



**Fermi National Accelerator Laboratory**

**FERMILAB-Conf-91/87**

## **Computing at Fermilab**

**Thomas Nash**  
*Fermi National Accelerator Laboratory*  
*P.O. Box 500*  
*Batavia, Illinois 60510*

**March 1991**

\* Presented at Computing and High Energy Physics Conference, KEK, Tsukuba, Japan,  
March 10-15, 1991.



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

## **Computing at Fermilab**

**Presented by**

**Thomas Nash**

**Fermi National Accelerator Laboratory  
Batavia, IL 60510 USA**

Through a long history, advances in the basic understanding of elementary particles and their forces have paralleled advances in technology. The technology made it possible to overcome existing barriers to experimental progress. Arguably, the dominant barriers to progress in experimental high energy physics are now associated with data handling and computing. Even in theoretical physics this has become the case in an important area of work. The Quantum Chromodynamics (QCD) theory of the strong interaction can only be calculated numerically using the Monte Carlo relaxation approach known as lattice gauge theory.

Both experiment and theory are computer technology limited: no one can identify a "requirement" on computing or data capacity that is independent of cost or other realities. This is not a result of greed. More computing/data capacity simply means that more science could be done, and so cost is the primary limiter. High energy physics (HEP) has been forced, therefore, to turn significant attention and resources to finding extremely cost effective solutions to its computing using whatever technology is available. This is goal driven computer science, integrating commercial solutions at the chip, board, and system level.

Fermilab's program centers on high luminosity fixed target and collider physics which intrinsically produces large numbers of complex events. Combine this with the large and broad based user community that Fermilab supports, and it is no wonder that computing demands are the largest. The data taking requirements of experiments at Fermilab are now discussed in units of tens of Terabytes. Computing to reconstruct raw electronic signals from ADCs and TDCs into physics parameters preparatory to the physics analysis is counted in units of a thousand VAX 11/780 years -- VAX Unit of Performance: VUPs. Nowadays, this type of computing is carried out on parallel farms of RISC servers reading 8 mm tapes. The tapes are written on-line by parallel "walls" of as many as 40 Exabyte 8 mm drives. The reconstruction software for a large experiment is prepared by as many as 100 physicists, at 30-50 institutions. Code size has passed 2 million source lines of code (MSLOCs) at the Collider Detector at Fermilab (CDF). Even larger numbers of physicists on individual experiments are involved in writing the analysis software and working with data to produce physics results. These large analysis efforts present new challenges in terms of data service and networking.

## **A Plan for Computing at Fermilab in the 90s**

On October 1, 1989 a new Computing Division was formed at Fermilab to address the increasing computing demands of the 90s. The emphasis was on communication with its clients, a new openness, and a conviction that technical agility was going to be the new order of the day. The fundamental quandary confronting the new organization was recognized as being the apparent contradiction between the demands of the physics for the latest technological tools and the requirement, driven by the huge scale of modern experiments, that these be brought to bear in an operationally smooth manner.

An ongoing program of "Strategy" meetings held on almost a weekly basis has covered a wide variety of topics. Recent meetings focussed on such topics as data bases, GOSIP, a Fermilab Fortran standard, 8 mm tapes, Unix support, software license tracking, and AIX. The most important series of meetings to date were held in spring of 1990 to develop a new "model" for Fermilab computing. As is the case for all strategy meetings, each of the 10 meetings was open and attended by over 40 people. A third of the attendance came from the community of computer users (CDF, D0, fixed target, engineering, etc.), a third from the Computing Division itself, and a third represented strong commercial participation (including Amdahl, DEC, IBM, SGI, and Sun).

The previous model of computing at Fermilab was developed around 1983. It has been referred to as the "pawn broker" model because it was drawn in terms of three balls, representing the three components of Fermilab computing it envisioned for the 80s. These were understood as a front end (VAX cluster), a general purpose batch engine (which was implemented as an Amdahl mainframe), and "a production engine for stable codes" (the ACP farms of micro processors, microVAX farms, and now Unix RISC farms). This picture served the Laboratory for most of the decade. One part of the picture was never implemented: a fourth ball at the center of the other three labelled "central file server". Based on prevailing technology predictions at the time which promised such capabilities soon, it was to allow transparent data access from all three or more flavors of computers. This kind of file service, now required on a more dedicated scale, remains one of our key needs.

The 1990 strategy team focussed on computing at the analysis stage. It was assumed, with general agreement, that the reconstruction phase is now being effectively treated using farms of Unix servers based on workstation technology (as we will describe later). Debate during these meeting was often intense. Sides formed around "big boxes" versus "small boxes", centralized versus anarchy. In the end a synthesis was developed that stepped back from the deep end toward the concept of dedicated workgroup computing. Using cost effective workstation technology as much as possible, the plan is to continue to centralize commonly needed support and high performance batch processing (farms). The key new element is to develop data and analysis servers to meet the specific needs of individual large groups or classes of users. These servers will be logically, and in many cases physically, local. Other conclusions were to focus on two operating systems VMS and Unix (down from the 5 that were being supported: Cyber-NOS, VM, ACP, VMS, Unix), and to put a heavy emphasis on robotics and on a reconfigurable, high capacity (FDDI at first) hub and spoke network based on fiber links to regional centers.

## **Hi Performance Batch Computing for Experiments: The Agriculture Business**

The technological approach to reconstructing raw data, both on and off-line, is now well understood in terms of using farms of small commercial computers picked from a market place where cost effectiveness is steadily improving. In operation at Fermilab now are something like 600 modules of the original Advanced Computer Program's ACP 1 farm design based on the Motorola 68020. The management issue has become how to "un-user" them. This phrase resulted from the recent two year effort to retire Fermilab's Cyber 875s. The last 875 went out the window, literally, by crane from the 8th floor last September, despite the entreaties of diehard users. A similar painful retirement of the ACP 1 systems is anticipated.

A second ACP effort developed a module based on the MIPS R3000. This project introduced RISC and the excellent MIPS compiler to the HEP community. The project also developed a basic farm support operating system called Cooperating Processes Software (CPS) which operates in Unix and VMS and which is ported to all types of farms at Fermilab and elsewhere (such as the SSCL simulation farm). It allows each node to communicate from its user program with any other node or I/O channel available to the farm.

It has now become more economical to build (integrate) farms out of commercial servers, most of which use RISC microprocessors. The day of home built processor modules built out of microprocessor and DRAM chips appears to be over, for off line computing at least. A microVAX farm at CDF was the first to be integrated out of commercial computers. Now the major new workhorses are Unix RISC farms. At this writing about 575 VUPs worth of farms based on Silicon Graphics, Inc. (SGI) servers are running, and the procurement of a 1750 VUP farm based on IBM RS 6000 servers is complete. This farm will be installed in May. (VUPS quoted here are as measured on Fermilab physics benchmarks and are 20-30% lower than vendor claims in most cases.) We have budgeted for an additional 1500 VUP farm for this fiscal year (ending in September). The competition in the recent open procurement was extremely severe.

### **Analysis of Data: Unsolved Problems**

Development of software and analysis of reconstructed data by large collaborations has for some years been carried out on huge VAX clusters. Fermilab operates two of the largest VAX clusters in the world. As the size of the software and the volume of data increases, this approach is coming under strong pressure. It is simply not financially possible to keep 50 TBytes -- or even one TByte -- of data spinning on rotating magnetic media. Other problems include systems and operation management of these centralized clusters. Presently, we are trying to find solutions that conceptually, at least, decentralize analysis and software development computing and focus it on the needs of a compatible community. For the large collider experiments, with collaborations of several hundred scientists each, such systems will consist of VMS/Unix compatible (we hope!) compute and file/data servers that use robotic and hierarchical storage techniques. All of this must be accessible by network to collaborators in the US, the far east, and Europe.

The traditional approach to supporting data analysis was to use mainframes. We still cannot afford to live without their environment of shared programs, shared data, high bandwidth I/O, and centralized tape handling. But we cannot afford to pay for increasing their capacity to meet the increased needs. VAX cluster upgrades cost \$650K a crack, and the complaining and saturation continue as if nothing was added.

The concept of a dedicated work group cluster is not new: the FNALD VAX cluster has been dedicated to CDF for some time and has been used for analysis. The wish is to provide such services with the increased kVUP and TByte demands of today, within an imaginable budget. Figure 1 shows, from the FDDI network perspective, what was meant by the strategy team in defining work group computer clusters. The distinction is that they are dedicated to one group or class of users. The system may all be located at the computer center (as is FNALD) or it may all be at a remote location (accelerator theory, engineering clusters), or, as will be the case for the collider experiments, it may be logically local, but physically distributed around the site. For sure, those parts of workgroup clusters that involve manual tape handling by operators will be located in the center, and it is likely that major robotic systems will be there as well, as will all farms, even if dedicated to one experiment for an extended time.

A study group has just been formed to define FY91-92 work group cluster requirements for CDF, and then, D0. The result is expected to be initial procurements for both experiments this summer. The group will try to merge requirements with affordable technology for analysis compute servers, data/file servers (spinning, robotic, etc.), networking (local and to collaboration), and distribution of data on media to the collaboration. In a few months we should start to know what a modern work group computer really looks like - - and whether the time is finally here for transparent Unix/VMS file service, the dream of the 1983 model, but now in a dedicated "center".

### **Towards an Open Environment: Unix**

In another presentation to this conference, Joel Butler describes the pleasures and perils of Unix. Yes, Unix is essential for an open and highly competitive and cost effective environment. But, which Unix did you mean? There is a multidimensional matrix of incompatibility that one has to deal with. It is not as bad as Unix vs VMS, not by a long shot, but with limited resources we cannot support every Unix system at a full level. An important conclusion of the 1990 strategy plan was to limit strong support to 2 - 3 Unix vendor lines. One is SGI, with farms and many users already in place. A second was to be the winner of the major '91 farm acquisition, now known to be the IBM RS 6000. A possible third will be chosen for features relevant to use on the desktop (graphics, human interface, etc.). The expectation was that we would possibly consider adding and/or subtracting one fully supported line to this list in future years. Basic support of other Unix systems will continue to be provided to allow evaluation of new products and improvements.

Fermilab gained experience with Unix in a rapid burst of effort over the last year. In addition to recruiting some first class Unix gurus and system managers, there were two crash Unix support efforts that were designed to develop experience much in the spirit of getting a new experiment up. One was a project to seed Unix workstations around the laboratory. Some 26

workstations of 4 flavors (DEC, Sun, SGI, IBM) were made available to 9 experiments (and the astrophysicists). This required a huge effort to support everything from reading 8 mm to running CERNLib and PAW on these systems. A second crash program was Fermilab's support of computing at the Division of Particles and Fields 1990 Summer Study on the mountain top at Snowmass, Colorado. Along with 20 VAX stations and 12 Macs, this project introduced 19 Unix workstations (SGI, Sun, DEC), all on a common network. The experience gained from these two efforts has been invaluable in developing a basis for supporting Unix at Fermilab. A growing package of Unix applications, tools and utilities tested on supported platforms is developing into a Fermilab Unix Library - - FUNLib - - which will be made available to the HEP community.

### **Support Required for Distributed Computing: System Management, User Support, MIS and Networks**

The heterogeneous distributed computing environment poses its most serious problem in support areas. The number of support personnel simply cannot scale with the number of nodes - - or even with the number of vendor flavors. Not only is this true for central computing support organizations, where budgets for this sort of thing are visible and well contained, but it is also true at the local level where the costs tend to be hidden. Not every user of a workstation wants to be - - nor should they be - - a Unix expert and system manager. Many local departments, each with private system management summed over a large laboratory, adds up to a significant total system support effort. Ideally, local system support should only be required at a basic level for such activities as booting systems and backing up data (if that is to be done locally). However, central resources, unless there are great advances in the tools presently available to do the job, are not ever going to be adequate to provide the level of service that would relieve most local departments and groups of significant responsibility for their systems.

Central support for distributed computing at Fermilab now involves at least 20 people full time (plus an estimated additional dozen from local departments). About 37 clusters with almost 400 nodes are supported at various levels. The planning at the departmental level sometimes takes into account the central and local support that is required - - and sometimes it does not: one pathological example with 19 nodes includes 7 flavors of Unix operating system. Centralized support includes many activities and functions. The following list is incomplete, but instructive:

Operating system: installation - upgrades - patches

OS related trouble shooting (often involving incompatibilities): peripherals - drivers - utilities-applications

Local system manager education & consulting

General user consulting and product support: distribution - documentation

Network : installation - upgrades - trouble shooting - address management

File System: maintenance - trouble shooting - backup

Distributed peripherals - - mainly printers in many flavors

## Technology tracking

Hardware and software inventory and management of maintenance agreements and licenses

The last item on the list represents one of several acute areas where MIS like data base tools have become a necessity to manage a complex distributed computing environment. A Computing Division "MIS Project" is being strongly staffed to address these needs and give us a chance to better understand what is being provided and supported and used, by whom, for whom, and at what cost.

Computers and peripherals have become like commodities in terms of price competition and profit margin. In such a market there may seem to be little place for those firms that formerly provided the "warm and fuzzy" environment of main frames and mini computers to exercise their experience and strengths. However, there is a big need for tools for centralized support of a *heterogeneous* distributed environment. This is crying for attention and is likely to be profitable since the need is so great that many will be willing to pay for it -- handsomely, if the product is right. There is some reason to be optimistic that this has been recognized, but it appears to be several years before real help will appear. Here, once again, high energy physics is pushing the technology and industry, driven by the goal of moving its science forward. Now this is happening in an era of large and distributed collaborations needing extensive distributed computing.

Essential to distributed computing is a reliable and effective network. Dedicated workgroup computer clusters are likely to be spread over several sites, as shown in Figure 1, with part of the system in the center for operator access and for flexibility in allocating resources among workgroups. The individuals in the workgroups themselves are often located in several office clusters. These are some of the factors that demand more bandwidth and motivate a strong emphasis on integrating new network technologies and effective topologies into the Fermilab network. FDDI has been selected as the main path to upgrade the system over the next two years. After that it may be necessary to consider even higher bandwidth protocols in certain areas, and possibly bringing fiber into every office. The latter would be driven either by a new emphasis on visualisation in HEP analysis or by increasing amounts of workstation capacity sitting on desktops that could be harnessed as a shared resource. In this context we do not yet understand whether desks are going to be populated more with high capacity workstations or dedicated screens (like X-terminals) accessing compute servers.

## **Research and Development for HEP Experiment Computing**

Research and development in HEP computing has traditionally focussed heavily on hardware development of special purpose processor systems and, in recent years, on parallel farms of microprocessors. Given the availability of highly cost effective commercial systems, and the willingness of vendors to integrate them into farms and other systems catered to the needs of HEP, it appears that the era of developing your own off line computers is ending.

The demand for R&D appears to have shifted toward software. We have identified two areas to emphasize. Analysis of data is the intensive and demanding activity that leads directly to the end purpose of experiments: physics results. As we improve the turn around time for passing through large data sets, we need also to increase the machine to human brain bandwidth with improved statistical visualisation techniques integrated into a common, all encompassing, and effective graphical user interface (GUI) accessing all analysis, data access, and software tools. A small group was formed to work in this area over the last year, and it has explored available GUIs and possible approaches. They are beginning to develop a GUI library, first with a Motif implementation of an interface for Isajet. Motif is being used for the present because of its portability. However, compared to NeXT Step, with it's broad, reusable object oriented tool kit, Motif is technologically weak. At least in the opinion of this writer, it is questionable whether Motif is a strong enough platform on which to develop an interface that crosses that non-linear acceptance threshold where everyone will jump to use it because it so obviously increases productivity. Nonetheless, because of the lack of better GUI portable standards, Motif is likely to be the standard for GUIs in HEP for the next few years.

A second area of research also demands attention: as collaboration counts push towards 1000 participants and large experiment software package pass 2 million source lines of code (MSLOCs), high energy physics is going to have to pay attention to modern ideas in software engineering. This refers in part to computer aided software engineering (CASE) tools, of course, but it means much more than that. Object oriented software, for example, and project management methodologies can increase productivity and reduce error rates. No huge improvements have been demonstrated anywhere (and shouldn't be promised for the future), but every little bit counts in this critical area. In HEP the research questions would be directed at refining and extending what has been learned in other contexts to our sociology and personalities, and to the flexibility and agility demanded in the short time constant environment of dynamic basic research on large experiments. Advanced software engineering research has the potential to address what may prove to be one of the most difficult technological barriers of the next generation of experiments, and HEP may once again be a pathfinder in an important technological area. We hope to be able to describe progress in this area at future conferences.

### **Fermilab Computing in 1991: A Quantitative Review**

In terms of raw computer power, Fermilab capacity has dramatically increased recently. Figure 2 plots total central capacity versus time in terms of Fermilab VAX Units of Performance (VUPs) based on HEP benchmarks. VUP ratings tend to be some 20-30% below manufacturer MIPs ratings. With the acquisition of a 1750 VUP Unix workstation based farm from IBM to be installed in May 1991 (Farm '91), the total capacity will pass 3000 VUPs. Another 1500 VUP acquisition (Farm 91.5) has been budgeted for this fiscal year (ends 9/30). Prior to this year over 600 VUPs of Silicon Graphics based farms were installed. A large capacity of computing also exists outside the center and will grow considerably in the new emphasis on distributed computing. We estimate Fermilab's total capacity (exclusive of PCs, MACs, and controls systems) will be about 6500 VUPs after the IBM farm installation. In

terms of floating point the total capacity lab wide will be 7200 MFLOPS of which 5000 MFLOPS belong to the lattice gauge processor discussed below.

We cannot live by MIPS alone, MBytes are essential also. Total memory in computers at Fermilab is about 10,000 Mbytes (central) and 14,500 MBytes (lab wide). Disk capacity is 420 GBytes (central) and over 600 Gbytes lab wide. In the computer center there are almost 200 tape drives (44 9 tracks, 138 8 mm, 12 3480 format including the 1 TB STK robot). There are over 230 known 8 mm tape drives throughout the lab. In peak months the number of tape mounts has passed 40,000 (almost 1 mount per minute), though this has fallen off recently with increased use of the robot and a temporarily lighter load.

### **Quarks, Galaxies, & Stars I: Lattice Gauge Calculations**

As we noted earlier, although the theoretical lattice gauge calculation is very different from experimental data reconstruction and analysis, it is also very successfully addressed using explicit parallel approaches and goal directed integration of large systems. Fermilab has developed a grid oriented parallel computer that is now running physics at 5 GFlops (peak). This machine is presently based on 256 processors using the Weitek XL chip set. The connectability (at 20 Mbytes/sec/channel) is denser than a hypercube. Cross bar switch back plane crates each contain about 8 processors. The crates are arranged in fully connected planes of 9 crates. The full system contains 4 planes connected to each other at the 9 points of the plane.

For such a machine to be truly productive, it is essential to develop software that makes the architecture of the powerful parallel computer transparent to its scientific users. Programming is in C, with explicit parallelism directives supported by CANOPY, a top level language that allows physicists to think in terms of sites, and fields on sites, which are then automatically mapped onto whatever hardware structure is being used. CANOPY has been ported to many platforms and is becoming a *lingua franca* of lattice gauge physics. Its applicability is broadly to all grid oriented problems. CANOPY has the same relationship to grid oriented computing that CPS has to farms: it allows the scientific user easy, transparent, access to parallel hardware.

Test versions of a new processor module have been running since early this year. This new module contains two Intel I 860s. The plan is to replace the Weitek based modules, plane by plane, to produce a 50 GFlop (peak) system this summer. The new machine will support all existing CANOPY based code without change.

### **Quarks, Galaxies, & Stars II: Observational Astrophysics & The Digital Sky Survey Project**

For a while, the theoretical interests of high energy physics have been close to astrophysics. Interests converge in the study of the very early universe. Major astrophysics observational projects have reached the stage where many of their technological needs are similar to particle experiments. Fermilab is forming an experimental particle astrophysics group to support its participation in the Digital Sky Survey Project (DSSP) in a collaboration with the University of Chicago, Princeton University, and the Institute for Advanced Study. The goal of this project is to produce a three dimensional map of 1 million galaxies across a quarter of the sky. Two orders of magnitude larger than previous surveys, this will allow the most extensive study of structures in

the universe to date. Such structures are important clues to events in the universe when temperatures were so high that the kind of elementary forces and particles normally studied in Fermilab's high energy accelerator experiments dominated the cosmos. In particular, Fermilab will provide support of much of the computing needs of this project based on its experience with its traditional experiments. From an experimental science standpoint, the experience of Fermilab physicists in selecting and analyzing large and complicated data sets is expected to be a major contribution. The requirements on computing and data acquisition of the DSSP are similar to that of a good sized fixed target experiment at Fermilab and will provide a similarly significant but not overwhelming challenge.

From a computing standpoint, this project may present the most interesting opportunities in areas having to do with software engineering and methodical system design. Computers and human beings will exchange data with each other in a complex decision making process both in real time (weather, seeing, and pointing decisions, to name a few) and over a long data reduction cycle. As innocents in observational astrophysics, we are trying to understand the data requirements and the inter relationship between system components, in various possible implementations, by using software engineering concepts such as object oriented system analysis, entity relations, and functional analysis. Operationally, we want to learn who (human/computer) does what, when, where, and how -- and how much does it cost? From a data standpoint, the questions are: What data is needed at each function? What are their relations and dependencies? What (hardware, software, people) does things to/with the data?

The goal is to develop implementation options for the collaboration in a way that identifies tradeoffs between costs and requirements and operational considerations (such as which activities are carried out on the mountain top at Apache Point, New Mexico and which at remote centers such as Fermilab). Just as we are modest about our knowledge of astrophysics, our astronomer colleagues, more than particle physicists in similar collaborations, show appropriate respect for the computing and DAQ requirements of the project. The result is that we are mutually very receptive to a systematic approach to design for the experiment's data aspects.

Perhaps, an important spin off from this astrophysics project will be a deeper understanding of the software engineering ideas and methods needed for systematically planning and designing complex hardware and software systems in HEP experiments. As we noted earlier in describing R&D priorities, nothing could be more important as we move toward the large scale experiments of the future at Fermilab and other laboratories.

### **Acknowledgements**

I would like to thank all the people who helped prepare this presentation, in particular, Lauri Loebel, Peter Cooper, Phil Stebbings, Mark Kaletka, and Luann O'Boyle. All the members of the Computing Division who have worked so hard to get us this far, and appear to be willing to continue the struggle, deserve special recognition and thanks from me. It is their work that was described in this presentation.