Fermi National Accelerator Laboratory

FERMILAB-Conf-90/79

Computing and Data Handling Recent Experiences at Fermilab and SLAC*

Peter S. Cooper
Fermi National Accelerator Laboratory
P.O. Box 500
Batavia, Illinois 60510

April 9, 1990

* Talk presented at the 1990 Conference on Computing in High Energy Physics, Santa Fe, New Mexico, April 9-13, 1990.

Operated by Universities Research Association Inc. under contract with the United States Department of Energy

COMPUTING AND DATA HANDLING
RECENT EXPERIENCES AT FERMILAB AND SLAC

April 9, 1990

Peter S. Cooper
Fermi National Accelerator Laboratory
PO Box 500, Batavia, IL 60510

ABSTRACT

Computing has become evermore central to the doing of high energy physics. There are now major second and third generation experiments for which the largest single cost is computing. At the same time the availability of "cheap" computing has made possible experiments which were previously considered infeasible. The result of this trend has been an explosion of computing and computing needs. I will review here the magnitude of the problem, as seen at Fermilab and SLAC, and the present methods for dealing with it. I will then undertake the dangerous assignment of projecting the needs and solutions forthcoming in the next few years at both laboratories. I will concentrate on the "offline" problem; the process of turning terabytes of data tapes into pages of physics journals.

INTRODUCTION

Computing has come a long way in high energy physics. The newly formed Computing Division at Fermilab (of which I am a member) commands 10% of the lab staff and more than 10% of the budget. Seventeen years ago we acquired our first "central" computer - as surplus from LBL! This history is well known, so I won't pursue it here. The remarkable fact remains that in a period when the cost of computing has literally dropped by orders of magnitude the fraction of available funds spent on computing and computing resources has grown substantially in high energy physics.

What I will concentrate on here is what this has lead us to. We face computing problems of enormous proportions. Furthermore, there is every reason the believe that the need for rapid growth will continue and accelerate. My topic is how we are handling these problems today and where they may take us tomorrow. I will focus on the "offline" half of the problem; the reduction and analysis of the data tapes. The scope of this paper will be my own laboratory (Fermilab) and SLAC. I wish to acknowledge and thank Chuck Dickens and Charlie Prescott of SLAC for re-educating me about SLAC's computing enterprise. The credit is theirs; the errors are mine.

WHAT'S THE JOB?

Let's begin with an overview of the problem. Most experiments today have three or more stages, or PASSes, in their analysis. PASS1 is a reconstruction of all, or at least most, of the events on a data tape. These jobs tend to be very CPU intensive. They also have the nasty habit of making the data set larger rather than smaller. Many experiments can't (or won't) throw much away at this stage, so they add the analysed quantities to the back end of the raw data and keep it all. PASS2 is typically a recompute and sort job. The inevitable last 10% of the main event analysis which either didn't make it into PASS1, or was done wrong and needs to be "better" gets done the second time through. Then, the various "physics streams" get sorted out in to as many as a dozen different overlapping data sets. Most of these are even smaller than the original! PASS3 is the "DST" (Data Summary Tape) stage where as many graduate students as there are on the experiment attempt to see how many tries it takes to remove the oxide from the magnetic recording material (either tape or disk) on which the DSTs have been stored. In the process they do physics, write papers and earn degrees. PASS3 may contain the production and processing of mini, micro and even nano DSTs. Formally these would be passes 4-6 but they all tend to have the same characteristics and so I will treat them here with PASS3 as a single entity.

PASS1's are rarely done more than once. No one can stomach another year - and a typical PASS1 takes at least that long. PASS2's, or at least the sort phase, may happen several times as people remake their DST's with better knowledge of their constants, cuts and algorithms. PASS3's and beyond get done regularly. The students will remake the mini and micro-DSTs every day if you let them.

These PASSes have significantly different characteristics as computing jobs. PASS1s are anywhere from highly to ridiculously compute bound. They tend to need mainly logical and integer arithmetic rather than floating point calculations. Floating point is required but MIPs are more important than MegaFLOPs. PASS2's have significant compute requirements (typically 10-20% of PASS1) and large I/O requirements as they try to sort thousands of input tapes into dozens of hundred tape piles. PASS3's and beyond tend to do large amounts of I/O; preferably from online storage. They also can be doing heavy calculations as the double precision matrices get inverted to do the complicated fits.

There are also the Monte Carlos in which people try to simulate at least there entire experiment (if not the entire world). These are completely CPU bound in and of themselves. Most groups take the sensible and prudent course of having the monte carlo write a data tape which they then feed to their PASS1-3 analysis chain re-invoking the entire monster once again.

Finally there is everything else - software development, code management, word processing, etc. This can be look on as "infrastructure" but, in fact, it is where most of the people spend most of their computing time. If you doubt how important all this is just try to change someone's favorite system for a "better" one.

To characterize the present loads and PASS1 requirements I have chosen experiments which have already taken data at each lab. These experiments (shown in Table 1.) are of different characters and each represents, in the case of Fermilab, a class of similar experiments. I've also estimated the total lab wide load for the 1988-1990 time period. At Fermilab this included the last fixed target run in which 16 experiments took data and the last collider run in which CDF and two smaller experiments ran. At SLAC this period saw the analysis of Mark III data from SPEAR, PEP data from several detectors and Mark II data from the first runs of the SLC.

EXPERIMENT	E731	E687	CDF	E769	FNAL Total	Mark II	SLAC Total
To Tape							
events/sec	300	100	1	500		2	
bytes/event	500	2.5K	100K	3K		35K	
total events	1500M	60M	7.5M	500M		5M	
Media							
type	9trk	9trk	9trk/8mm	9trk	9trk	3480	3480
number	5K	1K	5K/1K	10K	40K	800	2K
total bytes [Tb]	0.75	0.15	0.75	1.5	6	0.15	0.4
Reconstruction							
MIP-sec/event	0.1	15	200	20		15	
Instructions/byte	100	2400	750	2800	1500	1000	1000
Total [MIP-years]	5	10	50	320	600	2	5

Physics	Expt.	Spokesman
CP Violation in Ko decays	E731	Winstein
Photo-production of Charm and Beauty	E687	Butler
Collider Detector at Fermilab Pbar + P at $\sqrt{S}$ = 1.8 TeV	CDF	Shochet/ Tollestrup
Hadro-production of Charm	E769	Appel
Mark II at the SLC Physics at the Z pole	Mark II	Goldhaber/ Dorfan/ Feldman

Table 1. Typical Present Experiments

The units of Table 1 require comment. I've tried to present these data in a platform and media independent manner. MIPs are in the standard units used at Fermilab (1 MIP = a VAX 11/780). The measure of CPU boundedness is "Instructions/byte". This is the number of instructions executed per byte of data I/O (paging, swapping, constants, etc not included). Note that a job which is CPU bound on a 1 MIP VAX 11/780 at 50 instructions/byte will be I/O

bound on a 100 MIP processor (or farm of processors). This metric is a useful way to discuss the potential performance of the same job on different platforms.

HOW ARE WE DOING IT NOW?

The presently installed central computing capacities are shown for SLAC and Fermilab in tables 2 and 3 respectively. As the table shows, SLAC has about 150 MIPs of other computing including VAXes at experiments, workstations on peoples desks, CAD/CAM systems for engineering, etc. Fermilab is proportionately larger with 500-800 such distributed MIPs. The Fermilab table contains only the "central" computing systems. It doesn't include the distributed systems for lack of space. The Fermilab reorganization of computing has recognized this reality. One of the five departments in the new computing division is "Distributed Computing" with responsibility for the systems aspects of the central VAX clusters as well as network, hardware maintenance and the other functions necessary to keep such a large and far flung computing enterprise working.

Oct-89

SYSTEM	CPU [mips]	Tape Drives 9 trk T3480 8mm	Tape I/O [Mb/sec]	Disk [Gb]
CENTRAL COMPUTING				
IBM Systems	70	12	6.0	80
IBM 3090-200E	50			
IBM 3081K	20			
2-STK Robots		16	12.0	
DISTRIBUTED COMPUTING				
DEC Vaxes	44.2			28
1- VAX 8810	5			
2-VAX 8800	20			
1-VAX 8600	4			
2- VAX 785	3			
11- VAX 780	11			
2- VAX 750	1.2			
Workstations	71			
47-VS2000	47			
3-VS3200	15			
5-VS31000	9			

Table 2. SLAC Systems

SLAC's central computing is a homogenous one vendor shop; a style very consistent with their present needs. Their emphasis is on having all the data available all the time. They have concentrated heavily on robotics for tape mounting and run a "lights out" computer room operation. They currently have two STK 3480 tape robots with a total capacity of 2.4 Tb online and available with "seek" times of less than 20 seconds and transfer rates of 3 Mb/sec.

Fermilab is a collection of four major types of computing based upon the "pawnbroker" computing model (fig 1) the basis for which can be found in the Ballam committee report (ref 1) of 1983. The Fermilab implementation of this model has two very large VAX

clusters running VAX/VMS playing the role of the interactive front-end systems where program development, mail, word processing and the like are done. Typical loads are 400-450 users logged on in the afternoon. The numeric intensive "number cruncher" is the Amdahl 5890-600E system running VM/XA. This system has been in operation for just about two years. It is now heavily used for PASS2 and PASS3 analyses. It is mainly a batch engine with most job submissions coming from a few logged on interactive users or from the VAX clusters via a DECnet connection - as the model envisioned.

Apr-90

SYSTEM	CPU [mips]	Tape Drives 9 trk	T3480	8mm	Tape I/O [Mb/sec]	Disk [Gb]
CYBER 875 - gone 9/30/90	30	22			8.8	30
IBM 4381 - Business Systems	10	4			0.8	15
FNAL Cluster	76	8			6.0	85
FNALF 8830	18					
FNALC 8650	6					
FNALB 8650	6					
FNALG 11/780	1					
15-MV3100	45					
FNALD Cluster	109	12		12	12.0	50
FNALH 8800	12					
FNALI 8820	12					
FNALJ 8820	12					
FNALE 8600	4					
3-VS3200	9			12		
20-VS3100	60					
Amdahl 5890/600E	120	16	8	2	22.3	90
STK Robot			8		12.0	
ACP Farms	364	16		6	5.4	10.3
FNACP3 52 nodes	37	3			0.9	1.6
FNACP7 65 nodes	55	2			0.9	2.6
FNACP9 60 nodes	43	3			0.9	1.5
FNACPA 65 nodes	55	2		3	0.9	2.2
FNACPB 112 nodes	87	3			0.9	1.5
FNACPC 102 nodes	87	3		3	0.9	0.9
Silicon Graphics	640			27	3.3	15.3
4-4D240 1 is Physics Dept's	272			18	2.0	4.8
25-4D25S	300			8	1.0	6.9
1-4D240 Power Series	68			1	0.3	3.6

SYSTEM	CPU	Tape Drives			Tape I/O	Disk
CYBER 875 - gone 9/30/90	30	22	0	0	8.8	30
IBM 4381 - Business Systems	10	4	0	0	0.8	15
FNAL Cluster	76	8	0	0	6.0	85
FNALD Cluster	109	12	0	12	12.0	50
Amdahl 5890/600E	120	16	16	2	34.3	90
ACP Farms	364	16	0	6	5.4	10
Silicon Graphics	640	0	0	27	3.3	15
Fermilab TOTAL	1349	78	16	47	70	296

Table 3. Fermilab Systems

The third ball of the pawnbroker is "farms". These are collections of processors running "stable" computing bound production jobs (read PASS1). The processing is done in parallel with one event, or more typically one block of events, sent to each of up to 100 processors with the results of the calculation fetched

and written back to tape. CPU powers of up to 100 MIPs are achieved with parallelism operating on a time scale of minutes rather than nanoseconds. There are two types at the moment; ACP and RISC/UNIX systems. CDF also built a farm of 20 VAXstation 3100's to help with the completion of their PASS1. These systems have now been absorbed into one of the central VAX clusters (FNALD). The ACP farms are made of single board computers based on Motorola 68020 chips after a Fermilab design (ref 2,3). The host which controls the farm and does all the I/O is a MicroVAX 3200. These six systems did the lions share of the PASS1 from the previous round of experiments - both fixed target and collider. The RISC/UNIX boxes, presently mainly Silicon Graphics 4D/240 systems are in operation about 9 months. This class of computing clearly represents the "modern farm" on which the next round of experiments will do their PASS1's. The only tape I/O in use at the moment on these systems is 8mm. This is quite consistent with their future use at Fermilab.

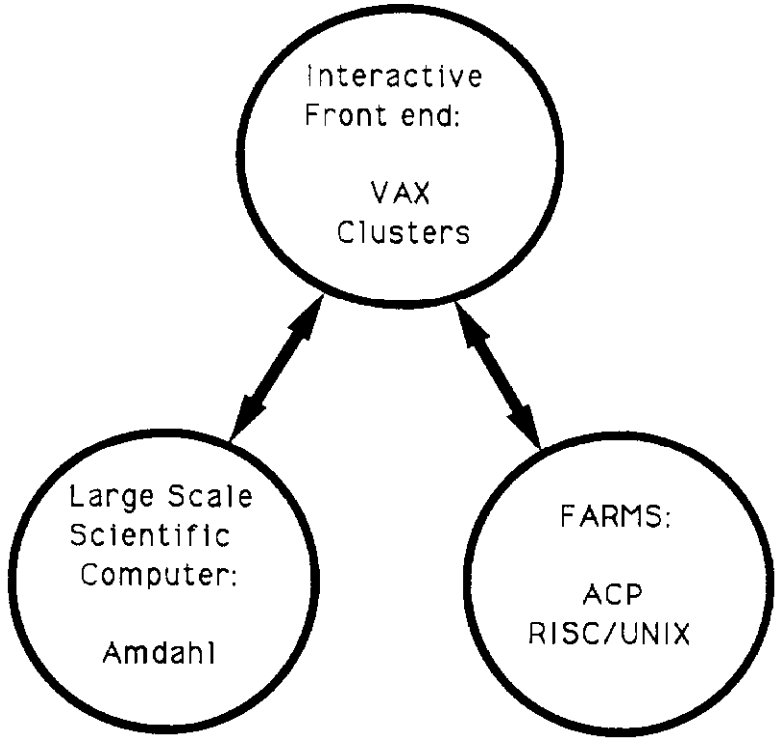

Figure 1 - The Fermilab Model of Computing

The growth profile of installed central MIPs at Fermilab is shown in figure 2. The trend is clear; in the 18 month period from the end of the previous fixed target run (April 1988) until the beginning of the new computing division (October 1989) installed MIPs doubled every six months! After a brief respite to catch our breath and reorganize we are off again on the exponential slope.

At SLAC the total load of PASS1 analysis is tractable. Electron colliders are cursed and blessed with the low event rates that the small cross-section provides. At Fermilab the total load

is dominated by the fixed target experiments - indeed half of it is from one experiment. The cross-section is for all practical purposes infinite. The problem is how to dig out the interesting signal. Some of this is done with fast triggers and online processors which are the topics of other papers. There still remains the offline part of the job. In the case of E769 which deliberately chose to record 1/4 of the inelastic cross section on tape, the rejection required to reduce their 500 million triggers to 10,000 reconstructed charm events is of order 1/50,000 or about one good event per raw 9 track data tape. The other experiments are not quite so extreme but collectively they sum to just as big a job.

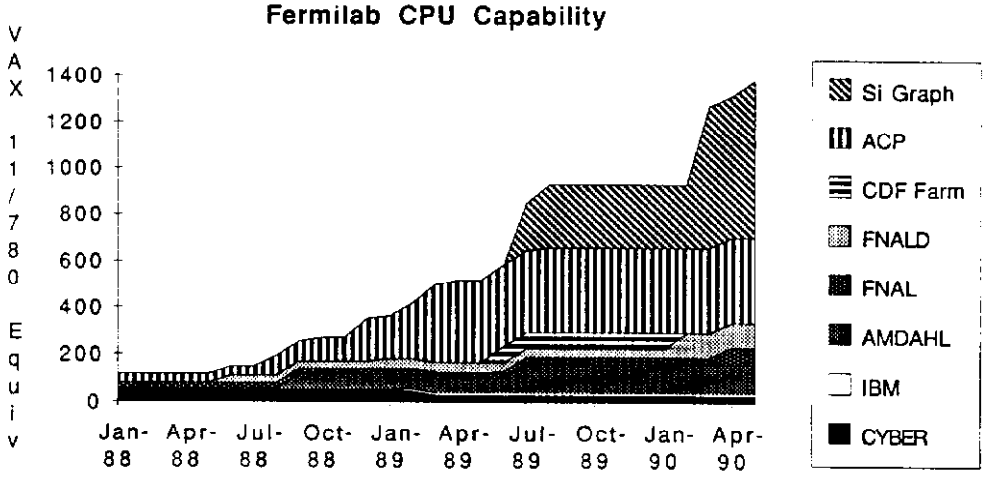

Figure 2 - Fermilab CPU Capability

The Pass2 load is an acute rather that a chronic problem. There is a substantial amount of tape handling but most experiments can get through this in a few months. A major new problem is multiple media. With 9 track, 3480/18 track and 8mm tapes all in active use at Fermilab presently, it always seems that the data is on the wrong medium for whatever needs to be done next.

The PASS3 loads are just seriously beginning at Fermilab from the previous data (and we have just embarked on another fixed target run in mid February). These jobs tend toward very heavy I/O usage; both tape and disk. SLAC solves this problem elegantly with their robots and homogenous system. At Fermilab with four different operating systems, three different kinds of tape media and 10 times the total data volume things are harder.

Figures 3 and 4 shown the pattern of delivered CPU and tape mounts at Fermilab over the past year. All but a few hundred of the 44,000 tape mounts last month were done by hand by operators. This activity requires a staff of 25 at a cost of just under 1 M$ per year. Our first big STK robot is scheduled for installation the week of this conference. As the growth curves clearly imply we are going to automate or die.

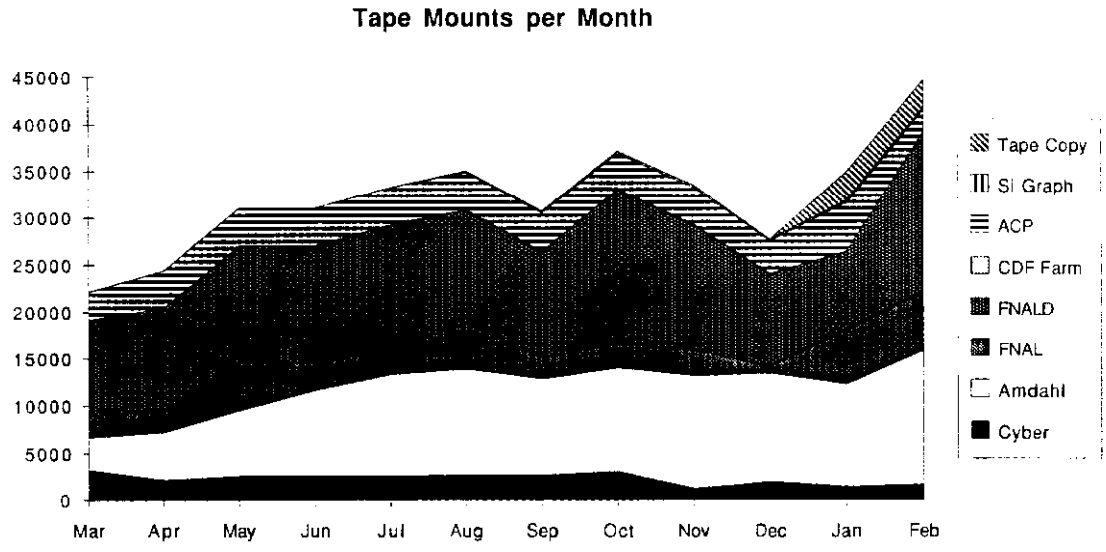

Figure 3 - Fermilab Tape usage per Month

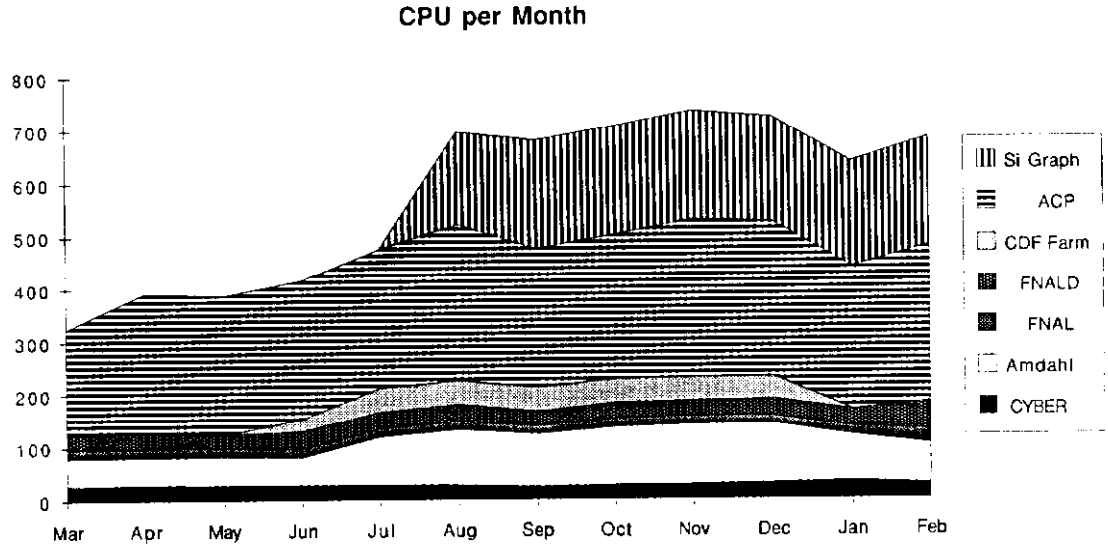

Figure 4 - Fermilab CPU usage per Month

SO WHAT'S THE PROBLEM?

In Table 4 I show a similar selection of experiments scheduled to be analysing their data in the next 2-3 years. These are analogous to those in table 1; in fact most are either the next run of the same experiment or the direct successor to a table 1 experiment. Data volumes are up a factor of 5-7 at both SLAC and Fermilab. The detectors are, if anything, more complicated so compute times grow somewhat. SLAC will need about a factor of two increase in CPU with proportionate growth in peripherals. At

Fermilab we are looking at growth of a factor of 5 in total capability, from 1000 to 5000 MIPs in terms of CPU, over this three year period. (Recall that the table shows only the PASS1 requirements.) The assumption made to arrive at these factors is that an experiment should be able to complete its PASS1 in of order one year. Any longer than that and last run's data (from two years ago) won't be done before this years new data arrives.

EXPERIMENT	E761	E687	CDF	E791	FNAL Total	SLD	SLAC Total
To Tape							
events/sec	1000	500	5	8000		1.5	
bytes/event	500	3K	150K	3K		25K	
total events	1200M	600M	40M	7000M		16M	
Media							
type	9trk	8mm	8mm	8mm	8mm	3480	3480
number	4K	1K	3K	10K	20K	5.5K	10K
total bytes [Tb]	0.6	2	6	20	40	1.1	2
Reconstruction							
MIP-sec/event	0.1	15	300	6		15	
Instructions/byte	200	2000	3000	2000	2000	140	140
Total [MIP-years]	4	320	375	1300	2500	10	20

Physics	Expt.	Spokesman
Hyperon Radiative Decays	E761	Vorobyov
Photo-production of Charm and Beauty	E687	Butler
Collider Detector at Fermilab Pbar + P at $\sqrt{S}$ = 1.8 TeV	CDF	Shochet/ Tollestrup
High statistics Hadro-production of Charm	E791	Appel
The SLD Detector at the SLC Physics at the Z pole	SLD	Baltay/ Breidenbach

Table 4. Typical Future Experiments

Its hard to change these estimates very much. At SLAC they will be driven almost completely by the lumonisity that the SLC can achieve. This effectively bounds the requirements from above. At Fermilab most of the data is coming from existing experiments with well known characteristics and relatively mature analysis codes. The physics goals dictate the requirement of high statistics and that in turn drives the computing load. Of course, estimates are always low so thing may be even worse.

FUTURE DIRECTIONS

So how are we to respond to continuing exponential growth? The first answer to this question is to plan - but not for too long! Both SLAC and Fermilab now have in progress committees charged with making plans for the next five years or so. At the coarsest level the goals of the two groups are quite similar. The basic end result

desired is a plan for incorporating the latest and most cost effective new technologies: workstations, RISC compute servers, UNIX, advanced networking, intelligent file serving, etc. At the next level the plans must necessarily diverge as they address the different situations at the two labs.

SLAC will likely wish to augment their present central facility with both more mainframes and at the same time begin to include compute servers, workstations and advanced networking (FDDI). Their real needs will be driven by the SLD data rate. This is highly uncertain at this time so I can't begin to guess the scope of what they really need to do. If it is only the factor of two I projected above then things are probably fairly straightforward. Their present model of computing can clearly be scaled a factor of two. If the real needs are much more than that they may need some more radical approach.

Fermilab does not share this luxury. Our present model of computing is just not capable of what we now require of it. It has served us relatively well for the past 5-7 years but it's earned it's gold watch. So what is the new model of Fermilab computing to be? This is a question we are actively involved in answering right now so I can't just write down the solution. Some pieces are quite clear others almost completely opaque. The overall organizing principle - the picture that replaces the "pawnbroker" is in the latter category.

Two thing are very clear. The medium of choice for data recording and the output of PASS1 analysis will be 8mm tape or some equivalent "cheap" recording medium. This is a decision driven strictly by economics. 40 terabytes of raw data plus a factor of 1.5 for PASS1 output (100 Tb total) would cost 5 - 10 M$ just for the media if we used either 3480 or 9 track tapes. There is also the small problem of what you do with 500,000 tapes. The media cost for 8mm is a factor of 10 less as is the number of individual volumes which have to be handled and stored. Jack Pfister will have significantly more to say on this subject in another talk at this conference (ref 4).

Likewise, the only apparent affordable compute engines for PASS1 processing are in the RISC/UNIX compute server class. These systems sell today for as low as $500/MIP barebones and about $1000/MIP with enough peripherals, software and maintenance to make them functional PASS1 compute engines. With PASS1 jobs running at more than 1000 instructions per byte either a few 8mm tape drives on each system or a network connection to an "I/O server" system provides adequate I/O capability for PASS1 work. How do organize and manage these new farms is more problematical. We have significant experience with managing ACP farms and are now in the process of adapting these techniques and software to these new systems (ref 5).

The hard part, at the moment, looks like the PASS3's. Understanding the real requirements for I/O, CPU, shared data, online vs almost online vs offline data storage, etc. is a challenging undertaking. Some will argue that the whole job can be done with a "loosely coupled" system of distributed workstations/compute servers with only a network to tie them together for common file access, backup, etc. Others favor a more monolithic architecture with a large "central file server" at the center and arrays of RISC/UNIX compute servers hanging off the networks at the back end.

Neither of the two descriptions I just gave are models of computing. They give, perhaps, a flavor for the kind of issues with which we are wrestling. Trying to sort this all out makes for a lively set of meetings and discussions. I truly wish that I could report the beginnings of the answer to you here at this conference - but I can't. If we can't begin to answer these questions in about six months you can visit our tomb. It will be just outside the new Feynman Computing Center at Fermilab and it will be constructed of unanalysed data tapes! Khruschev's old admonition of the sixties has taken on a new meaning ("We will bury you").

REFERENCES

1. "Future Computing Needs for Fermilab" Fermilab TM-1230 0062.000 December 1983 (unpublished)

2. "The ACP Multiprocessor System at Fermilab",I. Gaines, et.al, Comput.Phys.Commun 45,323(1987)

3. "Software for the ACP Multiprocessor System",J.Biel, et.al, Comput.Phys.Commun 45,331(1987)

4. "Recent Experiences and Future Expectations in Data Storage Technology",Jack Pfister, elsewhere in these proceedings.

5. "The Cooperative Processes Software Method for Multiprocessing Computing",C. Kaliher's,elsewhere in these proceedings.