THE COUNTING HOUSE

Brian Hayes

Every large scientific institution has a culture, an ethos, an indigenous style. At the Fermi National Accelerator Laboratory the style is utilitarian chic. Fermilab delights in the ingenious and elegant use of the found object, the industrial artifact converted to some higher purpose. In the 1970s, while a rival European laboratory quarried Carrara marble for shielding in a large detector, Fermilab filled a similar need by scrounging armor plates from two decommissioned battleships. The laboratory's auditorium is built out of concrete castings left over from the construction of a beam-line tunnel. These design choices were not a product of mere frugality; there is also an aesthetic principle behind them, a pride in turning to good account materials that others might have discarded or covered up.

This same aesthetic sense can be seen at work in decisions that have shaped the laboratory's computing program. A major consumer of computing resources at Fermilab is a task called event reconstruction, which was traditionally done by mainframe-class computers, such as the larger members of the Digital Equipment Corporation VAX series. As the volume of data increased during the 1980s, the computing requirements outgrew those machines. One might have expected them to be replaced by bigger and costlier supercomputers, such as those made by Cray Research, or by massively parallel machines, such as the CM-2 (with 65,536 processors) made by Thinking Machines, Inc. The Fermilab Computing Division took a different path: They created a "farm" of off-the-shelf workstations and developed the software needed to make hundreds of them cooperate. The individual computers are not very powerful, but they are cheap, and so a budget that would barely pay for air-conditioning a supercomputer can buy quite a number of them.

I recently spent a few days visiting Fermilab and talking with Thomas Nash, who was then the head of the Computing Division (he has since been named Associate Director for Scientific Technology and Laboratory Information), and with others who do various kinds of computing there. I shall briefly comment on four aspects of the Fermilab computing program: data acquisition, event reconstruction, analysis of experimental results, and computing in pursuit of theoretical understanding. A fifth activity in which Fermilab is participating, the Sloan Digital Sky Survey, deserves a column of its own.

Brian Hayes, former editor of American Scientist, *has been writing about computing for 10 years. Address: 211 Dacian Avenue, Durham, NC 27701. Internet: hayes@concert.net.*

Computing on the Prairie

Fermilab rises out of an Illinois prairie, now being overtaken by the western suburbs of Chicago. When the laboratory opened in 1971, its main instrument was a 200 giga-electron-volt (GeV) proton synchrotron, which was soon upgraded to 400 GeV. Today the original synchrotron serves as a "booster," supplying protons to the Tevatron, a new machine with a maximum energy of nearly 1,000 GeV, or 1 tera-electron-volt (TeV). The Tevatron can accelerate protons for collisions with fixed targets or can operate as a proton-antiproton storage ring, where matter and antimatter collide head-on. Two very large detectors, called CDF and D0, surround the interaction regions when the Tevatron is run as a collider; more than a dozen other experiments get a share of the beam when the machine is switched to fixed-target mode. During my visit in October the Tevatron was tuning up for a collider run, for which the physics community has high hopes: It is expected to turn up evidence of the top quark, the sixth (and almost certainly last) of the particles that form the substructure of protons and neutrons.

Computing is essential to the operation of all the Fermilab machinery. Of course statements of that kind have become a bland commonplace; one could say the same of a bank or an airline. In high-energy physics, however, computing has become truly central. The traditional limits on what could be learned in a particle-physics experiment were set by the energy of the accelerator and by its luminosity—how intense and concentrated a beam it could create. Now, for some experiments, the most important constraint is the rate at which data can be gathered and digested. In a 1992 collider run, for example, the CDF apparatus was exposed to 500 billion collisions, and yet some of the papers published after the run discussed just 28 events. Finding the 28 gems among all the dross is a formidable computational challenge.

Data Acquisition

The selection process begins as the data are gathered. A large modern detector has many parts: wire chambers that trace the trajectory of a particle, calorimeters that measure its energy, small silicon sensors that precisely locate the vertex where two paths diverge. Signals from all of these devices are combined to create an image of the fleeting interactions of invisible particles.

I was taken down into the detector pit of experiment E687, which examines the production of particles and antiparticles by high-energy photons. My guides to this underworld were Vicky White, deputy head of the Computing Division and a specialist in data acquisition, and Joel Butler, spokesman for the E687 collaboration, who has since become the new head of the Computing Division. In such a place the scale of the machinery makes an immediate impression; Butler pointed out that everything must break down into pieces of no more than 30 tons, since that is the capacity of the overhead cranes. But I was equally struck by the profusion of cables. They emerge from some parts of the detector in turbulent Medusa tangles; elsewhere hundreds of them lie side by side in smooth, laminar streams, or they are gathered into thick bundles like ships' hawsers. Most of this wiring consists of coaxial cable, the same kind that brings MTV into the living room, but here thousands of cables run in parallel. The data rate is immense.

The bundles of cables extend less than a hundred meters, from the detector into a nearby "counting house." The signals generated by the detector are in analog form—that is, they are voltages or currents that vary over a continuous range. They are digitized, and the digital information is stored temporarily in a buffer. Other processes withdraw the data from the buffer, assemble readings from various parts of the detector into a unified record of a single event, and pass the result on to a computer that records it on magnetic tape. The preferred tape format is an eight-millimeter tape cartridge called an Exabyte cartridge, which holds five gigabytes of data.

Even with high-speed electronics and capacious storage media, capturing all the events for perusal later is generally not an option. Instead, events are saved only when they meet certain predetermined criteria, in a selection process called triggering. Roughly speaking, the signals are put through a sieve, which allows the few interesting events to pass but blocks the more numerous routine ones. Most experiments employ a hierarchy of triggers. At the first level the response must be very rapid, and so the triggering logic is implemented in hard-wired circuitry. Only the simplest logical analysis is carried out, such as checking for the presence or absence of a certain combination of signals. The second-level trigger—which further filters those events that survive the first winnowing—is usually implemented in high-speed programmable controllers. More sophisticated criteria can be applied here, such as looking for the temporal coincidence of two signals. Some experiments include a third level of triggering, in which events are examined—and either saved or discarded—by a program running on a cluster of workstations.

Triggering algorithms are a critical element of the data-acquisition process. In some cases only one event out of every 100,000 makes it through all the levels of filtering to reach the data-logging tapes. If the triggering criteria are not defined carefully, and the wrong subset of events is saved, months of effort could be wasted. There is particular concern about third-level triggers, since they rely on much more complex software than the lower-level triggers.

Farming at Fermilab

Once the data have been captured on magnetic tape, the pressure to keep up with the ongoing torrent of events is relieved. Nevertheless, a great deal of computing still needs to be done before meaningful results can be extracted from the data.

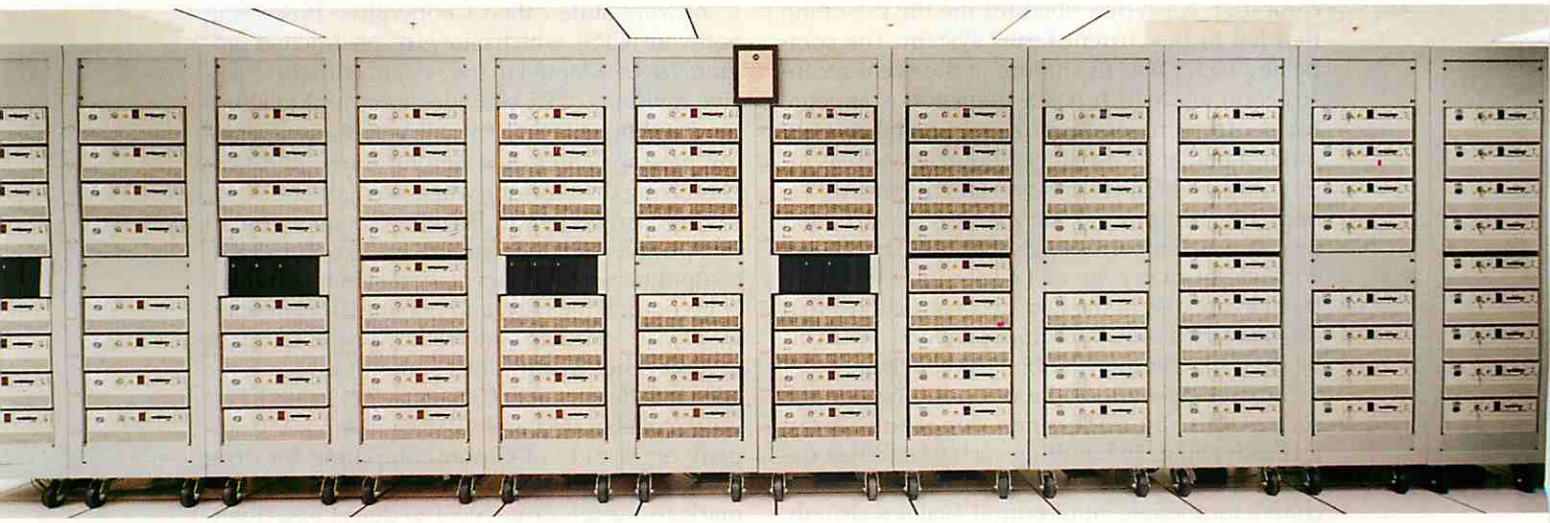

Figure 1. One hundred IBM RS/6000 workstations make up part of the "farm" computer system at Fermilab. (Photographs courtesy of Fermilab.)

The first task is event reconstruction, the purview of the processor farms mentioned above.

The raw data tapes do not describe particle trajectories or momenta. They merely say which detector elements were "hit" by particles and how much energy was deposited in calorimeters. Event reconstruction is the process that makes sense of these fragmentary reports, figuring out which hits go together to form the track of a single particle. It is a pattern-recognition task, similar to a connect-the-dots puzzle, except that the dots are not numbered.

The reconstruction process for one experiment, E665, was explained to me by Stephen Wolbers, who is both a member of the E665 collaboration and co-group leader (with Frank Rinaldo) of the Farms Supercomputing Group. E665 is a fixed-target experiment, studying the scattering of muons by other particles. In a data run that ended in January 1992 the group had recorded 150 million events on 1,500 Exabyte tapes. Before the processing of these tapes could begin in earnest, however, the reconstruction algorithms had to be adjusted to reflect the alignment of the detector elements, the various electric and magnetic fields that deflect particle paths, and the calibration of energy measurements made in the calorimeters. The "production run" with the full data set began in January 1993, and it was still under way during my visit, with completion expected in January 1994. Thus it is a year's worth of computing—but if the job were done on a single VAX 11/780, it would run for 500 years.

The impracticality of completing such tasks with a VAX or a similar machine was already apparent a decade ago. Fermilab's response was to form an Advanced Computer Program (ACP). The first solution to the event-reconstruction problem was a machine made up of some 300 circuit boards based on the Motorola 68020 microprocessor. Each board had its own memory, but disk and tape storage were centralized. This first ACP parallel computer went on line in 1986.

As the demand for computing capacity continued to grow, plans were laid for a new parallel computer. Nash described for me the reasoning that led to the current farms system. The compelling factor was the nature of the event-reconstruction job, which has a tremendous appetite for central-processor power but makes only modest demands on other resources, such as input-output channels and working memory. Perhaps most important, an event-reconstruction program is "embarrassingly parallel," because it consists of many small tasks—150 million of them in the case of the E665 reconstruction—that can be carried out almost independently, with little need for communication between them.

Supercomputers of the conventional kind—such as the Cray X-MP series—are at an economic disadvantage in handling such jobs. When the central processing unit is fully utilized, the machine's formidable input-output facilities are only lightly loaded, and thus the customer pays for capacity that is never used. In massively parallel computers such as the Thinking Machines CM-2 a large share of the cost is the elaborate web of high-bandwidth interconnections between the thousands of processors; again this hardware would be little used in event reconstruction.

By the late 1980s, processor chips such as the 68020 had fallen out of fashion, replaced by reduced-instruction-set-computer (RISC) designs. Accordingly, the ACP system was upgraded with new circuit boards based on a RISC processor. But a change in the computer marketplace soon suggested another strategy. Complete RISC workstations, with disk drives, power supplies, cabinets, keyboards, Ethernet connections and system software, were available off the shelf for less than the cost of building a custom processor board. Thus was born the idea of a processor farm. The term harks back to the days of disk farms, when much acreage in computing centers was given over to arrays of disk drives the size of washing machines.

Nash and his colleagues eventually selected two workstations for the farms: the IBM RS/6000 and the Silicon Graphics 4D. The manufacturers recommended their most powerful models, but the physicists were able to get higher overall performance by buying a middle-of-the-road model in larger numbers. Today the farms consist of 140 IBM and 180 Silicon Graphics machines, mounted in tall racks. Each node has a local disk and enough random-access memory (typically 16 megabytes) to run the reconstruction programs. The nodes are linked via Ethernet to one another and to additional workstations that act as input-output servers, providing access to some 70 tape drives and 60 gigabytes of disk storage. The total computing capacity is roughly equivalent to that of 10,000 VAX 11/780s.

The problem with a do-it-yourself processor farm is that it needs do-it-yourself software. It is particularly challenging to provide a uniform interface to a system made up of hardware from multiple manufacturers. Fermilab has developed a software suite called Cooperative Processing Software (CPS), which allows reconstruction programs to be adapted to the system without extensive revisions. CPS provides tools and facilities for distributing tasks to multiple processors, passing messages and data among them, and synchronizing their actions when needed.

The farm system at Fermilab seems to be almost universally accounted a success. Indeed, computing with clusters of workstations has lately become all the rage, and not only in high-energy physics. Similar systems have been set up at several other institutions, some of which have adopted the Fermilab CPS software. In 1992 Fermilab collaborated with IBM and Merck & Company on a study of cluster computing for drug design, and there have been inquiries from a market-research firm trying to digest vast quantities of data about supermarket purchases.

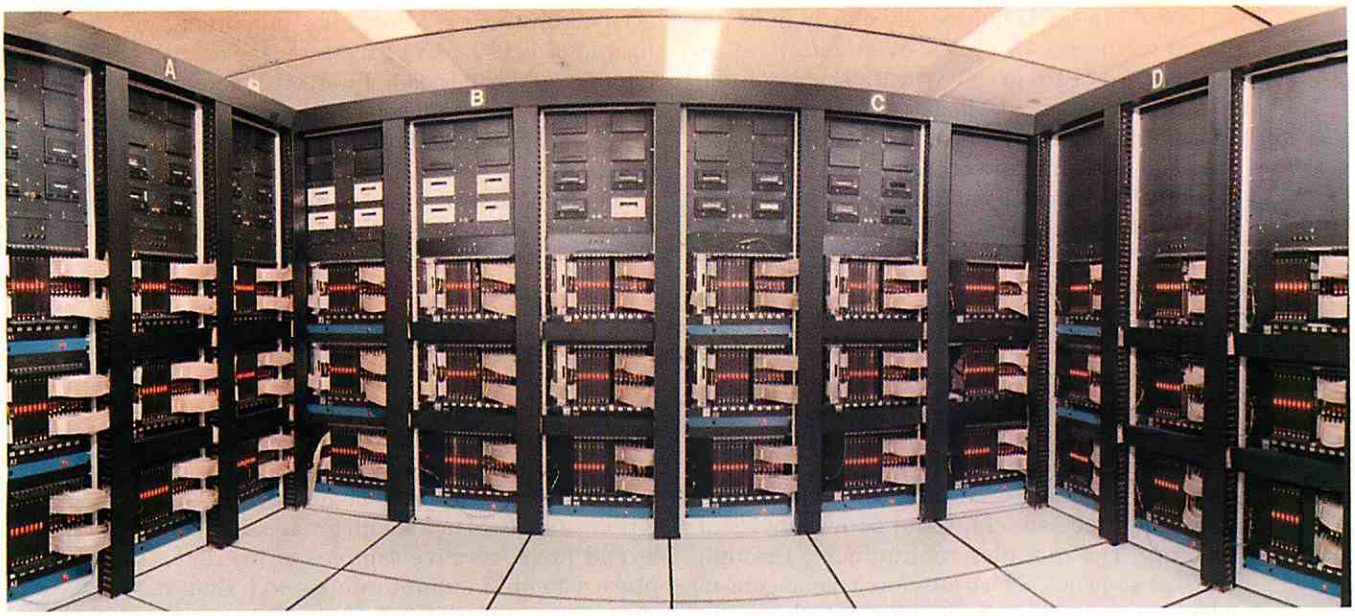

Figure 2. ACPMAPS, a supercomputer with 306 modules connected by a crossbar switch, is used for theoretical studies of quarks and gluons.

Harvesting the Physics

Event reconstruction does not reduce the bulk of an experiment's data set; on the contrary, because a copy of the raw detector data is generally saved along with the reconstructed event, the size of the tape archive actually increases. Thus, having gotten this far, the physicist still faces a daunting task of interpretation. How is one to make sense of a data set measured in terabytes? Clearly, some further sifting must be done. This is the first task of data analysis.

The selection process in data analysis is similar to the triggering process in data acquisition: Events are examined one by one and discarded unless they meet certain criteria. But during the analysis phase there is no need to make decisions in haste, before the next event arrives. Hence the criteria can be more complicated and the algorithms more thorough. Typically the process is an iterative one, with different "cuts" being tried until the results are satisfactory. The product of the first analysis run is a new set of tapes, called data-summary tapes, or DSTs. These tapes are then further distilled to produce mini-DSTs and sometimes micro-DSTs.

Analysis programs at Fermilab have been run on a variety of hardware platforms, including VAXes and an Amdahl mainframe computer. During my visit the Amdahl system was being replaced by a cluster of workstations. Some of the analysis machines are served by two remarkable "silos" (more farm equipment!) in which robot arms move in darkness to mount tapes as they are needed.

Conversations with physicists from the CDF and D0 collaborations suggested that data analysis is the phase of computing in high-energy physics that gives rise to the most frustration. The software tools in use for these tasks are apparently adequate to the need, but they were spoken of without affection or enthusiasm. One physicist offered to trade his high-performance workstation for a humbler Macintosh or PC, if it would provide a better user interface. It seems the trendier developments that have swept over other areas of scientific computing, such as exploratory data analysis and scientific visualization, have not caught on in high-energy physics.

Computing on a Grid

Another kind of computing done at Fermilab is not well suited to workstation farms: theoretical studies of the structure of matter, specifically calculations using a technique called lattice gauge theory. Lattice-physics programs perform poorly on workstation clusters because they require too much communication between processors. The theory group at Fermilab and members of the Advanced Computer Program therefore set out to build a machine dedicated to lattice physics. The first version, completed in 1991, had a peak performance of 5 gigaFLOPS, or 5 billion floating-point operations per second. The machine has since been upgraded to run at a peak rate of 50 gigaFLOPS. It is called ACPMAPS, or Advanced Computer Program Multiple Array Processor System.

Lattice gauge theory is an approach to studying the interactions of quarks and gluons (the latter particles bind quarks together inside nuclear particles). Straightforward approaches to this task do not work; the methods used to calculate the binding of a planet to a star or of an electron to an atomic nucleus break down when they are applied to the powerful forces that act among quarks and gluons. Lattice gauge theory finesses the problem by inventing a universe where spacetime is filled with a gridlike lattice, and particles can exist only at the vertices of the lattice. The real universe is obviously not like this, but solutions can be extrapolated to the real world by observing what happens as the lattice spacing is reduced to zero.

Lattice physics is notoriously greedy in its demand for processor cycles; even with a rather small lattice of 16^4 (65,536) sites, a single calculation could require 10^{12} floating-point operations. This prodigious need for computing capacity, and the geometric structure of the problem, make lattice physics a natural candidate for parallel processing, and a number of high-performance computers have been built just to run lattice programs. ACPMAPS is the largest of them. Nevertheless, Mark Fischler, a designer of the system, maintains that the machine's greatest distinction is not its speed but its flexibility: It can accommodate a variety of algorithms and has therefore become a test bed for new ideas in lattice physics.

In its present configuration ACPMAPS has 306 processor modules, each of which includes two Intel i860 processor chips and 64 megabytes of memory. The modules communicate through crossbar switches, which are closely analogous to telephone exchanges. Just as any telephone can reach any other telephone connected to the same system, any processor in ACPMAPS can gain access to the memory of any other module.

The operating software for ACPMAPS, called Canopy, is designed to hide the complexities of the architecture from the physicist writing lattice-gauge-theory programs. Programs can be written in terms of "sites" connected to form a "grid," with "paths" leading from one site to another and "fields" defined at each site on the grid. The mapping of these concepts onto the actual array of processors is left to the software system. Indeed, the number of processors allotted to a job is determined only when the program is run.

ACPMAPS is a new system, but it has already produced some noteworthy results. For example, it has yielded the first lattice computation of the strong coupling constant—the factor that defines the strength of the force between quarks and gluons—with well-understood error bars.

The Petabyte Problem

My visit to Fermilab came just a few days before the final Congressional vote killing the Superconducting Supercollider. I spoke with Irwin Gaines, a member of the Fermilab Computing Division who was part of a collaboration building a detector for the Texas accelerator. He had given much thought to the challenges of data processing at the new laboratory, where the data rate would have been at least an order of magnitude greater than it is at Fermilab. Although his analysis will not now be put to use in Waxahachie, the ideas should prove valuable whenever and wherever the high-energy physics community is able to move on to a new generation of accelerators.

A single experiment at the SSC would have generated a billion events per year and recorded a megabyte of information on each event; that adds up to a total data volume of one petabyte (10^{15} bytes) per year, enough to fill up 200,000 Exabyte tapes. The off-line computing capacity needed for event reconstruction would grow proportionately. Instead of 10,000 VAX equivalents, the new laboratory would have needed 100,000 or a million.

Although these numbers are impressively large, Gaines argues that the hardware requirements could easily have been met by the time the SSC was running. But software systems would also increase in size and complexity, and there the outlook is more troubled. Some experimental collaborations are already maintaining more than two million lines of FORTRAN software. Managing the complexity of still larger systems will require formal methods of software engineering. Certifying the correctness of the software will become essential, because the validity of the experimental results depends on it.

For data analysis, Gaines suggests that an entirely new approach may be in order. Instead of writing a series of tapes with progressively smaller and more selective samples of the data, the physicist could submit queries to a data base storing the entire data set. Each query would retrieve a subset of events in the same way that a query submitted to a library-catalogue data base retrieves a subset of the book collection. But there is a difference of scale: The petabyte-size database for a physics experiment is far larger than any library catalogue (records for at least 100 billion books would fit in a petabyte). Hence new methods of indexing, of hierarchical data-base storage and of parallel data access would need to be developed to make the idea practical.

It is interesting to observe that even as computing has moved toward center stage at Fermilab, the institution remains focused on physics. Before my visit I thought I might find, among all the physicists doing computing, a few computer scientists doing physics, but it appears the intellectual current flows in one direction only (although Vicky White is a mathematician). The physicists do their computing in a physicist's way, more interested in the results than in the neat hack. They do their computing in a Fermilab way.

Bibliography

Fischler, Mark. 1992. The ACPMAPS system: A detailed overview. Fermilab technical report TM-1780. Batavia, Ill.: Fermi National Accelerator Laboratory.

Fischler, Mark, Mike Uchima, George Hockney and Paul Mackenzie. 1990–93. *Canopy Version 7.0: Canopy Manual.* Batavia, Ill.: Fermi National Accelerator Laboratory.

Sullivan, Kevin Q., Matt Fausey, Dave Potter, Frank Rinaldo, Marilyn Schweitzer, Roberto Ullfig, Steve Wolbers and Bob Yeager. 1993. *CPS User's Guide.* Batavia, Ill.: Fermi National Accelerator Laboratory.

Gaines, Irwin, and Thomas Nash. 1987. Use of new computer technologies in elementary particle physics. *Annual Reviews of Nuclear and Particle Science* 37:177–212.

Nash, Thomas. 1991. High energy physics experiment triggers and the trustworthiness of software. Proceedings of the 1991 CERN School of Computing, Ystad, Sweden.

Rinaldo, Frank, and Stephen Wolbers. 1993. Loosely coupled parallel processing at Fermilab. *Computers in Physics* 7(2):184–190.

Rinaldo, Frank J., and Matthew R. Fausey. 1993. Event reconstruction in high-energy physics. *Computer* 26(6):68–77.