaling hypothesis. It should be noted that the dynamical range of the simulations is not large, and the observed deviation from scaling could well be a transient behavior. In any case, scaling is expected to establish eventually, when gravitational damping of small-scale wiggles becomes important.

The most promising line of attack in the quantitative study of string evolution appears to be a combination of analytic and numerical techniques. Kibble outlined the approach he developed with Ed Copeland, in which the string network is characterized by three distinct length scales: the first is related to the average string density, the second is the persistence length along the strings, and the third characterizes the small-scale structure. The system of equations for the three scales contains unknown parameters which are to be determined from numerical simulations. Work on this project is still in progress.

Strings formed at a phase trnsition can survive until present only if the phase transition occured after or near the end of inflation. However, it has been recently shown [7] that circular loops of string can spontaneously nucleate in de Sitter space. These loops are expanded by the subsequent inflation, and by the end of the inflationary era they have a spectrum of sizes extending well beyond the present Hubble length. If the loops were exactly circular, they would all collapse to form black holes. However, quantum fluctuations cause some deviation from circular shapee. To estimate the amplitude of these fluctuations and the corresponding probability of black hole formation, one has to develop a quantum field theory of linearized perturbations on strings in de Sitter space. This was discussed by Jaume Garriga [8]. The unperturbed world sheet of a nucleating loop is itself a $1+1$-dimensional de Sitter space, and it can be shown that the perturbations are described by a set of two tachyonic scalar fields in this background. Garriga estimated the number density of black holes resulting from nucleating strings and used the observational bound on the γ-ray emission by evaporating black holes to impose constraints on the parameters of the model. The excluded region of the parameter space comes close to, but does not overlap with the region where the loops may serve as seeds for structure formation.

3. Miscellaneous

Diego Harari [9] discussed possible observational consequences of a light pseudoscalar fiels ϕ coupled to electromagnetism through the Lagrangian $L_{int} = g\phi\mathbf{E}\cdot\mathbf{B}$ with $g = const.$ As electromagnetic waves propagate in an inhomogeneous ϕ-background, the direction of polarization is rotated by an angle $\Delta\theta = \frac{1}{2}g\Delta\phi$, where $\Delta\phi$ is the change of ϕ along the wave trajectory. If ϕ is a Goldstone boson field resulting from a global symmetry breaking at energy scale v, then ϕ can be expected to vary by $\Delta\phi \sim v$ on the horizon scale. One can then use the observed polarization properties of distant galaxies and quasars to impose a constraint on the coupling constant g. The constraint is circumvented in models where the symmetry breaking occurs before inflation. The field ϕ is then nearly homogeneous and $\Delta\theta$ is negligible.

The workshop was closed, quite appropriately, with the talk by Charles Misner about the future of the universe. He suggested the possibility that inflation, which presumably occured in the early universe, was one episode in a sequence of successive

inflationary periods at ever decreasing energy scales. According to this viewpoint, we are now living in a false vacuum with a very small eenergy density (that is, a small, positive cosmological constant). At some point in the future, when the temperature and density of the universe become sufficiently low, this vacuum will be destabilized, and after a period of inflation the universe will enter a new era characterized by much lower temperatures, slower processes, and weaker interactions. In the discussion that followed, voices from the audience called for an immediate ban of low-temperature experiments, which could trigger the fateful vacuum decay.

References

[1] L.P.Grishchuk & Y.V.Sidorov, Phys.Rev.**D42**, 3413 (1990); L.P.Grishchuk, H.A.Haus & K.Bergman, Phys.Rev.**D46**, 1440, (1992)

[2] G.Smoot *et. al.*, COBE preprint (1992)

[3] A.R.Liddle & D.H.Lyth, to be published

[4] L.Amendola, S.Capozzielo, M.Litterio & F.Occhionero, Phys.Rev. **D45**, 417 (1992)

[5] W.A.Hiscock & D.A.Samuel, Montana State University Preprint (1992); Phys.Rev.**D44**, 3052 (1991)

[6] S.Coleman & F. De Luccia, Phys.Rev. **D21**, 3305 (1980)

[7] R.Basu, A.H.Guth & A.Vilenkin, Phys.Rev. **D44**, 340 (1991)

[8] J.Garriga & A.Vilenkin, Tufts Preprint (1992)

[9] D.Harari & P.Sikivie, Phys.Lett. **289B**, 67 (1992)

Paper presented at the 13th Int. Conf. on General Relativity and Gravitation
Cordoba, Argentina, 1992: Part 2, Workshop Summaries

Observational and astrophysical cosmology

H. Quintana†[1]

† Astrophysics Group, Pontificia Universidad Católica de Chile
Casilla 104, Santiago 22. Chile.

1. Introduction

The subject has witnessed much advances, both in its observational and theoretical aspects. Some of the most important recent developments were covered by the invited speakers and are given elsewhere in this book. The single most exciting recent result has been the detection and implications of the microwave background fluctuations. In recent years a body of evidence pointing to the presence of ever larger structures has grown. The largest observed galaxy aglomerates reach a size of order 100 Mpc and it is likely we have not reached the limit. Some unexpected results on the large scale structure await confirmation, like the detection of periodicities in the redshift distribution of distant galaxies. Most of these results have been reported in other recent conferences. For this reason and also due to the time and geographical proximity of other astronomical related conferences, only a few results were presented here. We review the observational results first and follow with the theoretical presentations.

2. OBSERVATIONAL RESULTS

A review of the current status of faint galaxy evolution, galaxy counts and color distributions was presented by L. Infante (P.Universidad Católica de Chile). In the past 15 years several groups have obtained galaxy counts, initially from photographic plates and more recently from deep CCD images. An excellent mean relation for galaxy counts as a function of blue (J) magnitude over the sky was obtained by Infante by combining UK Schmidt telescope galaxy counts of Maddox et al (MNRAS 247, 1p, 1991), 3.6m CFH telescope counts of Infante and Pritchet (Ap.J. 1992, in press) and faint CCD counts either from Tyson (A.J. 96,1, 1988) or Lilly et al (Ap.J. 369, 79, 1991). These counts span over a factor of 10^5 in brightness. Analysis of the data show that significant galaxy evolution is needed to model observations. However, redshift surveys are not compatible with evolution. Color distributions of faint galaxies may help to understand the above apparent contradiction.

[1] E-mail: hquintana@astro.puc.cl

D. Garcia-Lambas (Observatorio Astronómico, Córdoba) and coworkers presented work on the statistical properties of the large scale distribution of galaxies based on numerical simulations of biased cold dark matter (CDM) Ω=1, Λ=0 cosmologies. They used different schemes for biased galaxy formation assuming a local nature ($0.5h^{-2}$ Mpc) for the processes that segregate luminous and dark mass. Biased and antibiased models for galaxy formation were consider by using simple prescriptions for the assignment of the galaxies, as well as models were galaxy luminosity depend on local density. For all the models where galaxies are more clustered than the mass they found that the angular two point correlation function has not enough power on large angular separations as compared with APM measurements. In antibiased models good agreement between the angular correlation function of simulations and the observations were observed. However, the galaxy peculiar velocities in the simulations exceeded by a factor $\simeq$1.5-2 the observed values. Similar results were obtained for the velocity dispersions of clusters of galaxies. In spite of these major problems, the cosmic Mach number, that is the ratio of systematic to rms velocity dispersion of large volumes (sphere of radius $\simeq 18h^{-1}$ Mpc), was in better agreement with observations that the standard biased model. Antibiased CDM models could be reconciled with observations if astrophysical processes that reduce the peculiar velocities of galactic halos were involved in galaxy formation.

The peculiar motion of the Local Group was analized in work presented by H. Jerjen and G. A. Tammann (Basel, Switzerland). The motion was determined from 15 clusters whose relative distances are known with minimum bias. The resulting local motion is 735 ± 184 km/s towards $l = 274°$, $b = 3°$. This 4σ signal is in perfect statistical agreement with the motion inferred from the dipole of the cosmic microwave background (CMB). The median distance of 6400 km/s of the 15 clusters sets an upper limit to the co-moving volume. The solution leaves so small velocity residuals that the peculiar velocities of the individual clusters must be small ($\leq$ 200 km/s). The data imply a local slow-down of the expansion field due to the Virgo cluster of $v_{VC} = 241 \pm 44$ km/s. An almost identical value, i.e. $v_{VC} = 233 \pm 44$ km/s, is *independently* determined using the relative distances of the Virgo, UMa, and Fornax cluster and of eight nearby supernovae of type Ia. These results do not require the adoption of any zero-point of the extragalactic distance scale.

Unusually large mass-to-light ratios in two groups dominated by a pair of interacting elliptical galaxies were reported by H. Quintana. The central members are dumb-bell galaxies with typical signs of tidal interactions, surrounded by many satellites. Both groups studied in detail have high relative velocity of the dumb-bell components (500-900 km/s). The group velocity dispersions are also exceedingly high for such poor agglomerates. If traditional virial methods are applied, large M/L ratios are deduced, implying dark matter proportions comparable to that present in the outer parts of galaxies. Alternative interpretations in terms of a finely tuned merger of two groups of galaxies, following the ellipticals which form the dumb-bell, could be also consistent with present data.

Radio observations at 1435 Mhz of the variability in flux density and polarization of the blazar PKS 0521-365 were presented by H. Luna and coworkers (Instituto Argentino de Radioastronomía). A periodic behaviour was observed, in correspondence with a quasilinear rotation of the polarization angle. The observations were interpreted using the model of Königl-Choudhouri. The equilibrium configurations of relativistic magnetic pression-dominated jets were studied by the use of a shock wave that accounts for the variability, obtaining a value for the radius of the jet and the distance from the emission

zone to the nucleus.

3. THEORETICAL RESULTS

B. Carter (Observatoire de Meudon, France) presentation on Superconducting Cosmic Strings provided an introductory overview of a new unified mathematical framework for the description of a wide range of macroscopic string and membrane type phenomena. It also described provisional conclusions that can be drawn from its application to the particular phenomenon of superconductivity in cosmic strings.

The subject of cosmic strings was originally founded by Kibble, who initiated a body of work whose most important conclusions are derivable from a simple gauge coupled (Higgs type) scalar field theory proposed as a useful toy prototype for more realistic theories involving the relevant kind of spontaneous symmetry breaking, leading to a topologically circular family of degenerate vacuum (ground) states with local vortex defects describable in the macroscopic limit as thin strings. Such a model gives rise to strings of the 2-dimensionally Lorentz covariant Goto-Nambu type in which (in units $\hbar$=c=1) the resulting string mass-energy density U and string tension T are equal to the square of a fixed mass scale m characterising the scalar Higgs field: $U = T = m^2$. Popular applications of this simple model were based on the supposition that it represented an approximation to a GUT in which the Higgs mass scale was given by $Gm^2 \approx 10^{-6}$, which implied that the gravitational field of the ensuing strings would have been sufficiently strong to be envisaged as possible seeds for galaxy formation.

The foregoing scenario has lost much of its original popularity for diverse observational reasons, but Carter pointed out that there are other, purely theoretical, reasons for thinking that if cosmic strings occur at all in nature, then the most likely scenario is characterized by a very much lower Higgs mass scale, in the range commonly envisaged for the various electroweak unification theories currently subject to experimental investigation, corresponding to $Gm^2 \approx 10^{-32}$. The reason for this radical reassessment is that whereas longitudinally Lorentz invariant Kibble type string loops must radiate away all their energy in the long run, it is becoming clear that in generic cosmic string models the Lorentz invariance will be broken by a mechanism first suggested by Witten, which allows the existence of stationary equilibrium states (Davis and Shellard, Phys. Lett. B207, 404, 1988). The ensuing distribution of (centrifugally supported) loop states would give catastrophic cosmological mass density excess if Gm^2 were anywhere near the order of magnitude of 10^{-6}. Although the efficiency of the process envisaged by Davis and Shellard is hard to evaluate, it would still be sufficient, even if it were very low, to undermine the viability of string forming theories with Gm^2 large enough for structure formation by individual strings to be significant. On the other hand if the efficiency is high, relic loops with $Gm^2 \approx 10^{-32}$ might collectively constitute a significant or even cosmologically dominant CHUMP type CDM constituent (Ann. N.Y. Acad. Sci., 1992, in press).

In order to obtain a more detailed understanding of the processes involved and, in particular, to tackle the stability problem whose solution will be needed before a conclusive picture appears, Carter has been developing a general purpose formalism for the mechanical description of p-branes of any dimension (using a nomenclature in which p=0 designates point particles, p=1 designates strings, and p=2 designates membranes

such as cosmic domain walls). The fundamental equation governing the extrinsic motion of any such system in the absence of external forces is always expressible (Class. Quantum Grav. 1992, in press) in the simple form $T^{\mu\nu}K^{\rho}_{\mu\nu}= 0$, where $T^{\mu\nu}$ is the relevant surface stress momentum energy tensor and $K^{\rho}_{\mu\nu}$ is the second fundamental tensor of the support surface, as defined in terms of the first fundamental tensor $\eta^{\mu\nu}$ by $K^{\rho}_{\mu\nu} = \eta^{\kappa}_{\nu}\eta^{\lambda}_{\mu}\nabla_{\lambda}\eta^{\rho}_{\kappa}$. The simplest example of the application of this formalism is to a point particle trajectory with unit tangent vector u^{μ} for which one has $\eta^{\mu\nu} = -u^{\mu}u^{\nu}$ and hence $K^{\rho}_{\mu\nu} = \eta_{\mu\nu}\dot{u}^{\rho}$, where $\dot{u}^{\rho}$ is the ordinary acceleration vector. In this case $T^{\mu\nu} = mu^{\mu}u^{\nu}$ and m is the mass parameter. The general dynamic equation written above just gives the familiar geodesic equation $m\dot{u}^{\rho} = 0$ together with mass conservation, $u^{\mu}\nabla_{\mu}m = 0$. In the case of a string world sheet, instead of a uniquely defined tangent vector u^{μ} one has a uniquely defined antisymmetric unit surface element tensor $\epsilon^{\mu\nu}$ in terms of which $\eta^{\mu}_{\nu} = \epsilon^{\mu}_{\rho}\epsilon^{\rho}_{\nu}$. For a simple Kibble type cosmic string one gets that the curvature vector $K^{\rho} = K^{\mu\rho}_{\mu}$ should vanish since in this case $T^{\mu\nu} = -m^2\eta^{\mu\nu}$. This particular dynamic equation, $K^{\rho} = 0$, is a covariant expression for the equations commonly obtained in coordinate dependent form from the Goto-Nambu (surface measure) action. In the more complicated generic situation when currents are present the surface stress momentum energy density tensor will be given in terms of its timelike and spacelike eigenvectors u^{μ} and v^{μ}, say, by $T^{\mu\nu} = Uu^{\mu}u^{\nu} - Tv^{\mu}v^{\nu}$ with T < U. The propagation speed C_E of extrinsic (transverse) perturbations is then given (Phys. Lett. 228B, 446, 1989) by C_E^2 = T/U, and it can be shown as a general theorem (Phys. Lett. 238b, 166, 1990; Class. Quantum Grav. 8, 135, 1991) that a string loop achieves stationary equilibrium when its longitudinal velocity is given by this same value C_E. The problem of stability of such equilibrium states has not yet been fully solved and its treatment will require knowledge of the speed C_L of longitudinal perturbations which is given by C_L^2 = -dT/dU. The linear approximation used in most earlier work implied $C_L > C_E$, but more accurate work by Peter (Phys. Rev. D45, 1091, 1992) shows that in fact $C_L < C_E$ in Witten's superconducting string model. In a different context (Phys. Rev. D41, 3038 and 3886, 1990) allowance for thermal or other noise (in the form of microscopic wiggles) leads to an intermediate model with $C_L = C_E$ whose dynamic equations are exactly integrable in flat space.

B.J. Carr presented work with J.H. MacGibbon (Queen Mary & Westfield College, London) on the quantum emission from primordial black holes (PBH) which may have formed from initial density fluctuations or phase transitions in the early Universe. In contrast to previous approaches to this problem, they assumed that fundamental quarks, gluons and leptons are emitted rather than composite particles, once the black hole temperature exceeds a few hundred MeV. The quark and gluon jets then fragment into the stable photons, leptons and hadrons. This radically changes previous estimates of the contribution of evaporating black holes to the observed cosmic-ray backgrounds. They find that the black hole density required to explain the observed gamma-ray, electron, positron and antiproton fluxes at around 0.1-1 GeV are all comparable providing the holes cluster inside the Galactic halo (as expected). This provides some support for the possible existence of PBH from the Galactic centre and their final explosions are also likely to be undetectable. Nevertheless, the possibility that cosmic ray observations could provide information about high energy particle physics and the early Universe is very attractive.

An analysis on the effects on cosmological large-scale structure (LSS) of adding non-Newtonian corrections to the usual cosmological fluid was presented by Caroline

Lewis (Center for Relativity, Austin). This work was done by making the anisotropic stress term in an imperfect fluid time-dependent and then computing the solutions to the density perturbation evolution equations that result from using the gauge-invariant, scalar perturbation formalism suitable for nonlinear dynamics set up by Bardeen (Phys. Rev. D 22, 1882, 1980). The non-Newtonian terms cause a nonlinear relation between stress and velocity gradients that causes an amplification of the density perturbations over the standard Hot Big Bang result. These fluctuations can then act as seeds for galaxy formation.

The motivation for this unusual viscosity treatment comes from the Einstein statistical mechanical result that the effective viscosity for a suspension of spheres in an otherwise uniform fluid medium is proportional to the concentration of the spheres. Bubble nucleation theory for a generic early universe, first order phase transition leads to an estimate for the anisotropic stress that acts as a source for the superhorizon sized density fluctuations in the Bardeen formalism. Depending on the equation of state, density contrast growth rates going as conformal time to the 17th power and more are observed during the bubble nucleation time. This amplification could lead to structure formation (the bubbles are not the LSS!) and is noteworthy in two respects: (1) Microphysics is affecting superhorizon-sized fluctuations. Lewis believes this unusual result to be due to the nonlinear and nonequilibrium nature of the situation. (2) Viscosity has been used to build up structure contrary to standard cosmological viscosity usage. Lewis thought this were dissipative structures. A covariant formulation of complexity theory would be a useful tool to investigate these matters in more detail.

S. Gottlöber (Potsdam) discussed nonflat perturbation spectra and Microwave Background (MB) temperature fluctuations (S. Gottlöber et al. Phys. Rev. D43, 2510, 1991.) Simple inflationary cosmological models predict a quasi-flat perturbation spectrum, whereas observations indicate extra power in the spectrum on scales larger than about 50 Mpc. Such non-flat spectra are predicted by models with two subsequent quasi-de Sitter stages with an intermediate power-law expansion of the scale factor $a \propto t^{2/3}$. Vacuum polarisation effects described by higher-order correction terms in the gravitational Lagrangian may drive a first inflationary stage. The amplitude of the perturbations then depends on the coupling constant α in the gravitational Lagrangian $R + \alpha R^2$. A coherent scalar field (which could be created during this first inflation) may drive a second inflationary stage. The height of the resulting step in the spectrum depends on the ratio of the effective masses of the scalaron $M = 1/\sqrt{6\alpha}$ and the scalar field. The scale of the break is determined by the energy density of the scalar field at the end of the first inflation. A rapid transition from the first to the second inflationary stage leads to a steep step typically accompanied by additional small oscillations in the spectrum. In a smooth transition a major part of the spectrum shows decreasing power for increasing wave numbers. Then the Harrison-Zeldovich spectrum is typical for very large scales (outside the horizon) and it is reached again after the transition to the second inflation, i.e. for scales smaller than 50 Mpc. Transfer functions describe how the initial fluctuation spectrum changes during the cosmological evolution up to recombination, when it gives an imprint on the temperature angular distribution of the MB radiation. The break in the initial perturbation spectra changes drastically the angular distribution of the temperature described by the multipole moments $< a_{lm}^2 >$. In particular, for $l < 40$ the product $l(l+1) < a_{lm}^2 >$ rapidly decreases with increasing l whereas it is approximately constant in the flat perturbation spectrum. Also, small oscillations of the initial

perturbation spectrum lead to effects in the calculated temperature fluctuations at the corresponding moments and scales. The angular correlation functions calculated for different spectra nearly coincide at large angles. At small and intermediate angles non-flat perturbation spectra lead to much smaller correlation functions than the flat spectrum. The temperature fluctuations are expected to be on the verge of detection, which could decide between different initial spectra and inflationary models.

A comparison of three orthogonally crossed wakes with the CfA large-scale structures was presented by T. Hara and S. Miyoshi (Kyoto Sangyo University, Japan). They performed a numerical simulation of the evolution of three wakes which cross each other in a flat universe dominated by CDM with open cosmic strings. The results of the simulation were compared with the CfA observations in cone diagrams. They showed good agreement between simulations and the chosen data; three filaments cross in the observed angles and a wall-like distribution of galaxies appears. It was claimed that the resemblance between the oberved and simulated distributions could be improved by considering more complex situations.

Ideas on a connection of an oscillating gravitational constant and biological effects were presented by P. Sisterna and H. Vucetich (Universidad Nacional de La Plata, Argentina). A periodicity in the galaxy distribution of 128 h^{-1} Mpc has been suggested. Soon afterwards it was noted that this apparent spatial periodicity could be naturally explained by a time oscillation of the gravitational constant G or the Rydberg constant. On the other hand, periodic growth features of bivalve and coral fossiles show a periodic component in the time dependence of the number of days per year. They propose that a time oscillating gravitational constant can explain such a feature. T. Hara, P. Mähöhen and S. Miyoshi (Kyoto Sangyo University, Japan) also reported investigations on the inhomegenity of DM and luminous objects due to the wake formed by open cosmic string and on the anisotropies of cosmic MB radiation due to open cosmic strings. S.K. Chakrabarti argued that in the early epoch of galaxy formation non-electrical forces in radiation pressure supported tori can drive a poloidal electric current and generate primordial magnetic fields.

The author gratefully acknowledges support from FONDECYT Grant 90-371. Miss Lewis presentation was partially supported by the International Physics Forum of the American Physics Society.

Paper presented at the 13th Int. Conf. on General Relativity and Gravitation Cordoba, Argentina, 1992: Part 2, Workshop Summaries

Gravitational wave experiments

William O. Hamilton†[1] and Ho Jung Paik‡

† Department of Physics and Astronomy, Louisiana State University, Baton Rouge, LA 70803, USA

‡ Department of Physics, University of Maryland, College Park, MD 20742, USA

1. Introduction

There were three oral sessions and one poster session for Workshop C1 on Gravitational Wave Experiments. There was also an informal experimental roundtable held one afternoon. The first two oral sessions were devoted mainly to progress reports from various interferometric and bar detector groups. A total of 15 papers were presented in these two sessions. The third session of Workshop C1 was devoted primarily to theoretical and experimental investigations associated with the proposed interferometric detectors. Ten papers were presented in this session. In addition, there were a total of 13 papers presented in the poster session. There was some overlap between the presentations in the third oral session and the posters since only two of the serious posters were devoted to technology not pertinent to interferometers.

In general, the papers showed the increasing maturity of the experimental aspects of the field since most presented the results of completed investigations rather than making promises of wonderful results some time in the future. Unfortunately, the limited time allotted to experimental reports made the session more like a session of contributed papers at a very large national meeting rather than a session where work could be presented and the merits of various approaches debated and discussed. Hopefully future GR meetings will have more time devoted to experimental aspects of General Relativity research.

2. Sessions I and II

The first paper in Session I was a report by Whitcomb on the status of the LIGO project. The five-year project received first-year funding in 1992 to begin design and construction of a two-facility observatory for the detection and study of gravitational waves from astrophysical sources. Sites have been selected (Hanford, Washington, and Livingston, Louisiana) and detailed design will begin soon. Under the present schedule, the facilities will be completed by 1997 and initial observations will begin in 1998. Abramovici followed

[1] E-mail: PHMLTN@LSUVM (BITNET) or 7620::HAMILTON (SPAN)

with a second paper which reported on the sensitivity of the LIGO 40 m prototype interferometer at Caltech. The lowest noise measured is 1.2 x 10^{-18} m/$\sqrt{\text{Hz}}$ at 1 kHz in displacement sensitivity. The noise spectrum is dominated by seismic noise below 120 Hz and by shot noise above 1.5 kHz.

After these two reports on the laser detectors, there were a series of reports on the progress of cryogenic bar detectors. First, there were reports on the detectors in operation. Hamilton discussed the ongoing experiment at LSU. The LSU 4 K detector has been in continuous operation since April 1991. Its noise level corresponds to an rms strain of 5 x 10^{-19} . This is the best noise level that anybody has ever reported, albeit by a small margin. The data is extraordinarily clean, showing fewer than 20 events per day which lie above the expected exponential distribution. He stressed the need for multiple detectors operating in conincidence, regardless of the sensitivity of individual detectors. Looking into the future, Hamilton discussed an interesting possibility of constructing a large 50 mK sphere antenna. This idea is being studied jointly by the bar groups in the U.S. The sphere is omni-directional and may reach a sensitivity of h $< 10^{-21}$ due to its large mass and multimode nature. Ricci discussed the experience obtained from long-term operation of the 4 K EXPLORER detector at CERN. After being upgraded, EXPLORER was in continuous operation from July 1990 until December 1991. Its noise temperature has been better than 10 mK and the duty cycle of data taking larger than 70%. The 6 months of coincidence data between the Rome and LSU detectors is still being analyzed.

Then there were reports on ongoing construction work for ultralow temperature bar detectors. Coccia reported the progress on the construction of the ultralow temperature detector NAUTILUS. A 2350 kg Al 5056 bar was successfully cooled to 95 mK at CERN in 1991. After this cryogenic test, the detector was moved to Frascati in the spring of 1992. NAUTILUS is being reassembled and is expected to go into operation at the end of 1992. A progress report on a second ultralow temperature detector, being assembled at Stanford, was given by Mann. The cryostat has been extensively redesigned and modified to cool a 1800 kg Al 5056 antenna to 50 mK. Work is nearly complete on the Paik transducer and associated commercial dc SQUID for the first phase of the detector operation. This detector is expected to go into operation in the first half of 1993. The senstivity goal of both the Rome and Stanford ultralow temperature detectors is h = 10^{-20}.

Session I ended with Schutz's reassessment of the reported correlations between gravitational waves and neutrinos associated with SN1987a. He pointed out that one of the statistical tests used to establish the reported correlations between the room temperature gravitational wave antennas and the neutrino detectors is seriously flawed, and most other were devised *a posteriori* and contain considerable freedom to make choices that strengthen the correlations. Schutz concludes that the claimed gravitational wave-neutrino correlations are likely to be due, not to any physical effects, but simply to chance. There was disagreement expressed during the discussion but it was decided to postpone the argument until everything had been published.

Session II started with Blair and Tobar's progress report on the niobium bar antenna at Perth. The highest Q achieved with the niobium antenna at 4 K is 2.3 x 10^8. The vibration isolation system for the Perth antenna has been completely rebuilt and the detector is undergoing reassembly. Blair then discussed the status of the AIGO project. It is undergoing a second review in Australia following a previous recommendation by

ASTEC that the observatory be 50% Australian funded if international funding could be secured for the balance. Nicholson presented a coincidence analysis between the two prototype interferometric detectors at Glasgow and the Max Planck Institute in Garching. The 100 hours of coincidence data was used to test automatic data analysis algorithms which search for signals of astrophysical origin. Progress on the intermediate size laser interferometer in Japan was reported by Kawashima. The construction of TENKO-100, a 100m 100 bounce delayline interferometer, will be completed by the end of the summer of 1992. The hard work of attaining high sensitivity will follow.

Attention was then shifted again to real data. Astone presented a new upper limit on gravitational waves obtained from the long-term operation of EXPLORER. The detector noise temperature was 10 mK and corresponded to h = 6 x 10^{-19} for short bursts of gravitational waves. The new limit (less than 0.1 events per day for bursts with h larger than 4 x 10^{-17}) represents an improvement by an order of magnitude over the previous limit obtained with the Stanford detector in 1982. In the next paper, Bertotti described the search for low-frequency gravitational waves from the tracking data for the interplanetary spacecraft ULYSSES. The data was taken in a four-week period starting February 20, 1992 but unfortunately the data analysis has not been completed.

The final two papers in Session II were on improvements on the laser interferometers. Newton reported some improvements to the Glasgow prototype interferometer. There was a lengthy discussion about the effects of relatively poor vacuum on the performance of mirrors. Drever then discussed an idea of the late Brian Meers, an "ultrasensitive configuration" for laser interferometers, using a double sideband recycling technique.

3. Roundtable

An informal experimental roundtable was held during one afternoon. The intent was to allow discussion, unconstrained by formal structure, about the experimental problems and the future direction of the field. One question, posed by Blair, occupied much of the discussion. Blair asked: "If we were to start anew with bar detectors, for what frequency should we build them?" Thorne answered immediately that they should be designed for the lowest possible frequency, whereupon Will suggested that 1.4 kHz would be a good place. Considerable discussion ensued with the final advice to the experimentalists being that no one really knows the answer so they should build what they can do best. We all will then wait to see what is found.

Pizzella also announced, for the Rome and LSU groups, that the first result of the joint coincidence analysis is that for the 6 months studied there were no coincidences with an energy greater than 200 mK. This corresponds to h = $3x10^{-18}$.

4. Session III

The first paper in Session III was presented by Cutler for the Caltech relativity group. He presented their estimates for the frequencies and the duration for the inspiraling of coalescing neutron star binaries. They have also estimated the parameters for neutron star-black hole and black hole-black hole binaries. Monte Carlo simulations show that

if such signals are detected, using undescribed pattern matching techniques, that the signal to noise ratio will be relatively high, enabling the masses and the distances of the sources to be measured. Lobo then presented a paper with Krolak and Meers which set a formidable technological challenge, showing that if an interferometer were to be run using dual recycling and if the recycling tuning could be *dynamically* tuned as the signal from the binary was detected, the signal to noise would be greatly improved. How such detection and dynamic tuning could be accomplished was left as an exercise for the audience. Krolak then discussed the reverse problem: How well can one determine the mass, range, and phase of a detected binary system? He used a network of 3 interferometers, the two LIGO detectors in the US and the VIRGO detector near Pisa. The results show that range is the most uncertain of the parameters, but the signals may be detectable at ranges as large as 200 Mpc.

Bender discussed the possibility of massive black holes at the center of galaxies and the gravitational radiation signature from neutron stars spiraling into such objects. Such signals are inaccessible to anything but a space-based system.

Passive and active seismic isolation systems occupied the remainder of the session. Several of the papers on seismic isolators seemed to revisit much of the territory previously explored by those designing resonant bar detectors. Shoemaker discussed the design, at MIT, of a passive seismic isolator for LIGO. He pointed out the limits to the performance of such isolators caused by internal resonances in the components of the isolation stack. Gonzalez discussed the work for LIGO of the Syracuse group and the new theoretical approach she has used to investigate the old problem of thermal noise in a pendulum. Her work does not directly involve a normal mode expansion and seems to give reasonable results. Saulson continued with a discussion of the effects, on a LIGO pendulum isolator, of the stored tensile energy in highly stressed wires. Since the stored energy is so large, he asked if its release could be a noise source which is much larger than the thermal noise. This is a problem which still doesn't have an answer, either theoretical or experimental. The question centers about whether the energy is released all at once or in small increments. Barone gave the results of a computer program designed to optimize the performance of a static steel and rubber stack and pendulum support. Its optimization criteria led to successively smaller masses as one goes up the stack.

Newell gave a progress report on the Colorado program to develop an active isolation system for the mirrors of a LIGO system. As with all experiments, the meeting came at an inconvenient time, and he had not been able to close all of the feedback loops before he had to leave to report on his results. The vertical motion loops had been closed,however, and the prototype system performed as expected. Notcutt then spoke about the seismic isolators at Perth which eliminate steel and rubber stacks and replace them with star shaped plates in flexure. Magnets are used on the plates to provide damping. These have been already installed in the redesigned Perth resonant bar detector which is set for a cryogenic run Real Soon Now and are also being used in the Perth laboratory interferometer prototype. Their isolation seems to be quite good and isolators of this type should also perform well if cooled.

5. Conclusions

With the emergence of the long-baseline laser interferometers, the number of papers presented in the Gravitational Wave Experiments Workshop has exploded. The future

of gravitational wave astronomy looks bright. The construction of ultralow temperature bar detectors is nearly complete. The LIGO project is moving ahead with the design and construction of two 4 km baseline interferometers in the U.S. and the VIRGO group will build at least one in Europe. These sensitive antennas should be operational by the turn of the century. By that time there may even be a few very massive sphere antennas carrying out an all-sky search for gravitational waves in coincidence with the laser interferometers.

It appears that another decade of hard work is ahead of us. But the payoff may not be that far away. The new detectors are now beginning to reach very interesting sensitivity levels. If Nature cooperates, who knows? One of us may be reporting a clear detection of gravitational waves before the turn of the century, perhaps at one of these GRG meetings.

Report on the Workshop 'Other gravitational experiments and observations'

Bruno Bertotti

Dipartimento di Fisica Nucleare e Teorica, Università di Pavia

The scene of Experimental Gravity has changed much since the beginning of this discipline in the 60's (see, e.g., *Experimental Gravitation*, Proc. Int. School of Phys. E. Fermi, Course LVI (B. Bertotti, ed.), Academic Press (1974).) Until a few years ago extensive experimental programmes were undertaken in a precise theoretical contest: to test well defined relativistic theories of gravity which had in common a metric framework and were precisely classified in a parametric scheme, the Parametrized Post-Newtonian (PPN) formalism; among them, of course, Einstein's theory of relativity, which corresponds to a particular point in this parameter space, had the *a priori* privilege of great simplicity and long tradition. Since the net result of this work was the falsification of most alternative theories of gravity, much of its motivation and impact is now on the wane; at the workshop only three papers (2. 3. and 5.) were explicitely devoted to measurements of the PPN parameters.

Among the current work of this kind in progress in Experimental Gravity, the Gravity Probe B (paper 5.) stands out not only because its main purpose is to measure directly a new force of nature, the "gravito-magnetic force" (which has never been done so far), but especially because the sophistication and the ingenuity involved in this space experiment, consisting in a gyroscope in an Earth orbit, is stunning. The paper is devoted to a problem which is crucial in every space experiment aimed at measuring a single, or a few parameters: since usually the measurement requires the whole mission and is performed only once, it is in principle impossible to assess the statistical error, and great care must be faced to avoid and to evaluate the systematic errors. For the Gravity Probe this will be done with a careful *a priori* analysis and by splitting the 16.5 months duration of the experiment in three phases, each characterized by different experimental conditions. The paper concludes "A demonstrated absence of systematic experimental error would allow an overall accuracy of the drift rate of 0.22 *marcsec/y*", which is an outstanding performance, especially if compared with the main effects on the spin axis of the gyroscope to be measured, the geodetic precession and the frame dragging; they amount, respectively, to 6.6"/*y* and 42 *marcsec/y*.

The other 11 papers address different questions; among them the most important, in my view, is the verification of the Principle of Equivalence in all its forms and implications. Since the mass of a body includes the contributions to the binding energy from all the relevant interactions (Will in 4. lists nine of them), the statement of equality of inertial and gravitational masses implies the equality of each contribution (albeit with less accuracy). Therefore the Principle of Equivalence is a very complex (and, in my view, *a priori* unlikely) statement, with bearings on all the forces of nature and their

quantum properties; any increase in the accuracy of its verification may lead to surprises. This general theme has been recently made more interesting by the proposal – and the extensive experimental programmes related to it – that in nature there is a new force (the "Fifth Force") between baryons; since the rest mass of a body is not quite proportional to its baryonic content, a violation of the Weak Equivalence Principle would ensue.

Unfortunately, contrary to the PPN related experimental research, we have at present no viable and complete theory outside the metric framework which violates this principle and we cannot predict definite thresholds for its violations, which makes the planning of these experiments difficult and moot. The main contribution along this line is the old work by A.P. Lightman and D.L. Lee (*Phys. Rev.* **D8** 364-376 (1973)), where they have constructed a generalized electromagnetism in which the dielectric constant ϵ and the magnetic permeability μ depend on the gravitation potential; then the electromagnetic interactions change from place to place and violate the Equivalence Principle. This theory, however, holds only for slow motion. In the paper 11., an interesting contribution along this line, the formalism of Lightman and Lee is extended to quantum mechanics and applied to a quark model of a nucleon.

To illustrate the gist of the problem, consider the electrostatic binding energy, which contributes to the mass of an ordinary body by $\approx 0.1\%$ and therefore is found to obey the Equivalence Principle to within a part in $\approx 10^8$. In the ordinary theory the formal reason for this is that electromagnetism is coupled to gravity only through the metric field $g_{\mu\nu}$, which takes up in an freely falling frame the Minkowsky form $diag(1,-1,-1,-1)$. Therefore this coupling is universal, and so is the corresponding binding energy. Different couplings and their effect on the motion may provide interesting insights and possible violations; for example, an additional term in the Lagrangean function for the electromagnetic field $F^{\mu\nu}$, proportional to

$$R_{\mu\nu\sigma\rho}F^{\mu\nu}F^{\sigma\rho},$$

where $R_{\mu\nu\sigma\rho}$ is the Riemann tensor, seems to be the simplest interaction, after the ordinary one

$$g_{\mu\nu}g_{\rho\sigma}F^{\mu\rho}F^{\nu\sigma};$$

it contributes to the binding energy a new term which depends on the place, the nature of the body (and its orientation with respect to the Earth if it electrically polarized!) However, the propagation of electromagnetic waves would be also affected and will subject the new interaction to experimental constraints.

These questions are, of course, related to the validity of Lorentz trasformation and special relativity; and, again, we have no complete, alternative theory with which to confront experiments. Interesting advances along this line have been presented at the Workshop by H. Vucetich and his group (paper 10.). On general terms, one can perhaps remark that, while excellent verifications of special relativity are available for laboratory conditions, we are in great ignorance when cosmic bodies are involved, in particular the Universe itself. Violations of local Lorentz invariance of the order of the ratio of the mass of the body to the mass of the Universe would certainly escape the present tests; and since cosmology provides a natural standard of rest one could certainly imagine models in which the Universe as a whole evolves as a threedimensional, riemannian manifold without being embedded in a pseudo-riemannian spacetime. In this case one should, of course, ensure the local validity of the traditional spacetime picture. One could recall

here that the full Lagrangean of general relativity is indeed uniquely recovered from the hamiltonian dynamics of a threedimensional metric *provided an embedding spacetime with the appropriate signature is assumed* (see S. A. Hojman *et al* , *Ann. Phys., NY* , **96** 88-135 (1976).)

It seems difficult to achieve in the laboratory a verification of the Weak Equivalence Principle with better accuracy than in the experiments carried out in the past by Dicke and Braginsky; space experiments seem to be the next step. The paper 4., presented by C.M. Will, was an interesting discussion of the theoretical implications of the great project STEP (Satellite Test of the Equivalence Principle), at present under consideration by the European Space Agency. STEP will consist of a drag-free satellite in a circular Earth orbit, carrying three pairs of accelerometers of a new type. They will use test masses coated with superconducting Niobium, suspended on magnetic bearings and coupled to SQUID detectors. At frequencies $\approx 0.1\ mHz$ they will measure relative accelerations with an accuracy of $10^{-14}\ cm/sec^2$. The final accuracy in the difference between inertial and (passive) gravitational masses is expected to reach one part in 10^{17}! As Will has noted, such an accuracy will enable, among other things, to test the equality between the inertial and the gravitational contributions due to other interactions, in particular the part of the weak force which violates parity.

Laboratory tests of the Equivalence Principle are certainly much less accurate, but available now. In the paper 14. B.R. Heckel, from the University of Washington in Seattle, has reported the new results of their torsion balance: for Be and Al, for example, the two masses are equal to within $4.9 10^{-12}$, similar to the negative outcome of Braginsky and Panov's experiment with Pt and Al, with an error of $0.9\ 10^{-12}$ at 95% confidence level (Soviet Physics JETP **34** 463-466 (1972)). One should also mention an improvement by one order of magnitude in the verification of the Equivalence Principle for photons, which, as well known, leads to the prediction of the gravitational frequency shift; this was done in a long experiment lasted ten years by means of Mössbauer effect and presented in 12. If spacetime is endowed with a non symmetric fundamental tensor, like in Moffat's theory (*Phys. Rev.* **D19** 3554 (1979)), light deflection due to a gravitational field would be dependent upon its polarization – an example of *gravitational birifringence.* This, of course, would also violate the Equivalence Principle for photons. In 6. a new limit on the critical parameter of Moffatt's theory has been deduced from the depolarization that this effect would produce on the Zeeman components of spectral lines emitted by magnetically active regions near the limb of the Sun (where the gravitational deflection is not negligible). For a criticism of Moffatt's theory, see, however, the recent paper by T. Damour, S. Deser and J. McCarthy *Nonsymmetric gravity theories: inconsistencies and a cure*, to appear in *Phys. Rev.* D.

Many people have speculated on particular violations of the Equivalence Principle based on anomalous spin-dependent effects in long range interactions, which can be tested in the laboratory. At the workshop, after a general introduction on these problems (paper 8.), new upper limits on a long range ($\geq 3\ cm$) interaction between polarized electrons and nucleons have been presented.

The paper 14. is also related, in a way, to the Equivalence Principle. Its naive interpretation suggests that a charge freely falling in a gravitational field should not emit electromagnetic radiation, just like a charge at rest in a flat spacetime; however, since radiation is a non local phenomenon, extending all the way to infinity through the future light cone, the application of the principle is not warranted. Indeed, theory

shows that the freely falling charge should radiate following ordinary electromagnetism, with the appropriate value of the acceleration. The paper 14. considers the radiation emitted by bunches of relativistic, freely falling charges, like one will have in the LEP of LHC machines, and addresses the problem of its detection. The main obstacle is the acceleration due to the magnetic field of te Earth, several orders of magnitude greater.

The extensive and interesting experimental work on the "Fifth Force" did not lead so far to any substantiated claim (see E. Fischbach and C. Talmadge, Nature **356** 207-215 (1992)), but, of course, at some level Newton's law of gravity must be violated! A very interesting experimental technique in this field is provided by three pairs of accelerometers placed along three orthogonal axes. Since a pair of accelerometers measures the second derivative of the Newtonian potential U, if U fulfils Laplace's equation, the readings of the three pairs must sum up to zero. This is the theoretical basis of the experiment 13., where superconductive gradiometers are used to test this feature over a range of a few meters. The paper gave a progress report, concentrated on the main remaining problem, removing the error due to misalignments of the accelerometers. The final accuracy in the Yukawa coupling constant α is expected to reach 10^{-3}.

In a poster paper A. Spallicci presented, also on behalf of six other authors, the plans of the COLUMBUS Metrology Science Team for gravitational experiments for Spacelab, the EURECA programme and the International Space Station Freedom. The main projects are two:

Flying two clocks based upon a hydrogen maser, a very accurate microwave frequency standard, as a test for future space experiments (in particular the Solar Probe, see, e.g., the paper by J.D. Anderson in *Relativistic Gravitational Experiments in Space*, R.W. Hellings (ed.), NASA Conference Publication 3046 (1989); for another technique involving frequency standards in space, see B. Bertotti and G. Giampieri, *Class. Quantum Grav.* **9** 777-793 (1992)). This project will allow time transfer with an astonishingly low error of 10 ps.

Secondly, a special box (the Pico Gravity Box) is planned to reduce to $10^{-10} cm/sec^2$ or lower the dynamical noise in a spacecraft between 1 Hz and 1000 Hz; this will allow some gravitational experiments, in particular a test of the Weak Equivalence Principle.

Experimental gravitation is promoting, and will use, immense progresses in the technology of spacetime measurements, in particular of time, frequency, position, angles and velocity. Fractional accuracies or the order of 10^{-9} or better are now common, and several projects plan to achieve much more. The excellent review paper 1. has given a quick and broad look on the most striking advances to be expected in the future in three areas: stable frequency standards, which are expected to achieve in the mHz band stabilities better than a part in 10^{16}; accurate standards of rest (accelerometers and drag-free systems) and very stable lasers; angular measurements in space in the μHz range with interferometric techniques (in particular, the POINTS project (Precision Optical INTerferometry in Space)).

The workshop, with its generally excellent papers, has demonstrated again the vitality and the urgency of Experimental Gravitation, in particular the research on the Equivalence Principle.

Appendix. List of the talks given in the workshop

1. R. Hellings *From Gedanken to Werken: Space-time measurements for the 21st*

century (Review).

2. C.M. Will *Is momentum conserved? A test in the binary system PSR 1913+16.*

3. R.B. Orellana and H. Vucetich *The Nortdvedt effect in the Trojan asteroids.*

4. C.M. Will and STEP theory panel *Theoretical significance of a proposed satellite test of the equivalence principle.*

5. G.M. Keiser *et al An update of the estimated accuracy of the Gravity Probe B gyroscope experiment and plans to test for systematic experimental error.*

6. M.D. Gabriel, M.P. Haugan, R.B. Mann and J.H. Palmer *Gravitational birifringence: a new test of the Equivalence Principle.*

7. J.M. Daniels, T. Jen and W. Ni *Experimental search for spin-dependent effects in gravity and other long-range interaction.*

8. R.C. Ritter *et al Test of Equivalence Principle violation by* $\sigma \cdot \mathbf{r}$ *interactions.*

9. S.E. Perez Bergliaffa and H. Vucetich *Limits on local Lorentz invariance. A new approach.*

10. E.A. Logiudice, S.E. Perez Bergliaffa and H. Vucetich *Limits on the violation of Einstein Equivalence Principle in a gravitationally modified standard model.*

11. M. de Francia, G. Goya, R.C. Mercader and H. Vucetich *Mössbauer null redshift experiment. Systems in a magnetic field.*

12. M.V. Vol Moody and H.J. Paik *Null test of the gravitational inverse square law.*

13. D. Fargion *Gravitationally induced synchrotron radiation by ultrarelativistic bunches in LEP and LHC.*

14. B.R. Heckel *New limits for the weak equivalence principle.*

Paper presented at the 13th Int. Conf. on General Relativity and Gravitation
Cordoba, Argentina, 1992: Part 2, Workshop Summaries

Quantum field theory in curved space–time

Mario Castagnino
Instituto de Astronomía y Física del Espacio
Casilla de Correo 67- Sucursal 28, 1428 Buenos Aires- Argentina

When we began to study, in the seventies, Quantum Field Theory in Curved Spacetime (QFTCST), we thought that we were constructing a fundamental theory, encompassing General Relativity and Quantum Field Theory (QFT). It was soon evident that QFTCST is only the semiclassical approximation to Quantum Gravity (QG) a yet unknown theory. After 20 years, normally a physical theory, if it is not fundamental, decays and dies. On the contrary, QFTCST is very much alive, as it is proved by the almost 50 abstracts submitted to this workshop. This is so because many people are trying to define and develope QFTCST in very interesting and important cases, such as Gott's space or non globally hyperbolic spaces etc. (as we shall see in section 1), because QFTCST is useful to study the quantum nature of Black Holes (BH) (as we shall see in section 2) and because this formalism is an esential tool for Quantum Cosmology; like in inflationary models (as we shall see in section 3).

So let us review the main topics of the Workshop, where we will find interesting contributions to all this of these lines of research.

1. The General Theory

In his talk, David Boulware proposed a QFT in Spaces with closed time-like curves, the Gott's spacetime, whith no anomalous stress- energy tensor. In the special case of a total deficit angle of $\frac{2}{\pi}$, it is possible to find a complete orthonormal set of eigenfunctions of the wave operator. From these, the QFT is constructed. The resultant interacting QFT is not unitary, because the field operators can create real, on-shell, particles, in the acausal region, which propagate for finite proper time accumulating an arbitrary phase, before being annihilated at the same spacetime point. As a result, the effective potential within the acausal region is complex and probability is not conserved.

Wai- Mo Suen posed the following question: "are perturbative constraints necessary in semi-classical gravity?" In fact, recently Simon and Parker proposed that those solutions of the semiclassical Einstein equation, which are not perturbatively expandable in powers of the Planck's constant, should be disregarded. This would then avoid the instability of flat space, and exclude pathological solutions of the semi-classical Einstein equation. Suen argued that such a restriction might not be necessary (at least for models involving only free quantum fields in which the semi-classical theory is exact to all matter loops),based on: (i) although an exact flat space is unstable with respect to infinitesimal perturbations, Robertson-Walker spacetimes with arbitrarily small(but finite) Hubble expansion are stable in semi-classical gravity, and (ii) for a

range of renormalization parameters, solutions of the semi-classical equation starting out dominated by the higher derivative terms in the equations (hence locally not perturbations of classical solutions) are automatically driven towards classical solutions by the backreaction from quantum fields. Hence the existence of our present universe is not necessarily in contradiction to the " unrestricted semi-classical theory".

Juan Pablo Paz raised some criticisms about this talk because, he said, Simon and Parker proposal was not well interpreted by Suen.

Sung-Won Kim introduced another question about stability: "are time machines unstable by the particle production?" Kim and Thorne tried to calculate the vacuum fluctuation of quantized fields [1]. It was shown that the vacuum fluctuations produce a renormalized stress-energy tensor (SET) that diverges as one approaches the Cauchy horizon, which will be cut off by QG. However, there is a controversy. Hawking [2] conjectures an observer-independent location for the breakdown in the semiclassical theory. In his talk, Kim studied this quantum stability problem using another method, namely particle production by an arbitrary gravitational field. When the wormhole forms in the infinit past, the result is finite, while it is divergent near the Cauchy horizon when the wormhole forms at a finite time. If we adopt Kim- Thorne's conjecture, then the divergence can be cut off by QG ; therefore, the total energy cannot prevent the formation of the closed timelike curves when one is within a Planck length.

Finally, Tevian Dray, Corinne Manogue and Robin W. Tucker studied an interesting problem both related to QFTCST and QG. They consider the metric $ds^2 = f(t)dt^2 + g(t)dx^2$ where g is everywhere positive and f has two roots, at both of which it changes sign, and where furthermore $f(t) < 0$ for $|t|$ sufficiently large. This corresponds to a spatially symmetric space-time which is initially and finally Lorentzian, with a Euclidean region in between. Since the massless scalar wave equation is conformally invariant in two dimensions, and since all 2-dimensional surfaces are conformally flat, it is easy (in principle) to solve the wave equation in any region of constant signature. The issue is how to match solutions when the signature changes. Specificaly, is there a physically reasonable prescription for matching solutions of the wave equation to solutions of Laplace's equation so that the resulting picture can be reasonably described as propagation?

A basic property of the usual theory of the scalar field, related to unitarity, is the existence of a conserved product on solutions, namely the Klein - Gordon product. Choosing the wave equation, so that there is a conserved Klein - Gordon product, implicitly determines the junction conditions one needs to impose in order to obtain global solutions. The resulting mix of positive and negative frequencies produced by the presence of Euclidean regions depends only on the total conformal width of the regions, and not on the detailed form of the metric. Calculating the change of basis (Bogolubov relations) between in and out plane-wave basis functions and ignoring some unimportant phase factors one finds that the resulting solutions satisfy [3].

$$u_k^{out} = cosh(k\Delta\tau)u_k^{in} + i\frac{|k|}{k}sinh(k\Delta\tau)\bar{u}_{-k}^{in}$$

where $\Delta\tau$ is the difference in conformal time τ between the two roots of f. It is interesting to note that this relation extends without modification to the case where there are several Euclidean regions; $\Delta\tau$ then denotes the total conformal width of the Euclidean regions.

The authors further show [4] that solutions and their canonical momenta are well behaved at (each side of) the boundaries between regions of constant signature even though the wave equation is not, so they propose a propagation rule on matching the canonical data at each boundary. The solutions obtained earlier satisfy this condition, and therefore solve the wave equation everywhere. This problem is related with the famous De Witt problem [5] and some related examples with the turnneling models [6][7].

In the poster section M. Ale and L. P. Chimento [8] presented an exact solution of the semiclassical Einstein equations and L. Rodriguez and F. D. Mazitelli studied the Quantum instability of Minkowsky spacetime. Both can be also considered contributions to the General Theory.

2. Stress - Energy Tensor and Black Holes

Paul Anderson, William Hiscock and David Samuel developed a method which allows for the computation of the exact expectation value of the quantum stress-energy tensor for free scalar fields in static spherically symmetric spacetimes. They presented some of their results for massless scalar fields with various couplings to the scalar curvature for Schwarzschild B Hs and various Reissner-Nordstrom B Hs. The ability to compute the exact expectation value of the quantum SET in arbitrary static spherically symmetric spacetimes allows then to address the semiclassical back-reaction problem in these spacetimes. The authors present a method for solving this problem for extreme Reissner-Nordstrom B H and a Schwarzschild B H in a box.

Larry Ford and Thomas Roman talked about constraints on negative energy fluxes seen by inertial observers falling into an evaporating B H. It is known that QFT allows violations of the classical energy conditions, in the form of locally negative energy densities and fluxes, if they are produced by quantum coherence effects. They must be followed by a more than compensating positive fux. Then QFT imposes restrictions on the magnitude and duration of the negative energy flux. In flat space time they obey "uncertainty-principle" type inequalities [10] of the form: $|E|(\Delta T) \leq h$. Here $|E|$ is the magnitude of the negative energy which can be absorbed in a time ΔT. More recently it has been shown by the authors [11] that similar inequalities hold for negative energy fluxes propagating on a $Q = M$ extreme, four-dimensional, Reissner-Nordstrom B H background. These inequalities prevent the unambiguous observation of the limited duration violation of Cosmic Censorship ("Cosmic Flashing"), which would otherwise occur before the arrival of the subsequent positive energy flux. An apparent counter-example to these "quantum inequalities" is the constant negative flux seen by a stationary observer near the B H. But such observer should also see acceleration (Unruh) radiation which mask the negative energy. For this reason, the authors choose to examine the negative energy flux seen by various inertial observes falling into a two-dimensional evaporating Schwarzschild B H. If there are no constraints on these fluxes, then the observers could (in principle) use them to produce, for exanple, gross violations of the second law of thermodynamics. A numerical analysis of the flux as a function of proper time indicates that here there exist quantum inequality-type restrictions on the magnitude and duration of the negative fluxes seen by inertial observers. That is, $|F|\tau^2 \leq h$, where $|F|$ is the magnitude of the negative flux and τ is its typical duration in proper time. In this case at least, quantum inequalities are

satisfied by inertial observes in curved spacetime.Consequently, such observers should not see gross, macroscopic effects due to negative energy.

Ted Jacobson proposed to use B H as microscopes. In fact, when we look at Hawking radiation emerging from the vicinity of a B H, we are registering quanta in outoing quantum field modes. If we assume with Hawking [12] that the field is non-interacting, then these modes can be traced backwards in time to ingoing modes far from the hole at frequency $\sim \omega exp(t/2r_s)$ in the asymptotic rest frame of the hole. Here ω is the outgoing frequency,r_s is the Schwarzschild radius, and t is roughly the time interval between formation of the hole and reception of the quantum. After a time $t \sim 2r_s ln(\omega_{Planck}/\omega)$, these corresponding ingoing modes exceed Planck frequency. He presented an alternative derivation for Hawking radiation, which avoids the role played by super-Planck frequency modes , imposing, in the frame of observers freely falling from asymptotic rest into the B H, that the Planck frequency modes are in the vacuum state. No assumption is made about the state of any other modes. Unruh [13] pointed out that this condition suffices to deduce the existence of Hawking radiation. It is shown that the emitted radiation is approximately thermal. The condition that the Planck frequency modes are in their ground state cannot be derived from any assumption regarding sub-Planckian physics in the free-falling frame. Thus Hawking radiation is a descendent of very short distance physics.

Roberto Camporesi and Atsushi Higuchi computed the SET for scalar and spinor fields with arbitrary mass in anti-de Sitter spacetime using the ξ -function technique. The results agree with those obtained by Pauli-Villars regularization. Also the trace anomaly for the Wess-Zumino model is studied in this space time and takes the "convencional" value. The same authors also calculated the spectral function $\mu(\lambda)$ and the zeta function $\xi(z)$ for a field of integer spin s on a N-dimensional (simply connected) hyperbolic space.

Finally M. Dorca and E. Verdaguer studied a Quantum Field in a Colliding wave Space-time.They study the quantization of a massless scalar field on a spacetime representing the head- on collision of two plane fronted gravitational waves. They consider a vacuum solution of Einstein's equations in which the two waves focus on a non-singular Killing-Cauchy horizon. The interaction region of the two waves is locally isometric to a region of the interior of a Schwarzschild black hole, with the Killing-Cauchy horizon corresponding to the event horizon of the black hole. In this case the colliding wave solution can be maximally extended through the Cauchy horizon, provided one of the transverse coordinates is made cyclic. The resulting spacetime represents the creation of a Schwarzschild B H by the collision of two plane waves propagating on a cylindrical universe [14]. That extension is, however, not essential for quantization of the scalar field.

In the colliding region two unambiguous and physically meaningful vacua can be defined. The"in" vacuum is defined through the positive mode solutions associated to the timelike Killing field in the flat region between the two plane waves before their collision, i.e. the ordinary flat spacetime vacuum. Also an "out" vacuum can be defined at the Cauchy horizon with the modes associated to the two null Killing fields of that horizon. These are easily identified by using a kind of Kruskal-Szekeres like coordinates to describe the interaction region; in these coordinates the metric looks flat at the Cauchy horizon. A particle detector in free fall near the horizon would not react if the quantum state is in the "out"vacuum. That vacuum corresponds to the Unruh vacuum [15] defined at the past horizon of the maximally extended Schwarzschild metric.

The authors propagate the "in" modes through - out the flat, plane wave and interaction region and compare with the "out" modes in the Cauchy horizon by computing the corresponding Bogoliubov coefficients. It is found that particles are created with a thermal spectrum, with a temperature that is proportional to the inverse of the focusing time of the plane waves, which is a measure of the energy of the waves. That is, the"in" vacuum contains a thermal distribution of "out" particles . The spectrum is not exactly thermal, in the sense that it depends on the momentum of the modes in one of the transversal directions rather than the frequency of the modes. This can be understood when comparing with B H radiation, in the sense that in the Schwarzschild B H the responsible for Hawking's radiation [12] is the infinite redshift at the horizon, whereas in the colliding wave problem it is due to the collapse of one of the transversal directions of the waves. The main result is exact even though the solution for the modes cannot be found exactly.

In the poster section a result on charged non-abelian BH solutions of the Einstein-Yang-Mills equations was also presented by D. V. Gal'tsov and M. S. Volkov [16].

3. Particle Production

Esteban Calzetta and Maria Sakellariadou showed how particle creation processes, that are normally used to explain the isotropy of the Universe, and the dissipation of its inhomogeneity, make also that inflationary models can be considered more natural than their Standard counterpart [17]. While the Standard Cosmological Model provides an appealing and so far unchallenged description of the evolution of the Universe from the Big Bang to its present configuration, it also raises the issue of the origin of the highly special initial conditions required for this kind of evolution. Inflationary models of the Universe have been suggested as an explanation for these initial conditions, but they too depend upon a kind of fine tuning of the initial conditions. Concretely, recent work has established that Inflation requires initial conditions to be homogeneous on scales in excess of one horizon length. This homogeneity cannot be explained through classical physical processes, and therefore, as far as classical cosmology is concerned, Inflationary models do not seem to be more "natural" than their Standard counterpart.

The possibility remains, that the picture changes when quantum cosmological effects are taken into account. For example, it is known that quantum particle creation processes could afford an explanation of the isotropy of the Universe, and in general would tend to favour dissipation of inhomogeneities. Thefore the authors explore the relevance of quantum effects to the onset of inflation, focussing on inhomogeneous but isotropic models with respect to a preferred point, and showed that these effects relax the requirements for succesful Inflation to the point where they become compatible with generic cosmic initial conditions.

Several contibutions were related to particle production in the poster section: L. P. Chimento, A.E. Cossarini and F. G. Pensa [18] studied a Spin 1 Field in Rindler Space and found a non-Planckian spectrum for the Rindler observer, A.Higuchi, A. Matsas and D. Sudarsky [19] presented a contribution about the Bremsstrahlung and Zero-energy Rindler Photons and U.Percoco, V.M.Villalba and P.Pujol [20] analyzed the vacuum effects asociated with the scalar and Dirac particles in non uniformly accelerated frames of reference.

Finally C. Loustó, J. J. Wheeler and G. Domenech, M. Levinas and N. Umérez presented interesting results in their posters [21], [22] and [23].

4. Conclusion

Unfortunately many authors that sent abstracts were absent at the meeting. Anyhow interesting subjects were analized and many interesting contributions were presented. But we were particulary impressed by three contributions, that we consider rather important: the study of David Boulware, about QFT in Gott's space time, was an excellent exploration of a new, difficult, and promising subject. The contribution of Paul Anderson, William Hiscock and David Samuel, about the exact computation of the expectation value of the SET, was a solution of an old and well known problem. The report of Esteban Calzetta and Maria Sakellariadou was an outstanding example of how QFTCST can help us to improve the inflationary model in the best tradition of this theory.

References

[1] Kim S. W., Thorne K. S., Phys Rev. D, **43**, 3929 (1991).
[2] Hawking S. W., Phys Rev. D, to be published.
[3] Dray T., Manogue C. A., Tucker R. W., Gen. Rel. Grav., **23**, 967 (1991).
[4] Dray T., Manogue C.A., Tucker R. W., The Effect of Signature Change on Scalar Field Propagation (in preparation)
[5] Anderson A., De Witt B., Found of Phys. **16**, 91 (1986)
[6] Hartle J. B., Hawking S. W., Phys Rev. D **28**, 2960 (1983)
[7] Vilenkin A., Phys. Lett. B **117**,25, (1982) and Phys. Rev., **27**, 2848 (1983)
[8] Ale M., Chimento L. P., Abstracts GR, 13, 575 (1992)
[9] Rodriguez L., Mazzitelli F. D., Abstracts GR 13, 612 (1992)
[10] Ford L. H., Phys. Rev. D., **43**, 3972 (1991)
[11] Ford L. H., Roman T. A., Phys. Rev. D, **41**, 3662 (1990)
[12] Hawking S. W., Nature **248**, 30 (1974) and Comm. Math. Phys., **43**, 199 (1975)
[13] Unruh W. G., Phys. Rev. D, **15**, 365 (1977)
[14] Yurtsever U., Phys. Rev. D, **40**, 360 (1989)
[15] Unruh W. G., Phys. Rev. D, **14**, 870 (1976)
[16] Gal'tsov D. V., Volkov M. S., Abstracts G 13, 596 (1992)
[17] Calzetta E., Sakellariadou M., Phys. Rev. D, **45**, 2802 (1992)
[18] Chimento L. P., Corssarini A. E., Pensa F. G., Abstracts GR 13, 588 (1992)
[19] Higuchi A., Matsas G. E. A., Sudarsky D., Abstracts GR 13, 601 (1992)
[20] Percoco U., Villalba V. M., Pujol P., Abstracts GR 13, 611 (1992)
[21] Loustó C. O., Abstracts GR 13, 607 (1992)
[22] Wheeler J. T., Abstracts GR 13, 614 (1992)
[23] Domenech G., Levinas M., Umérez N., "The ideal fluid as the classical limit of free quantum field and a model for an interacting fluid in Curved Space- Time", Preprint IAFE (1992)

LIST OF PARTICIPANTS

ARGENTINA

Abadi M.G.; Obs. Astronom., Univ. Nac. de Cordoba.
Alarcon, Alejandra; 25 de Mayo 467, Las Varillas.
Alvarez, Paola Elisa; IAFE, Buenos Aires.
Ameri, Oscar; Resistencia.
Anchordoqui, Luis Alfredo; Calle 3 N 131, La Plata.
Aquilano, Roberto; IFIR, Univ. Nac. de Rosario.
Barraco, Daniel; FAMAF, Univ. Nac. de Cordoba.
Birman, Graciela Silvia; Inst. Argentino de Matematica, Buenos Aires.
Caceres, Manuel O.; Centro Atomico Bariloche.
Calzetta, Esteban; IAFE, Buenos Aires.
Castagnino, Mario; IAFE, Buenos Aires.
Cebral, Raul; IAFE, Buenos Aires.
Chavez, Catalina; Resistencia.
Chimento, Luis P.; Depto. de Fisica, Univ. Nac. de Buenos Aires.
Cossarini, Adriana Edith; Depto. de Fisica, Univ. Nac. de Buenos Aires.
Diaz, Eliseo Narcizo; Cordoba.
Dagotto, Andres; FAMAF, Univ. Nac. de Cordoba.
Dain, Sergio; FAMAF, Univ. Nac. de Cordoba.
Dalvit, Diego; FAMAF, Univ. Nac. de Cordoba.
Deza, Roberto; Depto. de Fisica, Univ. Nac. de Mar del Plata.
Dominguez, Eduardo; FAMAF, Univ. Nac. de Cordoba.
Dotti, Gustavo; FAMAF, Univ. Nac. de Cordoba.
Edelstein, Jose D.; Depto. de Fisica. Univ. Nac. de La Plata.
Gaioli, Fabian; IAFE, Buenos Aires.
Garate,Alexandre; FAMAF, Univ.Nac. de Cordoba.
Garcia, Marta; FAMAF, Univ. Nac. de Cordoba.
Garcia Lambas, D.; Obs. Astronom., Univ. Nac. de Cordoba.
Gattoni, Alberto; FAMAF, Univ. Nac. de Cordoba.
Gleiser, Reinaldo; FAMAF, Univ. Nac. de Cordoba.
Guibert, Roberto; FAMAF, Univ. Nac. de Cordoba.
Hamity, Victor H.; FAMAF, Univ. Nac. de Cordoba.
Harari, Diego; IAFE, Buenos Aires.
Jakubi, Alejandro; Depto Fisica, Univ. Nac. de Buenos Aires.
Jofre,Oscar Emilio; IAFE, Buenos Aires.
Kandus,Alejandra; IAFE, Buenos Aires.
Kozameh, Carlos; FAMAF, Univ. Nac. de Cordoba.
Laciana, Carlos; IAFE, Buenos Aires.
Lamberti, Walter; FAMAF, Univ. Nac. de Cordoba.
Lara, Luis; IFIR, Univ. Nac. de Rosario.

Leguizamon, Enzo; FAMAF, Univ. Nac. de Cordoba.
Lehner, Luis; FAMAF, Univ. Nac. de Cordoba.
Levinas,Leonardo; IAFE, Buenos Aires.
Lewis, Pedro; IFIR, Univ. Nac. de Rosario.
Lombardo, Fernando; IAFE, Buenos Aires.
Macchi, Guido; Cordoba.
Maldacena, Juan Martin, Centro Atomico Bariloche.
Manias, M. V.; Cordoba.
Martin, Maria; Cordoba.
Moreschi, Osvaldo; FAMAF, Univ. Nac. de Cordoba.
Nagy, Gabriel; FAMAF, Univ. Nac. de Cordoba.
Nicasio, Oscar; FAMAF, Univ. Nac. de Cordoba.
Nicotra M. A.; Obs. Astronom., Univ. Nac. de Cordoba.
Nuñez, Carmen; IAFE, Buenos Aires.
Ortiz, Omar; FAMAF, Univ. Nac. de Cordoba.
Paganini, Victoria; FAMAF, Univ. Nac. de Cordoba.
Perez Bergliaffa, Santiago; Depto. de Fisica, Univ. Nac. de Buenos Aires.
Perez,Silvina; FAMAF, Univ. Nac. de Cordoba.
Portnoy, Guillermo Ruben; IAFE,Buenos Aires.
Raggio, Guido; FAMAF, Univ. Nac. de Cordoba.
Reula, Oscar; FAMAF, Univ. Nac. de Cordoba.
Rodriguez, Victor; Facultad de Filosofia, Univ. Nac. de Cordoba.
Romero, Gustavo; Depto. de Fisica Univ. Nac. de La Plata.
Sanchez, Cristian; FAMAF, Univ. Nac. de Cordoba.
Sautu, Sandra; Depto. de Fisica, Univ. Nac. de Buenos Aires.
Scochimarro, Roman; Univ. Nac. de Buenos Aires.
Sforza, Daniel; IAFE, Buenos Aires.
Sisterna,Pablo Daniel; Depto. de Fisica, Univ. Nac. de Mar del Plata.
Stoico, Cesar; IFIR, Univ. Nac. de Rosario.
Surpi, Gabriela; IAFE, Buenos Aires.
Thibeault, Marc; IAFE, Buenos Aires.
Tissera,Patricia; Obs. Astronom., Univ. Nac. de Cordoba.
Tomasini, Maria Cecilia; IAFE, Buenos Aires.
Trobo, Marta Liliana; Depto. de Fisica, Univ. Nac. de Buenos Aires.
Umerez, Norberto Daniel; IAFE, Buenos Aires.
Vega, Marcelo; IAFE, Buenos Aires.
Vazquez, Alberto; IAFE, Buenos Aires.
Vucetich, H.; Depto. de Fisica, Univ. Nac. de La Plata.
Zandron, Oscar Sergio; IFIR, Univ. Nac. de Rosario.

AUSTRALIA

Anderson, Malcolm; Dept. of Theor. Phys., RSPHYSSE, Australian Nat. Univ.
Benn, Ian M.; Dept. of Math., Univ. of Newcastle.
Blair, David G.; Dept. of Physics, Univ. of Western Australia.
Brewin, Leo; Dept. of Math., Monash Univ.
Chow, Ernest; Dept. of Math., Monash Univ.
Lun, Anthony W.C.; Dept. of Math., Monash Univ.
McIntosh, Colin B. G.; Dept. of Math., Monash Univ.
Notcutt, Mark; Dept. of Physics, Univ. of Western Australia.
Scott, Susan M.; Centre for Math. Analysis, Australian Nat. Univ.
Smrz, P. K.; Dept. of Math., Univ. of Newcastle.
Szekeres, Peter; Dep. Physics and Math. Physics, Univ. of Adelaide.
Tobar, Michael; Dept. of Physics, Univ. of Western Australia.

AUSTRIA

Tapia, Victor; Postfach 132, 5024 Salzburg.

BELGIUM

Ferraro, Rafael; Faculte des Sciences, Universite Libre de Bruxelles.

BRAZIL

Aguiar, Odylio Denys; INPE/ DAS, Av. D. Astronautas 1750 - C P 515.
Araujo, Marcelo E.; Depto. de Matematica, Univ. de Brasilia.
Arcuri, Regina Celia; DRP CBPF,Rio de Janeiro.
Bedran, M.L.; Instituto de Fisica UFRJ, Rio de Janeiro.
Calvao, Mauricio O; Univ. Federal de Rio de Janeiro.
Chan, Roberto; CNPQ - Obs. Nac., Rio de Janeiro
Da Silva, Maria de Fatima A.; CNPQ - Obs. Nac., Rio de Janeiro
Garcia de Andrade, Luiz, Inst. de Fisica UERJ, Rio de Janeiro.
Horvath, Jorge E.; Inst. Astronom. e Geofisico, Univ.e de Sao Paulo.
Lemos, Jose P. S.; CNPQ - Obs. Nac., Rio de Janeiro
Mendoca da Silveira, Francisco Eugenio.
Mondaini, R.P.; Instituto Politecnico de Rio de Janeiro, RJ.
Monteiro S., Waldemar; Inst. de Física, Univ. Fed. do Rio de Janeiro.
Pereira, Jose Geraldo; Inst. de Fisica Teorica, Univ. Estadual Paulista.
Pilotto, Fernando, Rau Maris e Barros 437, Porto Alegre.
Pinto Neto, Nelson; C.B.P.F., Rio de Janeiro.
Reboucas, Marcelo J.; C.B.P.F. - DRP, Rio de Janeiro.
Rodrigues, Ligia M. C. S.; C.B.P.F./D.R.P., Rio de Janeiro.
Romero, Carlos; Depto. de Fisica - UFPB.
Santos, Nilton Oscar; Depto. de Astrofisica, Obs. Nac., Rio de Janeiro.
Sivaram, Chandra.
Soares, Ivano Damiao; C.B.P.F./D.R.P., Rio de Janeiro.

Vasconcellos, Cesar A.; Inst. de Fisica, Univ. Fed. Rio Grande do Sul.
Vilar, Luiz C.Q.; Rua Pirassinunga 95, Jacarepagua.
de Souza, Mario Everaldo; Depto de Fisica, Univ. Federal de Sergipe-CCET.

CANADA

Abolghasem, Gholamhossein; Dept. Math., Statistics and Computer Science.
Blais, J.A.R.; University of Calgary.
Coley, Alan; Math. Dept., Dalhousie Univ.
Duggal, Krishan Lal; Dept. of Math. and Statistics, Univ. of Windsor.
Fee, Greg; Computer Science Dept., Univ. of Waterloo.
Hewitt, C.G.; Univ. of Waterloo.
Hobill,David; Dept of Physics and Astronomy, University of Calgary.
Kunzle, Hans P.; Dept. of Math., Univ. of Alberta.
Ludwig, Garry; Dept. of Math., Univ. of Alberta.
Mann, Robert; Dept. of Physics, Univ. of Waterloo.
Marleau, Francine; Dept. of Astronomy, Univ. of Toronto.
McLenaghan, R.G.; Dept. of Appl. Math., Univ. of Waterloo.
Reyes, Enrique; Dept. of Math., Univ. of Saskatchewan.
Strickler, Iris; University of Waterloo.
Torrence, Robert J.; Dept.of Math. and Statistic, Univ. of Calgary.

CHILE

Banados, Maximo; Centro de Estudios Cientificos de Santiago.
Contreras, Mauricio; Dep. de Fisica, Univ. de Chile, Santiago.
Cruz Marin, Norman; Depto. de Fisica, Univ. de Santiago.
De la Jara Fuentes, Jaime; Depto, de Fisica, Universidad de Chile, Santiago.
Del Campo, Sergio; Inst. de Fisica, Univ. Catolica de Valparaiso.
Gomberof,Andres; Dep. de Fisica, Fac. Ciencias, Univ. de Chile.
Infante, Leopoldo; Grupo de Astrofisca, Pontificia Univ. Catolica.
Lopez, Carlos A.; Depto. de Fisica, Univ. de Chile, Santiago.
Martinez S., Cristian; Depto. de Fisica, Univ. de Chile, Santiago.
Pena Alvarez, Leda; Depto. de Fisica, Univ. de Chile, Santiago.
Perez Ponce, Alejandro; Depto. de Fisica, Univ. de Chile, Santiago.
Quintana, Hernan; Grupo de Astrofisica, Pontificia Univ. Catolica.
Sanhueza, Magdalena; Depto. de Fisica, Univ. de Chile, Santiago.
Teitelboim, Claudio; Centro de Estudios Cientificos de Santiago.
Zamorano, Nelson; Depto. de Fisica, Univ. de Chile, Santiago.
Zanelli, Jorge; Depto. de Fisica, Univ. de Chile, Santiago.

FRANCE

Carter, Brandon; Dept. of Relat. Astrophysics, Observatoire de Paris.
Choquet-Bruhat, Y.; Tour 66 Mecanique, 4 Place Jussieu.
Elbaz, E.; Inst.des Sciences de la Matiere, Univ. Claude Bernard - Lyon 1.

Gourgoulhon, Eric; DARC, Observatoire de Paris-Meudon.
Mizony, Michel; Bat 101, I.M.I., Universite Lyon 1.
Munier, Alan; Centre d'Etudes de Limell-Velenton.

GERMANY

Brauer, Uwe; Max Plank Inst. für Astrophys., Garching.
Campanelli, Manuela; Univ.t Konstanz, Fakultat für Physik.
Danzmann, K.; Max Planck Inst. für Quantenoptic, Garching.
Ehlers, J.; Max Planck Inst. für Astrophys., Garching.
Frauendiener, Joerg; Max Planck Inst. für Astrophys., Garching.
Gottlober, Stefan; Astrophysikalisches Institut, D-O- 1591 Postdam.
Kasper, Uwe; Project group "cosmology", Univ. of Postdam.
Kramer, Dietrich; Theor. Physik. Inst., Friedrich-Schiller Univ. Jena.
Lousto, Carlos Oscar; Univ.t Konstanz, Fakultat für Physik.
Neugebauer, G.; Friedrich-Schiller-Univ.et, Theor. Physik. Inst.
Perlick, Volker; TU Berlin, Sekr Pn 7-1, Hardenbergstr. 36.
Pfister, Herbert; Inst. f. Theoretische Physik, Univ.et Tubingen.
Rendall, Alan; Max-Planck-Inst. für Astrophys., Garching.
Salazar, Humberto; Sektion Physik, FSU-Jena.
Salie, Nikolaus; Theoret. Physikal. Inst.,Univ. Jena.
von Borzeszkowski; Horst-Heino, Buckower Ring 37, 0-1141 Berlin.
Zimdahl, Winfried; Univ. Düsseldorf, Inst. für Theor. Physik II.

INDIA

Banerji, Sriranjan; Dept. of Physics, The Univ. of Burdwan.
Chakrabarti, Sandip K.; Th. Astrophys, TIFR, Homi Bhabha Road.
Dadhich, Naresh; IUCAA Post Bag 4, Poona Univ. Campus.
Dhurandhar, S. V.; IUCAA Post Bag 4, Poona Univ. Campus.
Iyer, B. R.; Raman Research Laboratory, Bangalore - 560 080.
Johri, V.B.; Dept. of Math. Indian Inst. of Technology, Madras.
Roy, Sisir; Indian Statistical Institute, 203 B. T. Road.

IRELAND

Dolan, Brian; Dublin Inst.for Advanc. Studies, School of Theor. Phys.

ITALY

Astone, Pia; Dip. di Fisica, Univ. di Roma.
Barone, Fabrizio; Dip. di Scienze Fisiche, Univ. di Napoli.
Bertotti, Bruno; Dip. di Fisica Nucleare e Teorica, Univ. di Pavia.
Coccia, Eugenio; Dip. di Fisica, Univ. di Roma "Tor Vergata".
Cognola, Guido; Dip. di Fisica, Univ. di Trento.
Di Bartolo, Cayetano.
Di Fiore, Luciano; I.N.F.N., Napoli.

Ellis, G. F. R.; SISSA, Trieste.
Fargion, Daniele; Dip. di Fisica, Univ. di Roma.
Ferraris, Marco; Dip. di Matematica, Univ. di Cagliari.
Francaviglia, Mauro; Istituto di Fisica Matematica, "J. L. Lagrange", Torino.
Giacconi, Paola; Dip. di Fisica, Univ. di Bologna.
Husain, Viqar; High Energy Group, ICTP, Trieste.
Immirzi, Giorgio; Dip. di Fisica, Univ. di Perugia.
Modena, Ivo; Dip. di Fisica, Univ. di Roma "Tor Vergata".
Nelson, Jeanette; Dip. di Fisica Teorica, Univ. di Torino.
Occhionero, Franco; Oss. Astronomico di Roma, Roma.
Piotrkowska, Kamilla M., SISSA/ISAS, Strada Costiera 11.
Pizzella, Guido; Dip. di Fisica, Univ. di Roma.
Recami, Erasmo; Dip. di Fisica, Univ. di Catania.
Ricci, Fulvio; Dip. di Fisica, Univ. La Sapienza.
Stella, L.; Oss. Astronomico di Brera, Milan.
Tartaglia, Angelo; Dip. di Fisica Politecnico.
Visco, Massimo; C.N.R./IFSI., Roma.

JAPAN

Hara, Tetsuya; Kyoto Sangyo Univ., Kitaku, Kamigamo.
Kawai, Toshiharu; Dept. of Physics, Osaka City Univ.
Kawashima, Nobuki; Institute of Space and Astronautical Science.
Kuroda, Kazuaki; Inst. for Cosmic Ray Research, Tanashi.
Morikawa, Masahiro; Dept. of Physics, Kyoto Univ.
Nakamura, Takashi; Yukawa Institute for Theoretical Physics.
Nishida, Shogo; Dept. of Astronomy, Univ. of Tokyo.

KOREA

Kim, Sung-Won; Dept. of Science Education, Ewha Womans Univ.

MEXICO

Camacho Quintana, Abel; Depto. de Fisica, UNAM-IZT.
Capovilla, Riccardo; Depto. de Fisica, CINVESTAV-IPN.
Alducin, Pablo; Depto. de Fisica, Univ. Auton. Metrop., Iztapalapa.
Gaftoi Negoescu, Ana Violeta; Urraca 79, Las Alamedas.
Garcia, Alberto; Depto. de Fisica, CINVESTAV-IPN.
German, Gabriel; Inst. de Fisica, UNAM.
Harleston, Hugo; ICN-UNAM.
Llorens, Marc Mars.
Lopez Romero, Jose Mauricio; Av. Republica Federal Sur.
Macias, Alfredo; Depto. de Fisica, Univ. Auton. Metrop., Iztapalapa.
Matos, Tonatiuh; Depto de Fisica, CINVESTAV-IPN.
Nahmad-Achar, Eduardo; ICN-UNAM.

Nunez, Dario; ICN-UNAM.
Obregon, Octavio; Depto. de Fisica, Univ. Auton. Metrop., Iztapalapa.
Rosenbaum P., Marcos; ICN-UNAM.
Ryan, Jr., Michael P.; ICN-UNAM.
Sinha, Sukanya; ICN-UNAM.
Sobczyck, Garret; Fac. de Estudios Sup. Cuautitlan, UNAM.
Socolovsky, Miguel; ICN-UNAM.
Sussman, Roberto; ICN-UNAM.

NETHERLANDS

Slagter, R. J.; Prinses Marielaan 16, 1405 EP Bussum.
Spallicci, A.D.A.M.; European Space Agency/ESTEC.

NEW ZEALAND

Kerr, Roy; Dept. of Math., Univ. of Canterbury.
Perjes, Zoltan; Dept. of Math., Univ. of Canterbury.

NORWAY

Andre, Georg; Rognersvingen 6, 7021 Trondheim.
Bakke, Finn; Theoretical Physics, Univ. of Trondheim - NTH.
Knutsen, Henning; Hogskolescenteret Rogaland, Ullandhaug.
Soleng, Harald H.; Dept. of Physics, Blindern.

PERU

Bernui, Armano; Fac. de Ciencias, Univ. Nac. de Ingenieria, Lima.

POLAND

Chrusciel, P.; Institute of Math., Polish Academy of Sciences.
Krasinski, Andrzej; N. Copernicus Astronom. Center, Polish Academy of Sciences.
Krolak, Andrzej; Institute of Math., Polish Academy of Sciences.

PORTUGAL

Bento, Maria da Conceicao; CFMC-INIC.
Bertolami, Orfeu; CFMC-INIC.
Da Costa, Jose Manuel C; Dept. de Matematica, Univ. de Madeira.

PUERTO RICO

Esteban, Ernesto; Physics Dept., University of Puerto Rico.

ROMANIA

Ceapa, Alexandru; P O Box 1-1035, 70700 Bucharest.

RUSSIA

Gal'tsov, D.V.; Dept. of Theoretical Phys., Moscow State Univ.
Grishchuk, Leonid P.; Sternberg Astronomical Institute, Moscow Univ.
Melnikov, Vitaly N.; NPO "Eltest".

SOUTH AFRICA

Bishop, N. T.; Dept. of Math., Univ. of South Africa.
Hellaby, Charles; Dept. of Applied Math., Univ. of Cape Town.
Krige, J.D.; Dept. of Math. and Applied Math., Univ. of Natal.
Maharaj, Sunil; Dept. of Math. and Applied Math., Univ. of Natal.
Nevin, Jennifer; Dept. of Math. and Applied Math., Univ. of Natal.
Taylor, David Roy; Dept. Comp. and Applied Math., Univ. of the Witwatersrand.
Xu, Chongming; Dept. of Applied Math., Univ. of Cape Town.

SPAIN

Carot, Jaume; Depto. de Fisica, Univ. Illes Balears.
Chamorro, Alberto; Depto. de Fisica Teorica, Univ. del Pais Vasco.
Chinea, F. J.; Depto. de Fisica Teorica II, Fac. de Ciencias Fisicas.
Herrera, Luis, Fisica Teorica, Fac. de Ciencias, Univ. del Pais Vasco.
Lobo, J. Alberto; Depto. de Fisica, Univ. de Barcelona.
Martin-Senovilla; Jose Maria; Depto. de Fisica, Univ. de Barcelona.
Mas-Franch, Luis; Depto. de Fisica, Facultad de Ciencias.
Pavon, Diego; Depto. de Fisica, Univ. Autonoma de Barcelona.
Raluy Tomas; Luis, Apartado Postal 258, C.P. 43500, Tortosa.
Rastall, Luisa R.; Apartado Postal 258, C.P. 43500, Tortosa.
Ruiz, Eduardo; Grupo de Fisica Teorica, Univ. de Salamanca.
Swider, L.; Apartado Postal 258, C.P. 43500, Tortosa.
Verdaguer, Enric; Dept. Fisica Teorica, Univ. Autonoma de Barcelona.

SWEDEN

Aman, Jan.
Andersson, Lars; Dept. of Math., Royal Institute of Technology.
Bengtsson, Ingemar; Inst. of Theoretical Physics, S-41296 Goteborg.
Edgar, Brian; Dept. of Math., Univ. of Linkoping.
Iriondo, Mirta; Dept. Theoretical Physics, Royal Institute of Technology.
Rosquist, Kjell; Dept. of Physics, Univ. of Stockholm.
Salinas, Ener; Univ. of Stockholm.
Widlund, Lennart; Univ. of Stockholm.

SWITZERLAND

Bestgen, Ruth; Dept. of Theoretical Physics, Univ. of Bern.
Held, Alan; Dept. of Theoretical Physics, Univ. of Bern.
Jerjen, Helmut; Astronomical Institute, Univ. of Basel.

Thomi, Peter; Hubelmattstr 14, 4500 Solothum.

TURKEY

Nutku, Yavuz; Dept. of Math., Bilkent Univ.

UNITED KINGDOM

Andersson, Nils; Dept. of Physics and Astronomy, Univ. of Wales.
Barnes, Alan; Dept. of Math., Univ. of Aston.
Bonnor, W. B.; 1 South Bank Terrace, Surbiton.
Carr, Bernard; School of Mathematical Sciences, Queen Mary College.
Clarke, C. J. S.; Faculty of Math. Studies, Univ. of Southampton.
Dickson, Christopher; Dept. of Physics and Astronomy, Univ. of Wales.
Dunbar, David C., DAMTP, Univ. of Liverpool.
Grant, James; Dept. of Applied Math. and Theoretical Physics.
Griffiths, J. B.; Dept. of Math., Univ. of Technology.
Hall, G. S.; Dept. of Math., Univ. of Aberdeen.
Kibble, T. W. B., Blackett Laboratory, Imperial College.
Koutras, A., 170 Kensington Park Road, London W1-2GR.
Liddle, Andrew R., Astronomy Centre, Univ. of Sussex.
MacCallum, M. A. H., School of Mathematical Sciences, Queen Mary College.
Marriott, Neal, The Institute of Physics, IOP.
Mason, Lionel J., The Mathematical Institute, Univ. of Oxford.
Newton, Gavin, Physics and Astronomy, Glasgow Univ.
Nicholson, David; Dept. of Physics and Astronomy, Univ. of Wales.
Penrose, R.; Dept. of Math., Univ. of Oxford.
Schutz, B. F.; Dept. of Physics, Univ. of Wales.
Skea, J. E. F.; School of Mathematical Sciences, Queen Mary College.
Tavakol, Reza; School of Mathematical Sciences, Queen Mary College.
White, S.; Institute of Astronomy, Univ. of Cambridge.
Wolf, Thomas; School of Mathematical Sciences, Queen Mary College.

URUGUAY

Gambini, Rodolfo; Instituto de Fisica, Fac. de Ciencias.
Griego, Jorge; Instituto de Fisica, Fac. de Ciencias.

USA

Abramovici, Alex; Gravitational Physics, Caltech.
Ashby, Neil; Dept. of Physics, Univ. of Colorado at Boulder.
Bacinich, Edward J.; 1048 S. Ocean Blvd., Palm Beach, FL 33480.
Bender, Peter; JILA, Univ. of Colorado at Boulder.
Berger, Beverly K.; Dept. of Physics, Oakland Univ.
Boulware, David G.; Physics Dept., Univ. of Washington.
Brill, Dieter R.; Dept. of Physics, Univ. of Maryland.

Brown, J. David; Dept. of Physics, Univ. of North Carolina.
Burd, Adrian; McDonnell Center for Space Science, Washington Univ.
Carinhas, Philip; Physics Dept., Univ. of Wisconsin-Milwaukee.
Carlip, Steven; Dept. of Physics, Univ. of California at Davis.
Carroll, Sean; Mail Stop 10, 60 Garden Street.
Christensen, Steven M., MathSolutions Inc., P.O. Box 16175.
Cutler, Curt J.; Div. of Physics, Caltech.
Daughton, Alan; Physics Dept., Syracuse Univ.
Diaz, Mario C.; Dept. of Physics, Mercyhurst College.
Dray, Tevian; Dept. of Math., Oregon State Univ.
Drever, R. W. P.; Div. of Physics, Caltech.
Dubal, Mark; Center for Relativity, Univ. of Texas at Austin.
Estabrook, Frank B.; Jet Propulsion Lab., Caltech.
Evans, Charles R.; Dept. of Physics, Univ. of North Carolina.
Finkelstein, David; School of Physics, Georgia Institute of Technology.
Ford, Larry; Dept. of Physics, Tufts Univ.
Friedman, J. L.; Physics Dept., Univ. of Wisconsin.
Garriga, Jaume; Dept. of Physics, Tufts Univ.
Geroch, Robert; Enrico Fermi Institute, Univ. of Chicago.
Goldberg, Joshua N.; Dept. of Physics, Syracuse Univ.
Gomberof, Andres.
Gonzalez, Gabriela; Dept. of Physics, Syracuse Univ.
Goode, Stephen W.; Dept. of Math., California State Univ.
Halliwell, Jonathan; Center for Theoret. Physics, MIT.
Halpern, Leopold; Dept. of Physics, Florida State Univ.
Hamilton, William O.; Dept. of Physics and Astronomy, Louisiana State Univ.
Harrison, B. Kent; Dept. of Physics and Astronomy, Brigham Young Univ.
Hartle, James B.; Dept. of Physics, Univ. of California at Santa Barbara.
Heckel, Blayne, Univ. of Washington.
Hellings, Ronald.
Higuchi, Atsushi, Enrico Fermi Institute, Univ. of Chicago.
Hiscock, William A.; Dept. of Physics, Montana State Univ.
Hobill, David; NCSA, 605 E. Springfield Ave.
Hu, B. L.; Dept. of Physics, Univ. of Maryland.
Isenberg, Jim; Dept. of Math., Univ. of Oregon.
Jackson, Martin; Dept. of Math., Univ. of Puget Sound.
Jacobson, Ted; Dept. of Physics, Univ. of Maryland.
Jantzen, Robert; Dept. of Math. Sciences, Villanova Univ.
Keiser, George M.; W.W. Hansen Lab. of Physics, Gravity Probe B.
Kidder, Larry; Dept. of Physics, Washington Univ.
Klainerman, S.; Dept. of Math., Univ. of Princeton.
Konkowski, Deborah A.; Math. Dept., U. S. Naval Academy.

Kuchar, Karel; Dept. of Physics, Univ. of Utah.
Lewandowski, Jerzy; Dept. of Physics, Syracuse Univ.
Lewis, Caroline; Center for Relativity, Univ. of Texas at Austin.
Mann, Laurence D.; High Energy Physics Lab., Stanford Univ.
Mather, John C.; NASA Goddard Space Flight Center
Meyer, David A., IPAPS, Univ. of California, San Diego.
Misner, Charles; Dept. of Physics, Univ. of Maryland.
Moncrief, Vincent; Dept. of Physics, Yale Univ.
Morse, Peter Andrew; Dept. of Physics, Univ. of California.
Newell, David Bryan; JILA, Univ. of Colorado at Boulder.
Newman, E. T.; Dept. of Physics, Univ. of Pittsburgh.
Paik, Ho Jung; Dept. of Physics, Univ. of Maryland.
Paz, Juan Pablo; Theoretical Astrophysics, Theoretical Division, LALN.
Price, Richard H.; Dept. of Physics, Univ. of Utah.
Pullin, Jorge; Dept. of Physics, Univ. of Utah.
Pyne, Ted; Mail Stop 10, 60 Garden Street.
Redmount, Ian H.; College of Letters and Science, Univ. of Wisconsin.
Ritter, Rogers D.; 117 Chestnut Ridge Rd, Charlottesville.
Rovelli, Carlo; Dept. of Physics, Univ. of Pittsburgh.
Saulson, Peter R.; Dept. of Physics, Syracuse Univ.
Seidel, Edward; NCSA Univ. of Illinois, 605 E. Springfield Ave.
Shahid Saless; Bahman, Jet Prop. Lab, Caltech.
Shoemaker, David; Dept. of Physics, MIT.
Simone, Liliana E.; Dept. of Physics, Washington Univ.
Smolin, Lee; Dept. of Physics, Syracuse Univ.
Sorkin, Rafael; Dept. of Physics, Syracuse Univ.
Sudarsky, Daniel; Enrico Fermi Institute, Univ. of Chicago.
Suen, Wai-Mo; Dept. of Physics, Washington Univ.
Taylor, J. H.; Dept. of Physics, Princeton Univ.
Teukolsky, Saul; 330 Newman Lab., Cornell Univ.
Thorne, Kip; Dept. of Physics, Caltech.
Vilenkin, A.; Tufts Cosmolgy Institute, Tufts Univ.
Wahlquist, Hugo D.; Jet Propulsion Laboratory.
Wald, R. M., Enrico Fermi Institute, Univ. of Chicago.
Wheeler, James T.; Dept. of Physics and Astronomy, Swarthmore College.
Whitcomb, Stanley E.; LIGO Project, Caltech.
Will, Clifford M.; Dept.of Physics, Washington Univ.
Wilson, Edward P.; Math. Dept., Univ. of Dallas.
Wiseman, Alan G.; Dept. of Physics, Washington Univ.
Witten, Louis; Dept. of Physics, Univ. of Cincinnati.
Woolgar, Eric; Dept. of Physics, Syracuse Univ.

USSR

Soloviev, Vladimir; Inst. for High Energy Physics, Protvino.
Zalaletdinov, Roustam M.; Inst. of Nuclear Physics, Uzbek Academy of Sciences.

VENEZUELA

Aragone, C.; Dept. de Fisica, Univ. Simon Bolivar.
Barreto, William; Univ. de Oriente, Nucleo de Sucre - Esc. de Ciencias.
Hernandez, Hector; Depto. de Fisica, Univ. de los Andes.
Khoudeir, Adel; Depto. de Fisica, Univ. de los Andes.
Melfo, Alejandra; Depto. de Fisica, Univ. de los Andes.
Nuñez, Luis A.; Depto. de Fisica, Univ. de los Andes.
Patino, Alberto; Depto. de Fisica, Univ. de los Andes.
Percoco, Umberto; IVIC, Centro de Fisica.
Rago,Hector; Depto. de Fisica, Univ. de los Andes.
Villalba, Victor M.; I.V.I.C. Centro de Fisica.

TOTAL NUMBER OF PARTICIPANTS: **425**

AUTHOR INDEX

International Scientific Committee

R M Wald (USA) (Chairman)
P Bender (USA)
D Blair (Western Australia)
M Castagnino (Argentina)
J Centrella (USA)
G F R Ellis (Italy)
Li-Zhi Fang (China)
V Ferrari (Italy)
G Gibbons (United Kingdom)
L Grishchuk (USSR)
A Held (Switzerland)
E T Newman (USA)
O A Reula (Argentina)
B Schmidt (Germany)
I Shapiro (USA)
L Smolin (USA)
K Thorne (USA)
W Unruh (Canada)
C V Vishveshwara (India)

Local Organizing Committee

C N Kozameh (Chairman)
M Castagnino
R J Gleiser
V H Hamity
P W Lamberti
O M Moreschi
J A Pullin
O A Reula