



Fermi National Accelerator Laboratory

FERMILAB-Conf-86/132

THE ACP MULTIPROCESSOR SYSTEM AT FERMILAB*

T. Nash, H. Aretí, R. Atac, J. Biel, G. Case, A. Cook,
M. Fischler, I. Gaines, R. Hance, D. Husby, and T. Zmuda

Advanced Computer Program
Fermi National Accelerator Laboratory
P.O. Box 500, Batavia, Illinois 60510, USA

September 1986

*(Presented at the XXIII International Conference on High Energy Physics, July 16-23, 1986, Berkeley, California)

THE ACP MULTIPROCESSOR SYSTEM AT FERMILAB

T. Nash, H. Areti, R. Atac, J. Biel, G. Case, A. Cook,
M. Fischler, I. Gaines, R. Hance, D. Husby, T. Zmuda

Advanced Computer Program
Fermi National Accelerator Laboratory
P.O. Box 500 Batavia, IL 60510, USA

The Advanced Computer Program at Fermilab has developed a multiprocessor system which is easy to use and uniquely cost effective for many high energy physics problems. The system is based on single board computers which cost under \$2000 each to build including 2 Mbytes of on board memory. These standard VME modules each run experiment reconstruction code in Fortran at speeds approaching that of a VAX 11/780. Two versions have been developed: one uses Motorola's 68020 32 bit microprocessor, the other runs with AT&T's 32100. Both include the corresponding floating point coprocessor chip. The first system, when fully configured, uses 70 each of the two types of processors. A 53 processor system has been operated for several months with essentially no down time by computer operators in the Fermilab Computer Center, performing at nearly the capacity of 6 CDC Cyber 175 mainframe computers. The VME crates in which the processing "nodes" sit are connected via a high speed "Branch Bus" to one or more MicroVAX computers which act as hosts handling system resource management and all I/O in offline applications. An interface from Fastbus to the Branch Bus has been developed for online use which has been tested error free at 20 MBytes/sec for 48 hours. ACP hardware modules are now available commercially. A major package of software, including a simulator that runs on any VAX, has been developed. It allows easy migration of existing programs to this multiprocessor environment. This paper describes the ACP Multiprocessor System and early experience with it at Fermilab and elsewhere.

1. INTRODUCTION

Few high energy physicists will argue with the statement that there is a severe computer cycle famine affecting several critical problems in the field. Progress is most dramatically impacted by long delays in reconstructing raw experiment event data because of the lack of adequate computing resources. Two other areas that are also hungry for large increases in available computing are theoretical lattice gauge calculations and accelerator simulations. These problems are obvious motivations for Fermilab's R&D effort in computer technology known as the Advanced Computer Program (ACP). The ACP has addressed them by developing a parallel multiprocessor system which has been described in detail previously.¹ For several months this system has been operated with the highest reliability at better than promised performance levels in the Fermilab Computer Center. Experience with the system is the main subject of this report.

Another class of computer related problems is, perhaps, less obvious, but equally critical. These involve the very poor intrinsic interface between a physicist (or any human) and a computer. The most dramatic manifestation of this is the extraordinary amount of talent and time that is presently sunk into experiment analysis programming. Additionally, the complex high level programmable triggers that will be an absolute necessity at SSC experiments imply that software must be verifiable in real time. Although this area of human interfacing has not to date been seriously addressed by the ACP or others, it must inevitably become an important activity in the future.¹

2. THE ACP MULTIPROCESSOR CONCEPT

In contrast to the requirements felt by computer science and commercial parallel processing R&D projects, the ACP has been fortunate in being driven by problems with clearly identified parallel structures. This has permitted the strategy of starting with the simplest parallelism in which each processor works independently, with no interprocessor communication. Such "trivial" parallelism meets the requirements of the event reconstruction problem, where, as shown in Figure 1a, the data from each interaction event can be processed entirely independently of others.²

This is the problem most acute in high energy physics. It is the problem for which the first large scale ACP multiprocessor has been configured and which is running in production at Fermilab. As time goes on, more generality is being added to this basic system, one step at a time, to allow high bandwidth interprocessor communication for lattice gauge calculations and cost effective direct I/O from the CPUs for data analysis.

It does not take a deep understanding of lattice gauge theory to recognize quickly the appropriateness of a grid architecture (Figure 1b) for that problem. Similarly, a "systolic" ring of processors (Figure 1c) clearly matches accelerator orbit simulations. An extensive study at Cal Tech³ has shown that hypercubes, in which each processor is connected to its neighbors in n dimensional space, are acceptably efficient architectures for essentially all "scientific" problems.

Spectrometer experiment E691 is finding that 53 68020 based processors process data at a rate nearly that of 6 CDC Cyber 175s. This corresponds to about 70% of a VAX per node including multiprocessor operation overheads. Other kinds of programs perform better: the CDF level 3 trigger software, which is under development, consists primarily of data unpacking at this stage. It runs at 97% of a VAX when a certain small CERNLIB routine available in assembly language on the VAX is also coded in assembly language on the 68020.

The cost of a single CPU varies with the market pricing of processors and memory. Including assembly and testing labor, the first 70 ACP 68020 CPUs were built for about \$2000 each at a time when the new processor chips had to be obtained at a premium.

4. A RECONSTRUCTION MULTIPROCESSOR

For event oriented problems such as reconstructing experiment data a tree structure is ideal (Figure 2). MicroVAXES are typically used as host computers handling all tape I/O. They communicate with the processors as masters on a high speed "Branch Bus" developed by the ACP. Up to 16 VME crates

are attached to this bus through a "Branchbus to VME Interface" (BVI) module, which in turn is a master on VME, where the CPU nodes reside. A "VME Resource Module" (VRM) handles arbitration in each crate.

Since the branch bus has been designed to handle very high speed on line requirements, off line limitations are determined by how many events can be handled by the host computers which manage which node gets which event, and, of course, the QBus I/O limits of about 0.5 MBytes/sec. A one MicroVAX system supports 25 events/second. For larger systems as shown in Figure 2, a 2 MicroVAX host will handle up to 100 events/second. Small physics data "logical" events may be combined into larger "physical" events, if required. The ACP software is designed in terms of VAX process modules and runs without change on a 1, 2, or 3 MicroVAX host. In fact, at substantially reduced performance, it will also run on any VAX through a commercial DMA interface in association with an ACP Branch Bus Controller (BBC). The processes in a multiple MicroVAX host communicate with each other through a shared memory in a VME crate to which each MicroVAX is connected via an ACP Qbus to VME Interface (QVI) two part module.

A branch bus crossbar switch is being built which will connect up to 8 "roots" to up to 8 "branches". The primary application of this is in on line triggers where several roots connected to an experiment data acquisition system can carry enormous data rates - up to 160 MBytes/second. By connecting 2 branch busses, in and out, to each crate, and carrying the out branch around to the input end of the switch, a powerful inter node connection system is possible for such problems as lattice gauge theory.

5. MULTIPROCESSOR SOFTWARE SUPPORT

An application program destined to run on the ACP Multiprocessor must be divided into two parts. One, running in the host, handles all I/O. The other, running in each of the nodes, contains the actual number crunching part of the program. The ACP provides system subroutines to communicate between the host part and the node part. Routines are provided to broadcast calibration constants at the beginning of a run and to sum statistics at the end. Individual events are sent, and the corresponding results retrieved, asynchronously as instructed by user called send and get subroutines.

It is very easy to convert an existing reconstruction or Monte Carlo program to meet the multiprocessor requirements. In a program of tens of thousands of lines, as little as 20 additional instructions are required. Recently, a visitor, totally inexperienced with the ACP system, was able to bring up the Lund Monte Carlo program in 2 days. More complex programs, of course, take longer, with time often spent finding existing program bugs exposed by the change of compiler and the larger amount of data being processed.

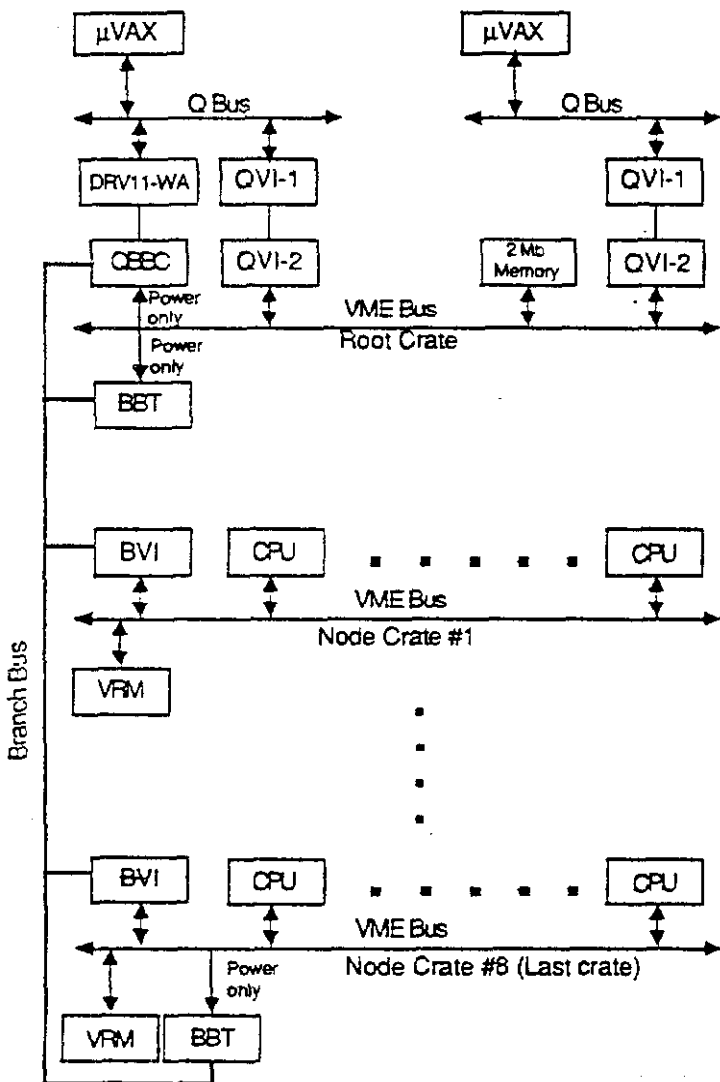


Figure 2: Block diagram of ACP Multiprocessor in the Fermilab Computing Center.

In addition to the application program's host part, ACP system resource management, error handling,

integration of memory and bus control circuitry will free up space and allow additional features, possibly including two processors on one module. The goal will be to dramatically increase the cost effectiveness over the present approximately \$2000/equivalent VAX.

A branch bus controller that resides in VME and allows any VME master to master the branch bus is presently in design. This can be used with VME tape controllers for direct I/O without the speed restrictions imposed by the MicroVAX and its QBus. This "VBBC" has important other applications in on line triggers. It will also make it possible to develop good intercrate communications, noted earlier, for problems like lattice gauge.

In collaboration with the Fermilab Theory Department, the ACP will develop and exploit this interconnection mechanism. Lattice gauge experience will first be gained at the single crate level where software protocols for communicating between node Fortran programs will be established. This protocol will remain the same as the interconnection mechanism becomes more sophisticated and the size of the multiprocessor being used for lattice calculations grows.

Much of this calculation involves a well defined kernel of floating point operations. Large performance increases are possible with a single board array processor coprocessor attached to each node. This device will be based on 32 bit floating point multiplier chips and will have performance in the range of 10 million floating point operations per second per node. Access to these microcoded modules will be through Fortran subroutine calls. Design work will start at the end of 1986.

So far, the emphasis has been on computing intensive problems. If anything, the delays incurred in passing huge numbers of reconstructed data summary tapes (DSTs) through a computer center for physics analysis has been an even more severe impediment to progress. A typical large experiment with 100 or more DSTs must wait over a week to see a new set of histograms incorporating new analysis cuts or variables if the full data sample is used. The opportunity exists to dramatically improve this situation by attaching a WORM (Write Once, Read Manytimes) laser disk to each multiprocessor CPU.

The cost of a WORM disk is rapidly falling below \$2000. Their capacity is of the same magnitude as a high density tape. The match appears good: the disk can be read through in less than a half hour which is about how long it takes a VAX class ACP CPU to process a typical experiment's analysis program for one tape. One DST disk would be attached to each of a hundred or so nodes allowing a complete experiment reconstructed data base to be

processed in well under an hour instead of over a week. A low cost mechanical juke box like contraption would support multiple sets of experiment disks.

The dramatic improvement in analysis turn around time that will result from such hardware would strongly motivate work on another problem, already alluded to in the introduction. It would no longer be tolerable for physicists to spend hours struggling with histogram packages, unpacking routines, and Fortran, each time they needed to change a few histograms or add variables. We hope that, in time, cheap work station/personal computers will carry physicists' analysis tools that can be controlled with a few clicks of a mouse. Physicists deserve the same level of "friendliness" that businessmen now find routine with their spreadsheets and data base programs on the better personal computers.

ACKNOWLEDGEMENTS

We would like to acknowledge the work of Chip Kaliher of the Fermilab Computing Department in developing the batch submission software and Chris Sliwa and his E691 colleagues for their efforts in bringing up the first production software on the ACP multiprocessor. We also acknowledge the important contributions of Steven Bracker during the conceptual design phases of this project. Thanks also to Terry Grozis for preparing this manuscript.

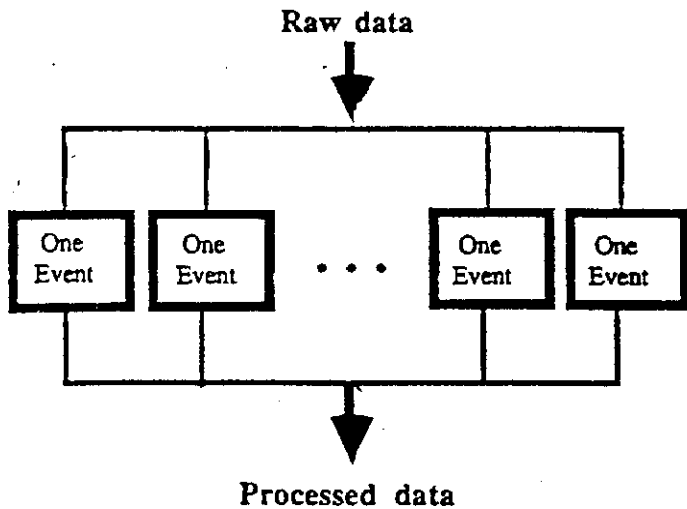
REFERENCES

- (1) T. Nash, et al., "The Fermilab Advanced Computer Program Multi-Microprocessor Project," conference proceedings, Computing in High Energy Physics, Amsterdam, June 1985 (North Holland), and references therein.

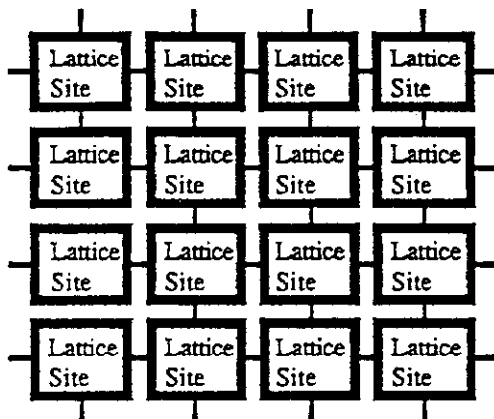
M. Fischler, "Software for Event Oriented Processing on Multiprocessor Systems", proceedings, Symposium of Recent Developments in Computing, Processor and Software Research for High-Energy Physics, Guanajuato, Mexico, p. 175, 1984.

High energy physicists with access to DECnet may access a complete list of ACP technical manuals at `BSNDBG::ACPDOCS_ROOT:[DOCS]DOCLIST.DOC`
- (2) We believe this concept was first brought out in Paul F. Kunz, "The LASS hardware processor", Nucl. Instr. Meth., pgs. 135, 435 (1976).
- (3) G.C. Fox and S. Otto, "Algorithms for Concurrent Processors", Physics Today, pg. 50, May, 1984.

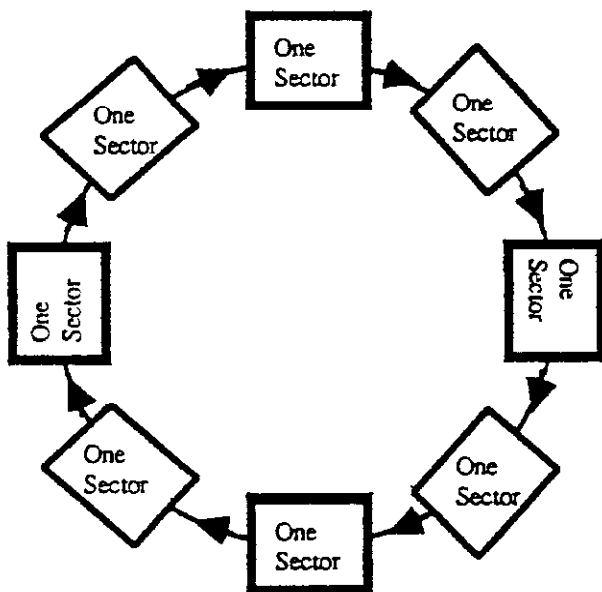
a) Experiment reconstruction:



b) Lattice gauge calculation:



c) Accelerator simulation:



In fact, there may be more appropriate bussed structures than those cited above as immediately obvious for lattice gauge and accelerator calculations. However, what is important, is that some arrangement of processors, each with its own "local" memory and capable of very cost effective computing in a high level language (Fortran), will be the computer of choice for these three, and many other, scientific problems.

Such computers are referred to in the jargon of computer science as general purpose, MIMD (multiple instruction, multiple data stream), explicitly parallel, local memory machines. The ACP is addressing the common requirements of such machines with design goals that emphasize cost effectiveness, user friendliness, and configuration flexibility.

The processor needs represented by the little boxes are essentially the same in each of the three architectures of Figure 1. They are met by maximally cost effective single board computer "nodes". Given the requirement of Fortran as the *lingua franca* of scientific computing, the most cost effective computing engines are the new generation of 32 bit microprocessors. These, supported by ample memory, form the basis of ACP nodes. In the future they will be augmented with single board array processors.

To complete the common requirements for a usable multiprocessor of this type, the ACP has developed a high speed interfacing bus system to connect a commercial host computer to the array of nodes. This hardware is accompanied by a complete package of appropriate system software. The software handles compiling, downloading, and debugging of node programs, host-node data communication, and (soon) node-node communication.

3. ACP CPU NODES

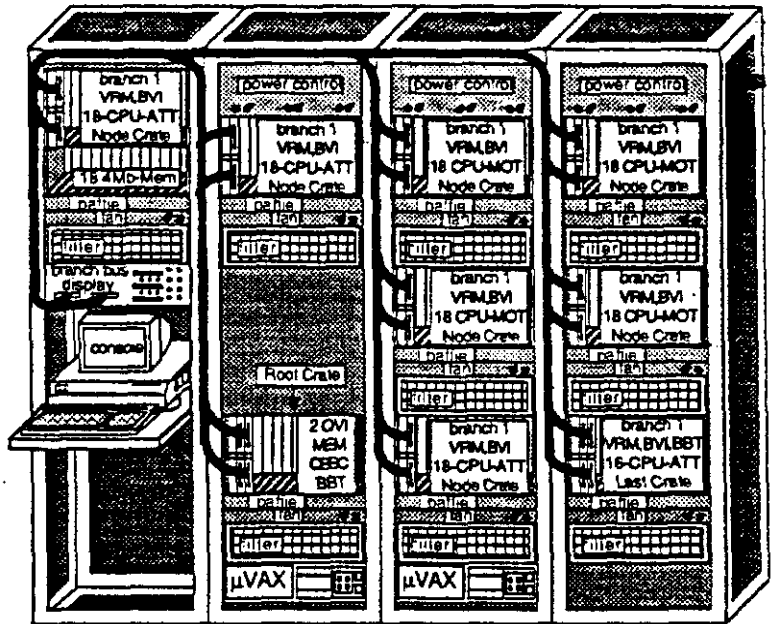
Two CPU designs have been produced, one based on Motorola's 68020 microprocessor, the other on AT&T's 32100. Each design includes the corresponding floating point coprocessor and 2 MBytes of on board memory. Both run at 16.6 MHz with one wait cycle memory reads and writes. The modules are in standard double high VME packages with full VME single word slave and master protocols. In addition, VME block transfers in slave mode at upwards of 20 MBytes/sec are supported for on-line trigger applications. A memory extension of 4 (or 6, if required) MBytes on a second board in a neighboring slot, or in a crate directly below, is available. The modules also have a daughter board coprocessor interface with several possible future applications.

The 68020 based CPU is supported by a full ANSI Fortran 77 compiler from Absoft Corporation. An AT&T compiler is presently used for the 32100 CPU. Performance for these processors is very program dependent. For track reconstruction code, it ranges between 60% and 90% of a VAX 11/780 with floating point accelerator and VMS 4.x Fortran. In real production experience, the Tagged Photon

Figure 1: The "natural" local memory, parallel, architectures for the three most important computing problems of high energy physics.

and I/O management reside in the host computer. Only a minimal operating system runs on the nodes. This is part of the philosophy of making it easy to incorporate new and competitive processors as nodes without a large new investment in software development.

Typically, the production host MicroVAX is connected via DECnet to a mainframe VAX which is referred to as the development host. While the production host is conceived of as a single user environment, running batch jobs requiring huge numbers of computer cycles, the development host is a multiuser environment where program development is accomplished. The large machine has the disks necessary to store programs. It is also the place where a full, multi process, simulation of the ACP multiprocessor is available for users to debug programs before bringing them up on real nodes. In addition, the development host supports batch queues of jobs. Users submit jobs into these batch queues which are automatically submitted over the network to the production host after a previous job is finished, notifying operators of tape mounts required.



140 Node
Single Branch
2 μVAXs

6. PRESENT STATUS AT FERMILAB

At Fermilab, a large VAX cluster is the development host. In addition to what will be a 140 processor production system, there is also a small development system with a few CPU nodes of each type for compiling and debugging node application software. This system is supported by its own MicroVAX.

At this writing (September 16), all 70 production boards based on the 68020 are built and running, as well as 4 based on the AT&T 32100. The remaining 66 of the original run of 32100 CPUs are being assembled and will be available during the next month or so. In addition, about a half dozen 68020 based ACP CPUs built by Omnibyte Corporation are in use at Fermilab, and several have been delivered elsewhere. Because of a heavy demand for these powerful new processors for testing, software development, and exploratory activities, only 54 CPUs have been allocated to the production system, including one 32100 CPU. As the remaining modules are completed, they will mainly be added to the production system.

The system in the Fermilab Computer Center is shown in Figure 3. With the exception of the remaining CPUs being assembled, all components of the system are in place. Operation was turned over to Computer Center personnel in early July, and the 53 68020 node system has been running since then, with essentially no down time, reconstructing data from the Tagged Photon Charm Production experiment E691. In the first month of running, the experiment completed as many tapes as in the previous 7 months during which it used an average of 30% of the Computer Center's capacity.

Several other experiments (and lattice gauge theorists) with large and urgent computing needs are preparing to run on the ACP system. An additional 30 CPUs are being ordered, and it is anticipated that a second full scale system with over 100 CPUs

Figure 3: A drawing of the first ACP Multiprocessor installation in the Fermilab Computer Department.

will be ordered and installed in the new budget year starting October 1. Outside Fermilab, a number of universities and laboratories in the U.S., Canada and overseas have serious intentions to use the system for on or offline applications. Omnibyte Corporation of West Chicago, IL, which is commercializing the system, has orders for over 100 CPUs, mostly from outside Fermilab.

An interface to the Fastbus standard has been developed in collaboration with the CDF group at Fermilab. This interface has been tested writing data without error for 48 hours at 20 MBytes/second through the branch bus and VME into an ACP node memory. In addition to the CDF trigger application, this Fastbus interface will be used for a high level trigger in Yale-Los Alamos MEGA experiment at LAMPF. The switch, described earlier, is in development.

7. FUTURE PLANS

The successful multiprocessor gives the ACP the opportunity to exploit this system's flexibility and apply it to important new problems. New technology will also permit improved performance for experiment reconstruction.

Over the next year a new CPU module will be designed. Its specifications will be determined after a study of what commercial new VLSI products will allow the best improvement to the existing designs. The new board may be based on higher speed versions of the microprocessors presently being used or on different families if Fortran benchmark tests warrant it. It is likely that the on board memory will be based on 1 Mbit DRAMs instead of the present 256 K, and this would result in a corresponding increase in available memory. The