Fermi National Accelerator Laboratory

FERMILAB-Conf-88/211

Multi-Processor Developments in the United States for Future High Energy Physics Experiments and Accelerators*

Irwin Gaines

For the Advanced Computer Program:
H. Areti, R. Atac, J. Biel, A. Cook, J. Deppe, M. Edel, M. Fischler,
R. Hance, D. Husby, T. Nash, T. Pham, and T. Zmuda
Fermi National Accelerator Laboratory
P.O. Box 500, Batavia, Illinois 60510

March 1988

*Presented by Irwin Gaines at the Adriatico Conference on the "Impact of Digital Microelectronics and Microprocessors on Particle Physics," International Centre for Theoretical Physics, Trieste, Italy, March 28-30, 1988.

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

Multi-Processor Developments in the United States for Future High Energy Physics Experiments and Accelerators *

Irwin Gaines

For the Advanced Computer Program
H. Areti, R. Atac, J. Biel, A. Cook, J. Deppe, M. Edel, M. Fischler, R. Hance, D. Husby, T. Nash, T. Pham, and T. Zmuda
Advanced Computer Program
Fermi National Accelerator Laboratory+
Batavia, Illinois 60510 USA

ABSTRACT

The use of multi-processors for analysis and high-level triggering in High Energy Physics experiments, pioneered by the early emulator systems, has reached maturity, in particular with the multiple microprocessor systems in use at Fermilab. It is widely acknowledged that such systems will fulfill the major portion of the computing needs of future large experiments.

Recent developments at Fermilab's Advanced Computer Program will make such systems even more powerful, cost-effective, and easier to use than they are at present. The next generation of microprocessors, already available, will provide CPU power of about one VAX 780 equivalent/$300, while supporting most VMS FORTRAN extensions and large (>8MB) amounts of memory. Low cost high density mass storage devices (based on video tape cartridge technology) will allow parallel I/O to remove potential I/O bottlenecks in systems of over 1000 VAX equivalent processors. New interconnection schemes and system software will allow more flexible topologies and extremely high data bandwidth, especially for on-line systems. This talk will summarize the work at the Advanced Computer Program and the rest of the U.S. in this field.

1. INTRODUCTION

High-Energy physicists have always wanted more computer power than they could afford to buy in the commercial marketplace. However, the natural parallelism inherent in the HEP computing problem (running the identical program on many millions of different events) suggests a simple parallel processing solution. The pioneering emulator work by Kunz *et al.* at SLAC demonstrated the feasibility of using multiprocessor systems to provide cost-effective computing. Indeed, one of the earliest nine processor 168E systems is still doing work today running Monte Carlo programs for the LASS detector, almost a decade after it was first commissioned.

More recently, the advent of powerful 32-bit microprocessors has allowed the development of even more convenient and cost-effective parallel processing systems. Such systems are now an acknowledged important component of computing in high-energy physics.

---

*Talk given by Irwin Gaines at the Adriatico Conference on the "Impact of Digital Microelectronics and Microprocessors on Particle Physics," International Centre for Theoretical Physics, Trieste, Italy, March 28-30, 1988.

+Fermilab is operated by Universities Research Association, Inc., under contract with the U.S. Department of Energy.

They have made crucial contributions to data analysis in a number of current generation experiments, and will be indispensable for the large experiments of the SSC era.

In particular, the systems developed by the Advanced Computer Program (ACP) at Fermilab are rapidly becoming a standard, with over 30 installations in universities and laboratories worldwide. The first generation ACP systems, which provide CPU power at a cost of less than $2500 per VAX 780 equivalent, were brought on-line two years ago.[1] The initial 110 processor system has been heavily used at Fermilab for physics event reconstruction during this period, while a number of additional systems for both off-line and on-line use are currently being installed.

In this paper I will discuss the new developments which will lead to ACP systems of close to an order of magnitude greater cost effectiveness in the next year. This second generation ACP project will also allow much higher bandwidth for both I/O and interprocessor communication, and will have software tools allowing almost any UNIX or VMS based processor to be used as a node in a multiprocessor ACP system. I will also describe several other U.S. processor projects, concentrating on off-line aspects, since on-line use will be covered in Conetti's talk at this conference.

2. NON-ACP PROJECTS

Although there are no plans for future generations of emulators, both the early 168E and current 3081E systems are in use for data analysis and Monte Carlo. 3081E production systems are running at SLAC, Santa Cruz, Harvard and Oklahoma.[2] They are also being used on-line in a FASTBUS environment for triggering the MARK II experiment at SLC.

The most promising commercial multiprocessor system is scheduled to be installed in May at the University of Florida for use by the CLEO collaboration. It will consist of 32 MicroVAX 3200's (without the keyboard, monitor and graphics boards), 6 full MicroVAX 3200 workstations, and 5 MicroVAX 3600's connected by Ethernet, with 8 MBytes of memory on each processor and 5 RA82 disks (a total of 3 GBytes). The system will be run as a loosely coupled multiprocessor, with VMS acting as a file server for the diskless 3200's, although ACP-like master/slave host/node software may also be written for the system. This system will provide over 100 VAX 780 equivalents of processing power in a convenient package, at a cost (based on a research agreement with Digital) of under $4000 per VAX 780 equivalent. Costs on the open market will likely be somewhat higher. Disadvantages of this system are the limited bandwidth provided by Ethernet and the relatively high cost, particularly compared with second generation ACP-like systems. It is nevertheless an important indication that a leading computer company recognized and is beginning to respond to the availability of low-cost high-power multiprocessors. Sun Microsystems is also apparently committed to low-cost farmlike architectures using their SPARC CPU.

An unusual and powerful approach to multiprocessing is the data driven hardware processor designed at Nevis Labs by Bruce Knapp and Bill Sippach. The processor consists of 380 boards of 40 different types and is programmed by the way the individual modules are cabled together and downloaded with constants. When configured to do track finding it can reconstruct 100,000 events per second, while the VAX 780 takes 0.1-1.0 sec. to reconstruct an 8 track event. This enormous increase in processing power comes from a combination of pipelining and parallelism (ordinarily all 380 boards are processing simultaneously), fast cycle times (one operation on each module every 25 nsec), and because the modules carry out powerful operations that are specialized for track finding. The processor was used during the summer of 1987 to process 10^9 events from 8000 data tapes from BNL E766, a heavy quark spectroscopy experiment, and will be used for on-line processing in E690 at Fermilab.

An interesting and ambitious new approach to on-line multiprocessor systems that has been proposed by the Computer and Electrical Engineering Departments at Fermilab goes by

the name of FUSE, for FASTBUS Uniform System Elements. The proposal is aimed at architectures for greatly increased data acquisition throughput for next generation fixed target and collider experiments. The proposal specifies a number of building block functional units, which can then be combined in a variety of ways to build FASTBUS modules optimized for different applications.

Presently identified system elements include:

1) FASTBUS Master Interface, which provides a high-speed FASTBUS crate interface allowing efficient execution of sequential sets of FASTBUS operations for flexible readout of data from multiple slaves;
2) FASTBUS Auxiliary Port Interface, which provides a second data port onto the module through the FASTBUS auxiliary connector to any of a number of buses, including Lecroy ECLine, RS485, SCSI, and ACP Branchbus.
3) Module Control Element, which provides centralized control of the entire module, including connection to a serial port and a local area network; and
4) Event Processing Units, which provide application specific processing for data formatting and compression and event monitoring tasks as part of the readout system.

A typical FASTBUS module will include several processing elements, which can be digital signal processors, floating point engines, or general purpose processors depending on the application. These uniform system elements can then be combined into readout controllers, event builders, and front-end processors for future high-performance data acquisition systems.

3. SECOND GENERATION ACP SYSTEMS - NEW CPUS

Future multiprocessor systems will clearly benefit from the increasingly powerful processors now becoming available. Besides the "trivial" speedups by using newer versions of the processors already in use (for example, 25 MHz 68030s replacing 16 MHz 68020s) there are entirely new families of chips that can be used. In particular, the popular Reduced Instruction Set Computer (RISC) architecture has led to several new processors. Also heartening is the trend that the new processors often have software (including FORTRAN and COMPILERS and UNIX operating systems) available for them even before the hardware is available.

New processors that will offer at least a factor of three, and in some cases a factor of ten, more performance than the first generation ACP processors include:

1) The R2000 RISC chip from MIPS
2) The Fairchild, now Intergraph Clipper chip set
3) The AM29000 RISC chip from AMD
4) The SPARC RISC chip from SUN
5) The T800 transputer from INMOS
6) The 88000 RISC chip from Motorola
7) The 32532 processor from National Semiconductor

Extravagant claims are made for all of these processors by their manufacturers, suggesting potential performance of up to 17 million instructions per second (MIPS). Such claims need to be taken with a grain of salt, as a single instruction does not do the same thing on different processors. For high energy physics use, a natural standard is to take the VAX 11/780 as representing 1 MIPS in performance, and to measure the new processors only in comparison to a VAX on high energy physics codes written in high level languages, thus evaluating both the hardware and the compilers. It matters not how many instructions per second the CPU can execute if the instructions are not useful in FORTRAN or C and if the compilers fail to provide sufficient optimization.

The ACP has performed such physics benchmarks on several of the new chips. Based on these results, we have chosen to design a VME CPU board based on the MIPS R2000 CPU. The standard ACP FORTRAN benchmark suite (consisting of three programs: a small Monte Carlo/track reconstruction package; an actual fixed target tracking code running on experimental data; and a floating point intensive theoretical calculation) was run on a 16 MHz MIPS system. For the three benchmarks, performances were 7.9, 6.4, and 7.4 times that of a VAX 780, and in each case more than a factor of 10 greater than the present generation ACP 68020 boards.

The ACP MIPS processor board (see the block diagram in Figure 1) consists of:

1) a 16 MHz MIPS R2000 CPU;
2) a 16 MHz MIPS R2010 floating point unit;
3) Four 16 MHz Write Buffers;
4) a 32 KByte instruction cache;
5) a 32 KByte data cache;
6) 8 MBytes of parity checking main memory, made up of 1 Mbit (100 nsec nibble mode access) DRAMs, expandable to 16 MBytes;
7) interval timers that can interrupt the CPU; and
8) a full function VME Master/Slave interface supporting 20 MBytes/sec block transfers.

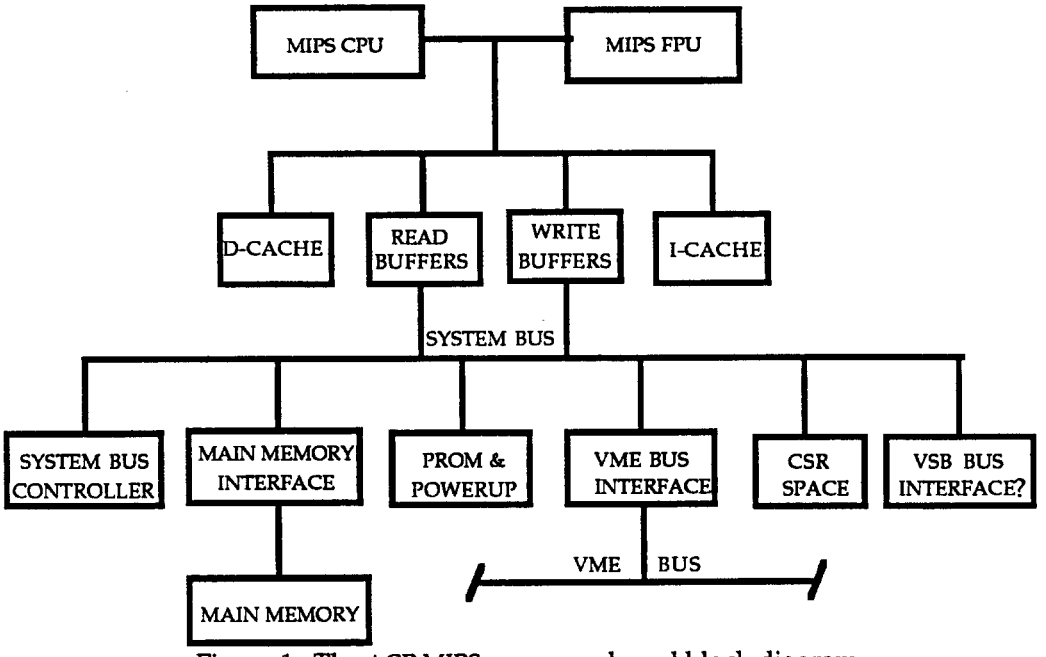

Figure 1. The ACP MIPS processor board block diagram.

Since the MIPS CPU chip has on-chip memory management, the board will be able to run the full UNIX operating system, booting either from a VME disk drive or using the Network File System (NFS) over the Branchbus.

This board will provide high-level language processing power with a cost effectiveness of roughly $300/VAX 780 equivalent, based on the physics benchmarks described and the features listed above. The FORTRAN compiler for these chips is the best we have encountered for a microprocessor, supporting many VMS extensions and comparing favorably with the VMS compiler in convenience and sophistication. Full UNIX program development tools are available. This processor will form the cornerstone of the second generation ACP systems. As

discussed below, other VME processors running UNIX can also be used in the system, giving an opportunity for single board computer manufacturers to compete directly at the board level. Whenever a system is assembled, the most cost-effective processors currently available can be used.

4. SECOND GENERATION ACP SYSTEMS - NEW INTERCONNECT TOPOLOGIES

The ACP Branchbus was developed to deal with the problem of linking several high performance local buses to a host and allowing high speed block transfers.[3] No commercial alternative was available. The original Branchbus is a 32-bit bus connecting a single master (QBus, Unibus or FASTBUS) to multiple crates (VME), supporting block transfers at up to 20 MBytes/sec. Three improvements to the Branchbus will allow higher performance and more complex interconnection schemes in future systems.

First, the Branchbus now supports multiple masters. A distributed arbitration scheme similar to that used on the SCSI bus allows up to 16 masters to share the bus. Existing masters can be used in multi-master systems by replacing their existing Branchbus Interface Daughter Board with a new Multi Master Branchbus Interface Daughter Board which handles the arbitration transparently to the user.

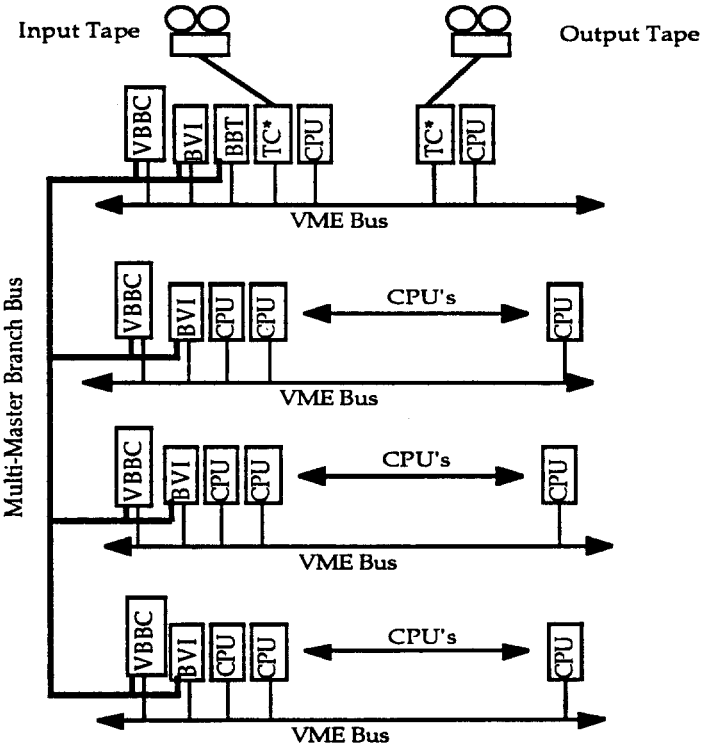

Figure 2. A High Performance Off-Line ACP System using the VBBC.

The VBBC (VME Branchbus Controller) allows any VME master, in particular any node in an ACP multiprocessor system, to act as a Branchbus master and read or write to any Branchbus address. This allows any processor in the system to communicate with any other processor without any host intervention, allowing the more elegant system architectures described below. (Figure 2 shows a block diagram of a high performance off-line ACP system using the

VBBC.) The VBBC is a VME slave and Branchbus master. It is a shared resource in the VME crate, allocated on a first-come first-served basis by a test-and-set bit in its control register. Once programmed with the Branchbus control words and address, the Branchbus cycles occur transparently to the VME master which simply reads or writes data to the VBBC. The design of the VBBC is complete, and first prototypes will be available in early summer.

The ACP Branchbus Switch allows full crossbar interconnection of up to 16 Branchbuses (or more using multiple switches). With this switch, any Branchbus master device can connect to any slave in the entire switch connected system. All channels of the switch can be active simultaneously. For example, eight of the Branchbuses could be connected to the other eight, all transferring data simultaneously giving an aggregate bandwidth of 8x20 MBytes/sec or 160 MBytes/sec (in addition to any local bus activity on any of the VME crates in the system). The Switch is based on the TI 74AS8840 16x16 four bit crossbar chip. The Switch is a backplane incorporating 14 of these chips. Modules may be plugged into the Switch Crate (see Figure 3) much as with VME. However, instead of the signals being connected in a bus structure, each slot in the crate is a crossbar switch point. Use of the Switch will allow future multiprocessor systems to obtain as much bandwidth for interprocessor communication as is required. The first two Switch crates are built and working.

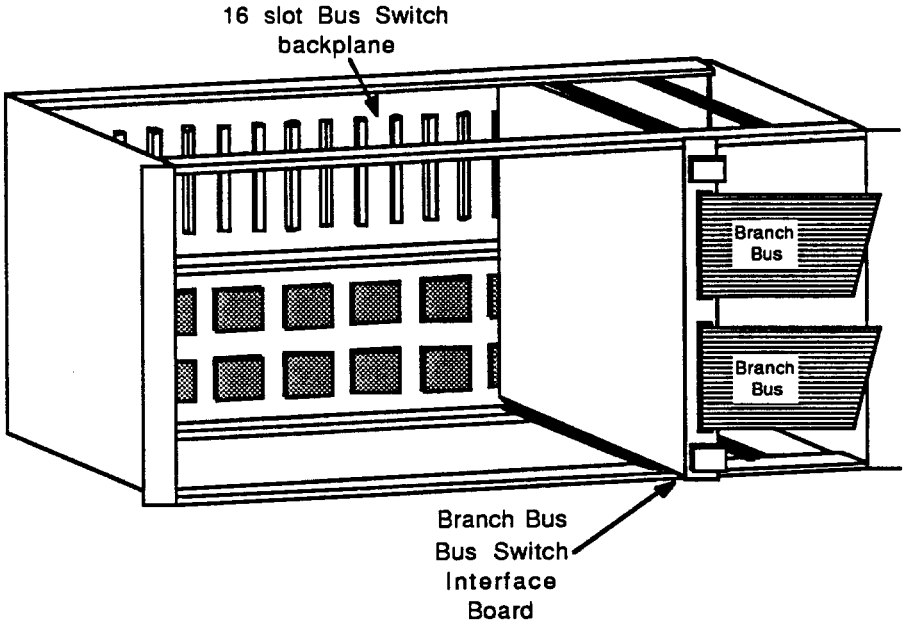

Figure 3. The ACP Branchbus Switch. The backplane uses single ended TTL Branchbus protocol.

5. SECOND GENERATION ACP SYSTEMS - NEW INPUT/OUTPUT

The increased CPU power available in future multiprocessor systems will require concomitant increases in input/output capabilities beyond the present standard of 6250 BPI magnetic tape. There will be a demand both for significantly higher density mass storage devices (to accommodate the ever-increasing amounts of data planned for future experiments, one Fermilab experiment is talking about tens of billions of events) as well as increases in I/O bandwidth.

Three developments will help confront the I/O problem. First is an increasing reliance on I/O devices that interface directly to the multiprocessor system bus. Unibus and QBus tape drives are being replaced by VME tape interfaces, such as the Ciprico TM3000, giving a bandwidth potential of the 20-30 MBytes/sec allowed by VME compared to the roughly 1 MByte/sec possible with minicomputer buses.

Second, new I/O devices are replacing magnetic tape. Optical disks, video tape cartridges (both VHS and 8mm formats) and digital audio tapehold out promise of much more compact (and cost effective) mass storage than conventional tapes. Current devices allow bandwidths comparable to those achieved with mag tapes, with further improvements expected in the next year. For example, the ACP group is studying the 8mm video tape device manufactured by Exabyte Corporation. Currently available devices store 2 Gbytes (as much as 12 conventional mag tapes) on a standard $10 tape cartridge. The drives cost less than $3000 and can deliver data at 250 KBytes/sec to VME (through a SCSI bus interface) or to QBus. Both density and speed are expected to double in the next year. No standard has yet emerged in this rapidly developing field, and the long-term reliability of these devices has not yet been established. However, it is clear that future systems can count on considerably better I/O performance than available in current systems.

Thirdly, this availability of cheap high-capacity mass storage devices which interface directly to the multiprocessor bus allows the possibility of parallel I/O. The multiprocessor system will have available to it many devices all reading and writing simultaneously, allowing the total I/O bandwidth to be increased to whatever level is required. In addition, the video tape cartridge devices in particular will allow cheap and simple mechanical loading devices to avoid waiting for human operators to mount tapes. We are developing such "juke boxes" with optial bar code labels to insure mounting of the correct set of tapes from a large data sample. With the system architectures discussed below, these three developments will insure that the voracious appetite of the new processors for data will be satisfied.

6. SECOND GENERATION ACP SYSTEMS - SYSTEM ARCHITECTURE

The next generation ACP system will use MIPS (or other) CPUs with at least an order of magnitude more processing power than the first generation 68020's. This increase in CPU power makes it imperative to provide a system architecture that removes the potential bottlenecks preventing the current systems from being scaled up by an order of magnitude. Such bottlenecks include the I/O bandwidth, interprocessor communication bandwidth, and CPU power available in the host processor. Moreover, we would like the software for the next generation system to provide greater flexibility with less complexity.

The second generation ACP system will meet these goals through a redesign of the system software. It will allow existing applications to run virtually unchanged, yet will provide a variety of powerful new features to allow users to realize the full potential of the new processors. (The new processors can also be used in first generation ACP systems alongside existing 68020 processors when the applications do not require greater performance than provided by the original systems which are limited by the MicroVAX host.)

In the second generation system *any node* can assume the functions previously exercised only by the MicroVAX host processor. In particular, any node in the system can do `send`, `get`, `broadcast` and `accumulate` operations to or from an individual node (chosen by the system software from a class or rank of nodes) or set of nodes in a given class or rank. As before, the system software will automatically find an available node for the user, freeing the user from the details of the hardware architecture. Also, any node in the system can do I/O, reading or writing data tapes and accessing disk files.

An example of a configuration for a reconstruction problem with multiple input tapes is shown in Figure 4. Note that this is a software configuration; the actual hardware connection of the nodes is over Branchbus via VBBCs and Bus Switches (if necessary) and is transparent to the programmer. Nodes in rank 1 read events from data tapes and pass them along to either class 2 or class 3 nodes, which process events of different trigger types. Nodes in rank 3 collect events from any nodes in rank 2, either class 2 or class 3, for output to tape.

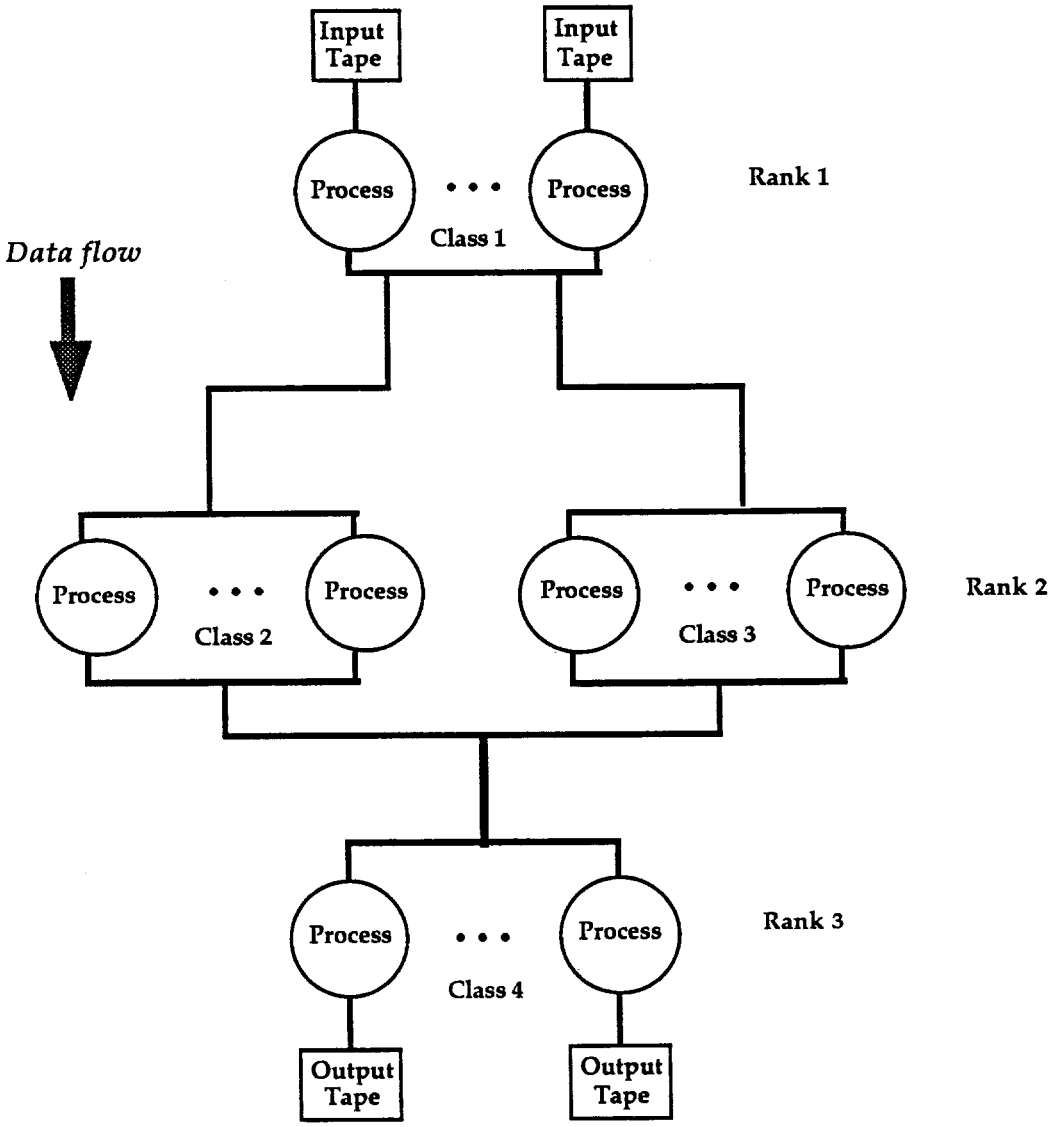

Figure 4. An example second generation ACP configuraton.

The important new features of the second generation systems include:

1) The system will no longer be configured as a single master host with many node slaves. Any node in the system can assume some of the host functions. (This eliminates host CPU power as a bottleneck, as well as naturally allowing the parallel I/O described in 2.)
2) Input and output can be carried out by many nodes simultaneously. This capability is enhanced by using low cost cartridge tape devices. (This eliminates I/O bandwidth as a bottleneck.)
3) Any node in the system can communicate directly with any other node without intervention by a host. (This eliminates internode communication bottlenecks, since the communication can now proceed in parallel without the system software overheads present in first generation systems.)
4) The nodes can be (logically) configured in ranks, with data flowing smoothly from each rank to the next, doing input at the first rank, output at the last, and processing in all intervening ranks. This provides a natural match to the structure of the HEP

computing problems as well as providing easier software development since the program in each rank will serve only a single function.

5) Any CPU running either VMS or UNIX and having a Branch Bus or Ethernet connection can be used as a node in the multiprocessor system, providing for great system flexibility as well as the ability to use existing hardware and software where appropriate.
6) Program development is done using the full set of UNIX (or VMS) compilers, linkers and debuggers for the class of nodes on which the process will run. The programs for each class of nodes can be developed independently.

Thus, the second generation system architecture provides a set of building blocks allowing a particular system to be matched to the set of applications it will run. Enough I/O devices and Branchbus interconnects should be provided so there are no bandwidth limitations. Enough standard nodes are added for the desired CPU power and any special purpose nodes (such as workstations for graphics) are also supplied. As each job is run on the system, nodes will be assigned to run particular user processes (input, output, or event processing) as appropriate. Not only the traditional compute bound event reconstruction tasks but also more I/O intensive data analysis jobs will find a home on these systems.

7. CONCLUSIONS

The use of multiprocessor systems for experimental event reconstruction is essentially a solved problem. Existing systems demonstrate that effective use can be made of systems with order 100 individual processors and up to 100 VAX 780 equivalents in total processing power. Improvements in system architectures and power of the processors in future systems will extend this to over 1000 VAX equivalents in a system.

8. REFERENCES

1] Gaines, I., Areti, H., Atac., R., Biel, J., Cook, A., Fischler, M., Hance, R., Husby, D., Nash, T., and Zmuda, T., "The ACP Multiprocessor System at Fermilab," Comp. Phys. Comm. 45 323-329 (1987); Biel, J., Areti, H., Atac., R., Cook, A., Fischler, M., Gaines, I.,Hance, R., Husby, D., Nash, T., and Zmuda, T., "Software for the ACP Multiprocessor System," Comp. Phys. Comm. 45 331-3379 (1987).

2] Kunz, P. F., Nucl. Instrum. Methods 135 435-440 (1976); Kunz, P.F., Gravina, M., Oxoby, G., Trang, Q., Fucci, A., Jacobs, D., Martin, B., and Storr, K., "The 3081/E Processor," Proc. Three Day In-Depth Rev. on the Impact of Specialized Processors in Elementary Part. Phys., Padova, 1983, 83-100 (1983).

3] Hance, R., Areti, H., Atac, R., Biel, J., Cook, A., Fischler, M., Gaines, I., Husby, D., Nash., T., and Zmuda, T., "The ACP Branchbus and Real Time Applications of the ACP Multiprocessor System" FERMILAB-Conf-87/76 (1987).