SSCL-N-691

REPORT OF THE SSC COMPUTER PLANNING COMMITTEE

January 1990

L. Price, Editor

Table of Contents

Committee Summary and Recommendations 3

Appendices

A. Committee Meeting Agenda 13

B. Parallel Computing at the SSC
P. Avery 14

C. Thoughts on a Balanced Model for SSC Off-line Computing
L. Cottrell 22

D. Networking at the SSC Laboratory
G. Brandenburg 48

E. Computing for Experiments at the SSC: An Initial Model for 1990–1992
H. Newman 56

F. Computing Requirements for Theory at SSC
F. Paige 90

G. Use of Commercial Computing Products Online
A. Lankford 94

1. Introduction

The SSC Computer Planning Committee met at the SSC Laboratory (SSCL) on December 12-13, 1989. Members of the committee were Joe Ballam (SLAC)–Chair, Paul Avery (Florida), George Brandenburg (Harvard), Les Cottrell (SLAC), Andy Lankford (SLAC), Harvey Newman (Caltech), Frank Paige (BNL), and Larry Price (ANL). The committee was asked to make recommendations for SSC computing, particularly from the point of view of the needs of outside users. We were asked to focus particularly on the next two years (FY 1990 and 1991) but to consider how the initial complement of computing will relate to the longer term. The writing and recommendations of the committee will be used by the SSCL staff in planning the SSC computing system and in responding to the DOE requirement of an "Information Technology Resources Plan."

This committee was chosen from the user community of experimenters and theorists interested in the physics potential of the SSC and the detectors required in order to explore physics in a new energy regime. Our recommendations and discussions focus on computing needed to support physics research at the SSC. Substantial computing will also be needed for design and implementation of the accelerator systems, for administrative purposes, and for other uses not in the scope of this committee's work.

The committee meeting approximately followed the agenda given in Appendix A. Informational presentations were made to the committee by M. Gilchriese (Associate Director of the Laboratory for Physics Research) indicating how the laboratory is organized and the position of computing in the organization; by P. Leibold (Director of Computing) on the current status and future plans for computing at the laboratory; by G. Chartrand (Networking Manager) on present and future networking; by L. Cormell (Physics Research) on plans for supporting detector simulation; by R. Hahn (Computing Department) on the long range plan for DOE; and by R. Talman and G. Bourianoff (Accelerator Division) on Accelerator Division-related computing needs. In addition, there were presentations by members of the committee on several aspects of computing at the SSC. Background reports by individual committee members are included in Appendices B through G.

In addition to the presentations, the committee spent about 5 hours in discussion, focussing on a computing plan for the next 2 years. The strong recommendation of the committee is that there is a great need almost immediately for computing cycles for detector design activities, chiefly simulations. These computing cycles cannot be supplied from existing computing in the high energy physics (HEP) community. They must be supplied by a new computing system

which should be located at the SSC. Therefore, the committee recommended that top priority be placed on the acquisition of computing engines. Some amount of support will be needed to make the system usable, particularly to use multiple processors in a single job and to make effective use possible by remote users across the network, but initially all but the necessary minimum of support may have to be sacrificed to the goal of getting the computers in the door. Broader support must, of course, follow as soon as possible.

The remainder of the main section of this report is a summary of the committee's discussions and recommendations.

2. Computing Requirements

The research program of the SSC will, of course, require computing for many different functions and purposes. In early times, these included simulation of physics events and of the response of a detector to them, mechanical design calculations for detectors, use of CAD and CAE systems, and theoretical computing. On-line computing, including all aspects of triggering the detector, will also be important, as will off-line analysis of experimental data and continued use of simulation programs as aids to analysis. Throughout the life cycle of the SSC, computers will be needed for "everyday" activities, such as word processing and editing, program development and small utility calculations, including use of spreadsheet programs or other commercial application programs. Theoretical calculations will be an important use of computing at the SSC, but, as discussed in Appendix F, will generally require modest resources.

During the next few years, all other computing requirements will be dwarfed by the need for detector simulation in support of the design of detectors for the physics program of the SSC. (On-line computing and triggering are areas that require substantial intellectual development in the near term, but do not require production hardware now. Some of the requirements are discussed in Appendix G.)

The committee used the estimated requirement for simulation computing made by the December 1988 Computing Task Force.[1] ,[2] It found a need after about three years for a total of 4000 VAX 11/780 equivalents (MIPS),[3] a number which can be compared with the contemporary estimate of less than 1000 MIPS installed in the entire US HEP program. (This number has undoubtedly risen in the intervening year, particularly as use of high-performance RISC-based workstations is beginning to be seen.)

Along with the computing power, a substantial amount of storage will be needed, both on-line disks and tertiary storage as might be provided by a jukebox

of multiple cassette tapes or optical disks. As discussed in Appendix E, Section 4, the 4000 MIPS of computing will need to be matched by at least 400 GB of on-line disk and 6 TB of tertiary storage accessible with a short delay time.

In order to model the concentration of computing power needed for a single user, we considered three characteristic jobs as discussed in Appendix E, Section 4. The result was the need to devote between 30 and 160 MIPS to a single user while the job is running. Since at least the higher of these numbers is larger than can be supplied by a single RISC microprocessor, which we assume to be the engine for most of the computing cycles, it will be necessary from the beginning to provide the means to share the computing of a single job between multiple processors. Considerations for providing parallel computing are discussed in Appendix B.

3. Model for Computing at the SSC

Recommendation: The SSC Laboratory should be a major resource for computing in the SSC physics program. To this end, SSCL should install, during the next few years, a distributed computing system consisting of computing engines, shared disk storage, central management services, such as file service and batch job scheduling, and workstations for program development and graphical display of results. A high-speed network should be provided to couple the components, which may be located at diverse places at the SSC Laboratory and, particularly for workstations, at other institutes across the US and around the world.

The initiation of computing at the SSC Laboratory marks a departure for computing in HEP in at least two ways. First, detector design and other requirements of the scientific program will quickly demand more computing than exists in the current HEP program by a substantial factor. Thus the scale of the problem is likely to require use of different models or at least modes of computing than the HEP community has employed before. It is a natural time to reevaluate the approach to computing used by the community.

The second consideration is the rapid move of the computing industry away from exclusive reliance on central computing facilities, whether mainframes or minicomputers. Many configurations of computing power are now possible, but increasingly some amount of local computing (in the form of workstations or personal computers) is being devoted to each user for control, display, program development and other functions. This local computer may or may not make

use of remote computers for computing cycles, file access, or other capabilities beyond those available from the local computer.

Both of the above considerations make it appropriate for the SSCL to evaluate the appropriate computing style for its needs without strong constraints of compatibility with previous operating systems used in HEP. It is likely, in fact, that significant differences in approach will be adopted compared to current practice. (The need to continue using much existing software will mean that some compatibility with current systems, in particular the existence of a good FORTRAN 77 compiler, must be maintained.) This potentially abrupt change, however, coupled with the need of a segment of the user community for early use of substantial computing, puts a burden on the SSCL computing staff to provide active leadership and support for the new mode of computing.

The SSCL has the opportunity to set the style of computing for HEP in the 1990s. It has the obligation to make a choice that will maximize user productivity on the new system and to provide the information and tools that will both allow early productive use and smooth user acceptance of a changed style of computing.

The SSCL computing staff outlined to the committee a concept for computing at the laboratory that used distributed hardware. Users would interact through workstations, gaining access to computing engines, storage devices and servers, batch job schedulers, and other service machines, which would all be connected by a high-speed network. Software would emphasize open systems, meaning that UNIX in one or more forms would be the normal base operating system on workstations and servers, and that communications would make use of international networking standards.

The committee in general endorsed the laboratory's plans for an open computing model, but noted several problems with a pure use of open systems. It is attractive to take advantage of the wide implementation of UNIX both in order to have a uniform environment and file system throughout the distributed computing system and to be able to move to different computing platforms based on performance and economic considerations and not on a commitment to a proprietary operating system.

This ideal of uniformity and flexibility is unlikely to be simply realized in practice, for several reasons. First, manufacturers find it necessary to modify UNIX significantly, both adding nonstandard enhancements on top of UNIX and modifying the lower levels to take advantage of unique hardware features. Second, UNIX has not been developed or optimized to serve a broad user base with a strong need for scheduling and allocating shared resources. As discussed in Appendices C and E, scheduling and allocating services will have to be added

to UNIX, either by the laboratory staff or by outside vendors, and may have to be done in ways that depend on the specific hardware chosen at the laboratory.

Graphics is another notorious area where the imperatives of performance routinely force nonstandard implementations which exploit specific hardware capabilities.

Appendices C and E discuss in some detail the combined use of parallel microprocessors and mainframes as the servers of computing and files in the system. In general, the mainframe or the equivalent is needed for high input/output (I/O) bandwidth functions, particularly file service, and for central coordination functions, such as batch-queue management. The cost tradeoffs of different solutions for computing and storage are summarized in Appendix C. It is clear that microprocessors, whether RISC or CISC, cost less per computing unit than mainframes or minicomputers, and furthermore their price is falling faster. For this reason, we expect SSCL to choose microprocessors of the type used in high-performance workstations for its main computing engines. In order to apply the necessary concentration of computing to the larger simulation jobs, it will be necessary to use multiple processors, i.e., parallel processing, as discussed in Appendix B.

4. Computer Acquisition Schedule

Recommendation: SSCL should acquire and make available to users 500 MIPS (VAX 11/780 equivalents) of computing power by October 1990. At the same time, 50 GB of disk storage and 1.5 TB of tertiary storage should be provided. Another 500 MIPS (with corresponding storage) should be provided by March 1991, and a total of 4000 MIPS (with 400 GB and 6 TB of storage, respectively) by March 1992. During FY 1992, a mainframe computer or other high I/O bandwidth centralized devices should be acquired for coordination of high-volume storage and file service, batch job scheduling, and other centralized services.

Much simulation calculation will have to be done for the detector proposals which are estimated to be due roughly at the end of 1991. This plan is consistent with the schedule of computing requirements for detector design suggested by the December 1988 Computing Task Force. Thus computing must be made available to users on an accelerated schedule. Because of funding and effort constraints, this means concentrating on the computing engine and storage parts of the system initially. Operation without the central services to be provided by the mainframe or equivalent will be less convenient and efficient, but it is

by the mainframe or equivalent will be less convenient and efficient, but it is important to get started with computing. We emphasize the importance of a second phase of providing those functions starting in FY 1992, whether literally through acquisition of a mainframe or by acquisition of more specialized servers for these central functions.

The schedule recommended above delivers an initial substantial increment of computing on the fastest possible schedule, given the need to benchmark and otherwise evaluate candidate computers. It is big enough that the computing staff must address multiprocessor issues from the start. Then it provides added computing at a rapid pace up to the required 4000 MIPS level. Of course, initial experience should be used to determine if the same type of processor should be added in the expansion or if a change in direction is desirable.

5. Wide Area Networking

Recommendation: SSCL should vigorously support high-speed networks to allow convenient access to computing and files by remote users. Support of the new style of computing, in addition to the noncomputing uses of networks, such as video conferencing, will require upgrade to T3 (45 Mb/sec) connections between SSC and other major HEP sites by FY 1992, and upgrade of connections to other HEP locations doing significant SSC computing to speeds greater than 56 kb/sec.

Although some guidance on network needs can be taken from the recent report of the Hepnet Review Committee (HRC), [4] it suffers from a) not having considered the SSC's program specifically and b) having excluded remote workstations and related aspects of modern computing that we expect to characterize computing for experiments at the SSC. However, we note that the HRC forecast a bandwidth requirement on the most heavily used link approaching T1 speed by 1991. The calculation for links to SSCL in 1991 will probably show fewer users, but each user requiring more bandwidth for support of remote workstations, leading to the conclusion that the upgrade to T3 speeds will be needed in 1992. (As discussed in Appendix D, it is planned that multiple T1 connections to SSCL will be provided by ESnet early in 1990. This upgrade from the present 56 kb/sec connections is crucial and needed immediately.) While not strictly computing, we also foresee significant use of the networking bandwidth for video conferencing by 1992.

Extrapolation of present uses of workstations suggests the need for dedicated bandwidth of at least 100 kb/sec for each intensive on-line user. We have not

attempted a serious estimate of the numbers of such users, but it seems likely that the number could reach 100 at peak times (in addition to many less intensive users) by 1992, leading to a peak bandwidth requirement of over 10 Mb/sec.

Availability of T3 connections by 1992 matches the upgrade planning of ESnet and NSFnet, who are the principal national suppliers of research networking. These upgrades of the agency networks are far from guaranteed, however. We stress that the bandwidth discussed here will be required for efficient use of the SSC (well in advance of machine turn-on) and should be provided by SSCL if the national networks cannot promise to provide it when it is needed. We also note that the approval and procurement cycle has historically taken 12 to 18 months after an upgrade decision has been made. Thus careful advanced planning is needed along with careful monitoring of the plans of outside network providers.

6. Personnel and User Support

Recommendation: Since SSCL will be forced by its requirements and limited budget to develop a style of computing not presently familiar to many high energy physicists, SSCL should provide adequate personnel to provide a friendly user interface and extensive documentation available to both local and remote users. In particular, it will probably be necessary to provide some local support for parallel computing solutions to concentrating computing power on high priority jobs. In the near term, the committee estimates that at least six people not in the present plan will be needed by October 1990, including two to three high-level systems programmers.

It is most probable that the SSCL will move rapidly to establish computing of a type with which most prospective users are not familiar. The new elements are likely to include the use of UNIX as the standard operating system, emphasis on parallel computing for computer-intensive applications, and extensive use of workstations and specialized server machines to produce a much more distributed computing environment than HEP has used to date. While there is extensive experience with these innovations in the computer science community, their use in HEP is confined to a few pockets. Aggressive support will be needed, in the form of documentation, on-line help facilities, and local software that makes the use of the new facilities as easy as possible. Compounding the problem, this support will be needed most at the beginning, when SSCL will be most concerned with bringing the system up in any operational mode.

A strong team of good people dedicated to making the system useful will be crucial to the success of this project. It will take several capable people focussed

on the needs of users to make this system a facility of general utility to the SSC physics community. Necessary activities will include utility software that hides the system's complexity where possible, documentation including both reference volumes and short, introductory manuals to make it easy to get started, and tutorials and demonstrations at the laboratory.

7. Standard Software Support

Recommendation: SSCL should provide software and documentation support in critical areas, such as graphics, where standards may lag behind the needs of the user community. It may also be necessary to supply extensive documentation for software developed at the SSCL or developed elsewhere, but in wide use at SSCL.

The radical change in computing environment will also eliminate the partial solutions that have been developed under VMS, VM, and other current operating systems for providing standard graphics environments, where commercial solutions have generally either been too expensive or not met the needs of HEP well. SSCL will need to work to replace (and hopefully improve on) provisions of graphics and similar specialized utilities. Possible solutions for SSCL include adoption of specific commercial packages (hopefully with program-wide licensing agreements), new software written for the purpose (by SSCL or by outside vendors), or locally written interfaces to multiple commercial packages. Early attention and planning should be given to these requirements.

8. Support Work by HEP Community

Recommendation: Since the SSCL computing staff will be hard pressed to acquire and make usable a substantial amount of new computing on the required time scale, use should be made where possible of outside groups who can provide needed services.

The committee identified the following areas as particularly suited to cooperative or collaborative effort involving outside groups and the SSCL computing staff: a) benchmarking computers for possible acquisition; b) development of high-priority application software; c) distribution and support of existing codes; d) multiprocessor job schedulers; and e) computer-aided software engineering. All such shared effort is seen as temporary, mostly appropriate to the near term while the SSC staff has not reached full strength.

9. Computing at Local Institutes

Recommendation: SSCL should develop recommendations for remote computing installations that can work well with the installation at SSCL, but should not develop its systems so as to impose a specific approach to computing on local institutes.

While computing at universities and other laboratories is not part of our charge, it seems clear that the optimum program of computing for the SSC program will include significant amounts of computing at local institutes, in addition to the concentration of computing at SSCL. It will be necessary for the SSCL computing staff to develop software and documentation to allow efficient interaction between local and remote computing. We expect that the local institutes will continue to make their own decisions about the type of computing they install. To the extent possible, the laboratory should make it possible for a variety of external computing systems to interact smoothly with the SSCL central computing system.

10. Computer Advisory Committee

Recommendation: A Computer Advisory Committee should be formed as soon as possible to advise the Laboratory Director on the computing needs of the community and computing policy at the laboratory. The committee should be composed of both outside users and SSCL staff members and should be chaired by an outside user.

The Computer Advisory Committee will provide a forum for users of the system to suggest and review the priorities as a function of time.

REFERENCES

1. M. Gilchriese (ed.), "Report on the Task Force on Computing for the Superconducting Super Collider," SSC-N-579, December 1988 (NOT FOR DISTRIBUTION).
2. See also S. Loken, "Proceedings of the Summer Study on High Energy Physics in the 1990s," p. 695, 1989.
3. For simplicity, we use MIPS as a symbol for the computing power of a VAX 11/780, including both integer and floating point performance. In fact, for simulation codes, floating point performance is probably the more important measure.
4. L. E. Price, *et al.*, High Energy Physics Computer Networking, Report of the Hepnet Review Committee, DOE/ER-0372, June 1988.

Computing Planning Committee
Agenda

December 12 (Tuesday)

1300	Welcome – M. Gilchriese
1310	Overview of SSC Computing Status – P. Leibold
1340	Networking at SSCL – G. Chartrand
1400	Simulation Software Goals – L. Cormell
1420	Long Range Plan for DOE – R. Hahn
1440	Overview – H. Newman
1530	Break
1600	Networking – G. Brandenburg
1630	Physics Simulation – F. Paige
1700	Computing for DAQ – A. Lankford
1730	Discussion
1900	Dinner – TBA

December 13 (Wednesday)

0830	Mainframes – L. Cottrell
0910	Software and Multiprocessor Systems – P. Avery
0945	Accelerator – Related Needs – R. Talman, G. Bourianoff
1030	Break
1100	Discussion – L. Price
1200	Working Lunch
1300	Discussion/Writing
1500	Adjourn

12/8/89

Computing Group Meeting

Parallel Computing at the SSCL

PAUL AVERY

January 1990

1. Overview

The particular structure of high energy physics data (independent events) is readily adaptable to a computing strategy in which the data stream, consisting of real or simulated events, is split into many small streams which are analyzed by different computers. This event level parallelism has been proven by a number of groups to be very cost effective, especially because of the advent in recent years of inexpensive but powerful 32-bit microprocessors, (e.g., the new RISC processors). The present generation of experiments (CLEO II, LEP, CDF, D0, etc.) make–or will make–substantial use of parallel processing for standard data reduction, detector simulation and even physics analysis.

Note that event parallelism is quite different from fine-grain parallelism and vectorization. Architectures based on event parallelism can be highly distributed since computers processing different events do not need to communicate with each other very often. The latter two techniques are typically used to solve problems involving large arrays or other repeating structures within a single subroutine or code fragment. They require computers possessing very tightly coupled processors and shared memory, such as is found in supercomputers or mini-supercomputers. High energy physics algorithms are just beginning to take advantage of these capabilities for shower simulations, tracking through detectors, and track finding, but their use is not widespread.

1.1 Need for Parallel computing at the SSCL

At the Superconducting Super Collider Laboratory (SSCL), where the combination of high statistics, event size and event complexity together pose a computing problem one or two orders magnitude larger than experienced today, the need for cost-effective parallel computing is even more acute (see below). Moreover, much of this computing power will be needed well before SSCL startup, since final SSCL experimental designs will need to be ready by the early 1990s, each requiring a massive simulation effort. The scale of these simulations will be greater in quantity and quality than anything previously attempted. This is due in large part to the increased size of SSCL detectors and the complexity and rarity of physics events at the SSCL. However, these simulations will also help determine whether a given design will even work.

1.2 Relation to the SSCL computing problem

It is important to understand the relationship between parallel processing and other important aspects of SSCL computing. In particular, the following general questions must be addressed.

1. What total capability is needed and how much of it should be in the form of parallel processing?
2. What fraction of this capability should be at the SSCL? Should there be regional centers having significant parallel capability?
3. What kind of network links are needed to support parallel computing?
4. Should several different architectures (including vector) be supported? Should a specific operating system (i.e., UNIX) be supported or favored for parallel work? How closely should the SSCL community work with vendors to develop new technology and software? At what level should high energy physics (HEP)-developed systems such as Fermilab ACP be supported?
5. On what time scale should parallel computing be acquired?
6. What problems of scale can be anticipated as parallel computing capabilities go from tens to thousands to hundreds of thousands of MIPS?

In addition, there are questions specific to the SSCL site itself.

7. How should the parallel computing be configured at the SSCL? Should there be several facilities? How closely should they be coupled to mainframes? What kind of input/output (I/O) capabilities (disk, tape, Ethernet, fiber optic) are required?
8. How many people are needed at the SSCL to support parallel computing?
9. What kind of software support should be provided by the SSCL personnel? Should the SSCL get into software standardization?

Obviously the answers to some of these questions require knowledge about conditions two or more years from now. At the same time, the computing needs of users and the available software and hardware are changing so rapidly as to invalidate all but the most general predictions (the computing estimates made at the 1985 Workshop on SSC Computing [1] optimistically assumed that processors having speeds of 8 VAX units with 16 MB of memory would be available in the early 1990s!). In this kind of environment it is probably necessary that the long-term SSCL computing effort be reviewed at least every two years.

2. How Much Parallel Capability is Needed?

To assess the computing needs for the SSCL, we define 1 VAX unit to be the speed of a VAX 11/780 when executing typical HEP codes. Sometimes the word millions of instructions per second (MIPS) is used as a synonym.

Parallel computing is needed for the following activities:

1. Accelerator physics simulations
2. Physics simulations (ISAJET, PYTHIA)
3. Radiation transport (CALOR89, EGS4, GEISHA)
4. Detector simulation (GEANT)
5. Standard data reduction (tracking, shower finding, particle identification)
6. Physics analysis (DST analysis, PAW, IDA)

Activities (2) through (4) are the primary concern of this report since accelerating needs are estimated separately and there will be no data to analyze for quite some time. It is interesting to note in passing, however, that the 1985 Workshop on SSC Computing [1] predicted that each large SSCL experiment will need approximately the off-line computing capability shown below as a function of trigger rate:

Trigger	CPU capacity
1 Hz	10,000 VAX units
10 Hz	100,000 VAX units
100 Hz	1,000,000 VAX units

These figures, which include data reduction, physics analysis and some Monte Carlo, were based on an extrapolation of UA1 made before the widespread use of GEANT for detector simulations and could be an underestimate (the extrapolation predicts a data reduction time of 1200 VAX-seconds for an SSCL event). A million VAX units of scalar processing certainly seems daunting by today's standards, but there are already commercial projects underway which could approach such levels when complete.[2]

The computing capacities needed for the simulations (2) through (4) were estimated [3,4] in 1988 to be about 1,000 VAX units for FY90 and 4,000 through 5,000 VAX units in FY91, and the authors believed that these numbers could easily underplay the true need. When coupled to the short timeline for presenting final designs, they demonstrate the critical importance of establishing a large computational infrastructure (computers, peripherals, networks, personnel and software) at the earliest possible date.

3. Some Recommendations for Parallel Computing

The following factors must be considered in establishing parallel computing guidelines for the SSCL: (1) total capacity needed per year, (2) timeliness in providing resources, (3) networking requirements, (4) utilization of existing resources, (5) distribution and number of personnel, (6) software needs, (7) flexibility of implementation, (8) growth, and (9) cost. We recommend that the following courses of action be taken.

1. Total CPU Capacity

A total of 1,000 VAX units of parallel computing should be obtained for FY90, with an additional 3,000 to 4,000 purchased for FY91. The machines should have memory sizes of at least 16 MB to accommodate the expected large size of SSCL-simulationed events.

2. I/O and Disk Space

Enough disk space should be provided to allow significant data sets to reside on disk (e.g., ISAJET or PYTHIA input files, small GEANT runs, debugging data sets). Based on the experience of current HEP groups (CLEO, D0), about 100 GB of disk are needed to support the 1,000 MIPS activity and perhaps 300 to 400 GB will be needed at the 5,000 MIPS level. This latter number will be better known after some experience with the initial disk storage. The 8-mm tapes should be used for data exchange and enough tape drives should be provided for this purpose. Personnel should pay close attention to integrating the parallel machines with computers on which users develop code, e.g., VAX clusters and mainframes.

3. Networking

High-speed networking should be implemented (at SSCL and other places) as quickly as possible so that (1) users can access significant parallel computing without the delays and slow response times that characterize today's networks, (2) small data sets (several to tens of megabytes) and GEANT geometry files can be transferred, (3) graphics images can be tranported quickly, and (4) software distribution can be improved. These speeds probably require T1 links (1.5 Mb/sec) from the SSCL to a few other major nodes initially, with later upgrades to T3 (45 Mb/sec) at a later time (possibly 2 years).

4. Location of Resources

At least 1/2 to 2/3 of the parallel computing capacity should be concentrated at the SSCL. However, for flexibility in responding to the needs of SSCL collaborations–many of whom will want some local computing to get started–the

possibility should be considered of having some significant computing located outside the SSCL as discussed below.

5. Computing Outside the SSCL

There is some urgency in providing parallel computing quickly enough for groups to carry out their detector simulations. A good way to exploit the good computing infrastructure already in place at several universities, national laboratories and supercomputer sites around the country might be to provide funds to a few (through a proposal mechanism) to carry out simulation tasks for particular collaborations or regions. This funding should be regarded as temporary and its necessity should be studied at each long-range SSCL computing review.

6. Operating Systems

The commercial RISC computers that make today's computing so cost effective, use the UNIX operating system without exception. It therefore makes sense, at least for the near to medium term, to adopt UNIX (in spite of its demonstrated shortcomings) for all SSCL parallel computing. The SSCL should, of course, follow closely developments in the computer industry towards highly parallel computer architectures and determine their suitability for high energy physics.

7. Software

The SSCL should provide good software support for users needing access to parallel processing resources. This support would take the form of making available software to allow simulation codes to run coherently on multiple machines, researching new ideas in parallel processing algorithms and architectures (in collaboration with vendors), providing new simulation codes for use by other SSCL groups, and disseminating HEP standard software to run on a variety of machines.

4. Comments on Localization of Resources

The question should be carefully considered of whether or not all SSCL parallel processing resources be concentrated at the SSCL, at least initially. In a centralized model, users would first develop and debug their simulation programs at their home institutions and then make production runs on the parallel processing facility at the SSCL. The generated events could be mailed back on 8-mm tape or else be analyzed directly at the SSCL and the results (graphical images, etc.) sent back over the network.

From the point of view of management, a centralized facility is certainly easier to administer, and economies of scale and good network connections might make

it work. It is wise, however, to consider the impact such a decision would have on the user community–most of whom have followed the industry trend of having significant local computing–and what resources are needed in the short run for the simulation effort needed to arrive at final detector designs.

For instance, there may be difficulties in the near term because the model presupposes the existence of an established computer center with very fast and reliable network connections, something that may take a while (likely more than a year) to establish. A significant delay in acquiring the needed facilities could adversely affect the simulation program. Also, it would not take advantage of the established computing infrastructure that exists at many universities, national laboratories and supercomputer centers (some of which are also exploring novel parallel architectures and other useful work). Incremental upgrades at these places can take advantage of existing facilities and personnel. Finally, it is a fact of life that simulation codes require a substantial amount of running to debug them, and are constantly changing as the authors try to put in new capabilities. This sort of activity is far easier to perform locally.

References

1. See the report by the Off-line Computing/Networking Group in the "Proceedings of the Workshop on Triggering, Data Acquisition and Computing for High Energy/High Luminosity Hadron-Hadron Colliders," Fermilab, November 11-14, 1985.

2. For example, the Intel Touchstone project, as discussed in Andy Lankford's report.

3. S. Loken, "Computing for High Energy Physics in the 1990s," 1988 SSC Workshop at Snowmass, Snowmass, Colorado.

4. Task Force on Computing for the Superconducting Super Collider, M. Gilchriese (ed.), December 1988.

Thoughts on a Balanced Model for SSC Off-line Computing

LES COTTRELL
ASSISTANT DIRECTOR SLAC COMPUTER SERVICES

January 1990

1. Disclaimer Notice

This document and/or portions of the material and data furnished herewith, was developed under sponsorship of the US Government. Neither the US nor the US DoE, nor the Leland Stanford Junior University, nor their employees, nor their respective contractors, subcontractors, or their employees, makes any warranty, express or implied, or assumes any liability or responsibility for accuracy, completeness or usefulness of any information, apparatus, product or process disclosed, or represents that its use will not infringe privately owned rights. Mention of any product, its manufacturer, or suppliers shall not, nor is it intended to, imply approval, disapproval, or fitness for any particular use. The US and the University at all times retain the right to use and disseminate same for any purpose whatsoever.

2. Notice

There are several trademarks mentioned in this document: Amdahl and UTS are trademarks of Amdahl Corporation; Apple, AppleTalk, AppleShare, and Macintosh are registered trademarks of Apple Computer Inc.; CCC is a registered trademark of Softool Corporation; DEC, VAX, MicroVAX, and VMS are trademarks of Digital Equipment Corp.; IBM, IBM-PC, OS/2 and PS/2 are registered trademarks and AIX, CMS, MVS, Systems Applications Architecture, SAA, VM, and VM/XA are trademarks of International Business Machines Inc.; Imprimis is a trademark of Imprimis Technology, Inc.; Jnet is a trademark of Joiner Associates Inc.; MIPS is a registered trademark of MIPS Computer Systems, Inc.; Multinet is a registered trademark of TGV Inc.; NeXT is a trademark of Next Inc.; Oracle is a trademark of Oracle Corporation; SAS and SAS/GRAPH are registered trademarks of SAS Institute Inc.; StorageTek and Nearline are registered trademarks of Storage Technology Corporation; Sun, NFS and SPARCstation 1 are trademarks of Sun Microsystems Inc.; UltraNetwork is a trademark of Ultra Network Technologies, Inc.; UNIX is a registered trademark of AT&T; X Window System is a trademark of the Massachusetts Institute of Technology; 20/20 is a trademark of Access Technology, Inc.

3. Introduction

This report is in response to a request to look at mainframe computing for the Superconducting Super Collider (SSC) Computer Planning Committee. It attempts to show some of the major trends occurring in cost/performance in the field of computer hardware. It then compares the offerings from the two major mainframe suppliers and briefly discusses the use of supercomputers in high energy physics (HEP). Using the trends shown earlier, it attempts to predict what is possible for the workstations of the future. In the modern era, data storage requirements and management have become a critical challenge in HEP computing so the report next looks at the area of tertiary storage (tapes). A brief view is taken of the possible operating systems in common use in HEP before considering the network possibilities to provide workstation support. Finally a possible model for experimental HEP off-line computing at the Superconducting Super Collider Laboratory (SSCL) is described including the various computing and network components and expected network performance.

4. Trends in Computer Hardware

4.1. CPU Performance Trends

Figure 1 shows the progress of processor performance over the last couple of decades. The mainframe prices are the list prices* of a minimal configuration. The mainframe curve shows that the dollar cost of MIPS† is falling consistently at about 28% per year since the early 1960s, also the residual values of used mainframes are falling at 40 to 45% per year.[1] The minicomputer prices are similarly chosen and their performance curve roughly tracks the mainframe though being slightly lower. The price performance for IBM-PCs, (and more recently) clones, PS/2s, and for Reduced Instruction Set Computer (RISC) workstations since the introduction of the IBM-PC/RT has been falling by 50% per year and is already two orders of magnitude better than for mainframes.

⋆ No attempt has been made to adjust for inflation or for the improved performance and reliability.

† The millions of instructions per second (MIPS) rating used herein is roughly based on 1 VAX 11/780 unit of processing per second (1 VUPS). Roughly speaking for HEP off-line computing an IBM 3081K = 24 MIPS and a DEC 8820 = 12 MIPS.

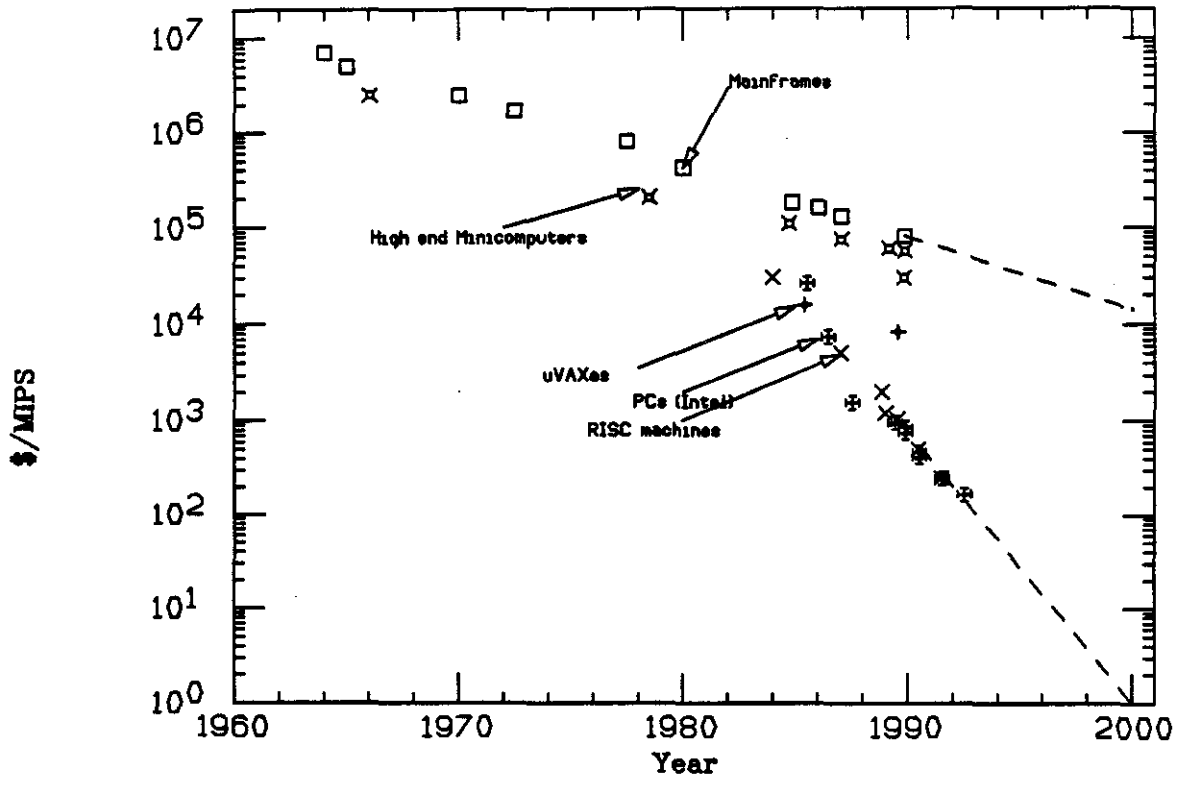


Figure 1. Trends for computer hardware cost performance in US dollars/MIPS for various classes of computers. The points beyond 1989 are industry projections, the dashed lines are straight line extrapolations.

The dramatically better price performance for RISC machines and PCs enables the rapid growth of these items and the reduced emphasis on mainframes for many functions. Such growth is not without attendant problems, particularly in the areas one took for granted on the mainframe. Such areas include system management, archiving, backup, naming conventions, sharing of data, communications, etc. The trick will be to take the best features of both centralized mainframes and distributed systems (e.g., user graphical interface, response time to locally executable commands).

The fact that the RISC and the PC/Intel Complex Instruction Set Computer (CISC) type machines have similar dollars/MIPS performance presumably means that it is not the chip CPU architecture that dictates price performance, rather it is dictated by issues, such as packaging, interconnections, housing, power supplies, documentation, distribution, advertising, vendor competition and marketing. This in turn means other factors should be taken into account when deciding on what type of machine to use. These include the application software availability, the familiarity and training available in schools and colleges, network support, standards support, interfaces and drivers to support new devices, such as optical disks, scanners, FAX machines and so on.

4.2. Memory Performance Trends

Figure 2 shows similar curves for primary (main), secondary (disk), and tertiary (tapes) memory. It can be seen that primary memory prices are falling at roughly the same rates (about 45%) for mainframes, minicomputers and the chips (dynamic RAMs) themselves. The offsets are due to the numbers of units sold, the performance (access time,* access path width and interleaving) and the infrastructure (chips must be mounted on boards, etc.). Increasing complexity of successive generations of dynamic RAMs (going from 1 kb/chip in 1973 to today's 1-Mb and 4-Mb chips) is primarily responsible for cost reductions, though less complex dynamic RAMs continue to decrease in price.

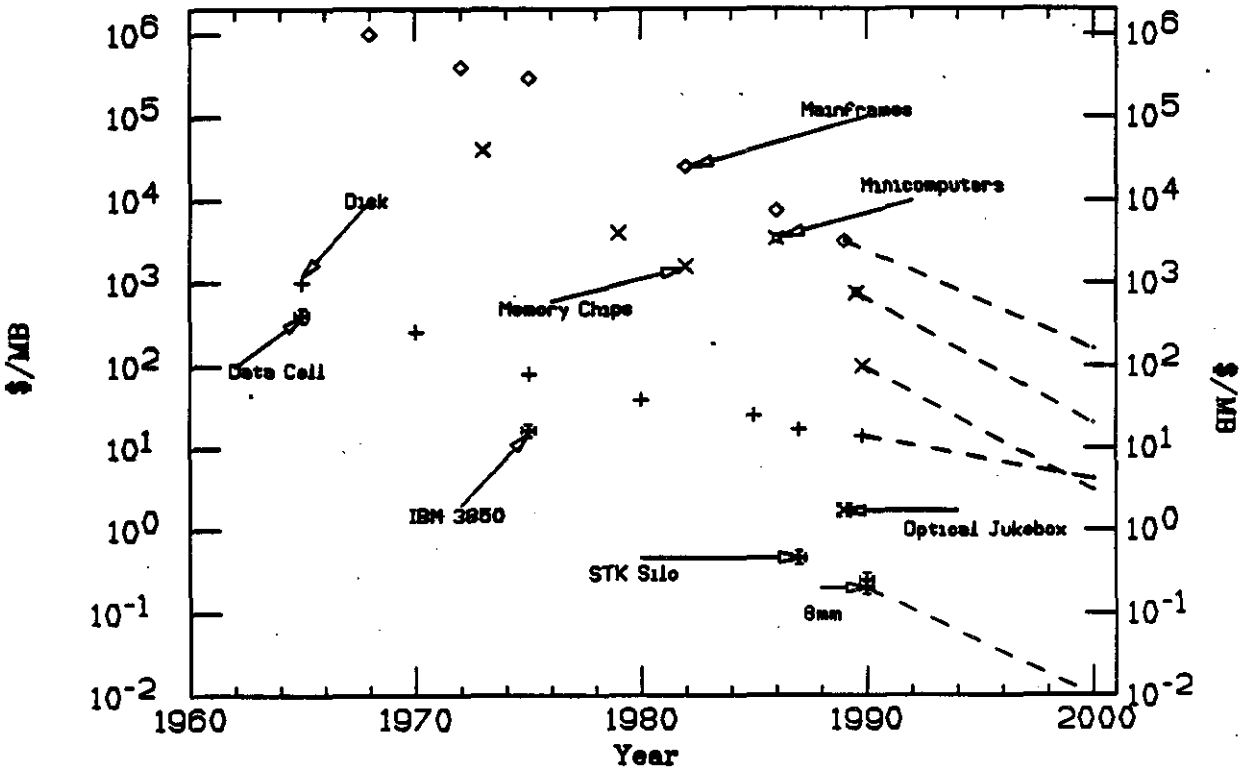


Figure 2. Trends in memory costs in US dollars/MB for various classes of memories.

Figure 2 also shows that magnetic disk storage is decreasing in price at only 20% per year. However, industry demand for on-line disk capacity has been increasing at close to 45% compounded annually. Typically users today have about 40 MB of disk space of their own. At today's prices, disks are still over five times cheaper than dynamic RAMs/MB. However if the slower decrease in disk prices compared to internal memory continues,† it will lead to a reduced dependence on magnetic disk memories for paging, swapping

---

⋆ For example, DRAMs with twice the speed are twice the cost.

† There are indications[2] that gains in this area may increase to 30% per year during the coming decade.

and temporary data storage in favor of RAM disks and larger central memories. Moving the data from disk to primary memory will enable a dramatic reduction in the input/output (I/O) process. Also as the cost of MIPS continues to fall faster than disk costs, there will be more and more cases where it is not necessary to save the intermediate data, rather recalculating it when necessary.

Finally, Figure 2 shows the dollars/MB for several automated tertiary storage devices. These include the IBM 2321 Data Cell, the IBM 3850, today's StorageTek Nearline Automatic Cartridge Store (ACS), and an 8-mm tape jukebox that is projected to be available soon.

4.3. Today's Price Versus Performance

If instead of looking at the trends with time, we look at a snapshot in 1989 of where the various classes of computers stand today in terms of cost for the major components, we get the results shown in Table 1.

Table 1: Representative 1989 prices for various classes of computers.

Function	Main-frame	High-end Mini-computer	Micro-computer	RISC Server	RISC Work-stations	PC/PS Clones
Computer	IBM3090	VAX6000	μVAX	MIPS	MIPS	Intel
Model	600J	410	III	6280	RS2030	386/25
$/MIPS	80,000	60,000	8000	3000	1400	1000
$/MB Primary mem.	3,000	760	250	400	300	100
$/MB Secondary mem.	16	22	22	10	10	5

It can be seen that the major price differences come in the CPU and primary memory components. Some of this is due to the increased performance of the devices one puts in the more powerful computer classes. In the case of primary memory, this includes error correction, cycle times, access path widths, and interleaving. It also has to do with how much one is stretching technology when the product is first announced. Probably the major effect, however, is the number of units sold.

5. IBM and DEC Mainframes

5.1. Comparison of Mainframe Cpu Prices

Table 2 shows the December 1989 list prices for equivalent configurations of IBM 3090J and DEC 9000 mainframes. In order to make comparisons meaningful, we have configured the mainframes with equivalent amounts of memory and similar numbers of channels (for IBM) and hierarchical storage controllers (for DEC). For IBM memory, we have used both central storage and extended storage. Some comments are appropriate on the equivalence of the configurations.

1. The single CPU DEC 9000-410 is rated at about 30 VUPs. SLAC has not benchmarked the CPU to know whether this is accurate for SLAC's HEP computing. DEC's benchmarks show the VAX 9000/IBM 3090-180 = 12.3/10.3 = 1.2 for double precision floating point applications. The IBM 3090-180/3090-180J cycle times are 18.5/14.5 = 1.27. Thus one might expect the VAX 9000/ IBM 3090-180J to be 1.2/1.27 = 0.94 or the VAX 9000 is slightly (6%) less powerful. The IBM 3090-180J cycle time is 14.5 ns, the DEC 9000 CPU cycle time is 16 ns, which is a 10% difference.
2. IBM extended memory is a fully supported integrated architectural component of IBM's mainframe architecture, does not require external I/O to access, and is much closer to central memory in its use and performance than say a RAM disk. The DEC 9000-440 can support up to 512 MB of main memory, the IBM 3090-400J can support up to 512 MB of central storage plus 4096 MB of extended storage.
3. Both IBM channels and DEC Hierarchical Storage Controllers (HSCs) are used to interface devices such as disk and tape to the mainframe. We have used HSCs rather than KDM70s (at $26K each rather than $71K for an HSCs) since, as described later in this report, we expect to share the disks across multiple mainframes. Each IBM channel can connect up to 256 devices and supports an aggregate bandwidth of 4.5 MB/sec. Each DEC HSC can connect up to 32 devices and plugs into a Cluster Controller (CI) which supports an aggregate bandwidth of 4 MB/sec. The IBM 3090-400J can support up to 128 channels, the DEC 9000-440 can support 150 HSCs.

Table 2: List prices for similarly configured IBM and DEC mainframes.

Make	Model and What It Contains	$K	MIPS
DEC	9000-410 with 256 MB (base = $1690K) + 2 CIs[α] (@ $36K each)		
	+ 16 HSCs (@ $71K each)	2898	30
IBM	3090-180J with 32-MB central storage		
	+ 16 channels (base = $2572K)		
	+ 256-MB extended storage ($885K)	4086	32
	+ 3092 ($278.6K) + 3097 ($111K) + 3089 ($39.9K)		
DEC	9000-420 with 256 MB (base = $2220K) + 4 CIs (@ $36K each)		
	+ 32 HSCs (@ $71K each)	5636	59[β]
IBM	3090-200J with 64-MB central storage		
	+ 32 channels (base = $4711K)		
	+ 192-MB extended storage ($700K)	5991	63[γ]
	+ 3092 ($278.6K) + 2 * 3097 ($222K) + 2 * 3089 ($79.8K)		
DEC	9000-440 with 512 MB (base = $3920K) +8 CIs (@ $79.8K each)		
	+ 64 HSCs (@ $71K each)	8742	117
IBM	3090-400J + 128-MB central storage		
	+ 64 channels (base = $6554K)		
	+ 512-MB extended storage ($1625K)	8839	123
	+ 3092 ($278.6K) + 2 * 3097 ($222K) + 4 * 3089 ($159.6K)		
DEC	There is no six processor DEC mainframe	N/A	N/A
IBM	3090-600J with 128-MB central storage		
	+ 64 channels (base = $12314K)		
	+ 512-MB extended storage ($1625K)	14599	180
	+ 3092 ($278.6K) + 2 * 3097 ($222K) + 4 * 3089 ($159.6K)		

[α]The CI performance of 4 to 6 MB/sec is probably a bottleneck to servicing the disks on the HSCs so more CIs may need to be added.*

[β]The performance of multi-CPU DEC 9000s comes from [3].

[γ]The performance of multi-CPU IBM 3090s is presumed to be similar to that of DEC 9000s.

* Depending on the number of disks to be supported and the number of paths required to the disks from the cluster, a more reasonable configuration might decrease the number of HSCs and increase the bandwidth by increasing the number of CIs.

It can be seen that price-wise the DEC and IBM CPUs are similar. The maximum cluster of DEC 9000s offering balanced I/O throughput would be four DEC 9000-440s or about 460 MIPS. The maximum such IBM cluster would be four IBM 3090-600Js or about 720 MIPS.

It is hard to know what discounts to expect. An IBM plug compatible machine (PCM) may have an advantage since there are three vendors [IBM, Amdahl, and HDS (formerly NAS)] competing for the market place, but only DEC makes a DEC/VAX mainframe.

Annual hardware maintenance for IBM mainframes and disks is about 2% of the list price. For DEC mainframes, it is about 35% more or 2.7% of the list price.

5.2. Comparison of Mainframe Disk Storage Prices

Table 3 compares in some detail the cost of disk drives from IBM and DEC. In both cases, we have taken high performance drives currently available from both manufacturers. Also included is a new high performance disk from Imprimis targeted at smaller machines. These prices include the IBM (3990) and DEC (HSC5X) interface costs.

Table 3: Comparisons of the top of the line IBM and DEC disks.

	3380 Std.	3380 E	3380 K	3390 Mod1	3390 Mod2	RA82	RA90	Imprimis Elite
GB/Device (volume)	0.63	1.26	1.89	0.94	1.89	0.855	1.6	1.5
(formatted)	0.53		1.59		1.64	0.622	1.21	1.3
Data Rate (MB/s)	3.0	3.0	3.0	4.2	4.2	2.4	2.8	3.0
Latency (ms)	8.3	8.3	8.3	7.1	7.1	8.3	8.3	5.56
Average seek (ms)	16	17	16	9.5	12.5	24	18.5	12
Formatted GB/sq ft.	0.2	0.4	0.6		2.0	0.45	1.7	
List Cost ($K/GB)	2.5*	7.7*	10.8*		14.5	16*	19.6	3
(incl. Interface)	2.6*	7.8*	11*		16.2	19*	22	4
Announced	1980	1985	1987	1989	1989	1988	1989	1990

*The costs for the pre-1989 announced disks are expected secondhand prices.

It is evident that the DEC drives are more expensive and have lower performance.† It is apparent that one gets good deals in the used disk market, with prices at 60% of original list within 3 years of

† A 1988 DECUS survey of large systems users resulted in an average VMS utilization of disk space at 75%. This compares well to SLAC's VM utilization of around 66%. Though this factor should also go into an exact calculation of relative GB/dollar prices, IBM's Shared File System (SFS) available in a year for VM/XA (well before the SSCL needs to make a decision) should reduce this disk utilization difference.

announcement and down to 5 to 10% of list in 7 years. Even on new disks much better prices can be obtained by competitive bidding since there are multiple vendors selling the disks, for example in a recent SLAC Request For Proposal (RFP) the bids quoted prices of better than 50% of list.

5.3. Mainframe Software Costs

In order to compare software prices, we will take the costs for software for the two mainframe software systems currently in use for interactive and batch use in HEP. These are the IBM VM system and the DEC VMS system. Though the IBM MVS system is in use at DESY, its use is the exception in the HEP community due to VM's improved ease of use, application development, interactive timesharing, and reduced need for resources. VM's main problem for HEP use, compared to MVS, has been its lack of a production batch system, but that has been largely solved with the introduction of the SLAC-developed SLACBATCH system.

IBM 3090/VM software can either be purchased on a one-time basis or it can be paid for on a monthly basis in which case it is typically owned after 48 months. DEC 9000/VMS software usually includes an initial license fee, a one-time distribution media cost, and then monthly software maintenance costs after the first year. IBM separately prices the operating system from the initial hardware cost. DEC, on the other hand, bundles the operating system and several utilities (e.g., DECnet, assembler, sorting) into the purchase cost of the VAX 9000.

To compare IBM 3090/VM with DEC 9000/VMS software costs, we will use the cost of ownership for a 5-year time frame. The products that were included in the comparison are: the operating system VM/XA versus VMS; the following languages: Assembler, FORTRAN, PL/I, C (Waterloo C versus VAX-C); ISPF under VM (required for the FORTRAN interactive debugger); Real Time Monitor/XAMAP/XAMON versus SPM; DIRMAINT versus Authorize bundled with VMS; RSCS and PassThru versus Jnet and DECnet; CCC versus CMS/MMS/DTM; SyncSort versus SORT bundled under VMS; XMENU versus FMS; Waterloo Script versus RunOff bundled under VMS; 20/20; SAS and SAS/GRAPH; SQL/DS and QMF versus RdB; IBM TC/IP versus Multinet TCP/IP; and a user-driven archive system. The prices are for an IBM 3090-180 and a *DEC VAX 9000-410*.

Not included are products that do not exist on both platforms. In particular, these include a production batch system, data staging, and support[‡] for a StorageTek ACS which do not exist on VMS, and a language sensitive editor which does not exist on VM. We have also not included a relational database system that will run on multiple vendor operating systems and hardware (e.g., Oracle).

With the above assumptions, the 5-year costs are about $635K for IBM/VM/XA and $580K for DEC/VMS.

‡ About $450K for five years for software.

5.4. Summary of Relative Strength of DEC/VMS and IBM/VM

Hardware IBM still has larger symmetric multiprocessing mainframes (six versus four CPUs), cheaper disk space that occupies less floor space, cheaper maintenance, and more competition from PCM vendors. They have been building the high-performance hardware longer (so there is more cheap secondhand equipment), they have access to ACSs, and they support the emerging standard ANSI X3T9.3 100-MB/sec High Performance Processor Interconnect (HPPI) between CPUs.

DEC, on the other hand, does not require watercooling, the mainframes require less floor space, has more advanced clustering, and access to more exotic devices, such as CAMAC, Fastbus, and optical jukeboxes.

Software IBM comes out ahead in VM's support for multiple guest operating systems, which is useful for testing new operating systems or new releases. It is more powerful than dedicating a member of a DEC cluster to such functions (presuming there are multiple VAXes clustered together) and provides more powerful tools to assist in testing and installing new operating system releases and the dependent applications. This leads to less impact on users in the form of reduced length of outage for the cutover and almost eliminates dedicated systems time. IBM is also ahead in tape management, archiving, backup, runs the SLACBATCH/HEPVM production batch services (including job allocation, scheduling, prioritizing, accounting, together with scheduling of scarce resources, such as tape and cartridge drives and setup support) software. IBM also has the REXX command language which is superior to the DCL command language.

DEC/VMS is more consistent and requires less tailoring by the customer, it runs on a wider range of hardware[§] (from a 1-MIPS μVAXstation to a 117-MIPS VAX 9000-440) and provides better user access to multi-tasking. It is also better in the area of application software development. In particular this includes the language sensitive editor; support for long file names; hierarchical file directories; better granularity of file access protection; cross-language system service calls, library support and full-screen cross-language language-sensitive debugger; and a better code management system.

Networking DECnet is much more prevalent in the HEP community than IBM's System Network Architecture, and DECnet is better integrated into the VMS operating system than TCP/IP. There is at least one implementation of DECnet phase IV for VM so this problem can be alleviated though at some extra cost (on the order of $100K). There are also indications that DEC may be ahead of IBM in providing integrated support for the ISO/OSI standards.

Caveat It should be borne in mind that both vendors and second-party vendors are trying to fix the deficiencies. DEC is getting into the mainframe market in a big way and must have projects to address the deficiencies, such as connectivity to ACSs, develop larger mainframes, add support for production batch services, etc. At the same time, IBM is developing its Systems Applications Architecture (SAA) which will address many of the applications development environment criticisms.

If one goes to UNIX as opposed to VM or VMS, then the differences are much more cloudy in the software area (see Section 9, Operating Systems).

§ IBM does market a VM workstation and the IBM 9370 series can also run VM. However they will not run VM/XA, nor are these machines in widespread use in HEP.

6. Vector Supercomputers

The main advantage of vector supercomputers is their ability to execute vectorizable code at great speed. Unfortunately, despite considerable efforts, there has been only limited success in vectorizing experimental high energy physics codes. With CPU cycles to perform Monte Carlo calculations being a major requirement for HEP these days, much of the effort is in trying to vectorize Monte Carlo codes, such as the detector simulation program GEANT. As a recent CERN/IBM publication[4] says, "Much more effort is required to determine the Monte Carlo computational nuclei; the final goal, a Monte Carlo Subroutine Library, is far from being complete." Even then, initially at least, the library will only exist for a single architecture, hence reducing one's flexibility in future purchases. The increase in speed that CERN/IBM[5] are expecting for GEANT is a factor 1.5 to 2.0 on average on an IBM 3090VF. We suspect some sizable fraction of this improvement comes simply from rewriting the code, and thus will also be in the scalar version.

The way MIPS are coming down in cost on scalar high-power RISC machines (see Figure 1),¶ we suspect it would be better to learn how to utilize these rather than concentrate the intellectual energy on vectorizing. We think that the vectorizing code should be a background activity and not drive any procurements. Basically it is not a big ticket item to add a vector unit to an IBM 3090 or a DEC/VAX 9000, if people have the intellectual time and desire to tackle this.

7. Workstations

It is obvious that workstations will play an increasing role in HEP computing in the future. Where today there is a personal computer or dumb terminal for each person at a laboratory, tomorrow there will be a workstation or X windows terminal for each person. The choice between X windows terminals and low-end workstations is not currently clear. The decisions will probably be based on support issues, bandwidth requirements, and host impacts.

Using the data in Figures 1 and 2, one can make some predictions as to what the workstations of the future may look like. For this we have taken as base systems, a low-end workstation (actually an Apple Macintosh II CX with a monochrome monitor) costing about $5K (the current DoE limit for operating versus equipment money) at the local bookstore and a higher end one (a Silicon Graphics Iris 4D/25 Turbo) costing about $35K with typical discounts. If we extrapolate according to the trends shown in Figures 1 and 2, we can predict what one may get for similar prices in the years 1995 and 2000.* The results are shown in Table 4.

¶ We would guess that supercomputers probably have a similar slope to the mainframes in Figure 1.

* We have assumed constant monitor prices (but the monitors will be better, higher resolution, more colors, less desk space, etc.).

Table 4: Price projections for workstations that store most of their data locally.

	1989 Low Cost	1995 Low Cost	2000 Low Cost	1989 Higher Performance	1995 Higher Performance	2000 Higher Performance
MIPS	2.5	25	60	16	96	300
MFLOPS	0.1	1	2.5	1.6	9.6	24
Primary Store	4 MB	40 MB	100 MB	10 MB	60 MB	150 MB
Secondary Store	40 MB	400 MB	1 GB	500 MB	3 GB	7.5 GB
Price	$5K	$5K	$5K	$35K	$35K	$35K

The configurations for the 1989 workstations are typical of workstations which make only limited use in sharing data over the network. In particular, they are configured with sufficient secondary storage to be fairly autonomous. Simply scaling the 1989 configurations to 1995 and 2000 as is done in Table 4, however, results in workstations whose major cost component is in the secondary storage. For example, over three-fifths of the cost of the workstations in 1995 is in the secondary storage, and by 2000 this has risen to over four-fifths.

Thus there will be pressure to reduce the amount of secondary storage on individual workstations and hence allow an increase in the money put into the other components. This in turn will increase the demand to store data elsewhere, sharing it over networks using file servers and distributed databases. One has to be careful not to go too far in this direction due to the impact on network traffic. In fact, disk-less workstations are already notorious devices to have as neighbors on a network. In Table 5, therefore, we have configured the workstations to have a more modest amount of secondary storage. This will be used to store the operating system, the major applications, and data that will never need to be shared. We have not projected out to the year 2000 in Table 5 since we do not feel comfortable with extrapolating the current styles of work or the current technology trends that far.

Table 5: Price projections for workstations that access most of their data from network file servers. The prices do not include the network connection costs.

	1989 Low Cost	1995 Low Cost	1989 Higher Performance	1995 Higher Performance
MIPS	2.5	50	16	480
MFLOPS	0.1	2	1.6	48
Primary Store	4 MB	80 MB	10 MB	300 MB
Secondary Store	40 MB	200 MB	500 MB	500 MB
Price	$5K	$5K	$35K	$35K

Comparing Tables 4 and 5, it can be seen that one can more than double the performance of the workstation itself by storing the bulk of the data elsewhere on shared disks and tertiary storage. There is no free lunch of course since the data will have to be stored elsewhere on file servers which cost money and utilize network resources to provide the services. The trick will be to ensure that the gains are not for naught due to network bottlenecks, that there are not unnecessary multiple copies of the data, that disk space is not wasted due to fragmentation by residing unnecessarily on separate workstations, etc.

8. Tertiary Storage

The increased emphasis on file servers and distributed databases means that we must pay careful attention to how we are going to store data and make it accessible at minimal cost. We also need to look carefully at using tertiary storage to minimize the requirements for relatively expensive disk storage.

There are three main types of secondary and tertiary storage in use today: magnetic disks, magnetic tapes and, more recently, optical disks. Since people costs* are going up while computer costs are decreasing, it is important to be able to access data from secondary and tertiary storage without requiring human intervention. Table 6 shows some typical characteristics of various automated (i.e., the data is accessible without human intervention) storage media. The optical disk parameters are for a DEC RV64 optical jukebox. The 3480 tape parameters are for an StorageTek Nearline ACS. The 8-mm tape parameters are for an Exabyte jukebox, that is rumored to be available soon.

* Today industry figures it costs about $2 per manual tape mount.

Table 6: Characteristics of various storage media. The cost/GB is for a complete system. The prices are list prices.

	Optical Disk	3480 Tape	3390 Disk	8 mm	6250 Tape
Burst read speed	1.33 MB/s	4.5 MB/s	4.2 MB/s	?	0.8-1.2MB/s
Sustained read	262 kb/s	3.5 MB/s	1.4-3.5 MB	500 kb/s	500-800 kb/s
Write speed	262 kb/s	4.5 MB/s	4.2 MB/s	500 kb/s	0.8-1.2 MB/s
Seek time	0.15 sec	17 sec	0.012 sec	?	2 min
Mount method	pick	silo	—	jukebox	ATL robot
Mount time	10 sec	11 sec	—	?	45 sec
Media lifetime	30 yr	$\leq$10 yr	—	?	$\leq$10 yr
Media cost	\$200/GB	\$35-\$70/GB	—	\$4.5/GB	\$66/GB
Capacity/volume	2 GB	0.2-0.32 GB	—	5 GB	0.15 MB
Capacity/box	128 GB	1200-2000 GB	52 GB	500 GB	$\leq$450 GB
Cost of box	\$200-\$300K	\$350-\$500K	\$760K	\$100K	\$0.25-\$1M
Cost/GB	\$2K/GB	\$0.3K/GB	\$15K/GB	\$0.2K/GB	\$2K/GB
Storage size	200 sq ft/TB	100 sq ft/TB	500 sq ft/TB	?	1200 sq ft/TB

It is apparent that 6250-BPI tapes are no longer competitive in this area. At the moment, the 3480 ACS appears to have the edge over optical jukeboxes. The ACS is cheaper per gigabyte, requires less space per gigabyte, can support larger capacities (one ACS can support up to 20 TB today), and has faster read and write speeds. The optical jukebox has an edge if the data is very sparse and stored over very many volumes. Optical disks also store 3 to 10 times more data per volume than a 3480 cartridge, and yet like a 3480 cartridge the optical disks are removable and can be carried to another system. However, unlike 3480 cartridges which have a standardized format, most company's optical drives cannot read data written by another company's drive. Attention should be paid to this area in the future since the optical drive technology is making great strides and is very attractive in the workstation and PC marketplace.

The 8-mm jukebox is a new player in this market. It looks very attractive on paper. There are, however, still some unresolved questions, such as how does it interface to a mainframe, what is the reliability of the robotics and the recording. Also the largest system is rather small by HEP requirements and it is doubtful if volumes can be moved from one jukebox to another without manual intervention. Not the least problem however, will be the software to support the device. It has to worry about error recovery (cannot find volume, something already in supposedly empty slot, my arm fell off, etc.), optimizing the arm movement, integrating into the operating system and the batch and device scheduling, etc. For example, the StorageTek silo support code under VM is comparable in size to the VM/CMS operating system itself; there are over 800 modules, it runs in a 5-MB virtual machine, it requires over 20 MB of disk space to store the code, its

executing code space is about 2 MB, there are over 170K lines of code and an estimated 40 to 50 man-years of effort involved. In addition the extra effort to transport the software from MVS to VM was 45K lines of code and 8 man-years of effort. The effort to integrate it into HEPVM was about 6 man-months.

9. Operating Systems

Operating system choice has a larger impact than hardware choice in the long term. This is due to the more stretched out development cycle and lifetime of the operating system and the infrastructure that gets built on top of it, including items like user training, applications, etc. There are several operating systems in common use in HEP today. The most important are VMS, VM, UNIX, PC/DOS and the Macintosh operating system.

Outside the HEP community, PC/DOS is by more than an order of magnitude more popular (in terms of number of units in place) than UNIX, VM and VMS all lumped together. However neither PC/DOS (and its successor OS/2)* nor the Macintosh operating system run on mainframes as a native operating system. They are also proprietary to IBM/Microsoft and Apple, and the source code is not available. Also these operating systems are not usually heavily used for HEP outside simple applications like word processing, spreadsheets, and foil preparation. UNIX (AIX) is generally recommended by IBM sales representatives for IBM PCs, PS/2s and clones if the environment already has UNIX, is a heterogeneous environment, and/or computer-intensive price/performance is required.

It would be nice to have a single operating system that spans all the machines from low-end workstations to mainframes. Today the only operating systems that can make such a claim are VMS and UNIX. VMS is well documented, has a consistent command structure, has well publicized application program interfaces, and seldom needs changing by the customer. VMS, however, is proprietary and does not run on today's hottest workstations.

UNIX, on the other hand, is non-proprietary, and it runs on virtually all manufacturers' hardware. In particular, it will run on RISC, Intel (PCs), Macintoshes and NeXT machines, and mainframes, and the source code is generally available. However neither Ultrix (DEC's UNIX) nor AIX (IBM's UNIX) will run as a single system image on the multiprocessor mainframes from DEC or IBM at the moment. Hopefully this will be fixed in the next couple of years. Amdahl's UNIX (UTS) does support multiprocessor PCMs in a single system image mode. There are several other features currently lacking in mainframe UNIX operating systems that will need to be addressed before it can take the place of something like HEPVM. These include (see also[6]): a batch job scheduler (such as SLACBATCH); remote job entry; checkpoint/restart capability; support of an automated cartridge store, full tape volume label support; accounting and reporting; files should be able to span volumes, a user driven archiving system, a commitment to support future storage

* Microsoft is reported to be working on a portable version of OS/2 which could become the standard for desk-top computing in the 1990s. In such a case, there would be a case to be made for running OS/2 on the mainframe (e.g., as a guest under VM).

devices, data integrity and hierarchical storage management features such as VMSTAGE,† FATMEN,[7] IBM's System Managed Storage;‡ and multiple levels of privilege to avoid superuser password proliferation. Hopefully these issues will be addressed in future releases of Ultrix, UTS and AIX.

Another problem is that not all UNIXes are identical. In order to make the operating system useful, manufacturers add their own features (for example, Amdahl has added over 1M lines of code to the regular UNIX to create UTS). Further, there are two regular UNIXes emerging, the one being developed by AT&T and Sun and the one being developed by the Open System Foundation (OSF). Unfortunately, it is looking as if HEP may have to choose both since the AT&T/Sun version appears to be winning for RISC workstations, whereas both DEC and IBM are members of OSF.

HEPVM solves many of the problems that UNIX and to a lesser extent VMS (e.g., no native support for an ACS, no production batch system, no tape-to-disk staging support, etc.) have for mainframes. However it is proprietary (IBM), the CMS component of VM is not multi-tasking which makes it a poor match for supporting an X-windows client, and its file system is rather limited in ways that affect its use as a distributed file system. These include: limited length file names; no file hierarchies; no read, write or execute protection at the record or even file level (only at the minidisk level and even then no execute protection), and no multi-write sharing. Some of these issues are being addressed in the new VM SFS which provides file sharing, hierarchical file directories, and improved disk space utilization. Even then, however, VM does not run on workstations, PCs or Macintoshes.

As the operating systems become more and more hidden from the user by windowing systems, distributed database access, client/server network paradigms, applications that span or run on multiple platforms, data format conversion tools to allow applications to exchange data, etc., it may be that the user will not have to worry as much about the underlying operating systems on the various machines being different. What will become more important is that users will be able to access functions and data while staying within their own familiar environment. For some time, however, there will still be a need for systems specialists to configure and customize the workstations and the various servers.

† The effectiveness of using VMSTAGE and an ACS at SLAC can be gauged by the fact that the number of manual tape mounts would be a factor of 10 greater but for these two tools. Staffing to support such a tape mounting load would cost roughly an extra $500K/year.

‡ Without such facilities historically it has taken one person to manage each 10 to 15 GB of on-line storage.

10. Network Support

10.1. High-End Workstations

It is expected that some high-end workstations will require file access rates exceeding those available from today's Ethernet technologies (less than about 500 kB/sec). Such workstations can be directly connected to 100-Mb/sec Fiber Distributed Data Interface (FDDI) networks, which can support 1 MB/sec today for a single workstation. If this is inadequate, they can instead be connected to 100-MB/sec HPPI-type network. Such networks can today support better than 4 MB/sec to the memory of VME-based workstations.

10.2. Low-End Workstations

It is expected that, for some time to come, low-end workstations (< $10K RISC workstations, PCs and Macintoshes) will not be able to afford expensive ($12K today) direct connections to high-speed FDDI or HPPI-type networks. For such workstations, there exist low-cost Ethernet connections which can support file transfer and access rates of a few hundred kilobytes per second. It is expected that for sometime such rates will be acceptable for such workstations.

File transfer rates for Ethernet-connected computers are expected to be between 50 kB/sec (for a Macintosh II) and 500 kB (for a Sun SPARCstation 1 in binary mode) depending on the computer, the Ethernet loading, the type of file transfer, etc. Workstations which are connected via AppleTalk-type networks appear to support file transfer rates of 10 to 20 kB/sec.

In order to sustain such performance to multiple workstations simultaneously, we need to address the issues of dividing up the Ethernets so traffic is to a large extent local to the individual Ethernets. This is done by separating the Ethernets by means of routers into subnets. These routers can, in turn, be connected to an FDDI backbone so that there are only two hops between any two subnets. Each subnet supports groups of users with similar data accessibility and communications needs. In this case instead of all traffic being seen on all Ethernets, most traffic is localized to the subnet and does not cross the router boundaries. Thus the aggregate traffic carried by all the subnets is increased over what can be carried by a single (non-subnetted) Ethernet.

The next step is to provide each of the subnets with individual file servers so that no one file server is overloaded and so that to first order, at least, file server traffic does not have to cross subnet boundaries. The master copy of the physics data is kept on the mainframe (on the disk farm and ACS). Conceptually what we need to do is to provide copies of at least some parts of this data. Rather than make physical copies we can use the disk sharing ability and clustering of multiple mainframes to provide multiple file servers. Thus connected to the large master mainframe by channel-to-channel (CTC) connectors will be one or more similar or smaller mainframes sharing the master mainframe disks. These secondary mainframes will be equipped with Ethernet interfaces and can hence act as file servers. The impact on the master mainframe should be minimized since the data does not have to be copied to the file servers, the file servers can offload the CPU load required to do the protocol handling and I/O, and control information and inter-mainframe communications can be passed via the CTCs.

The number of Ethernet interfaces that a single smaller mainframe can support will probably be limited by the CPU load[8] imposed by the protocol loading and the aggregate performance one can tolerate. Probably a single IBM 3081K class machine running VM/XA or a DEC VAX 6000 running VMS/Multinet could support two Ethernet interfaces with aggregate transfer requirements of 1 MB/sec or more with more limited requirements.

Even if the larger mainframe does not support UNIX, these smaller mainframes could support a UNIX-type file system by running UNIX on them (since they do not have to support the production computing of the larger mainframes). This might provide some performance improvement and also may enhance the ease of use as seen by the workstations since the file server environment will be more similar to that of a workstation running UNIX. However two copies of some of the data may be required in this case, one in the format of the larger mainframes operating system (VM or VMS) and one in UNIX format.

10.3. Other Network Servers

In addition to the centralized file servers given above, there will be many more specialized file servers. These will support logical groups of users. For example, there might be AppleShare file servers for Macintosh users to provide copies of the latest application and system software and mail servers, or a mechanical engineering group may have a server for CAD/CAM data.

There will also need to be other network services provided such as name servers and authentication servers. Some of these will be highly critical (i.e., users will be severely impacted if they do not exist) and so will need redundancy to provide reliability.

11. Possible Model for Computing

A straw-man model of computing for the SSCL that one might build towards is shown in Figure 3. This model attempts to address most of the concerns and opportunities mentioned above. In particular it attempts to take advantage of the increasingly low cost per MIPS for RISC machines; the improved information viewing tools for workstations; and the existing mainframe support for shared disks, hierarchical storage management, access to high performance cartridge and disk drives, and access to automated high performance tertiary storage devices. There are several major computing components shown in Figure 3. These include existing workstations, newer high-speed workstations, farms of computer servers, distributed group file servers, and the centrally managed mainframe data server supporting shared disk farm access and a large on-line tertiary storage. These are glued together by networks of various performance and costs.

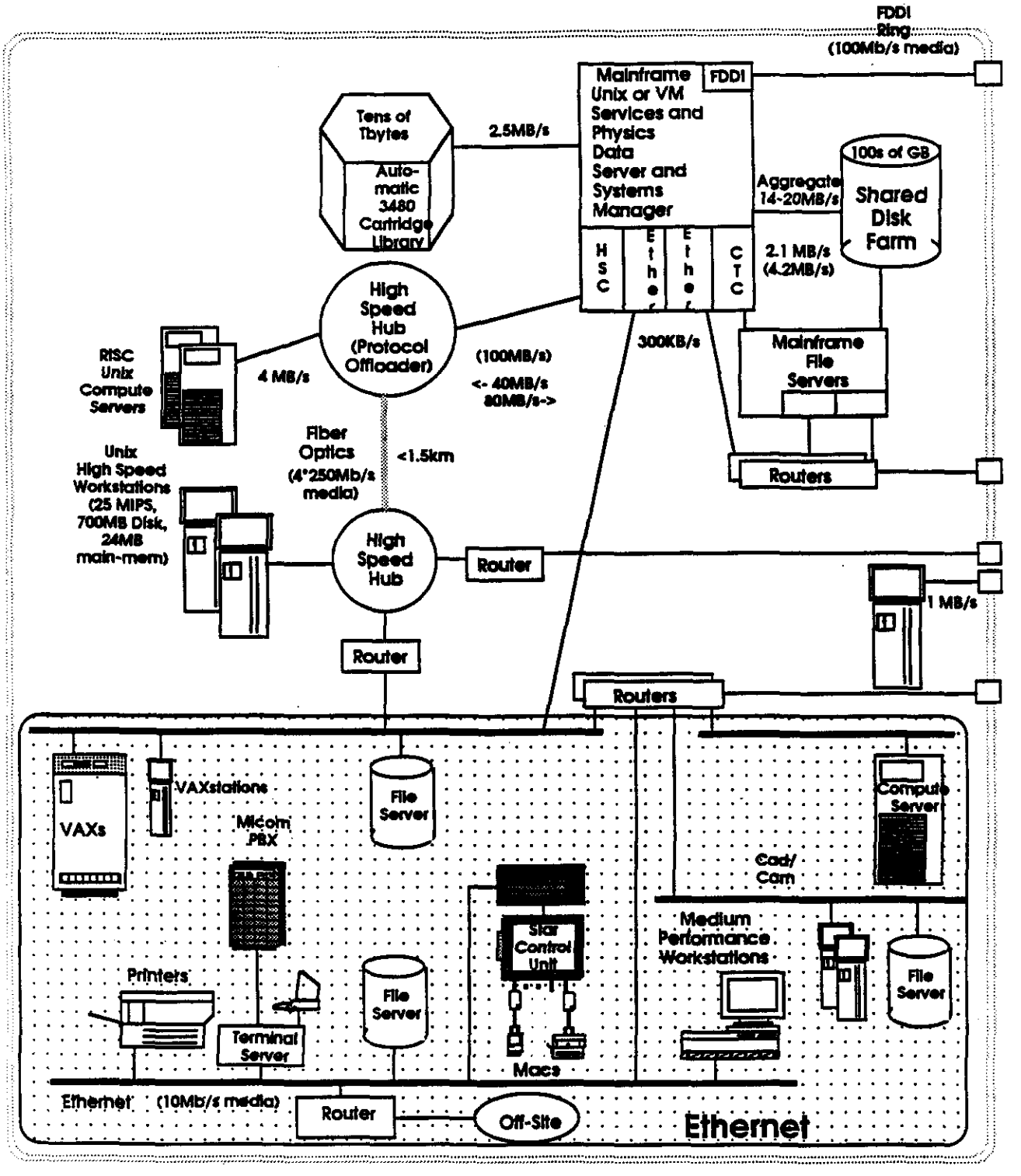


Figure 3. A straw man model for SSC off-line computing in the 1992 time frame. The numbers in parentheses are the media bandwidths, the rates not in parentheses are the expected performances for data transfer.

11.1. WORKSTATIONS

These exist today and are a major component of any future computing strategy. They need to be centrally planned and supported. By this we mean they need to be networked together, at any given time a limited set of devices, configurations and application software should be recommended, and support and coordination provided. They will hopefully run the same version of UNIX and support many common applications. They will span the range of low-performance workstations (today's PC and Macintoshes) to high-performance graphics engines, probably a ratio of around 10 low-performance workstations being bought for each high-performance workstation, though this ratio will vary by group. There will always be more of the former than the latter and the boundaries between the two types will move with time, the high-performance workstation of today ago will be a medium-to-low performance workstation of 2 to 3 years time. This rapid depreciation and obsolescence of technology will be an issue needing careful consideration.

11.2. COMPUTER SERVERS

Due to the increasingly attractive price performance of commercial RISC-based computers, it makes sense to use these as computer server farms. In order to reduce the management and pathological problems, it is probably advisable to use a few high-power computer servers rather than many low-power computer servers. Using idle workstations to provide computer cycles is questionable due to the social problems, the increased management problems, and the increased coupling of the specifications of the workstations and computer servers: Work will need to be done to extend the current HEP batch production system to provide distributed computer server support. Fermilab has done extensive work in this area.

Assuming that the typical SSCL Monte Carlo detector simulation takes $5 * 10^3$ MIPS/sec and generates a 2-MB event, we can use Ethernet-connected computer servers of 100 MIPS each with average file transfer rates of 2 MB * 100 MIPS/5000 MIPS/sec = 40 kB/sec. This average rate will not stress the Ethernet. Presumably the machine can be working on the next event while transferring the just analyzed event so it is not dead during the 2 MB/(40 kB/sec) = 50 sec that the file transfer is going on. In order to provide some margin of safety between the time taken to generate the event and the time to analyze it, a faster file transfer rate is desirable, say 100 kB/sec, so that the transfers will take on average 25 secs and the generation 5000 MIPS/sec/100 MIPS = 50 secs each. So every 50 secs, each farm machine will need 100 kB/sec of the Ethernet, and assuming a dedicated Ethernet can support 500-kB/sec aggregate we can support 50 sec (interval)/25 sec (transfer time) * (500 kB/sec)/(100 kB/sec) = 10 such farm machines in this fashion.

11.3. DATA SERVERS

The need to store yearly several tens of terabytes of off-line data generated at the SSCL by simulation, real data taking, reconstruction, etc., will require a centrally managed automated tertiary storage device. Today's front runner to provide this service is a 3480 ACS device connected to an IBM or PCM. Such devices are already fully supported and integrated into HEPVM. The performance characteristics of this device are given in Table 6. This device can be used to store the raw data, the reconstructed data, the simulated data, and the master DST.

The mainframe will also have access to a large (many hundreds of gigabytes) disk farm. The performance characteristics are given in Table 6. The file transfer rates can be increased by data stripping, however this will not help aggregate file transfer rates. The aggregate (for multiple simultaneous disk I/Os) data rates will depend on the number of paths to the data, and a reasonable configuration might yield rates of 15 to 20 MB/sec. The disk farm can be directly shared at channel speeds by up to four mainframes of the IBM 3090, 308x and 4381 varieties. These mainframes can also each have their own connection to the ACS and the mainframes can be interconnected by CTC connections which can be used for intercommunicating and control information. The shared disk farm will be used to store the smaller and more frequently used subset DSTs, the others being kept in the ACS. It will also be used to store mini-DSTs. Copies of smaller mini-DSTs may also be kept on distributed file servers with a few tens of gigabytes of disk space each.

Each of the mainframes can have multiple direct Ethernet connections via channel-attached interfaces (e.g., the IBM 8232 and, more recently, the IBM 3172, or the Bus Tech Incorporated Ethernet Link Controller). Via routers, these interfaces can be connected to an FDDI ring. By the time the SSCL is ready to install the mainframe, it is expected that the mainframe will support a direct FDDI connection. Conservatively we might expect this to support 1-MB/sec file transfer rates.

It appears that more and more the driving force in computing is the user requirements as opposed to the enabling technologies. The users see the workstations more than the data server and, hence, care more about how it appears than say what system the data server is running. Thus the workstations dominate the computing requirements and since they will be running UNIX it would be nice if the mainframe were also to run UNIX. A possible strategy for the mainframe procurement thus would be to request a generic UNIX-driven mainframe data server. The RFP would define the instantaneous (i.e., single CPU) and the symmetric multiprocessing MIPS available; the amounts of primary, secondary and automated tertiary storage; the I/O bandwidth and paths to the data; the network protocols, interfaces, drivers, performance and CPU loading acceptable; the software functions and applications required; and include the requirement for all the large system features already available and in use by the HEP community (some of these are mentioned in Section 9, Operating Systems). There are at least four mainframe vendors (IBM, Amdahl, HDS, and DEC) and many peripheral vendors who could bid on such a request, so you should expect aggressive pricing especially given the SSCL prestige and its leading role in the worldwide HEP community. The fallback situation would be to run either VM or VMS (presuming they have the requisite tools and functions), until all the tools and functions are available and then migrate to UNIX.

11.4. Network

As mentioned in the section on networking, the Ethernets should be subnetted in order to localize the traffic and, as standard FDDI routers become available, they should be evaluated and used to create a backbone to interconnect the Ethernets. The speed of off-site connections for some time will be limited to a fraction of Ethernet speeds so that the routers to support the off-site connections will be connected to the Ethernet. Care will be needed to provide high availability to services by techniques, such as redundancy and uninterruptible power sources, for critical components.

As higher speed networks become necessary, the SSCL can investigate using the 100-MB/sec HPPI hub-type fiber optic networks, like the UltraNetwork Ultra. These network interfaces also provide for off-loading of the protocols from the host. This is done utilizing the lower layer OSI protocols, provided both ends of the connection have the appropriate vendor supplied interface. With such a network, memory-to-memory speeds of 40 MB/sec have been measured between IBM mainframes and up to 4 to 6 MB/sec for VME-based RISC machines.

There must be strong support for wide area networking (WAN) to allow physicists and support personnel to efficiently communicate, compute, share data, etc., in order to enable effective collaborations from remote sites. This will require high-speed WANs, supporting the protocols and applications commonly used in HEP.

Table 7 shows some of the data flows and transfer rates that may be expected before 1992.

12. Future Challenges

There are many challenges facing HEP computing in the future, below we mention a few of those we feel will be more important:

1. Recognizing which of the enticing new technologies to invest in, at what stage in the evolution to step in, and determining how to integrate them seamlessly.
2. Managing the data, in particular deciding how to parse the data into the various storage hierarchies, providing easy-to-use caching/staging, shadowing, high availability, high-speed access, ease of access and distributed automated backup, and user-driven archiving.
3. Network, system management and environment monitoring for lights-out operations, providing high-speed, highly available connectivity.
4. Managing and coordinating the distributed environment both at the laboratory and worldwide.

ACKNOWLEDGEMENTS

We would like to acknowledge many useful discussions on DEC and IBM product offerings with Ted Johnston, Bill Weeks, John Halperin, Charley Granieri, Mike Sullenberger and Teresa Downey, and Rich Lynn of DEC. We obtained considerable help from Greg Mushial on the trends in IBM/Intel PC prices, Lois White provided information on the effectiveness of the ACS and tape staging on reducing manual tape mounts at SLAC, and Dennis Wisinski provided help with workstation prices and performance. Chuck Dickens and Terry Schalk provided encouragement and ideas to pursue. Chris Jones of CERN provided the data point for the IBM 3850 in Figure 2 and Harry Lichtbach of IBM provided the Data Cell and some of the early disk data points in the same figure.

Table 7: Data set transfer times for the typical size data sets stored in the data hierarchies used in the model. Both the source of the data and the destination are shown. Typical expected transfer rates are shown for both multiple simultaneous file transfers (aggregate rates) and for individual file transfers (single). Typical data set sizes are given. Examples of the data sets would be the master DST (4 TB), smaller subset DSTs (100 GB), larger mini-DSTs (10 GB), smaller mini-DSTs or a micro-DST (200 MB), single-reconstructed events (2 MB), a sample of say 10K events from a micro-DST (400 kB).*

Source	Destination	Method	Rates Aggregate (Single)	Typical Data Set Size	Transfer Time
ACS	Mainframe memory	Channel	20 MB/s (2.5 MB/s)	4 TB	55 hrs
Disk farm	Mainframe memory	Channel	20 MB/s (2.1 MB/s)	4 TB	55 hrs
Mainframe disk	File-server disk	High Speed Channel	20 MB/s$^{\alpha}$ (2 MB/s)	100 GB	1.4 hrs$^{\beta}$ 5.5 hrs
Disk farm	File-server disk	FDDI	10 MB/s (1 MB/s)	10 GB	17 mins$^{\beta}$ 2.8 hrs
Disk farm	Workstation disk	Ethernet	500 kB/s (100 kB/s)	200 MB	2.5 hrs$^{\beta}$ 0.5 hrs
Disk farm	Computer server mem.	Ethernet WAN	(100 kB/s) (10 kB/s)	2 MB	20 sec 3.3 mins
Disk farm	Workstation disk	Ethernet WAN	(100 kB/s) (10 kB/s)	400 kB	4 sec 40 sec

$^{\alpha}$This rate is limited by disk performance.

$^{\beta}$To achieve these rates will require multiple paths, such as disk stripping or multiple interfaces.

* These data set size estimates are from a private communication from Harvey Newman of Caltech and are for the proposed L* detector for the SSCL.

REFERENCES

1. *Computer Economics*, Volume 10, Number 3.
2. See for example: Parr, Gary L., "What's Next on the Computer Hardware Roller Coaster?," *Research & Development*, p. 8, November 21, 1989.
3. "VAX 9000 Series," *Product Insight*, p. 4, published by DEC, November 1989.
4. Antonelli, Francesco, "Bit Vectors on IBM 3090/VF Processor," DD/89/32 CERN, November 1989.
5. CERN Computer Newsletter 196, July–September 1989, p. 31.
6. Robertson, L. M., "CERN's Requirements for AIX on Large 3090 and Successor Systems," March 23, 1989.
7. "The FATMEN Report File and Tape Management: Experimental Needs," CERN, April 1989.
8. Foster, David, "TCP/IP Performance," CERN Memorandum, July 1989.

NETWORKING AT THE SSC LABORATORY

G. BRANDENBURG

January 1990

1. INTRODUCTION

Computer networking is now a fact of life for high energy physicists. In a field where the activities are concentrated at a half dozen major laboratories around the world, and where the participants come from several hundred far-flung institutions, excellent communications facilities are absolutely essential. Add to this the fact that all high energy physics (HEP) activities are highly dependent on computers in one way or and another and the need for effective networking of these computers is immediately clear. This subject was addressed in depth by the HEPnet Review Committee (HRC) Report issued in June 1988; a section of this report is included as an appendix.

The Superconducting Super Collider Laboratory (SSCL) is the newest HEP laboratory in the United States; and over the time it takes to complete its construction, it will become the largest. Computers are already playing a role at the SSCL headquarters, and this will increase dramatically as the staff grows. During the initial period, computing will be important for the design of the machine, simulation of the detectors, basic physics calculations, document generation, and many other uses. For each of these, it is important that there be a carefully planned internal network for those machines at the laboratory, and that there be effective links to the other HEP laboratories and the HEP world at large.

2. FUNCTIONALITY

The functionality that is required of an effective computer network is documented extensively in the HRC report. For both the internal and external networks, the traditional required functions are electronic mail, file transfer, remote login, and remote task entry. These functions are all supported by the network protocols that are currently available, namely DECnet, TCP/IP and, to a lesser extent, BITnet. (It should be noted that the problem of different networking protocols, which has been a confusing issue in the past, is gradually being resolved by establishing dual-protocol networks, where the user has freedom of choice. In the long-term future, migration to ISO protocols will provide a single interface for all users.)

In addition to the functions described above, there are numerous higher level functions that will be required. A few of these are task-to-task communications, distributed file systems, remote graphics displays, and network-wide information services. These typically require specialized software and/or hardware at each end of the connection. They also require significant network bandwidth. Many of the functions in this category have in the past been found only on local area

networks. However in the near-term future, the distinction between local area and wide area networking will vanish. The prime example are physicists who have high-powered workstations at their home institutions. To work effectively, they will need access to databases and code libraries that exist on machines at the SSCL and at other HEP sites. They will need to be able to compute on their workstations as if they were a part of local area networks that are separated by great distances!

Finally, the network connections should be able to provide bandwidth for non-computer functions. For example, video conferencing may be the only way to conduct business in a far-flung international collaboration. The bandwidth required for these purposes is a significant, but would be only a modest, addition to the total required for computing.

3. BANDWIDTH

For the internal networking at the SSCL, a single laboratory-wide Ethernet (10 Mb) should suffice for the short term. It will be necessary to separate some sections with heavy, localized traffic from the rest of the ethernet by bridges. In the medium term future, there should be a high-speed backbone (ca. 100 Mb) connecting localized ethernets at several different SSCL locations.

The bandwidth needs for external connections need to be viewed in the larger context of the entire HEP program. This is not just because the sharing of resources is sensible, but more importantly because the participation of all of HEP, including the other national laboratories, is essential for the success of the SSCL. These needs have also been studied in the HRC Report, although some would argue that it is already out of date. Based on this report and more recent experience, it is recommended that the major HEP sites need to be connected in the short term (1989-1990) by a T1 (1.5 Mb) network. The important links in this network should be upgraded to T3 bandwidth (50 Mb) as early as 1991. This is not the place to detail the network topology, but obviously the most important links will be those connecting the SSCL to the other HEP national laboratories.

Network connections from the US to Japan and Europe have basically the same bandwidth requirements as the domestic links. However, because of the higher cost for transoceanic cable, it is prudent to recommend roughly half the bandwidth in these cases relative to the most important domestic connections. This is true even if the cost is shared with the other end. It is important to specify that the major international connections must be routed over terrestrial cable because the delays that are intrinsic to satellite connections are unacceptable.

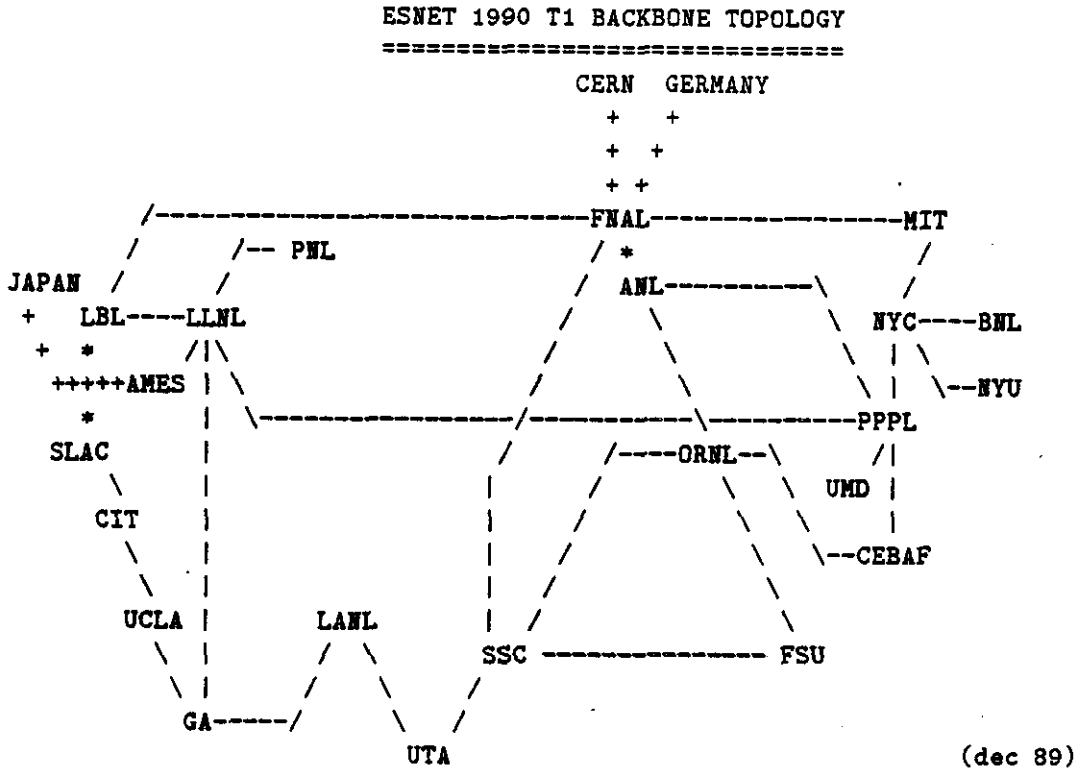
4. EXISTING NETWORKS

It is possible that most or all of the external networking needs of the SSCL can be met by networks that now exist or are currently being implemented. The SSCL is currently connected to the HEPnet DECnet via a 56-kb leased line to Lawrence Berkeley Laboratory. It is also connected to BITnet and NSFnet via the University of Texas at Austin. This means that it is currently possible to obtain all the traditional network functions referred to above. However, these connections should quickly become inadequate as activity in Dallas increases. Login response to and from the other HEP laboratories is already becoming sluggish.

By early next year, the SSCL will be installed as a node on the T1 ESnet backbone. This backbone will carry both DECnet and TCP/IP (and possibly also X25) and will connect all DOE Energy Research (ER) laboratories. It will also connect to several regional networks and to the principal foreign links. A map of this network is shown below. There are plans on a two-year time scale to upgrade the most heavily used links in this network to T3 lines. The bandwidth and costs of this network are shared by the HEP, SSCL, and other ER divisions of DOE. If this network is successfully implemented and upgraded on a timely scale, it should be able to meet most of the networking needs of the SSCL.

It is also recommended that the SSCL be connected as directly as possible to the NSFnet backbone. This network is complementary in purpose to ESnet, but is much larger in its constituency. It reaches all the regional networks in the US, may soon have European connections, and will have T3 service on its main backbone as early as next year.

Finally, the service that is provided by the above networks must be closely monitored. Pressure should be brought to bear on ESnet to upgrade or reconfigure those connections that are saturated. The eventual possibility of dedicated links from SSCL to the other HEP laboratories must also be kept as a contingency.



Notes: 1. All 27 lines shown are full-channel T1, excluding the international connections (+).
2. * = private microwave link

5. HEPNET

One of the principal recommendations of the HRC Report was the establishment of a HEPnet management office headed by the HEPnet Manager. This office is to coordinate the infrastructure that we know as HEPnet together with the help of the HEPnet Technical Coordinating Committee (HTCC). HEPnet in this sense includes not just the networks that provide the connections, but also the way in which they serve the HEP field as a whole. The HEPnet management is now centered at Fermilab and the HEPnet Manager is Phil Demar. It is crucial that the network experts at SSCL work closely with Phil and the HTCC to successfully integrate the SSCL into HEPnet. It is also essential that the SSCL play an active role in the management of ESnet through their representative on the ESnet Site Coordination Committee (ESCC) and the HEP representatives on the ESnet Steering Committee (ESSC).

APPENDIX
"Importance of Networking to HEP" from HEPnet Review Committee Report

Wide-area computer networking is a relative newcomer to the apparatus of high energy physics. One of the first tasks of the committee was to understand how much importance to attach to this new capability. It is possible to argue that in the absence of networks, physicists would continue to do effective research, coping as always with the difficulties of working on the frontier. While this view no doubt contains much truth, the committee has come to believe that, in fact, without the growth of HEPnet over the last decade, the style of HEP research would have taken a much different direction. Lacking the widespread network, collaborations would not have grown to include so many institutions and much more travel would be needed in order for the smaller collaborations that would exist to work together effectively. The most eloquent testimony to the vital role of wide area networking in HEP today is the individual decisions made by essentially all experimental collaborations and research groups, with their tight research budgets, to lease telephone lines and buy the hardware necessary to join the network. There are now well over 100 lines leased for high energy physics, each of which was installed for a particular research need. Groups today find that they literally cannot function as effective members of collaborations without network connections to their collaborators and to their experiment. Yet for this absolutely essential function, the DOE HEP program spends only 0.6% of its funds on wide area networking.

Looking ahead, it is clear that the successful mounting of experiments for the SSCL, and indeed the design of the SSCL itself, will involve even wider collaborations than exist now. It may not be too strong a statement to say that accomplishment of these tasks (barring a wholesale reorganization of physics employment) will not be possible without computer networking that is considerably enhanced over what is available now. It is already true that the R&D that has been done for both the SSCL and its experiments has relied implicitly on HEPnet and other networks. The Central Design Group coordinates the work of many researchers at their home institutions with a large part of the communication using HEPnet and other networks. While initial thinking about experiments has been centered at workshops and summer studies, even more advance and follow-up work has been (and is being) done by international groups at widely separated institutions, using the network for communication and for computing. These far-flung collaborations made possible by the existence of the network will rapidly increase their activities, and their support will necessitate early establishment of

the SSCL site as a major networking (and computing) center once its location is settled.

Experiments in high energy physics have long been intimately connected with computers. Computers control experiments and record data. Both in real time during data taking and during later analysis phases, specially written programs turn the high volume of raw data into forms that can be compared with the abstractions of theoretical understanding. Modern experiments are the result of extended collaborations, involving groups of researchers at institutions across the country and often on two, or even three, continents. The codes and data generated by these separated groups must be combined in a continuing process to design the experiment, operate it for data taking, and analyze the resulting data. Time and distance scales are such that the communication needed for such collaboration can only be accomplished by establishing wide area networks, tying together the computers used by the various groups, and giving remote researchers access to laboratory computers comparable to that available on site. We note that installation of high-performance local area networks has been a high priority at all of the accelerator laboratories in the last several years and that their existence is now taken for granted as a necessary part of the experimental (and accelerator operation) programs. The wide area networks between computers also provide enhanced written communication between separated collaborators through the medium of messages and computer mail. The paper(s) reporting the results of the research will most likely also be prepared collaboratively on several computers in the collaboration.

Already now and to an increasing degree, theoretical work in HEP also depends heavily on large computers for evaluation of theories that cannot be solved analytically and for symbolic manipulation. Such theoretical work also often relies on networks, either for reasons of remote collaboration as discussed above or because the necessary supercomputers are few in number and may not be located at a given researcher's institution. Design of detectors and accelerators already rely heavily on a scale of computing that can only be accessed over networks for most of the researchers involved. Whether using supercomputers or not, theorists have also benefitted from the ability of networks to provide communications for long-distance collaboration, enabling collaborators to work closely together who could not otherwise do so.

Thus a wide area computer network is an essential facility for most work in high energy physics. Unlike many other facilities, such as the detector for an experiment, it is a facility that by its nature serves the entire field and not a single collaboration or even a single laboratory. Since a given institution will often be involved in more than one collaborative effort; networks set up for individual

collaborations will inevitably merge into one national, and by the same process international, network. Thus the usual model of funding facilities for a particular piece of scientific work after its approval by a laboratory-based program committee becomes unwieldy when applied to computer networking.

In fact, the present HEPnet has grown in just this way, with lines and other equipment being installed to meet the needs of a particular research effort and then often being used almost immediately by other projects that need connectivity between the same two points. The result is that the bandwidth installed for one purpose is often inadequate for the shared use and the responsibility for funding and management of the network is blurred. The network has now grown to the point that rationalization is needed. The question must be faced whether or not a national (or international) network for high energy physics is a facility that must be provided centrally for the use of the whole field.

The committee was asked in its charge to determine the appropriate priority of networking relative to other needs of the high energy physics program. Our major conclusion, as outlined above, is that a capable and widespread computer network is a necessity in order for researchers to make efficient use of the other facilities provided. As such, it must be given equal priority with the other components of the program. A new accelerator, detector, or computer facility needs a corresponding level of networking in order to accomplish the research task for which it was intended. In this light, the 0.6% of DoE HEP funds currently spent on wide area networking seems small indeed. The analysis in this report shows that for a network that would better match the needs of the field, the funding should be approximately doubled in FY 1989 and approximately tripled in FY 1991.

Computing for Experiments at the SSC: An Initial Model for 1990–1992

HARVEY B. NEWMAN

January 1990

Computing for Experiments at the SSC: An Initial Model for 1990–1992

Harvey B. Newman
California Institute of Technology
Pasadena, California 91125

ABSTRACT

We discuss a model off-line computing environment at the Superconducting Super Collider Laboratory (SSCL) for the early developmental stages of experiments during 1990–1992. The transition to larger scale computing support for the startup of SSCL experiments is also briefly discussed. The environment is initially based on a balanced network of workstations, which have adequate computing power, disk space, input/output (I/O) and networking capabilities, and the full range of interactive graphics, computing and software development tools, to support the initial simulation-dominated computing task. Once terabyte samples of simulated events have been stored on tapes, the emphasis will shift to the development of reconstruction programs and databases, and to more realistic physics analysis studies with simulated events. A shift to a more centralized and managed environment with a very large data handling capability will then be needed. By 1992, the primary focus of the environment should be a central computing facility at the SSCL site, complemented by high-speed wide area networks to support preparations for the SSCL experiments on a nationwide and a worldwide scale. The central facility will be the primary file server, event server, database server, data communications engine, data and software repository, and coordinating center for the majority of activities related to computing for SSCL experiments.

1. INTRODUCTION

The off-line computing task for SSCL experiments presents a number of new challenges to the physicists planning experiments and to the SSCL. Early estimates have been made [1,2] of the computing power required, by scaling up from experiments at the Fermilab and CERN $p\bar{p}$ colliders and at LEP, and more recent studies have limited themselves to stating some of the individual near-term problems, to a discussion of the simulation tools [3], or to specific proposals to solve the most computer-intensive tasks [4]. In this report, we attempt to present an initial model of a computing environment for the startup phase of the Offline Computing Task at the SSCL. The model is based largely on the experience of the LEP experiments at CERN [5,6], and more specifically on the experience gained by L3 [7,8,9] in building up its computing and software systems since 1981, during a period of rapid changes in computing technology and in physicists' working methods.

Following the discussion of the Computing Planning Committee at the SSCL on December 12-13, 1989, we present a model computing environment which is initially dominated by workstations. The workstations will provide both the large central processing unit (CPU) power, and the necessary support for interactive graphics and physics analysis during 1990 and 1991, when the principal tasks are (relatively few) large-scale simulation activities. Once large data samples are stored, and even before the development of realistic reconstruction codes, databases, and software bases begin, the balance of computing tasks will shift. Data handling and I/O capacity for reading and writing samples of simulated events, and for serving the events and database parameters across site-wide and world-wide networks will soon be on the critical path, along with computing power, as discussed in the following sections.

The need for handling large data samples is expected to start in 1990, when experimental designs for SSCL proposals, and the determination of the critical tracking and calorimetric parameters which are needed to extract the physics sig-

nals under the demanding working conditions at the SSCL, are underway. The acquisition of a flexible central computing facility at the SSCL, with a very large I/O-handling capacity, which can be integrated with a large set of workstations over local area networks (LANs) and wide area networks (WANs), will be required no later than early 1992. One natural choice for the central facility, given the operating system requirements, and the long-term product stability requirements for serving the SSCL collaborations effectively, is a mainframe. The mainframe's I/O handling, data management, data communications, and task scheduling capabilities may be complemented by the high CPU power per unit cost available in high-end workstations, some of which may serve as dedicated computer-servers.

The degree of system integration achievable through high-speed connections between the mainframe and the computer-servers, and over to the LANs and WANs to support interactive computing, software development, and graphics, will be a determining factor in the overall system's effectiveness. The criticality of system integration, and the efficiency and degree of transparency of the network server software, as well as the network bandwidths, will increase as the size of the SSCL collaborations, the number of users, and the scale of data storage and handling, increase.

The transition from the initial workstation-dominated, simulation-dominated phase to the later phases leading to SSCL startup is assumed to be continuous. This means that the initial choices of workstations and software tools will have a substantial influence on future acquisitions on the basis of backward compatibility in the long term. The use of commercially available computing systems and manufacturer-supported hardware and software is therefore to be preferred over special-purpose computing devices.

2. PHASE I STARTUP ENVIRONMENT: 1990–1991

The model startup environment is shown schematically in Figure 1. The environment includes:

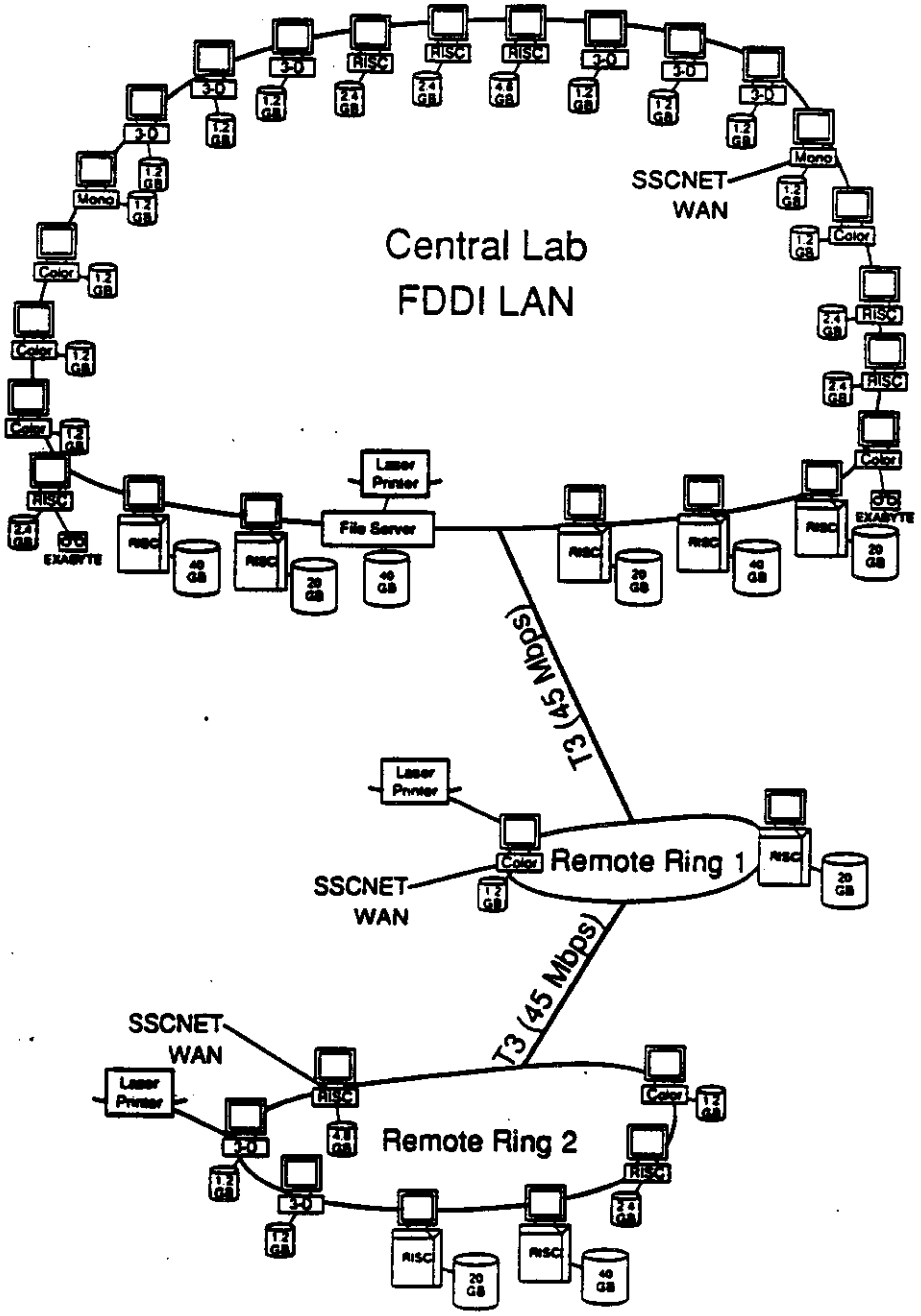


Figure 1. A schematic view of the model Phase I (1990-1992) computing environment for the SSCL. A range of graphics workstations, including simple monochrome, color, three-dimensional hardware-assisted, and high-end RISC stations are grouped into Fiber Distributed Data Interface (FDDI) rings. The rings are interconnected by fiber optic links running at T3 speeds. Each ring has a fast link to a nationwide wide area network (WAN) labeled SSCnet. The number of stations and the distribution and quantity of disk space is only meant to be an illustration. The many low-end stations which will be in use throughout the SSC site are omitted for clarity.

- A range of high-end (RISC) and medium-speed personal workstations to provide the full range of tools for interactive high energy physics (HEP) computing, software development and graphics, as well as CPU power. Additional low-end stations may provide each physicist with local software development and some local computing capability at low cost, along with the ability to execute jobs remotely on more powerful workstations.

- Arrays of Winchester disks directly attached to each of several high-end stations, where the station attached to an array acts principally as a source of CPU power.

- A fast LAN directly connecting the workstations. Direct connections of the workstations to the 100-Mb/sec FDDI (see [10]) are preferred, to provide sufficient bandwidth (1 MB/sec and up) for interactive and network file access applications.

- A fast WAN connection to a nationwide and worldwide network supporting interactive file access between remote sites and the SSCL at high speeds. Principal links in this network should be at T1 (1.5 Mb/sec) by 1990 and higher bandwidths by 1992. Individual users on remote workstations should be able to obtain sustained bandwidths of at least 0.1 Mb/sec by 1990 and substantially higher bandwidths by 1992.

This environment is designed to be cost-effective in satisfying the needs for CPU power in the near-term, principally for simulations. At the same time it should provide a sufficient range of software, graphics, and networking facilities, making it a sound basis for preparation of the software for SSCL experiments over the next several years. Given the rapid rise of the data handling problem, expected to cause a shift in the environmental architecture (to Phase II) no later than 1992, the Phase I environment has been chosen to consist of processors and systems that can be used as an integral part of the Phase II environment. It is particularly

important that the working methods used by the physicists in their daily work can continue, without complete discontinuity during the Phase I to Phase II transition.

The proposed Phase I environment, which is dominated by workstations, satisfies these criteria. Integration of workstation LANs into an overall hybrid environment consisting of workstations and mainframes, is now underway at CERN, for the L3 collaboration.

3. WORKSTATION CHARACTERISTICS

The workstations in the environment are full-fledged computers, mainly (but sometimes not exclusively) for single users. These workstations are to be the principal working tools for physicists involved in computing for SSCL experiments. They must be capable of running the largest HEP FORTRAN programs, including those used for detector simulation and design studies, and later for production reconstruction and physics analysis. The workstations must provide a full set of software tools so that the mainstream development of simulation and reconstruction codes, physics analysis strategies, interactive graphics display programs, and new menu-driven applications can be produced efficiently for each SSCL collaboration. The typical workstation characteristics (ca. 1990) which are required for HEP applications are (also see [10]):

- A high-resolution color graphics screen (typically 1024 × 1280 pixels now; possibly 1200 × 1600 pixels by the end of 1990). Additional display manager hardware and/or firmware support to speed low-level graphics (pixel) operations to the screen. Three-dimensional graphics performance in the range of 100K (medium range) to 1M vectors (high end) transformed and clipped per second.
- Operating system support for a large number (32 or more) processes running simultaneously, each (optionally) associated with a text or graphics window

on the screen. The ability to open a window associated with a process running on a remote node (network-wide computing). The ability to access files transparently anywhere on the network (network-wide file access).

- Computing power for single CPUs in the range of 4 (medium speed) to 30 (high end) MIPs (VAX 11/780 equivalents), where CPU power is measured for relatively large HEP codes written in FORTRAN. Availability of multi-CPU stations (typically four to eight) to provide total CPU power in a single computer-server of up to 200 MIPs.
- Requirements of 8 MB of memory for medium-speed stations; 16 MB for single-user high-end stations; 32 to 64 MB for computer-servers (8 to 12 MB per CPU).
- The ability to support one or more large local Winchester disks over a standard high-speed interface (e.g., Synchronous SCSI over VME at 4 MB/sec). A typical disk volume is 0.6 GB (formatted) in 1989, and it is expected to be 1.2 GB by the end of 1990. Medium-range stations should have one large disk, high-end single CPU stations should have one to two disks. Computer-servers with multiple high-end CPUs should have local disk concentrations of 10 GB and up (limited by available products).
- The ability to support one or more local tape drives. Exabyte drives with large capacity (to 2 GB) but slow speed (typically less than 200 kB/sec attached to workstations) are available. Direct connections to the new, compact 3480 cartridge tape drives at 1 MB/sec or more are expected in 1990 or early 1991. Direct connection of multiple tape drives to the CPU servers in the environment will be a distinct advantage, if possible.
- An excellent FORTRAN compiler, which takes full advantage of the pipelining available in the RISC processor architecture. An excellent debugging environment in which the user has multiple windows in which to view simultaneously: (1) the current program source line being executed, (2) the program text out-

put, (3) one more graphics output stream, and (4) commands to the debugger entered in text from the keyboard or by clicking on a menu with a mouse.

- An integrated set of high level three-dimensional graphics tools. An important feature for HEP applications is the availability of highly structured graphics objects (e.g., the PHIGS standard or APOLLO GMR3D) to match the use of data structures in HEP code.

- Menu-driven applications with user input from a mouse. The availability of a meta-language to construct new menus for specific applications. HEP applications, as in the graphics programs used for scanning and interactive reconstruction, are highly complex. In addition to multiple viewports and the option of several symbolic representations of reconstructed data, hundreds of menu panels and subpanels may be used (logically structured, several layers deep).* Each subpanel corresponds to selection of a program option or displays a piece of numerical information.

Leading (ca. 1989) workstations which satisfy most of the above requirements are manufactured by APOLLO (DN10000, DN4500 Series) or Silicon Graphics (IRIS Power and 4D Series). Other RISC workstations with performances at or near the top of the range are manufactured by MIPS Computer Systems, Inc. (RS2030, and more recent products based on the R6000 chip), Digital Equipment Corporation (DECstation 3100), Data General (DG Avion), Everex (Model 8820), HP (Model 9000-835), and Sun SPARCstation [11]. IBM has also announced a PC/RT which will provide 25 MIPs (probably not VAX 11/780 equivalents), with a $10K base price, and with the option to have up to 8 CPUs. However, APOLLO and Silicon Graphics are currently the only choices which offer the necessary performance/cost combined with the full range of features required for the working environment described above. As the field of workstations is progressing extremely

* L3's interactive graphics system, running on APOLLO workstations and based on DOMAIN DIALOGUE, is an example of an application of this type.

rapidly, a review of APOLLO, Silicon Graphics and alternatives, including a detailed consideration of processor, networking, peripherals and software products expected to be available in 1990, will be needed at the time of the first acquisitions in 1990.

From the above discussion, it is also clear that the workstations concerned are to be distinguished from Macintoshes, PCs (even if 80486-based), or NeXT computers. Each of these has it own attractions, as text and graphics processors in the case of Macintosh II's, or as a development platform for new object-oriented HEP applications in the case of NeXT [12]. However, it is my view that near-term acquisitions, leading to the principal computing tools for SSCL collaborations, should concentrate on the more powerful and HEP field-tested RISC and top-end CISC-based workstation architectures.

4. PHASE I COMPUTING AND NETWORK REQUIREMENTS

CPU Power Requirements

Initial estimates for the Phase I requirements for the SSCL are discussed in the December 1988 Computing Task Force Report [13]. The Phase I estimates are centered around event simulations for physics and detector design studies by a small- to medium-size user community (10 FTE in FY89; 40 FTE in FY90; 80 FTE in FY91). At the December 1989 meeting, a discussion of these and other estimates, and of the economically feasible near-term options, led to the following targets for CPU power to be installed at the SSCL:

(1) 500 MIPs by 10/90,

(2) 1000 MIPs by 4/91, and

(3) 4000 MIPs by 4/92,

where a MIP is defined as a VAX 11/780 equivalent. These targets are achievable through the installation of existing high-end RISC workstations, as discussed below. This level of CPU power will support a limited number of runs of the following example jobs:

(1) 10^6 fast (parametrized) shower events, possibly including tracking. Each event requires 400 MIP/sec (i.e., 400 sec on a VAX 11/780), with a maximum turnaround time of one month. This leads to a peak CPU need of 150 MIPs for this job alone.

(2) 10^5 fast shower events, with a maximum turnaround time of two weeks, leading to a peak CPU need of 30 MIPs.

(3) 10^4 full simulation events. Each event, with a typical energy deposited in the calorimeters of 2 TeV, is estimated to take 12 hours to complete on a VAX 11/780. Requiring a maximum turnaround time of one month leads to a peak computing need of 160 MIPs for this job alone.

The computing time estimates given for example job (3) agree approximately with the results of scaling up L3 simulation timings for LUND events at LEP Phase I. Bootstrap methods which use pregenerated low-energy showers to complete a high-energy shower which is terminated once relatively high cutoffs are reached [14,15], may yield speedups of up to a factor of 10. Greater speedups, as needed for examples (1) and (2), therefore require idealized detector geometry and/or parametrization of at least parts of the electromagnetic and hadronic showers.

It should be possible to exploit the timing/accuracy tradeoff in the early design phases of an experiment in order to meet the CPU time-per-event targets. The acceptable limits of loss in accuracy, and the optimum tradeoffs, will require careful study. This study is, in itself, a highly CPU-consuming activity with bounds that cannot be precisely determined in advance. Viewed in this light, the scenario of computing needs given above is quite restrictive.

The scenario is also restrictive in that a very limited number of jobs may be done within the total CPU power to be provided. It is particularly confining in the number of full simulation events that may be done, as some estimates [13] give the number of simulated background events which are required in the range of 10^5 to 10^6. Studies of signals faked by the pile-up of multiple background events, of rare shower configurations, and by the overlap of events and background from the SSCL machine will therefore have to be limited in scope, or eliminated entirely from the initial studies.

As SSCL startup draws nearer, larger data samples, more extensive physics studies, and a more accurate picture of the true capabilities of SSCL detectors will certainly be required. The restricted scope of the Phase I targets will then have to be expanded to progressively provide the full support for the preparation of the SSCL experimental program.

Estimated Data Volumes: 1990–1992

The data flow from the example jobs given above has been estimated in Ref. [13] and in specialized studies. Since the CPU power which is foreseen results in turnaround times of one week to several months for significant samples of simulated events, a large data volume must be stored. (If an order of magnitude more CPU power were available, the tradeoff in cost between data storage and resimulation of event samples would be different).

The event formats for early simulation studies must therefore contain a description of the energy flow in showers, at a sufficient level of granularity so that the simulated events may be used and reused, for more than one specific detector design and more than one running condition. The lengths of these events thus have little to do with the fully digitized data structures, typically estimated at 0.5-2.0 MB per event, that will be written by a production simulation program for an experiment at SSCL startup. However, the storage of enough information to allow an accurate

simulation of energy hits in calorimeter cells, as well as tracks, for a typical event with 1 TeV in the detector, still requires on the order of 1 MB or more. This estimate is supported by Ref. [13], where an estimate of the tracking and (simple) parametrized calorimetric information leads to a lower-bound estimate of 0.3 MB. Example job (1) above could thus easily write 1 TB of data over the course of a month. Several tasks like example job (2) above, running at the same time, could produce a similar volume of data.

Assuming full exploitation of the available CPU power, and long-term storage of accurate (but still approximate) energy flows and event tracks, the total rate of writing data would exceed 30 TB/year by the end of 1990. (Note that this data volume would represent only $\sim$1 simulated event of 1 MB/sec.) The rate of writing data would rise to approximately 250 TB/year by the end of 1991. If this straightforward strategy were followed, the data to be stored would be the equivalent of 30K Exabyte tapes of 5 GB each (expected to be available in 1990), or 10 to 20 times that number of 3480 cartridges (0.2 to 0.4 GB each) by the end of 1991. This is of the same order as the stored data volume expected to result from the first several years of LEP running. The expense of storing, managing and distributing copies of portions of this data would not be cost effective, relative to the provision of additional CPU power to resimulate some of the event samples. The associated disk space required for staging files in and out from tape and for short-term storage for daily interactive computing with files on local disks would also be prohibitive.

On the basis of this discussion, it is clear that a maximum typical event size on the order of 100 kB is required to keep the data volumes resulting from the first simulation studies within reason. This will undoubtedly require some care in data compaction, and in the choice of the data to be stored. Limited samples of $\sim 10^5$ accurate stored events (0.1 TB) may still be stored on 300 to 600 3480 cartridges or on 30 to 60 Exabyte tapes, along with one or more samples of 10^6 events of 10^5 bytes each. The time to read back and process (e.g., reconstruct) any of these

samples is cumbersome, but not out of the question. By using several tape drives and several CPUs in parallel to reanalyze the events, the time to reprocess the sample for a new study may be as little as a few days, if little CPU time per event is required.

It is obvious that event sizes well below 100 KB per event are highly desirable, as in an efficient DST format. The usefulness of these events is limited, particularly when the nature of the detector, the background and pile-up problems, and the effect of the details of the resolution on the separation of the signals from the background are under study.

Magnetic Disk Storage Requirements

The following requirements are based on being able to use simulated event formats which are typically 100 kB per event, as described above.

Based on recent experience, we will assume a minimum requirement of three days worth of data on disk (several gigabytes or several typical files) for staging operations to and from tape. In addition we will assume that one to two typical data files of 1 GB are required on disk at any one time for each FTE [3,10]. Because of the benefit in working efficiency and the relatively low cost (approximately $3K), each user's general disk space for software and a variety of small files is assumed to be 0.3 GB in 1990 (one medium-sized disk) rising to 0.6 GB in 1992. Several gigabytes of disk space will also be needed for distributed system software and utilities.

These estimates lead to an overall estimated disk space requirement of:

(1) 100 GB by 4/91, rising to

(2) 400 GB by 4/92.

Local Area Network Bandwidth and Dedicated Links

The speed of the local area network that connects the workstations in the proposed Phase I environment will be critical. The bandwidth of the network must be sufficient to provide remote file access and support for running many processes on remote nodes. However, the speed of token rings and the basic effective speed of Ethernets have increased little over the last 5 years, while workstation processing power has increased by 1 to 2 orders of magnitude. Once several high-end workstations are connected together, it is easy for users to bring the relatively robust token ring network (by 1989 standards) to its knees by running a few large HEP programs which read from or write to remote disks. File transport to and from remote disks must therefore be limited in order to maintain overall interactive responsiveness.

The origin of the disparity and a partial near-term solution is the FDDI local area network standard. Many of the leading workstation manufacturers decided to wait for FDDI, as a widespread standard that is expected to be used as widely in the future as Ethernet is used today. After 10 years of standards development, FDDI is only available in a few implementations, most of which function as the means of coupling (bridging) two Ethernets. The leading workstation manufacturers have direct FDDI connections in beta-test, with product releases expected to start in the first half of 1990.

FDDI can provide 1 MB/sec today [10] for a single workstation, in contrast to 50 to 100 kB/sec for lightly loaded Ethernets, or 150 to 250 kB/sec for the fastest lightly loaded token rings. It is therefore important to obtain direct FDDI connections for the high-end computer-servers and single-user stations in the SSCL environment, if possible at reasonable cost in 1990. The use of a limited number of high-speed channel (HSC) connections, with a raw speed of 100 MB/sec and a throughput of 4 MB/sec to VME-based stations (see Ref. 10) is also worth exploring.

In order to get a feeling for the expected capability of an FDDI LAN, one can scale from the current L3 APOLLO token ring LAN. Routine file copying of 10-MB

files is possible, as is occasional file copies (typically one hour) of 100 MB files. However, copying a few of these files simultaneously can slow down the network very noticeably for all users. For a performance increase of a factor of 5 to 10 with FDDI, one might expect to occasionally be able to copy a 1-GB file across the network. But such files cannot be copied often, by any user on demand, without degrading the overall interactive response experienced by other users.

A more effective solution may be the development (at SSCL or at another HEP laboratory) of a high-speed link running over an HSC between two workstations. Heterogeneous IBM-APOLLO and VAX-APOLLO links have been developed, over an IBM channel (4.5 MB/sec), with throughputs to workstation memory of 1.9 MB/sec and to disk of more than 1 MB/sec. The IBM-APOLLO link has the peculiar property that it can work between the IBM and an APOLLO workstation other than the station containing the VME-to-channel interface, resulting in file transfers over the APOLLO token ring at speeds up to 600 kB/sec. The use of dedicated VME-to-VME links, for example, is likely to be very effective over a high-speed path, such as an HSC. Such links also have the advantage of relatively little protocol overhead (a factor of 20 less than TCP/IP, in the case of the L3 link). The development of such links would also be of lasting benefit, since they could be used to integrate parallel processing resources into the central data handling facility, which is one of the main elements in the Phase II SSCL environment.

Reference 10 discusses the possible use of bridged Ethernets for SSCL computing. The experience at CERN is that the proliferation and isolation of Ethernet segments became a major manpower-intensive activity in the years before LEP startup. Keeping the Ethernets alive was and is a running battle in which new bridges have been installed at the rate of more than one per month. The Ethernets are alive today, with typical throughputs across multiple segments of 5 to 20 kB/sec. Connections between the CERN Lab I (Meyrin) and Lab II (Prevessin) sites, where Ethernet is the only choice at present, are not very effective in supporting work at Prevessin, since the principal software base, stored data files, and computer-servers

are in Meyrin. The Ethernet speed is simply too slow when heavily loaded, which occurs naturally during periods of greatest urgency. Sole reliance on bridged Ethernets alone is therefore not recommended at the SSCL, even for Phase I. Vigorous investigation of high-speed LAN alternatives, including FDDI as soon as possible, and the installation and/or development of dedicated fast links at critical points (e.g., between the computer-servers) is strongly recommended.

The limitations of LAN throughput in the foreseeable future also leads to the recommendation that computer-servers are coupled to relatively large local concentrations of disk space, as discussed above. The importance of this aspect of the workstation-dominated environment should not be underestimated.

Peak LAN loads can also be smoothed over by event serving software. While an interactive application is running on a workstation, event servers can work asynchronously to move blocks of a few events at a time to buffer areas, on the disk(s) attached locally to workstations. This arrangement is particularly effective for scanning applications, where human response times to scan events in detail are long (tens of seconds to minutes), and where scan lists of selected events can be prepared by a reconstruction program in advance. Event servers will also work well for long running applications, such as full detector simulations, where the ratio of CPU power to I/O is high. The number of servers running at one time must, of course, be of the same order as the number of active workstation users, or fewer, if the network load is to remain light.

Scaling from experience at the LEP experiments shows that workstation bandwidths of 30 to 50 kB/sec will often be sufficient for applications in which the main I/O is served events, as just described. For quick access to a data file in real time and for overall responsiveness in a wide range of applications, the typical range of required bandwidths is 100 kB/sec to 1 MB/sec.

Wide Area Networks and Remote Computing

Wide area networks will play a crucial role in providing remote access to the SSCL computing facilities from remote sites serving the SSCL experimental collaborations. The installation of major computing resources at the SSCL with top priority is dictated by the limited budget resources, and by the lack of major pre-existing computing resources (well above 100 MIPs) elsewhere in the U.S. which are available for SSCL simulations and detector design-related computing. This degree of centralization, necessary as it is, must be complemented by full involvement of a much larger sector of the HEP community than will be resident at the SSCL during 1990–1992.

In setting the scale of the network connections between the SSCL, some other HEP laboratories, and to selected sites which serve as focal points for the emerging SSCL collaborations, it is important to take a forward-looking view of the physicists' working methods. This means:

(1) Remote computing will be done increasingly on workstations. This trend should be reinforced strongly, for reasons of compatibility, graphics requirements, and the necessity of doing some of the computing locally at the remote site.

(2) Static models of network demands, expressed in terms of (terminal-oriented) characters per second, or fixed amounts of data to be transferred per month, are not relevant. The June 1988 HEPNET Review Committee (HRC) Report specifically considers this sort of static work load and avoids modern workstation-oriented applications [17] in its estimates. The HRC Report does not intend to encompass networking for SSCL experiments, and it is not recommended as the basis for any of the SSCL's network needs estimates.

(3) The interactive working methods which are in use over LANs at the SSCL should be extended (at necessarily lower speeds) to physicists working at

remote sites. In order to make this feasible, a target of 100 kb/sec of available bandwidth for each remote FTE should be installed by the end of 1990. With the aid of event servers, such as the ZEBRA data structure server now being developed at CERN [18], a remote user on a workstation would then be able to use databases and small event samples (up to tens of megabytes) which are available at the laboratories.

In order to support remote interactive computing and the corresponding bandwidths, the principal links connecting the SSCL to remote sites should be at T1 (1.5 Mb/sec) by 1990 and at higher speeds by 1992. The bandwidth should be guaranteed, dedicated bandwidth for SSCL applications, or else remote computing would be frustrating, and largely nonproductive. If existing national efforts (NSFnet, ESnet) cannot provide the necessary bandwidth, with guaranteed speed of service, then the SSCL should vigorously pursue other possibilities of obtaining its own dedicated links.

In order to allow physicists at remote sites to work efficiently, the installation of compatible workstation clusters with significant computing power (100 MIPs and up) is recommended. Such clusters are already being installed to support some of DHEP's major programs (examples are at Caltech and MIT for L3 at LEP). It would be appropriate for the SSCL Division of DoE to make similar initiatives, on a somewhat larger scale, to maintain an effective balance of on-site and off-site computing. The necessity of maintaining this balance, for technical as well as sociological reasons in the HEP community, is well established and should not be controversial.

5. PHASE I CONFIGURATION ELEMENTS

Workstations for the CPU Requirements

The October 1990 target of 500 MIPs can be achieved with three to six APOLLO DN10000 Series or Silicon Graphics Power Series stations used as computer-servers.

In addition, several single CPU high-end stations ($30K to $50K) and a greater number of medium range ($10K to $20K) stations should be purchased in 1990, so that approximately one station per 1.5 FTEs is available. As demand for the stations (inevitably) rises, lower end stations ($10K or less) can be used by physicists not heavily involved in simulation and analysis, and for some software development tasks. An adequate set of stations to satisfy the 10/90 CPU-power goals should be obtainable for approximately $1M, but it will important to (a) consider staged delivery of new (1990) products and (b) obtain competitive bids and/or academic discounts.

In order to meet the 4/91 goal of 1000 MIPs, it will be highly desirable to use some stations with a total CPU power in the range of 200 MIPs and up, in order to keep the network traffic associated with the writing of data files to disk or tape, during their creation, down to a manageable level (as discussed further below). Stations with this CPU power should be available as standard market items by this time. Acquisitions during this period should be scheduled carefully to take advantage of newly appearing products.

The 4/92 CPU-power goal of 4000 MIPs may require the acquisition of an additional 10 to 20 computer-servers. If the exponential fall of the price of workstations continues, the acquisition cost of these servers, and a complementary set of single CPU high-end and medium-speed stations, may be obtainable for a price in the range of $2M to $3M. One might expect to obtain a total of 40 to 50 workstations for this price, not counting low-end stations. These are only rough guesses. A review of the market situation and the size of the SSCL physicist user community needs to be reviewed in 1991 to determine the optimum purchase.

Once the line(s) of supported workstations are established, volume purchase agreements involving installations for experimental groups at many universities and other HEP laboratories, as well as the SSCL, and joint (R&D) projects with man-

ufacturers, may be an important factor in obtaining the maximum configuration within a given budget.

Applicability of Special Computing Systems

By early 1992, when the CPU requirements are climbing towards the projected 4000 MIPs, it will be appropriate to carefully consider the role of special purpose computing systems, such as the ACP II farm of VME-based processors. The evaluation of the cost effectiveness of such systems in real terms relative to commercially available products will require an analysis of the following factors:

(1) The fraction of the full range of computing tasks that will be supported by the special system.

(2) Compatibility of the program code, and the level of manpower needed to keep the programs and databases current and valid.

(3) Manpower required to support the system hardware and software, develop special utilities, and handle the data. Since the workstations are not supplanted by this system, its support must be provided in addition to the workstation support.

(4) Time (advance in cost) effectiveness of the special purpose system, over commercial products. Because of rapid advances in commercial RISC-based computers, especially in workstation form, one must consider when the present generation of the special system will be overtaken by commercially available systems with a higher overall level of functionality. The advantage of using a special computing system will be unclear if (a) the time until it is overtaken is less than 2 to 3 years, (b) the special system cannot use the latest RISC generation technology, because the manufacturer is not making the latest generation technology immediately available, or (c) the system is not fully mature, in hardware or software, resulting in too large a burden in physicist and laboratory staff manpower.

(5) Outlook for useful product life. The special purpose system must be supported for 4 to 5 years, if training of the SSCL collaboration members is to be warranted, and if use of the system is to have a substantial net positive impact on preparing an SSCL experiment. It is clear that the special system will have to undergo at least one major upgrade during this period, if it is to remain cost effective.

In summary, only a fully mature, highly compatible system, with an excellent outlook for support, and vigorous development over the 1992–1997 period, should be seriously considered. A detailed analysis of benefits versus costs should be made to determine if adopting the system is worth the risks.

From the above discussion, it should be clear that the issue of large-scale usage of special systems is largely confined to Phase II. A key issue for Phase II will be the degree of integration achievable with the central data handling facility which should be installed in 1992.

Winchester Disks

The short-term needs, for 100 GB of disk space by 10/90 and 400 GB of disk space by 4/91, are expected to be satisfied by arrays of Winchester magnetic disks. Erasable optical disks are not yet competitive in performance/price as an on-line secondary storage medium, and the predominance of online magnetic disks is expected to continue at least until 1992. It also needs to be emphasized that many of the disks discussed here must be concentrated and directly attached to the computer-servers used to simulate events. In the context of the SSCL Phase I environment, the LANs connecting the workstations will not be able to support reading and writing of data files across the network in a free and unrestricted fashion. Attempt to use remote file access in this fashion, in present-day experiments, quickly overwhelms the fastest token rings, and is simply not feasible on a significant scale over Ethernets. The use of FDDI rings, as soon as available, will increase the

flexibility of remote file access, but will not change this requirement (as discussed further below).

The cost of 100 GB, in the form of 80 Winchester disks with controllers and interfaces, is expected to be approximately $0.5M (cf. [13], but keeping current disk prices for high-end workstations in mind). The 400 GB of disk space, consisting of 200 drives of 2 GB each, is expected to cost approximately $1.5M by early 1992. In order to keep to these figures, competitive bids and/or academic discounts from workstation manufacturers may be required. Third party vendors should be considered as a low cost way to obtain the large numbers of Winchester disks required, if the reliability of the disks can be adequately demonstrated.

These disk space requirements are therefore feasible for the SSCL. The systems problem of supporting so many small disks on workstations is nontrivial however, and it merits further study. Given the current limitations (ca. 1990) of approximately 10 GB of attached disk space on a high-end workstation, special configurations may be required from the manufacturers to meet the SSCL Phase I needs. The outlook for this is optimistic, if only for the reasons of the high visibility and long-term sales potential of SSCL-related computing.

Tapes

Exabyte 8-mm helical-scan tapes have gained great popularity among part of the HEP community, particularly at Fermilab. Their data density per unit physical volume is very high (2 GB now, up to 5 GB starting in Fall 1990) and the cost of the drives is very low (typically $5K or less including interface). Their reliability as a backup medium over the short term has been shown to be high at Fermilab and at CERN [16], principally because of the extensive error recovery features provided in the drive firmware. Many HEP laboratories, including CERN, have expressed reservations about Exabyte tapes, and remain committed to 3480 cartridges because:

- The obtainable reading and writing speed is an order of magnitude or more slower than 3480 tapes. Startup times (e.g., tape retensioning) also slow down the elapsed time to read or write a file. The use of Exabytes as a principal medium at a computer center could greatly slow down operations.
- They are a backup medium and cannot be used to read or write records to tape directly. They therefore have limited flexibility.
- The drive hardware has been judged not to be sufficiently reliable by some manufacturers (most recently by Hewlett Packard).
- The 4-mm Digital Audio Tape (DAT) is expected to surpass the 8-mm tape technology in data storage capability per unit cost in 1990. Long-range support for DAT appears to be more likely than for Exabyte tapes.
- Increases in the density of 3480 cartridges have been long awaited, and are still expected. The 3480 cartridges are the current de facto standard tape medium. New compact tape drives from Storage Technology STC and from Fujitsu offer very attractive performance and price (down to $12K per drive from STC, when bought in pairs).

Because of the tradeoff in data density and cost per unit of data stored versus speed and flexibility, it is clear that both Exabytes and 3480 cartridge drives will be needed in the SSCL environment. The acquisition of eight Exabyte drives and six 3480 cartridge drives, for example, should be obtainable in 1990 for an approximate cost of $120K. Convenient (but slow) handling of 100 to 200 GB data samples at relatively low cost can also be done with an Exabyte tape library.

The Role of Standards and Open Systems

Given the long time scale for the preparation and execution of the SSCL experimental program, it is natural to emphasize the use of standards and open systems (UNIX, TCP/IP, ISO, etc.). A commitment to use UNIX, or other standards exclusively, has been stated on occasion, including at the December 1989 meeting. The

usefulness of standards to allow the physics groups collaborating in SSCL experiments to choose among alternative manufacturers of computing equipment, or to simultaneously satisfy their computing needs for the SSCL and for ongoing experiments, is an important consideration. At the same time, it is equally important to realize the limitations of standard operating systems (e.g., UNIX), graphics software (GKS, PHIGS), and network protocols (e.g., TCP/IP). This commitment is often expressed in simplified terms, since exclusive use is often inefficient or impractical.

Examples which preclude an exclusive commitment to UNIX are:

(1) The major operating systems in use today at the HEP laboratories, and by all large HEP experiments, are IBM/VM and VAX/VMS. UNIX is used increasingly, in a strictly non-exclusive fashion, by physicists on workstations.

(2) UNIX is not an operating system which is designed to serve a medium to large user community which shares a limited set of centrally sited resources. It does not have a large set of user and task priorities. The concept of task and resource usage scheduling is not native to UNIX, and system tools to provide these services have been provided by some of the major workstation manufacturers.

    The transition from SSCL Computing Phase I to Phase II, where some computer center concepts are needed, will mean that major system additions will be needed, to be provided by third party vendors. or perhaps to be developed by SSCL personnel.

(3) The interactive environment provided by workstations is built up of a set of tools built underneath, as well as on top of UNIX. The tools are often hardware-specific, to provide optimum performance/price, and are not transportable. The user thus is not provided with a uniform interface in practice, even though he uses UNIX.

Other examples precluding the exclusive use of standards are:

(1) The fact that maximum graphics versatility and performance requires the use of manufacturer-specific graphics software, which is often specifically tailored to take advantage of associated hardware and firmware. APOLLO, Silicon Graphics, and many graphics superworkstations (e.g., STARDENT) are examples.

An attempt to use GKS at CERN, as the unique supported graphics protocol, was unsuccessful. From the outset, it was clear that GKS does not allow multilevel structuring of data, so that matching the graphics software with the rest of the (ZEBRA-structured) reconstructed data object was going to be difficult or impossible. The trial implementations (by a third party vendor) for APOLLOs were unusably slow. Given the advantages in working efficiency and the options available through APOLLO-specific graphics software, the main thrust of effort (in L3, for example) was put into the use of GMR3D and DOMAIN DIALOG. The result was a graphics program [19] that successfully met the challenge of describing the L3 detector and its reconstructed data, with any level of required detail. It is expected that facilities of a similar (manufacturer-specific) type could be successfully employed for the SSCL.

(2) The use of fully standard network protocols often leads to poor performance, because of protocol overheads. Simpler protocols over reliably dedicated links, can often lead to major increases in transmission speed and/or reduced CPU overheads. As discussed at the December 1989 meeting, interfaces for TCP/IP (from BTI) are available that can support mainframe-to-workstation transfers over Ethernet at speeds of up to 300 kB/sec. The main penalty to be paid is that a 3090 CPU will be completely saturated when supporting three to four of these links. As discussed above, a dedicated IBM workstation link has achieved more than 1 MB/sec of throughput with a 5% CPU load.

INTRODUCTION to PHASE II:
A Mainframe-Oriented Environment

It is clear from the discussion above that the rapidly rising data volume during Phase I will soon begin to strain on the limited data handling capabilities of the local area network linking the workstations and the workstation-attached disks and tapes. Coordinating the computing for an increasing large user community, which will grow as the SSCL collaborations grow, will become problematic in a fully distributed environment.

It is therefore recommended that by 1992 the SSCL should make a transition to a more centralized, managed computing environment, as illustrated in Figure 2. The elements of the environment, and a conservative estimate of the high end characteristic available as system building blocks in 1997 are summarized in Figure 3. The environment is focused around a central data handling facility. A natural choice for the central facility would be a Parallel Integrated Computing System (PICS) [7,8,9], in which a series of relatively low-cost computer-servers are closely coupled with a general purpose mainframe system with a large set of peripherals and a very large I/O handling capacity.

The use of a mainframe is required to satisfy the diversity of computing needs, as well as the large volume of simulated data to be handled. As the software bases, databases of parameters, and the simulated event file bases are developed, and as physics analyses become more realistic as well as diverse, the mainframe system will be the principal means of providing access to the data. Centralization of this access, to a single master site with a full time operations and systems staff, is necessary because of economies of scale in data storage, speed of access, the need to schedule, and prioritize users jobs. Central siting of a large part of the data handling devices also has advantages in optimizing the use of operations and systems staff manpower. Tight coupling (i.e., high-speed connections) between some of the principal sources

of CPU power and the central data handling facilities is most easily achieved over relatively short distances.

A crucial aspect of the environment is a high degree of integration between the mainframe and the set of attached computer-servers, and with the workstations which provide the majority of the functionality for software development and physics analysis. The problem of mainframe-computer server integration, for production data processing, has been addressed over the last three years by the Parallel Processing Project (L3P3) [7,9]. The principal goals of L3P3 are:

(1) The integration of commercially produced high-end processors with a mainframe over a high-speed communications channel.

(2) The creation of software tools which provide application programs with the capability of parallel execution on dynamically managed attached processors.

(3) The creation of system software to manage the attached processors automatically and dynamically, with a level of sophistication similar to the management of mainframe CPU resources by a modern batch system.

Goal (1) has been achieved between APOLLO DN10000's and an IBM 3090/180E mainframe, using the fast link developed in the L3 collaboration, which has been described above. This development was based on earlier work at SLAC, at CERN, and in L3 using IBM 3081/E emulators. Goal (2) has been fully implemented, with static allocation of processors, with 3081/E emulators in L3. Single- and multiple-user usage of the attached DN10000 resources from the IBM 3090, and access to database and other mainframe resources from the DN10000's has been implemented, and will be used in production starting with the next run at LEP (Spring 1990). The resource management system [goal (3)] has been designed, based on the SLAC Batch Monitor System [7]. Implementation is expected to begin later this year.

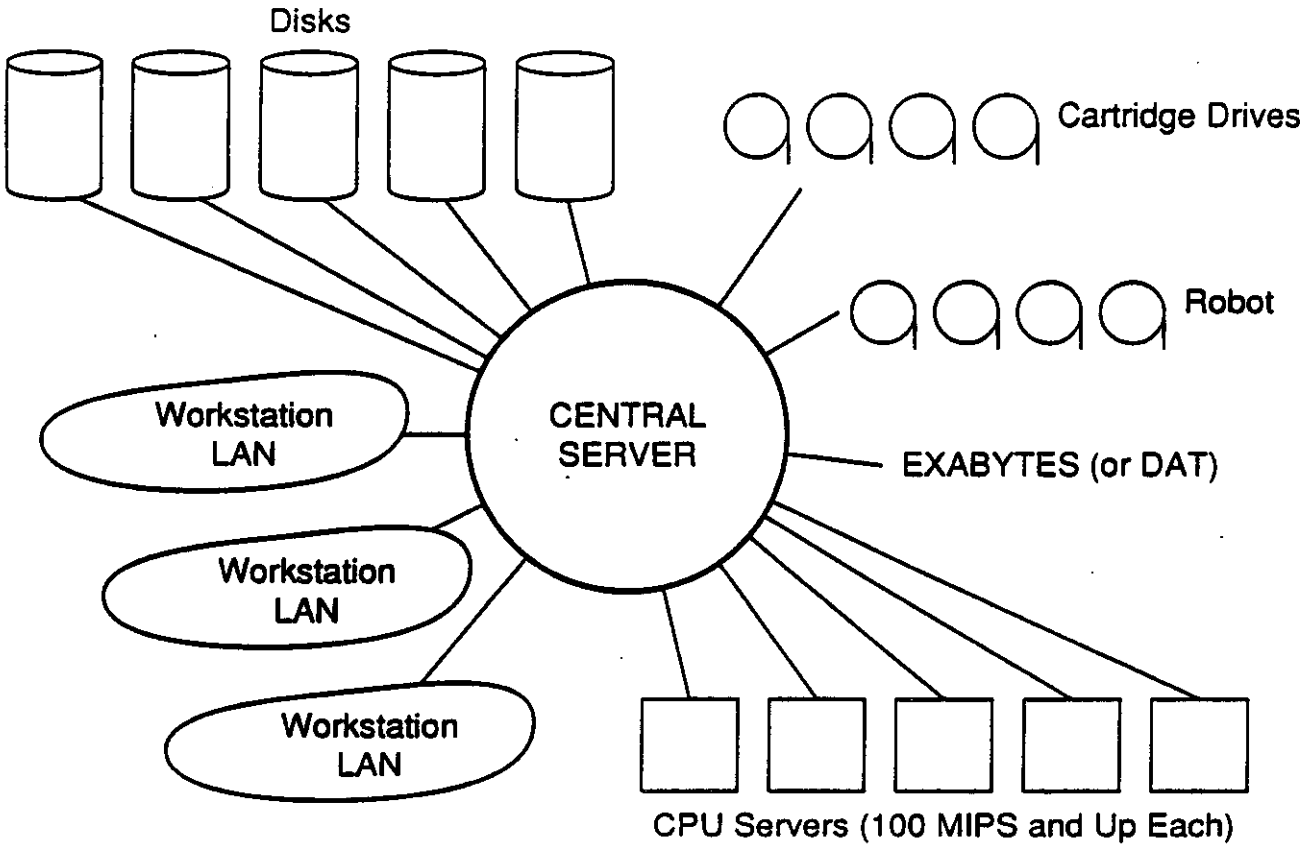


Figure 2. A schematic view of the Phase II environment (ca. 1993–1997). The workstation LANs are expected to evolve from the type of configurations illustrated in Figure 1. The LANs are likely to start as FDDI, and may be replaced by a higher speed LAN (in the GB/s range) at a later date. A natural choice for the central server is a mainframe, based on long-term trends in data handling, as discussed in the text. The connections between the central server and the CPU servers may be via an HSC (or its successor). The helical scan devices may be replaced by a faster technology, with higher densities than the 8-mm or 4-mm devices which are available in 1990, at an early stage in Phase II.

HEP COMPUTING ENVIRONMENT

"MODERN" Components (ca. 1997)

- MAINFRAMEs
  - — HOST: AP parallel integrated system
  - — HOST: To 120 MIPS per CPU; 1000 MIPS per system. Multiple HSC I/O Capability: To 100 MB/sec
  - — APs: Superworkstation technology with fast single-channel I/O capability
- SUPERWORKSTATIONS
  - — Low-End: 10 MIPS+ per CPU; for software development, simple analysis and graphics
  - — High-End: To 150 MIPS per CPU; 1000 MIPS per system. For real-time 3-D graphics and full interactive reconstruction.
- STORAGE MEDIA
  - — Erasable Optical or Magnetic Tape Cartridges: Need 1 GB or more per volume.
  - — Disks: 5 GB/disk for small systems; need 50 GB/disk for large systems.
- SITE-WIDE NETWORKS
  - — FDDI: 100 MB/s optical fiber ring
  - — Need for a 1-GB/s LAN by 1995
  - — Point-to-Point Links (HSC Technology): 100 MB/sec raw speed; >10 MB/sec throughput
- WIDE AREA (WORLDWIDE) NETWORKS
  - — Multiprotocol including ISO (to 2 MB/s)
  - — End links at T1 (1.5 MB/s) and up
  - — T3 Trunks (45 MB/s): Dedicated or shared
  - — U.S. Government and Research Networks: 1000 MB/s and up

Figure 3. Elements of the Phase II environment. The estimated characteristics of the computing and networking subsystems are based on conservative extrapolations from current (1990) technology, up to the period just before SSC startup (ca. 1997).

The general concept of this design, which is adaptable to different mainframe-attached processor combinations, should offer a viable path toward integration of the computing systems at the SSCL, starting with the transition from Phase I to Phase II.

Conclusion

The initial computing Phase I needs for the SSCL, and the SSCL physics program, may be met with a series of high-end workstations, complemented by adequate, concentrated disk space and high-speed local area network and wide area network facilities. Once the SSCL computing task broadens from the narrow focus of early simulation studies—to more realistic studies, to detailed detector design, and to the development of actual production codes for the experiments—a transition to an environment focused around a central data handling and data processing facility will be needed. The central facility, which will need the I/O handling capacity and management capability of the largest mainframes, could be implemented in a cost-effective fashion by coupling the mainframe to commercially available high-end RISC computers. The initial workstations, and a growing number of CPU servers in workstation (or compatible) form, will have to be integrated with the mainframe through a parallel processing software system.

In spite of advances in computer technology, resources for SSCL computing will remain scarce, in terms of data handling, as well as for CPU power, relative to the experiments' needs. The challenge of managing the resources, and of optimizing their shared use, will be at least as great for the SSCL as for the current generation of large collider-based HEP experiments.

Acknowledgments

We wish to thank Richard Mount and Fridolin Dittus for their contributions to the ideas and many of the computing system concepts presented in this report.

Discussions with Les Cottrell, and his contributions to this report have, as always been very valuable. We are also grateful to Joe Ballam, Gil Gilchriese and Roy Schwitters for the opportunity to contribute to the SSCL's plans and for their hospitality.

REFERENCES:

[1] J. Ballam (Chairman), "Computing for Particle Physics," Report of the HEPAP Subpanel on Computer Needs for the Next Decade, DoE/ER-0234, August 1985.

[2] H. Newman, "Computing for HEP Experiments 1987–1997," in Proceedings of the 1987 International Conference on Computing in HEP, Computer Physics Communications **45**, 27 (1987);
H. Newman, "Computing and Data Communications for High Energy Physics Research 1987–2000." Report to the National Academy Panel on Information Technologies and the Conduct of Research, 1987.

[3] S. Loken *et al.*, "Computing for High Energy Physics in the 1990's," from the Snowmass Summer Study on High Energy Physics in the 1990's, Snowmass, Colorado, 1988.

[4] T. Nash, "Computing Possibilities for the Mid-1990's," Talk at *Future Directions in Detector R&D for Experiments at pp Colliders*, Snowmass, Colorado, July 1988.

[5] D. O. Williams (editor), "The MUSCLE Report: The Computing Needs of the LEP Experiments," CERN/DD/88/1, January 1988.

[6] The CERN Green Book II: "Computing at CERN in the 1990's." Editorial Board chaired by D. O. Williams. CERN, July 1989.

[7] L3 Offline Computing Status Report, May 1989. This report is extracted from the Caltech HEP Group's 1989 Annual Report to the DoE (pp. 203-293).

[8] R. Mount, "Overview of the Essential Tools," Invited talk at the Conference on Computing in High Energy Physics, Oxford, April 10-14, 1989.

[9] F. Dittus, "Parallel Processing with Attached Processors in a Computer Center Environment," talk given at the Oxford Conference, 1989;
F. Dittus *et al.*, "LEPICS-P3 Architecture: Requirements and Envisaged Solution," Report to IBM prepared as part of the IBM-L3 Joint Study on Parallel Processing, CERN, November 1989.

[10] L. Cottrell, "SSCL Computing," Contribution to the December 1989 Meeting of the SSCL Computer Planning Committee.

[11] MIPS Magazine, Volume 1, No. 7, July 1989, p. 49. Many other comparisons of high-end RISC, CISC, and other high-end workstations (e.g., Intel 80486 or N10 based), may be found in this magazine.

[12] Work by W. Atwood and P. Kunz. W. Atwood, private communication.

[13] M. Gilchriese (editor), "Report of the Task Force on Computing for the SSCL," SSCL-N-579, December 1988.

[14] These techniques have been implemented as an option in the L3 detector simulation. For electromagnetic showers: work by M. Maire (Annecy), L. Luminari (Rome), and A. Gurtu (Bombay) in L3. For hadronic and electromagnetic showers combined: H. S. Chen (Beijing) in L3.

[15] H. S. Chen, talk presented at the Oxford Conference on Computing in HEP, April 1989.

[16] A. Petrilli, Proposal to the MEDDLE Committee, and CERN DD Division for the development of a workstation-based Exabyte to 3480 cartridge tape copying facility, October 1989;
A. Petrilli, "Helical Scan Devices and Their Possible Usage in HEP," CERN Report CERN-DD CO/1, January 1990.

[17] L. Price (Chairman), "High Energy Physics Computer Networking," Report of the HEPNET Review Committee, June 1988.

[18] Under development by R. Brun, B. Segal, R. Mount *et al.* at CERN.

[19] H. Stone *et al.*, L3 Collaboration, private communication.

Computing Requirements for Theory at SSC

FRANK E. PAIGE

January 1990

Computing Requirements for Theory at SSC

Frank E. Paige
Physics Department
Brookhaven National Laboratory
Upton, NY 11973
January 1990

The computing needs for most theorists can probably be satisfied by any reasonable Superconducting Super Collider (SSC) computing environment, but special facilities are needed for higher order QCD calculations, for lattice gauge theories, and for event simulation.

The majority of theorists write small programs over limited time spans to solve particular problems, e.g., to integrate cross sections numerically. FORTRAN is the standard language. Central processing unit (CPU), memory, and disk space requirements obviously vary but are generally modest compared to other SSC requirements. Since development takes most of the time and effort, good facilities are essential. This includes a good symbolic debugger, something which does not exist on IBM mainframes and on some UNIX systems. Standard libraries, such as IMSL, NAG, and SLATEC, are needed for special functions, numerical integration, solutions of differential equations, and other standard mathematical tasks. A high-level graphics library is required with on-line display comparable to a Tektronix 4010 and printed output for immediate use. The graphics library should also give publication quality output. Three-dimensional graphics workstations have not been extensively used in theory.

Most theorists are currently accustomed to working on a VAX running VMS. A VMS system should be maintained at the SSC for the foreseeable future to allow visitors to work efficiently.

Higher order QCD calculations and other complicated perturbative calculations are dependent on symbolic algebra programs both to do the Dirac algebra and to manipulate the result into a useful form. To be useful a symbolic algebra program must be reliable for complex calculations. A fast processor with adequate memory to prevent excessive paging is needed; to set the scale, the MACSYMA calculations for the one-loop corrections to

heavy quark production took about a thousand hours of VAX 11/780 time.[1] Symbolic algebra programs can also be used for a variety of other problems. The UNIX version of MACSYMA is the most commonly used general-purpose algebraic program. It lacks a standard package for Dirac algebra; the SSC should obtain and support one capable of working in n dimensions. MATHEMATICA should also be considered; it is relatively new, but it is reputed to be reliable and easier to use.

Lattice gauge theories are very regular and so adapt well to vector supercomputers or to massively parallel machines, such as the Connection Machine. Currently sustained speeds approaching 1 Gflop and memories of > 100 MB are available.[2] This may be sufficient to obtain some results for QCD, but one to two more orders of magnitude are needed both in speed and in memory. This may be attainable with special purpose computers. Since the lattice requirements are specialized, they should probably be dealt with separately and not considered in the general SSC computer discussion.

Event simulation does not benefit from a vector supercomputer; an attempt to vectorize ISAJET produced negligible gain.[3] Event simulation does suit a farm of microprocessors, such as will probably be used for detector simulation and analysis. An SSC event typically takes 1 to 10 sec to generate on a VAX 11/780, and samples 10^6 events are needed for background studies with reasonably good statistics. An event typically requires 30 kB of disk space, so large disks and tape backup are needed. Large production jobs require a support staff to manage them.

Event simulation programs typically contain 10 to 20 K lines of code; this is small compared to analysis programs, but still requires a code management system. An ideal system should provide a portable way for handling common blocks and machine-dependent code and good management facilities. No suitable system now exists. The VMS Code Management System (CMS) provides adequate history and control functions, but it has no built-in facilities for constructing program releases or for determining dependencies, it does not easily handle machine-dependent code, and it is not portable. PATCHY handles

[1] S. Dawson, private communication.
[2] A. Kennedy, talk at the SCRI User's Group.
[3] R. Holmes, IBM (Kingston), private communication.

common blocks and machine-dependent code in a portable way, but it is extremely clumsy to use as a code management tool. HISTORIAN was found by the UA1 collaboration to be too limited. UNIX can handle more dependencies, but it is not at all automatic.

Development of code management tools is vital for large experimental codes, and it should be a high priority item for the SSC.

Theorists spend a relatively large amount of time writing papers. Many theorists know TEX and prefer to write directly in it. In any case a good environment for document preparation, including intermixed text and graphics, should be supported.

Use of Commercial Computing Products Online

A. LANKFORD, SLAC

January 1990

1. Introduction

Commercial computing products will serve several functions in on-line triggering and data acquisition systems. The largest scale use of commercial processors will be for high-level event selection by a processing farm. Microprocessors will be found embedded in special-purpose low-level trigger processors or in data preprocessors. Processors will also serve as hosts for the system as a whole and for each detector subsystem, and workstations will be used to interface the physicists to the on-line system. In addition, commercial mass storage devices will also be used to record triggered events and commercial network or bus products may be used to provide interconnection among distributed processors. Finally commercial software products from operating systems to databases to software engineering tools will be used.

2. Processing Farm

The highest level of event selection, frequently called Level 3, is generally conceived as occurring in a farm of many microprocessors, as currently performed by ACP processors for CDF and as planned with MicroVAX processors for D0. This farm may be characterized by its input and output bandwidths, its processing power, and its software environment. The input and output bandwidths and the processing power are usually conceived as being provided by many parallel data links and by many parallel processors. In fact, the numbers of links and processors are not of principal importance. The aggregate bandwidth and processing power are important.

2.1 Farm Requirements

The required input bandwidth to the farm is dependent upon the physics goals of the experiment and upon the deployment of trigger selection criteria between prompt trigger processors and Level 3. The aggregate bandwidths most often discussed range from 10 GB/sec to 100 GB/sec. The 10-GB/sec rate arises from a conservatively designed data acquisition system for a high-P_T experiment with a prompt trigger rejection of 10^4, i.e., 10^4 event/sec $\times$ 1 MB/event = 10^{10} bytes/sec. Clearly, an experiment with a prompt trigger rejection of 10^5 to 10^6 would require less input bandwidth. On the other hand, a B-physics experiment operating at $\mathcal{L} = 10^{32}$ cm^{-2} sec^{-1} with a prompt rejection of about 10^2 would require input bandwidth of 100 GB/sec.

The required output bandwidth from the farm to mass storage is 10 to 100 MB/sec based upon writing 10 to 100 events/sec at 1-MB/event or 1000 events/sec at 100 kB/event.

The aggregate processing power of the farm is usually described as being between 10^5 and 10^6 MIPs. These numbers are loosely based upon needing 100 seconds on a 1-MIP machine to perform final event selection rejecting another factor of 10^2.

Although the internal hardware characteristics of a farm are not critical, the software environment which it provides is important. The farm must execute code which runs in off-line processors, which implies that the farm processors must have high-quality compilers compatible with those used offline. It must also offer a code development environment, or be compatible with such an environment on another machine, which facilitates production and initial debugging of new code. It must also offer adequate tools for *in situ* debugging of code during operation, i.e., debugging of code executing on any node in a multiprocessor system. Code running on such a powerful machine will require new levels of reliability. In addition, the operating system must provide tools for data transfer to and from processors and for control and monitoring of processors. In short, the farm must

provide a software environment as comfortable as provided by today's popular minicomputers.

Additional requirements upon the farm include the need to allow testing of new trigger code, verification of event selection processing, and monitoring detector performance as background tasks to the event selection process. These requirements demand the ability to share events or data among tasks, or potentially among nodes in a multiprocessor farm.

2.2 Options for Use of Commercial Computing Products in the Farm

Commercial products could be used as part of the farm in at least three different ways. Commercial microprocessors could be implemented on custom processor boards, the approach chosen by ACP I and ACP II. Commercial single board computers could be implemented as processing nodes, the approach chosen by D0. Finally, a commercial multiprocessor system could be implemented as a solution to the whole farm problem.

Vendors do offer powerful single board computers (SBCs) in popular board formats. VME has been a particularly successful standard within high energy physics. Presently, at least Motorola and CES are developing RISC-based VME SBCs. The Motorola module, for instance, will contain four 88100 processors with from 16 to 64 MB. On the SSC timescale, FUTUREBUS+ is likely to be the commercial bus standard for SBCs. Scalable Coherent Interface (SCI) may also be an approved and established standard on that timescale.

Industry has also become interested in large-scale applications of parallel processing for scientific computing in general. For instance, both IBM and Intel discuss multiprocessor systems with thousands of loosely coupled RISC-based nodes utilizing message passing in a two- or three-dimensional mesh. The Intel Touchstone Project funded by DARPA targets providing $\sim 10^5$ MIPs with 2^{11} i860 processors in a system by 1991. Systems with such high-performance nodes

require input/output bandwidths comparable to our needs. They may also provide the software environment which we need, and connections to host/control processors and to workstations.

Although the basic parameters, bandwidth and processing power largely define our farm requirements, another possible requirement often discussed is the need for an open architecture. A truly open architecture would allow one to exploit the most cost-effective microprocessor at the time of system implementation, instead of at the time of system design. This point of view is reinforced by the tendency to employ as much computing power as is available and the frequent need to expand computing power. The use of commercial products in the farm should conform to the standard of open architecture or should provide (guarantee?) a cost-effective solution to system upgrades and expansion.

2.3 Comparison between On-line and Off-line Farms

An on-line farm is quite similar to an off-line farm used for physics analysis and Monte Carlo event and detector simulation. As yet a systematic study of the similarities has not been performed. Input to the on-line farm is on high-speed data links to the DAQ system at potentially higher bandwidths (1 Gb/sec or 1 GB/sec) than an off-line farm, which is fed from data links to tape drives which operate at less than about 50 MB. Processing times for event selection online will be no longer than offline times for event reconstruction, and are likely to be somewhat shorter. Memory requirements per node offline and online should be roughly similar. Whereas more calibration constants may be required online, programs will be larger offline. Output bandwidth requirements per processor are less online than offline because only a fraction of event candidates processed online are output to mass storage. Overall processing power required online may exceed that needed offline because more event candidates are processed online than events are processed offline, even allowing for multiple offline passes through data. Additional processing power offline for Monte Carlo simulation may restore the balance of power between offline and online. Both on-line and

off-line farms have roughly similar needs for control monitoring, and collecting data (histograms) from nodes.

2.4 Timescale for Development

Although the on-line farm is not the only application of commercial computing products to the on-line environment which requires R&D, it is the application which requires the most development. We must define our requirements for bandwidth and processing power, but perhaps, most importantly, for software tools. We must monitor trends in the computer industry in order to choose at which level to apply commercial hardware, i.e., as chips, as single board computers, or as entire systems. We must also monitor trends in the computer industry to determine if it will provide the tools for the software environment which we need. If adequate software tools are not commercially available, then we must develop them. If major system software developments are necessary, then that will be the longest lead-time development for the farm.

If an open architecture is adopted, then the hardware development of the farm could conceivably be postponed until a small number of years before detector commissioning. Meanwhile, the eventual hardware development will benefit from experience gained by current and near-future experiments implementing farms of smaller scale, by development of large-scale farms for detector simulation and design, and by commercial large-scale multiprocessor development. These benefits would be enhanced by participation of developers of on-line farms in these other projects.

With an open architecture, development of software tools can precede the hardware design. In fact, the development of software tools may need to precede hardware development by about two years because of longer lead time.

If an open architecture is not adopted, then hardware development or selection must be done sufficiently early to afford time for software development.

Again, the time required will depend on the amount of software development required by the chosen architecture.