Fermi National Accelerator Laboratory

FERMILAB-Conf-89/120

Event Parallelism: Distributed Memory Parallel Computing for High Energy Physics Experiments *

Thomas Nash
Fermi National Accelerator Laboratory
P.O. Box 500, Batavia, Illinois 60510 U.S.A.

May 1989

* Invited talk presented at the 1989 Conference on Computing in High Energy Physics, Oxford, England, April 10-14, 1989.

Event Parallelism: Distributed Memory Parallel Computing for High Energy Physics Experiments

Thomas Nash

Advanced Computer Program
Fermi National Accelerator Laboratory*
Batavia, IL 60510
USA

Abstract

This paper describes the present and expected future development of distributed memory parallel computers for high energy physics experiments. It covers the use of event parallel microprocessor farms, particularly at Fermilab, including both ACP multiprocessors and farms of MicroVAXES. These systems have proven very cost effective in the past. A case is made for moving to the more open environment of UNIX and RISC processors. The 2nd Generation ACP Multiprocessor System, which is based on powerful RISC systems, is described. Given the promise of still more extraordinary increases in processor performance, a new emphasis on point to point, rather than bussed, communication will be required. Developments in this direction are described.

1. Introduction

The main computing activities of experimental high energy physics are, by good fortune, trivially parallel. Interaction events are independent of each other. The digitized raw data from different events essentially remains independent and can be reconstructed into physical parameters (individual track momenta, vertices, etc.) that can be analyzed on an event by event basis. This situation is likely to change dramatically in future very high luminosity hadron collider experiments, where multiple unresolved events are expected from a single beam crossing of a few nanoseconds. But even there, individual crossings will be independent.

Reconstruction of data, although requiring extraordinary amounts of computing, is managed by taking advantage of this intrinsic parallelism in what have come to be called microcomputer *farms*. Single events are sent to individual processors in the farm for reconstruction. The reconstructed data is written to Data Summary Tapes (DSTs) as soon as the processor is done. The processor *node* then starts on another event. This technique can be used either off line, reading events from raw data tapes, or on line to an experiment as a high level filter of events selected for writing on tape. Physicists study detector systematic errors by running extensive Monte Carlo programs to simulate events. This simulation is also an obviously event parallel problem which has been effectively run on multiprocessor farms.

* Work supported by the U.S. Department of Energy under contract #DE-AC02-76CH03000.

The first to take advantage of event parallelism in high energy physics was Paul Kunz at SLAC when he built bit slice microprocessor devices that emulated the IBM 360 instruction set[1] . These emulators and their successors have been heavily used by experiments since their introduction in the late 1970s.

Once the data is reconstructed and simulated, the real physics starts with an intensive statistical analysis of the DSTs. Kinematic quantities are computed for each event and used to select events for inclusion in a variety of one and two dimensional histogram and "lego" plots, and in regression and likelihood fits to theoretically proposed models and functions. This process is also event parallel. However, until recently it has been primarily an I/O intensive problem and carried out on mainframes and super mini computers. This situation is changing because of the increasing complexity of events, which now require more computing during analysis, and the impact of workstations and very low cost video technology magnetic storage devices. There is also an increasing realization that the line between reconstruction and DST analysis is not all that clean: analysis often involves re-reconstruction (and re-simulation) of parts of events as algorithms and calibration are refined.

The most important recent technological advance has been in reduced instruction set computers (RISC). They make possible extremely cost effective CPUs for both on line and off line multiprocessors and for workstations. The marriage of workstations, open system farms, and networks will have an important side impact: confronting experimentalists with rather urgent decisions about moving to UNIX, an operating system environment which they have, to date, essentially ignored.

Event parallel multiprocessors are a subset of the distributed memory, explicitly parallel class of computers. Examples of these include hypercubes and other machines which have resonated so well with the needs of theoretical physicists and others with site oriented differential equation computations[2] . In the following, I will describe the present status of event parallel computing in high energy physics and the most likely near and long range future direction for this kind of explicit parallelism.

2. First Generation ACP Multiprocessor Farms at Fermilab: Successes and Difficulties

Since 1984, Fermilab's Advanced Computer Program (ACP) has been developing highly cost effective parallel computers for high energy physics experiments. The First Generation ACP

---

1 . P. F. Kunz, Nucl. Instrum. Methods 135 (1976) 435.

2 . G. C. Fox and S. W. Otto, Phys. Today 37)5), (1984) 50.

Multiprocessor System has been described extensively elsewhere[3] . These systems generally consist of a single MicroVAX host managing a farm of up to over 100 single VME board computers based on Motorola's 68020/68881 microprocessors. The host accesses what can be very large disk and 6250 BPI tape drive banks. It passes data to and from the nodes over an ACP developed cabble protocol known as Branch bus. Branchbus interconnects the host and VME crates containing the 68020 based processors. In on line high level filter systems, it connects the VME processors to Fastbus based data acquisition systems at a band width of 20 MBytes/sec. ACP modules are commercially available through Omnibyte, Inc., West Chicago, IL. The original cost of crates full of processors was about $2500 per node, each with 2 MBytes of memory, at lab/academic prices. To this must be added the host plus peripherals (up to $50,000). 4 MByte expansion memory is also available. Systems purchased during the DRAM drought of '88 were considerably more expensive.

Around the world, there are at least 30 institutions with ACP multiprocessors, in Europe, Japan, Brazil, and North America. The largest systems are at Fermilab where there are 6 production systems with a total of 410 nodes and five software development systems with another 45 nodes. The total production capacity for typical high energy physics code is around 400 VAX MIPS. This is almost double that of the rest of Fermilab's multi million dollar general purpose computer center that includes 65 MIPS in VAX clusters and 120 MIPS in Amdahl mainframes, as well as older CDC machines now being phased out.

The original user of these systems at Fermilab was experiment E691 on a 55 node installation in 1986. This successful experiment increased the statistics of charm particles by more than an order of magnitude, allowing it to end an ongoing controversy about charm quark life time and decay properties. The ACP system reconstructed E691 data in a few months instead of the several years otherwise anticipated. Now, some six fixed target experiments, along with the big Collider Detector at Fermilab (CDF), are competing vigorously for time on the systems. One experiment (E769, a descendent of E691) requires 3 million node hours. Another requires 400,000. Some of this demand is being taken up by systems outside Fermilab at Los Alamos and the University of Chicago. However, the demand crisis has led to a decision to purchase two UNIX RISC based systems. This is influenced strongly by what is being planned for the 2nd Generation ACP Systems, described below.

The colliding beam detector experiment, CDF, is the other heavy user of ACP systems. It's highest level on line trigger is based on 55 ACP nodes, and two of the ACP systems in the Fermilab Computer Center are assigned for CDF use. One with 56 nodes is hosted by a MicroVAX 3200 with has had as many as four tape drives and five disks with 2 GBytes total storage. The other, a 65 node system, is Ethernet clustered to a VMS Version 5 VAX cluster. This allows direct

3 . I. Gaines, H. Areti, J. Biel, A. Cook, M. Fischler, R. Hance, D. Husby, T. Nash, and T. Zmuda, Comp. Phys. Comm.45, (1987) 323 and 331.

access from the farm to the tapes and disks on the big VAX cluster and only became possible last summer under Version 5. Most important is the direct access to CDF's huge Data Base Master of calibration constants which is maintained on that cluster. In particular, this means that the daily updates to CDF's calibration files do not have to be moved across a network as is the case on other systems. After a major effort by CDF, Computing Department, and ACP people, this huge code (1,300,000 source lines) can now run on these two systems at the experiment's data taking rate of 190 tapes/week.

It is clear that this approach to meeting the large computing needs of experiments has been successful. The systems and computing requirements are well matched: more than half the experiments are so compute bound that they can make effective use of single host systems with over 100 nodes; the rest use at least 50 nodes easily. Nonetheless, the heavy usage has exposed difficulties. In the case of CDF, with its special and large requirements, the difficulties have been rather serious. Some of the problems involving software development turnaround time have been resolved by the Computing Department's ACP support group and the ACP. They have provided: a Fermilab written linker that now handles CDF's more than 1000 entry point code in 5 minutes, compared to a prohibitive 6 hours previously; trace backs and post mortem dumps; and utilities to allow parallel compilation of a large number of subroutines on available compile nodes.

The more fundamental difficulties have fallen in two categories: delays caused by the new requirements these systems have put on a computer center whose experience has been with main frames; and the issues associated with conversion of a large collaboration's software package which had been designed for, and maintained in, a homogeneous VAX VMS environment. The new requirements for the computer center had to do with the procurement, configuration, setup, and maintenance associated with microcomputers and their peripherals. These requirements will become more familiar to large computer centers as they move from centralized mainframe dominance to distributed, networked, workstation environments.

CDF put in a huge effort to convert its code to the ACP systems. This was made more difficult by hardware support and ACP system utilities not yet up to meeting their requirements, and was finally successful, as we have noted. However, an experiment inevitably needs to make changes to their code. CDF's rate of changes has slowed down from almost daily updates to the production code. Developmental changes are still submitted by approximately 100 physicist programmers at 30 institutions around the world and distributed to the collaboration nightly over the network. The relatively unstructured code is maintained in a homogeneous VMS environment. It has become an unacceptably large burden to convert to another operating system -- recompile and link, retest and debug -- as often as required. Given the extraordinary pressures on this large collaboration to produce physics results on an urgent time scale, this is not a good time for them to

move their operations to a more open environment. Their conclusion is that they need immediate access to VMS based farms, which now are as cost effective as First Generation ACP Systems.

3. VMS Farms

VMS MicroVAX farms over the last few years have been advocated by the Brown University group on the D0 experiment at Fermilab, where they are to be used in the high level trigger with a special high speed dual port memory. A more conventional VMS farm is in the Aleph experiment trigger at CERN. At the University of Florida , a MicroVAX farm is under construction for off line use. For the reasons noted above, CDF at Fermilab experimented with a 10 node VAX Station 3200 farm at the University of Chicago last fall. They now have nearly operational a 20 VAX Station 3100 system at Fermilab.

Given the extensive VMS expertise available in the high energy physics community, these systems have proven reasonably easy to setup and operate. For experiments with a large number of collaborators writing software in a homogeneous VMS environment the convenience of VMS farms is strongly felt. They argue that the personnel costs of converting software and managing unfamiliar operating systems will overwhelm the claimed cost effectiveness advantages of competitive alternatives in the UNIX environment.

One internal CDF document, which was widely circulated, compared various 60 MIPS UNIX farm alternatives with a VMS farm[4] . A $600,000 charge was made to the UNIX systems for conversion and maintenance personnel costs compared to no charge for the VMS systems. Under such handicapping, the UNIX alternatives appeared to be 25%-70% less cost effective. From this perspective, anticipation of future performance gain in VMS micro computers based on past experience appears encouraging, as high as 10 VAX MIPS in 1991. The most reasonable VMS advocates argue that it may be possible to hang on with VMS without losing too much cost effectiveness until vendors provide a painless transition to the more open UNIX environment. Other VMS advocates can see no end in sight to the viability of what has been a comfortable closed environment.

4. The Revolution: Open Systems, Workstations, RISC, and UNIX

For those who aren't involved with large software investments under closed operating systems, there is a widespread sense that a revolution is in progress. This has been apparent in Wall Street activity and in the press. Typical are the headlines that appeared in the New York *Times* (April 4, 1989, page 1) , "Mainframe Computers May Be Near Extinction", and *Time*

4 . D. Quarrie, An Evaluation of the Cost_effectiveness of UNIX Machines for CDF, Feb. 20, 1989, CDF Note 871 (unpublished).

Magazine (March 13, 1989, page 55), "Where the Action Is... In computers, workstations are the workhorses of the future".

The workstation market increased at a rate of over 50% last year, and well over 80% of that was under the open and portable UNIX operating system. Much of the reason for this activity is the major technological advance associated with RISC architectures. RISC chips from a variety of manufacturers are now crossing the maximum computing performance levels of the largest mainframe machines. These chips bring mainframe performance to desk top workstations and office environment compute servers, at amazingly low cost. The mainframe computers are still able to handle many more peripherals and much more I/O capacity than the usurpers. Even this situation may change soon with distributed network file service and uniform file protocols.

The RISC technology is based on the recognition by computer scientists in the 1970s that many complex instructions in traditional (CISC) machines with large instruction sets, like the IBM 360s and DEC VAXes, were rarely used. These instructions effectively reduced the performance for all instructions because they mandated extensive microcode. RISC machines are generally pipelined with no microcode and very fast instructions. They defer complex instructions to software. The number of instructions per clock is improved by five or more, yet the corresponding increase in instruction count is small. The result: a big net win for RISC.

Companies responsible for RISC (or RISC like) chip sets at this time include AMD, Inmos, Intel, Intergraph, MIPS, Motorola, and Sun. The important MIPS and Sun SPARC architectures are available from as many as five semiconductor manufacturers. Workstations based on RISC are available from Apollo and HP (now joined), DEC, Data General, IBM, Intergraph, MIPS, Motorola, Silicon Graphics, Sun, and others. RISC performance on the most advanced chips is now at about 20-30 VAX MIPS.

Pipeline scheduling and other optimizations are relegated to compilers in RISC systems. This provides an opportunity for optimization beyond that possible in microcode. It also means that an excellent optimizing compiler is essential. Not all RISC family compilers are equally good in either robustness or performance at this stage. One suspects that the semiconductor houses have still not learned the lesson of how important compilers are for chips that have mainframe performance. Compute server/workstation inspired product lines have shown a greater sensitivity to the compiler issue, and one expects that soon several RISC families will be supported by high performance, robust compilers. At this time, the compiler for the MIPS family (used in the UNIX DEC Station 3100, in Silicon Graphics machines, and in the new ACP VME module described below) has been demonstrably successful in handling huge high energy physics codes in a high performance and relatively bug free fashion.

Semiconductor vendors have been showing future product performance curves that typically promise 160 VAX MIPS in 1991. This is 15 times higher than the most optimistic CISC projections for VMS microprocessors, cited earlier. Essentially *all* important RISC chip and workstation activity is under the UNIX operating system. This combination of facts is forcing high energy physicists to pay attention to UNIX for the first time and explains the new directions that have been taken by the ACP at Fermilab.

5. The 2nd Generation ACP Multiprocessor System

The emphasis these days at Fermilab's Advanced Computer Program is on open systems and an open network environment. The 2nd Generation ACP Multiprocessor System allows processors to run under either UNIX or VMS. They may be connected either through Ethernet or the backplanes of VME crates which are interconnected by the ACP's 20 MBytes/second Branchbus (Figure 1). Branchbus supports the high band width required in real time applications and Ethernet permits open and flexible access to whatever may be the most cost effective compute engines available. Open system network protocol standards (NFS, TCP/IP, UDP) are supported over all links.

In the original ACP system a single host with access to tape drives fed and retrieved data to and from a single rank of node slaves. In the new system, each CPU has full access to its own tapes and disks as well as those that may be attached to other CPUs (Figure 2). The new ACP multiprocessor software has the recursive name *ACP Cooperative Processes*. Under this system, all individual CPUs are full participants with a full (VMS or UNIX) operating system and complete inter-CPU communication.

Each CPU runs one or more processes. This is a particularly important feature that means a program with many processes, destined for a large multi-CPU machine, can be thoroughly debugged on a small (even a single node) system. If the small system uses the same chip family, the debugged multi process program executable will run without change on the large system.

The support software is layered: a full (VMS or UNIX) operating system on the nodes, networking tools (NFS, TCP/IP, UDP) , and higher level service routines tailored to high energy physics needs. The higher level support includes routines with functions and names based on the first generation ACP system, such as those that transfer blocks of data between processes: `acp_send` and `acp_get` . In the new system, all processes and CPUs can call these routines, not just the single host. It is also possible to call a remote subroutine on another process or CPU: `acp_call` . In order to allow more control (and less sub surface magic), there are explicit routines to establish queues and to queue, dequeue, and synchronize processes: `acp_queue`, `acp_dequeue`, `acp_synch` . Message passing, useful for data acquisition and other purposes is also available: `acp_transmit_message`, `acp_receive_message`.

The system, as Figure 2 suggests, can support all kinds of complex, multi ranked, heavily interconnected, distributed memory architectures. The general topology can be reduced to two simple ones that are well suited to reconstruction and DST analysis requirements, Figures 3 and 4. The first shows an input process with a tape drive, a multi-CPU, multi-process, event reconstruction rank, and an output process. In the first generation ACP systems, both the input and output activities were generally combined into a single process in the host. This resulted in synchronization complications that are avoided by the topology of Figure 3 allowed by the new system.

The analysis topology of Figure 4 is made possible by the availability of very low cost video technology devices such as Exabyte's EXB-8200 8 mm cartridge tape subsystem. The EXB-8200 can store 2 GBytes of data on a standard 8 mm cartridge and has a standard SCSI interface. Present versions operate at rates comparable to conventional 6250 BPI tape drives. Higher density and speed are expected as well as relatively low cost robot cartridge changers. The topology shown in Figure 4 could prove to be no less than revolutionary in its impact on experiments. Until now experiments that need to analyze all of hundreds of DSTs have had to wait many days or weeks for the tapes to be passed sequentially through a computer center mainframe. The new approach allows such a data base to be read in parallel through as many as a hundred input CPUs and cartridge devices. Instead of weeks, the massive data base will be scanned in less than an hour with the data sent directly to a single Rank 2 process, most likely in one or more desk top UNIX workstations.

With such rapid turn around, it will be necessary to pay attention to the human interface aspects of physicist analysis and software engineering tools. The requirement on such tools is that they should make it possible to iterate analysis and reconstruction software changes rapidly and to display statistical results in a visually dramatic and rapidly understandable form. Having made some initial investigations, we expect a significant effort on such workstation tools to begin in earnest at Fermilab soon.

I have emphasized the software and systems aspects of the 2nd Generation ACP Multiprocessor System, because of our intention that this be an open system with procurement competition for the Ethernet compute servers or VME CPU modules that will act as nodes in the new farms. At present there are products from DEC, MIPS, Silicon Graphics, and Sun, that are potentially relevant to this application. The ACP has developed a VME module based on the MIPS R3000 RISC chip set that will help to encourage strong competition in off line farm nodes. For on line applications that require the 20 MByte/sec I/O capability of these modules, there may be no other choice. The module is shown schematically in Figure 5. It has been licensed for commercial sale to Omnibyte Corp., West Chicago, IL.

Benchmarks on E769 reconstruction software indicate that this module will perform at 20 VAX MIPS for such code. The module consists of two boards that fit in one VME slot. One board plugs into the VME backplane and contains 8 MBytes of memory and a 20 MByte/sec, read and write, block and sequential, VME interface. The daughter board includes the CPU and 32 KBytes each of data and instruction cache. A special external bus (through the VME P2 connector) allows connection of additional memory boards, up to a maximum total of 32 MBytes. (Hopefully, that will keep our memory greedy colleagues quiet for a couple of years.) Access to extra memory is also possible, of course, through VME. The special external bus can also be used as a "vertical" bus bringing data in from other VME crates in on line data acquisition systems.

6. The Future of Explicit Parallelism for High Energy Physics Experiments

The industry seems unanimous about projecting extraordinary future performance for each of the RISC families through 1995. One, therefore, can suggest, subject to the usual *caveat* about peering into crystal balls, that this technology, and the strong industrial competition, will produce RISC processors at and beyond 500 VAX MIPS. It is interesting and important to consider the consequences for high energy physics of working in such a technological dreamland, apparently not all that far away.

Until recently dramatic increases in memory size have been more common than in processor speed. Although memory improvements continue, this may be happening at a slower rate than CPU performance --- certainly slower than the growing appetite for memory of high energy physicists, now acculturated to assuming endless increases in memory availability. It is not unreasonable to expect that 500 VAX equivalent nodes will have 128 MBytes of memory in 1995. Not bad. But the ratio to processing power is reduced somewhat below where it is now. The processor and memory technology environment is changed. Experimentalists will no longer have their memory demands easily satiated while they hunger desperately for computer cycles. It will be important to take advantage of incredible micro computer performance levels and cool it on memory profligacy.

The implications of the new processor environment for high energy physics computer designers will also be stimulating, to say the least. With 33 MHz RISC processors, we are already in a cache crisis. 20 nsec static memory is too slow and it is a struggle to find acceptably fast and large parts. Up to now designers of machines for the lattice gauge calculations of our theoretical colleagues could avoid cache by matching DRAM speeds to the relatively slow floating point chip clocks. The next time around, cache will be necessary, but theoretical problems make such regular accesses to memory that their cache miss rates are unacceptably high in standard cache. Industry is sensitive to both aspects of the cache crisis and seems to be moving toward fast on chip

and multi level cache. An alternative is some form of anticipatory cache which requires hardware and software, through compiler directives, to request memory in advance.

The second technical crisis resulting from the new processor environment is the obsolescence of computer busses for multiprocessor communication. Never mind the territorial religious wars about busses that have swept all high energy physics labs. No matter if we have crusaded for Fastbus, VME, Multibus, Nubus, or a home brew, your or my favorite bus cannot handle the communication requirements of 100 VAX power processor nodes. Even for event oriented reconstruction and trigger processing, just moving data in and out will saturate local crate busses.

The answer to the bus crisis is point to point communication between processors. Two variations on this solution are being explored by high energy physicists and others. There is considerable interest in Europe in the INMOS Transputer architecture. These microprocessors incorporate several communication ports on each chip that support direct links to neighboring Transputers. These devices will be used in the second level trigger and event builder subsystems of the ZEUS detector at the German high energy physics lab, DESY, in Hamburg. The architecture of the Transputer is certainly very appealing. However, it has not yet been picked up by other semiconductor companies, and Transputer processing performance has lagged by over an order of magnitude behind the performance of leading chips. This situation may change as other manufacturers recognize the limitations of conventional communication mechanisms.

An indication that attention is being focussed in this direction is the serious effort on a point to point multiprocessor protocol. This is known as the Scalable Coherent Interface (née SuperBus) Project and sanctioned as IEEE P1596. The working group is chaired by David Gustavson of the Stanford Linear Accelerator Center and includes participants from others in high energy physics as well as the likes of HP, Apple, Signetics, National Semi, MIPS, Motorola, Sun, Dolphin Server Technology (Norway), etc. The emphasis is supposed to be on point to point, switched systems, but there has been a lot of recent attention given to compatibility with ring architectures. At any rate, the requirement is for 64,000 nodes and 1 GByte/sec band width *per* node (!). The idea is for the SCI protocol to be defined in silicon and in standardized software modules, rather than in the traditional incomprehensible, legalistic document which ends up being interpreted differently by each designer who tries to follow it. One may be hopeful that SCI will be an ideal environment for future distributed memory parallel computer systems, in particular for high energy physics on and off line applications.

The ACP has developed a crossbar switch with impressive performance -- by today's standards-- and which offers a technological stepping stone to the future switched environment of SCI. The ACP Branchbus Switch is a 16 X 16 programmable crossbar, implemented as the

backplane of a 6U by 280 mm Eurocard crate[5]. Its 32 bit wide communication channels operate at 20 MBytes/sec. All 16 ports can be active simultaneously. Connected in pairs, this gives an aggregate band width of 160 MBytes/sec. Processor modules may be plugged directly into the crate, much as for a conventional bus crate like VME. However, instead of the signals being connected in a bus structure, each slot in the crate is a crossbar switch point. The ACP Switchcrate thus has the modularity and convenience of commercial bus standards without the limiting effects of bus saturation.

The switch backplane uses a single ended TTL version of the ACP Branchbus protocol described earlier. A module exists to interface the switch backplane to the differential RS485 Branchbus cables that have traditionally been used to interconnect VME and Fastbus crates. In this way, we could switch several Branchbus systems of VME crates, for example in experiment data acquisition systems. More importantly, Switchcrates may be interconnected in a variety of topologies by putting a few Branchbus interface modules in Switchcrate slots and using Branchbus cables for inter Switchcrate communications. For theoretical physics, we are presently assembling a 32 crate, 256 node system (Figure 6) in this way[5]. Here the crates are connected in a hypercube topology ideal for lattice gauge calculations.

The switch is based on thirteen Texas Instruments 16 X 16 X 4 crossbar chips (TI 74AS8840) as the main switching elements and a programmable read only memory which contains routing information. The routing PROM can be changed when appropriate for different topologies of multi crate systems. The time required to reconfigure a switch is roughly half a microsecond, and reconfiguring does not affect communications along any other channel than the ones being opened or closed. A system consisting of just point to point connections may have up to 2048 processing nodes.

A 20 MegaFlop (peak), C or Fortran programmable, floating point array processor (FPAP) module has been designed for the big theoretical physics computer. It is the first processor that plugs directly into the switch. The big machine will have 8 of these FPAPs per crate and a total peak performance of 5 GigaFlops. A single crate with 15 nodes and one port to the outside world performs in practice like a Cray XMP and costs well under $100,000. At Fermilab, some 35 nodes in 4 crates are running physics code now, with 64 planned for completion by July. The full system will be completed in fall with funds from the new fiscal year.

---

5 . T. Nash, H. Areti, R. Atac, J. Biel, A. Cook, J. Deppe, M. Fischler, I. Gaines, R. Hance, D. Husby, M. Isely, E. Miranda, E. Paiva, T. Pham, T. Zmuda, E. Eichten, G. Hockney, P. Mackenzie, H.B. Thacker, D. Toussaint, High Performance Parallel Computers for Science..., in Proc. Workshop on Computational Atomic and Nuclear Physics at One Gigaflop, Oak Ridge, TN, April 14-16, 1988, FERMILAB-Conf-88/97, and references therein.

I described earlier how the ACP's new VME CPU module consists of two boards, a mother board with memory that plugs into VME, and a daughter board with a CPU, FPU, and cache. This modular design will permit us to upgrade to new processors and additional memory as soon as appropriate. It is likely that a next step, very soon, will be to design a new version of the memory board that plugs into the Branchbus Switch Crate rather than into VME. This would quickly bring RISC processing into the point to point Switchcrate environment. Subsequent processors would then be immediately available in both the VME and Switch configurations, and one might dream of processors common to the needs of both high energy experimenters and theorists. Another likely improvement for the Switch would provide a direct interface to VME, as transparent as possible, so that VME I/O devices would be readily accessible by Switch based processors. These projects are relatively simple, and, although no commitment has yet been made to them, they do indicate the direction distributed memory parallel processing in high energy physics may take.

7. Conclusion: A Need for Software Engineering

It is clear that event parallel computers, and somewhat more sophisticated distributed memory, explicitly parallel computers, are going to be an integral part of high energy physics computing for some time to come. That's an easy conclusion to make. Another conclusion, perhaps not so pleasant, is that the scale of our computing problem has reached such a level that it cannot be dealt with simply by assembling the most powerful parallel machines possible. High energy physics will simply have to come to grips with what is commonly referred to as *The Software Crisis.*

Programs of over a million source lines and over a thousand entry points, written by over 100 people at over 30 sites, cry out for attention from modern software engineering. This means more than just the latest Computer Aided Software Engineering (CASE) tools, more than SASD bubble diagrams. It means understanding how software engineering ideas such as information hiding, object oriented programming, and, even, formal methods, could be applied to an environment like that found in experimental high energy physics. It also means a new discipline: walking a tight rope between the traditional anarchy of high energy physics software development (no requirements documents, no formal reviews, physicists testing their own code, etc.) and the excessive bureaucracy of, say, DoD requirements, where producing 2167A mandated paperwork often substitutes for accurate documents and penetrating reviews. Both discipline and new software engineering ideas are clearly required as we move on to the SSC era with experiments dwarfing by an order of magnitude the ones we are now doing.

8. Acknowledgements

The ACP developments described in this paper is the work of my colleagues, H. Areti, R. Atac, J. Biel, A. Cook, J. Deppe, M. Edel, M. Fischler, I. Gaines, D. Husby, M. Isely, M. Miranda, E. Paiva, T. Pham, and T. Zmuda. I acknowledge with pleasure their strong efforts and talents.

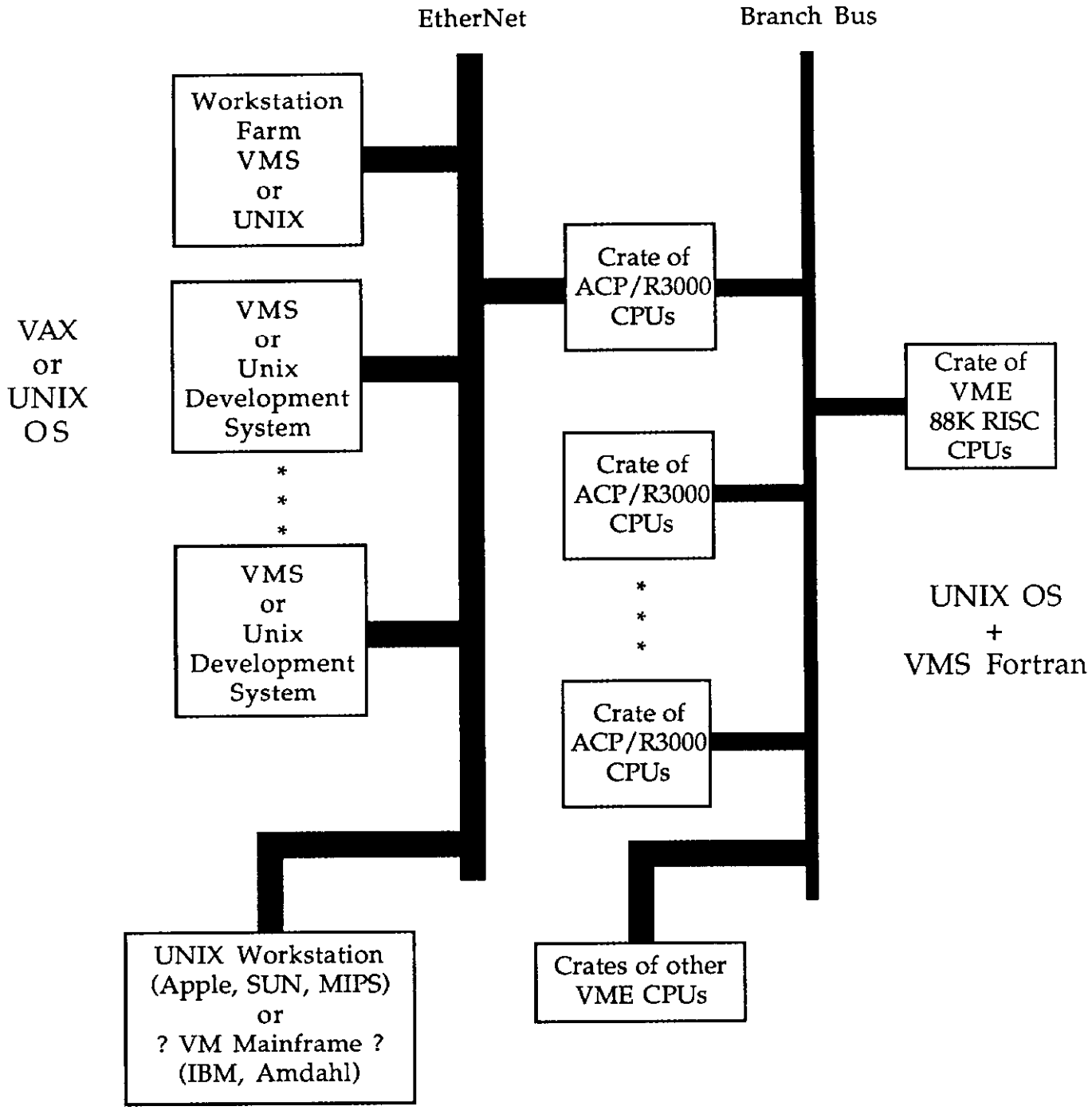
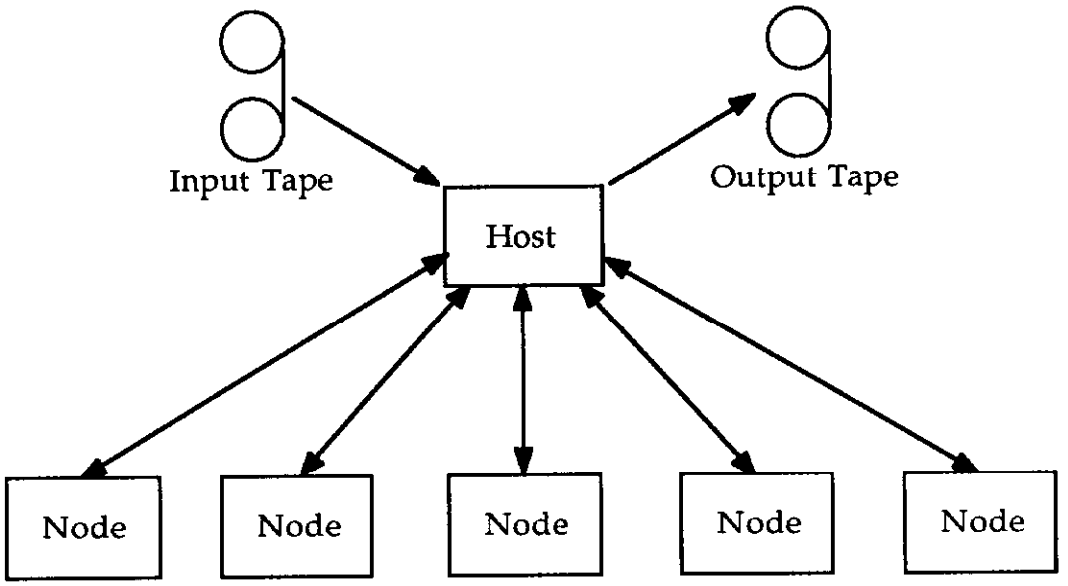


Figure 1. The Second Generation ACP Multiprocessor software allows a variety of computing engines on the dual Ethernet and Branchbus backbones.

First Generation



Second Generation

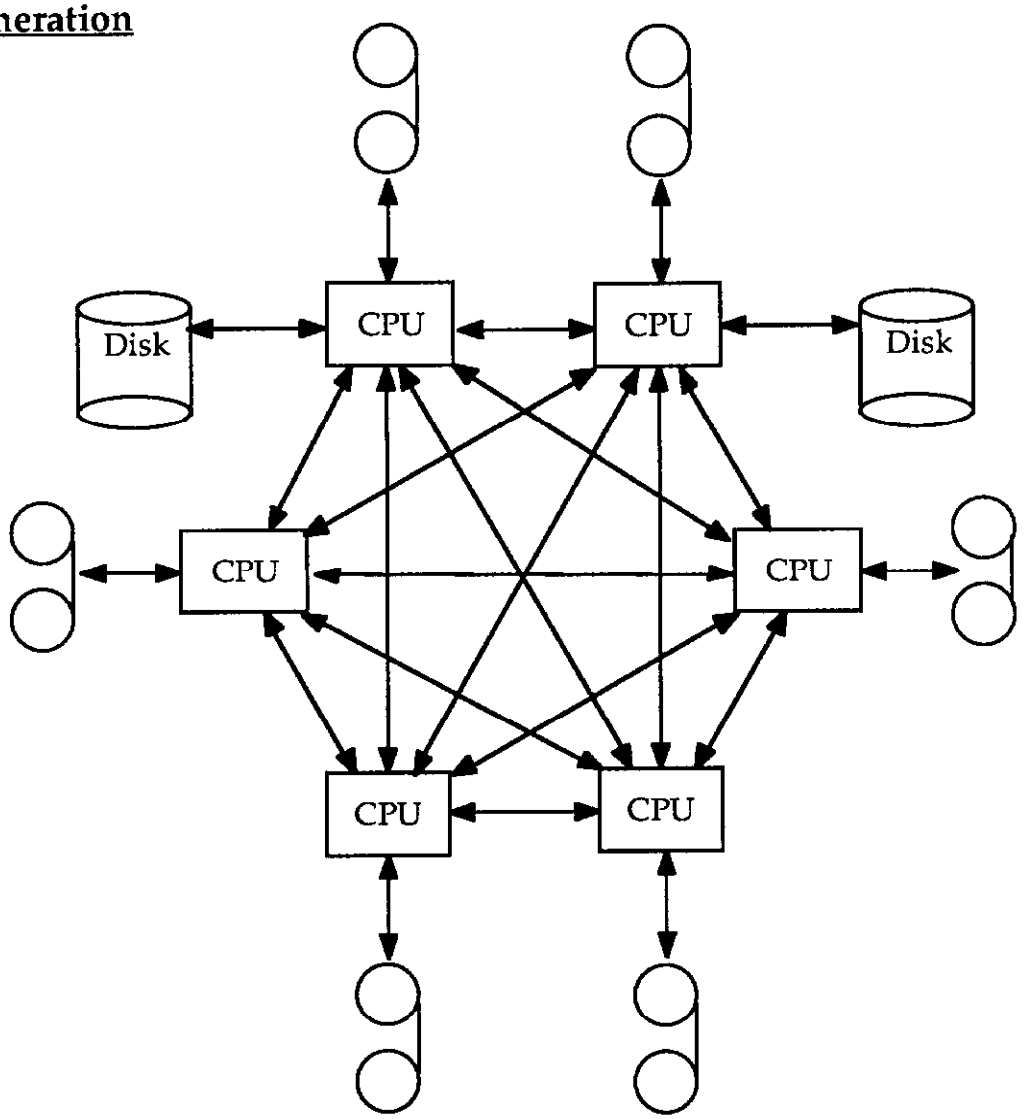


Figure 2. A comparison of communication and I/O options in the first and second generation ACP multiprocessor systems.

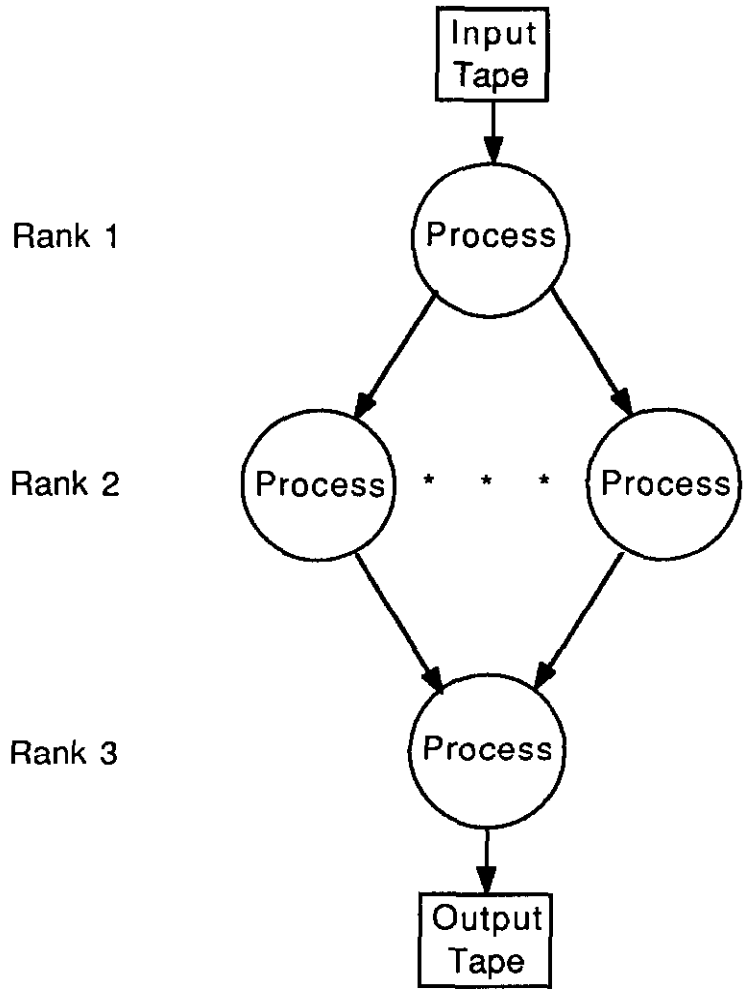


Figure 3. A simple topology suited to the event parallel experiment reconstruction task.

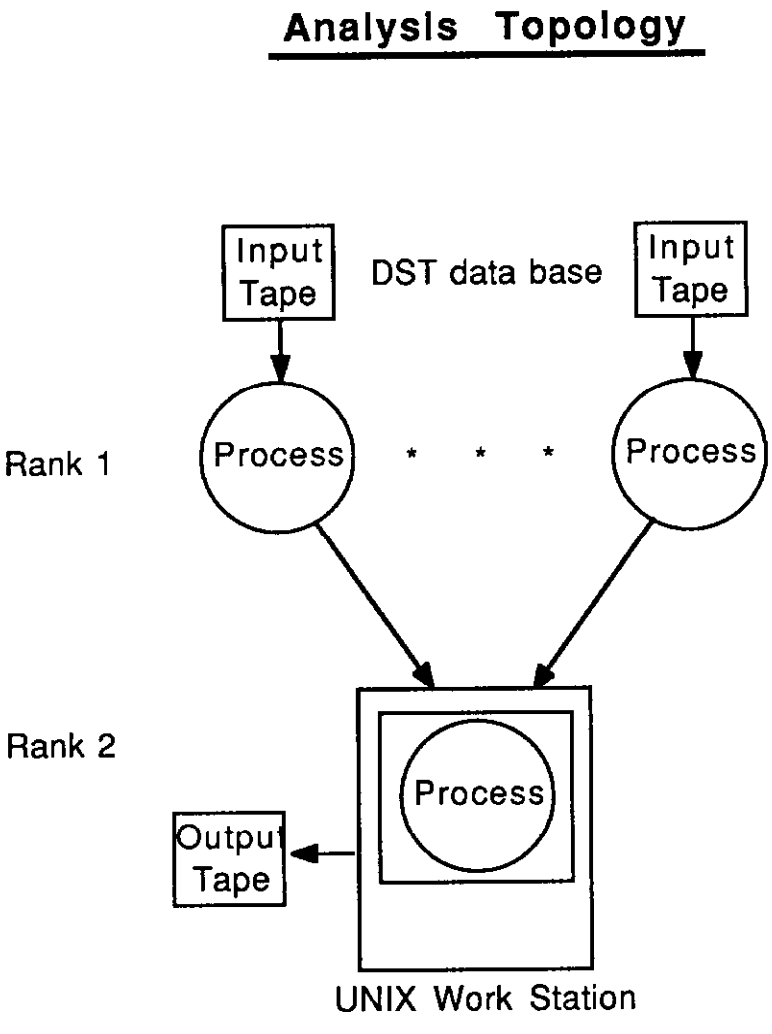


Figure 4. Another simple topology that permits large data summary tape collections to be analyzed rapidly in parallel directly into a work-station with its visualization capabilities.

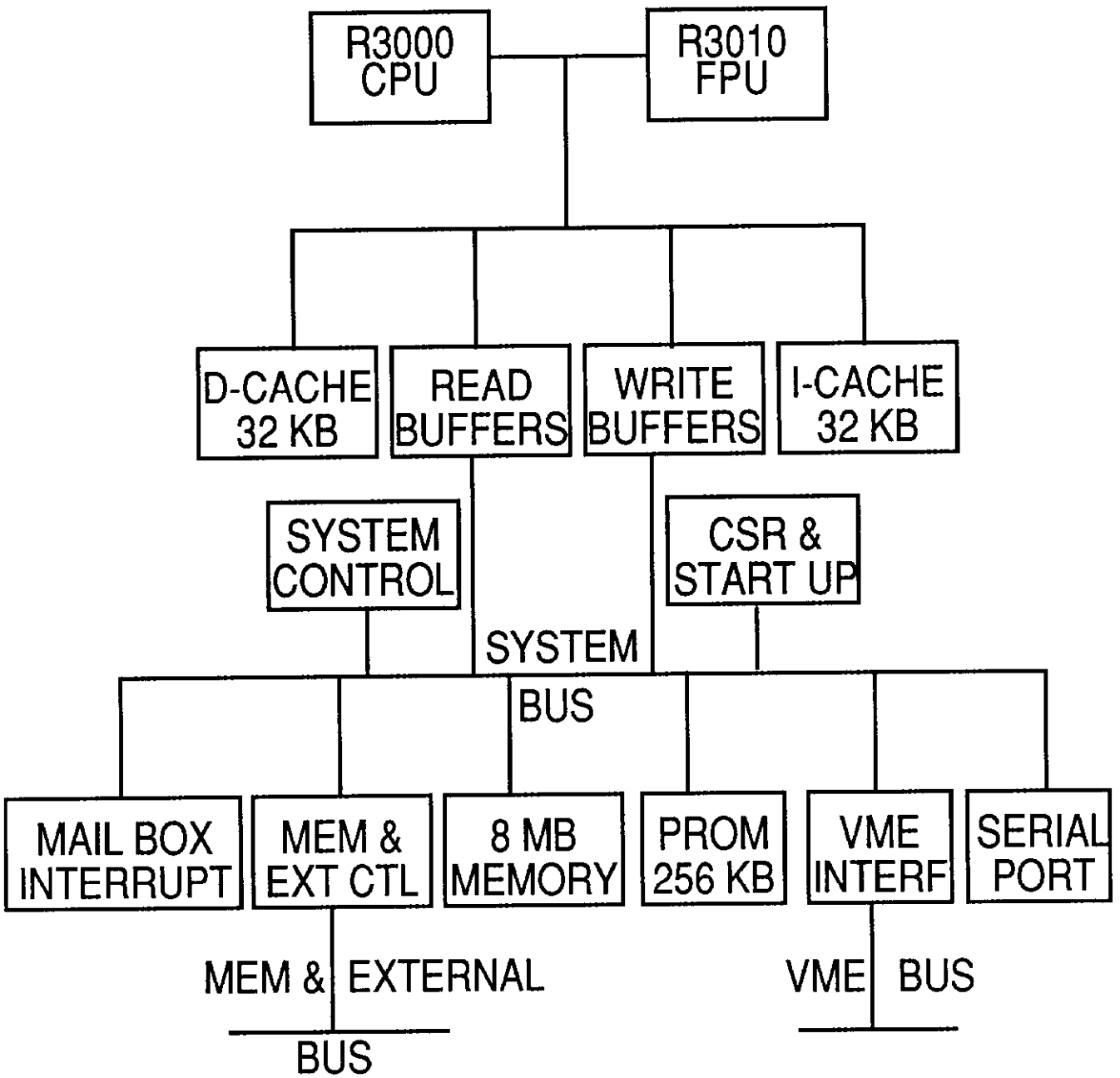


Figure 5. A block diagram of the ACP MIPS VME module.

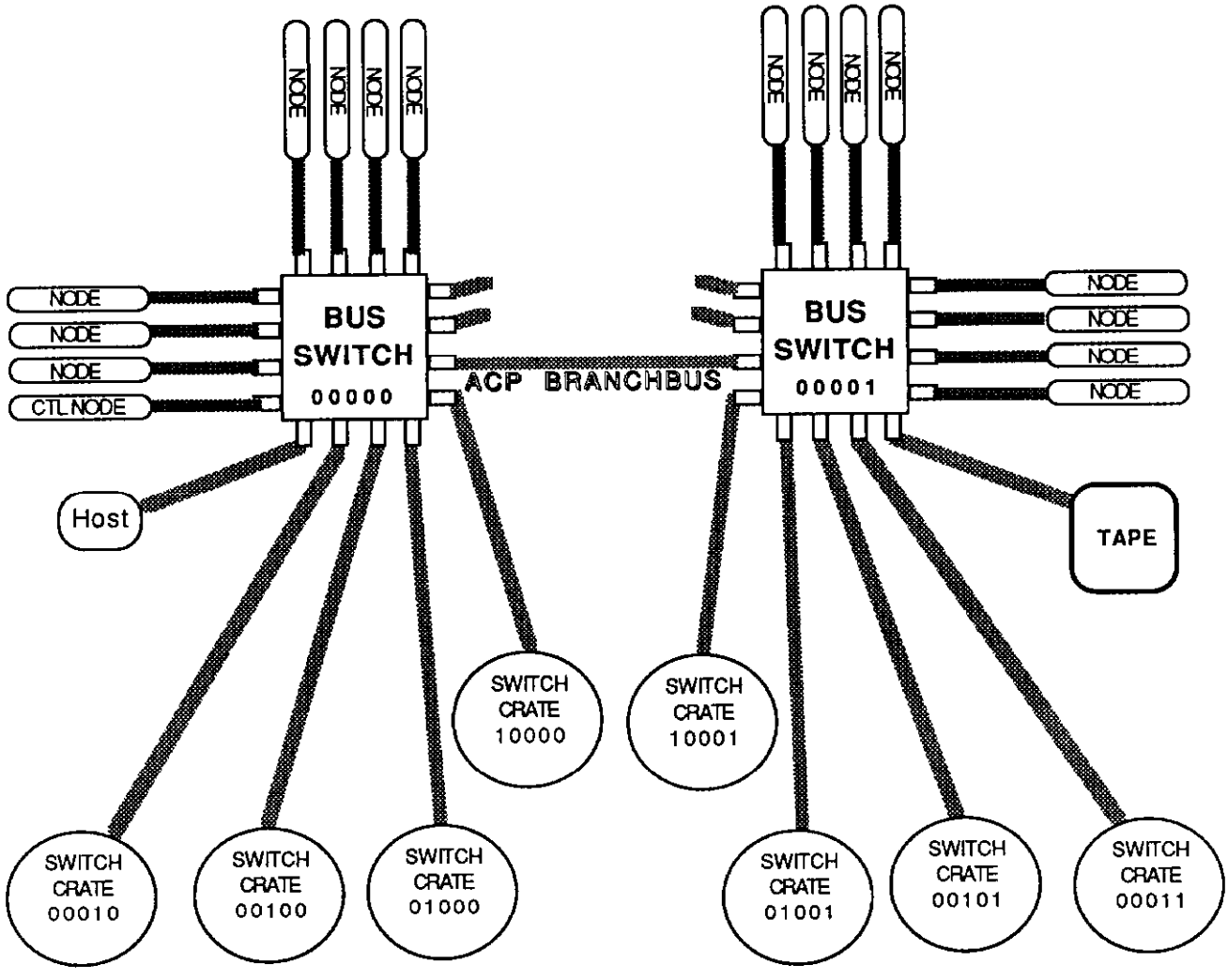


Figure 6. A diagram of the 5 Gigaflop ACP Multi Array Processor System under construction at Fermilab.