Fermi National Accelerator Laboratory

TM-1642

Future Microprocessor Farms: Offline and Online

H. Areti
Fermi National Accelerator Laboratory
P.O. Box 500
Batavia, Illinois 60510

January 12, 1990

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

Future Microprocessor Farms: Offline and Online

H. Areti
Fermi National Accelerator Laboratory
P. O. Box 500
Batavia, Il. 60510

Abstract

Microprocessor farms have been successfully employed in high energy physics for both offline analysis and online triggers. As the experiments continue to grow in size, so do the demands for processing power. The preliminary indications are that the large collider experiments will require at least a million VAX-11/780 equivalents of processing power for online trigger decisions and offline event reconstruction. This paper examines the current technology trends and projects the processing power that may be expected with the current farm architectures.

Introduction

The computing requirements of high energy physics experiments have grown steadily over the past decade. Until recently, the data taking phase of the experiments relied on a two level triggering scheme: the fast logic triggers, or level 1 triggers and high speed dedicated level 2 triggers, which implemented some type of filtering algorithm in hardware (ref. 1). Typically, the level 2 triggers reduced the rate at which data was taken by an order of magnitude. Additional suppression of trigger rate can be obtained by partial or full event reconstruction, the sophistication of which requires machines that are programmable in high level languages (Fortran, in almost all cases). These level 3 trigger processors are typically microprocessor farms, executing code that is closely related to offline event reconstruction programs (ref. 2). Each of the microprocessor nodes executes a copy of the filtering code with a computational power of one to a few VAX 11/780 equivalents. The data transfer rate into the online farms is a few tens of megabytes per second and this rate is more than adequate for offline event reconstruction. Thus, the same farm architecture meets both online and offline needs and provides a few hundred mips of processing power.

The future collider experiments will push both online and offline computing needs dramatically. The current projections for SSC experiments are for a million VAX 11/780 equivalents (ref. 3). This phenomenal demand requires a close look at the technology trends, both hardware and software in order to determine its viability. It

also raises the question whether the I/O architecture of the online and offline farms should be the same. The assumption here is that farm or farm-like objects have the best chance of meeting the computing requirements of the next generation of collider experiments.

Microprocessor Farms

A microprocessor farm is a collection of CPUs, with some way of interfacing to the outside world (fig.1).

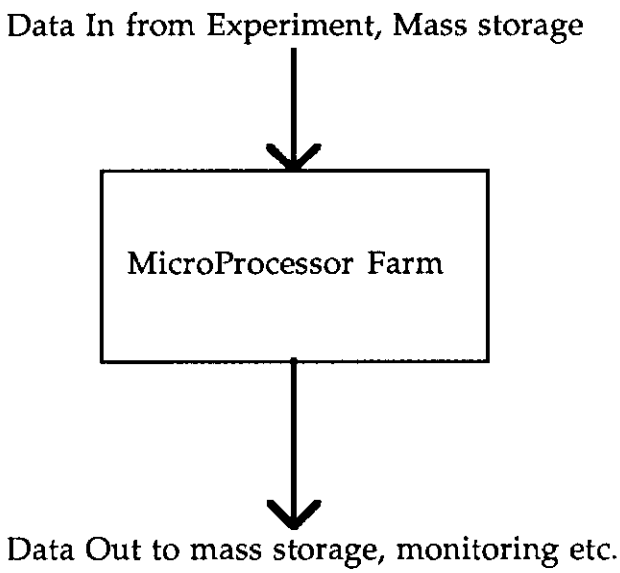

Figure 1.

The data input to the farm is from the event builder in the online environment and from storage media in the offline environment; the data output is usually to storage media. Interprocessor communication is not crucial and no great effort is expended in implementing high speed communications between processors. The arrangement of the CPUs inside the box is deferred to a later section of this paper. The heart of the farm is a node, which consists of a microprocessor, (usually a 32-bit machine), support chips (e.g. Floating Point Unit, Memory Management Unit etc.) and a memory system. The CPU is a 32-bit commercial microprocessor and is supported by high level language compilers and debuggers. The current trend in high performance microprocessors is towards Reduced Instruction Set Computers (RISC). State of the art RISC computers achieve a few tens of million instructions per second. This performance is a result of a combination of hardware and software features, whose continuing improvement is critical to the future of high speed computing. The memory system may be simple, consisting of dynamic rams only or hierarchical, consisting of high speed cache memories and a larger dynamic ram

memory. The heavy demand for performance does require hierarchical memory architectures.

Scope of the problem

It is instructive to look at the goal of a million VAX mips farm with the current technology. At 20 mips/node, the farm will require 50000 nodes. If the nodes are VME bus size cards, 60 of these nodes may be housed in a standard 19 inch rack, the farm will require more than 800 racks which translates into a minimum of 8000 square feet of floor space. (Floor space for mass storage is yet to be added to this). Further, if a cost of $2000.00/node is assumed, the system cost will be $100 million. For a billion dollar detector this represents a ten percent cost. To be manageable, a microprocessor farm needs nodes which are an order of magnitude higher in performance and should require no more than 1000 square feet of floor space. Thus, the first question to be addressed is whether a 200 mips node can be a reality in the time frame of SSC experiments.

Technology Trends

In the last few years, RISC processors have dominated the price/performance market. The primary performance goal in RISC design is to achieve one instruction execution/cycle. In order to realize this end, the RISC machines depend on pipelined instruction fetching and very high speed cache memories, implemented with static RAMs. Assuming 32-bit data paths, and an additional 32-bits for cache tagging, a 20 mips RISC machine requires a bandwidth of 160 megabytes/second. This number is an under estimate since, in reality, the current RISC machines do not achieve the one instruction execution per cycle. A more realistic number is 200 Megabytes/second. A naive linear scaling projects that a 200 mips machine requires 2 gigabytes/second data rate between the cache and the processor requiring static RAM cycle times of 4 ns. One of the techniques to ease the need for ultrafast caches, is to double or quadruple the data path size. The increase in data path size, while reducing the cache access times, increases the cache size and is likely to require that the CPU have an integrated, sophisticated cache on chip. This wide data path requirement fits in nicely with the trend toward processors that can execute multiple instructions/cycle (e.g. intel 80860, IBM America). There is every indication that the packaging technology will keep pace with the pin count of the resulting package. Of no less importance is the main memory bandwidth, since a cache miss stall of the CPU does adversely affect its performance. Even here, dynamic rams with access times under 30 ns will be available in the near future. Thus, it appears likely that RISC processors that can provide 200 mips will exist in the time frame of the SSC experiments. Figure 2 shows one of the possible board level implementations of a node.

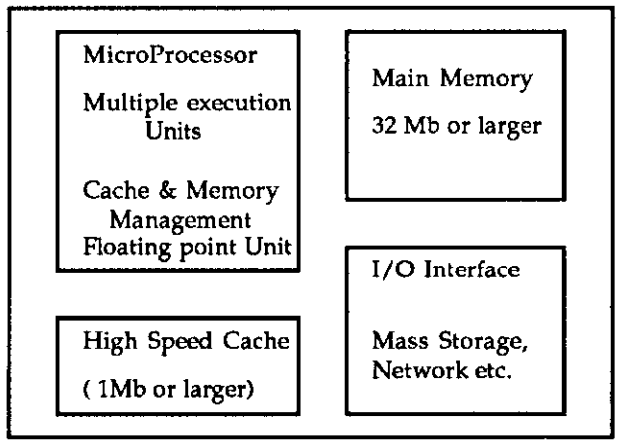

Figure 2.

The board level realization of a node using such high speed parts will require great care from the designers. It is imperative to keep the data paths short and lay out the boards to reduce noise and cross talk. The CAD systems will have to simulate electrical as well as logical design of the nodes. One of the developments that will help the designers is the advent of fast logic with output swings between 0 and 3.5 V. , which should reduce the noise level and power consumption.

In the past, it was normal for the hardware and software development of the microprocessors to be decoupled; the hardware was designed to allow operating systems, compilers etc. to be run on the processor. As a result, the compilers, especially the Fortran compilers, were very inefficient and the full potential of the hardware was seldom realized. The multiple execution units in the future microprocessors will require sophisticated parallelizing compilers that can fully exploit the low level parallelism present in the applications. Thus, the development of high performance processors needs a 'systems' approach, forcing a very close collaborative effort between hardware and software engineers.

Farm Architectures

There are many ways to embed the nodes in a farm, the simplest being the scheme shown in figure 3.

In this scheme the nodes act as data processors and the flow control resides with some intelligence outside the farm. It is assumed that each node contains enough memory to hold the incoming data, the code and output buffers. Calculating the data rate into the farm, however arbitrary the numbers may seem, is useful. The SSC event size is expected to be 1 Megabyte (ref. 3) and probably will require 100 VAX mips for online reconstruction, which implies 0.5 secs on a 200 mips processor. Thus, the data rate into each processor is 2 Mb/s, which is not excessive for one processor and may be supplied through traditional bus structures or fast serial links.

The 5000 processors that comprise the farm can handle nearly 10 Gb/s data, if this could be provided (ref. 3).

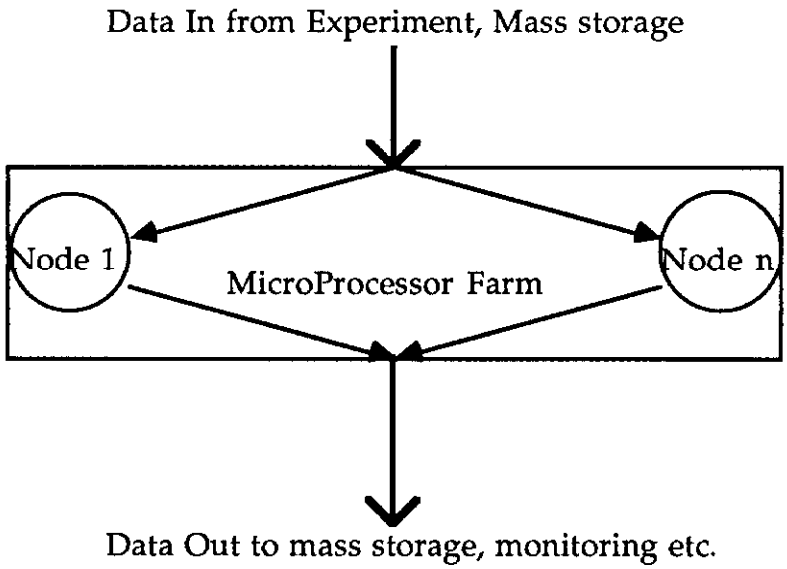

Figure 3.

It is tempting to implement multiple nodes on one board, given the degree of integration expected from the microprocessors, cache memories, main memories and I/O processors. Again, it is illustrative to deal with some concrete goals and attempt to accommodate 5 nodes on a board for a computational power of 1000 mips/board (figure 4).

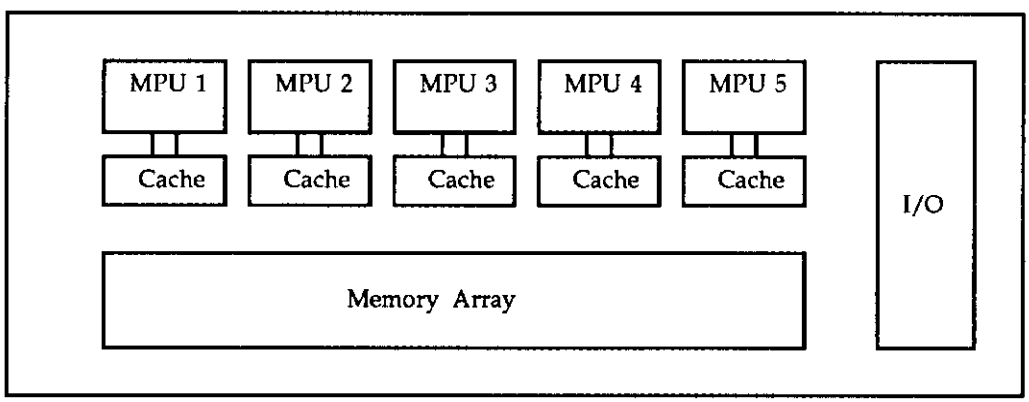

Figure 4.

In this scheme, each microprocessor has its own cache. The main memory system could be as simple or sophisticated as required, as long as it is accessible to the event builder for online usage and some mass storage controller for offline usage. The board space required is dominated by the size of cache and main memories. It is estimated that a node with 1 Mb of cache and 64 Mb of main memory will require less than 30 square inches of board space. This assumes that the static rams comprising the cache have 64k x 4 organization and the line size is 144 bits. (The cache line size is four 32 bit words and parity requires 16 bits). The dynamic rams comprising the main memory are assumed to have 4M x 4 organization and the data path size is 152 bits, including the bits needed for error correction. The memory controller and glue logic are all implemented in Application Specific Integrated Chips. Given that a fast bus size eurocard has greater than 200 square inches of usable space, five nodes may be contained on one of these cards.

The memory controller interfaces the CPU, the outside world and I/O controller to the main memory. Its exact implementation requires further thought. In its simplest form, it partitions the main memory into five separate and private memories, each assigned to one processor.

Like the memory controller, the I/O interface serves all five processors. This may contain serial ports, interfaces to mass storage (disk, tape etc.) and event builder and allow networking of the nodes.

The idea of implementing multiple nodes per board has several advantages. The first and foremost is the savings in real estate. A standard rack can accommodate 40 Fastbus size cards in two crates, including all the necessary cooling, allowing 40000 mips of processing power per rack. A million mips, then, take 250 square feet of room. In addition, there needs be only one data channel to serve 5 nodes and the five nodes share the cost of the interface. The disadvantage here is that this channel needs to be at least 5 times faster than one serving a single node and the software on the nodes should be able to share the resource. With the numbers given above, the data transfer rate requirement is approximately 10 Mb/s, which is definitely not excessive.

Implementation of the farm

Of the many ways of implementing microprocessor farms, only one possible scenario is discussed here, in which there are five nodes per board. Nodes on each board can communicate with each other through shared memory. Nodes on different boards may communicate with each other through a network such as Ultranet, which can also be used to download programs, data or test code. One or many nodes may act as server nodes and aid the other nodes in diskless boot operations. Each node runs an operating system and demands data from either the event builder or the mass storage. The processed data is written to a storage medium through the I/O interface. The medium of data transfer in all cases is a

single cable, capable of transporting greater than 100 mbits/second (high speed serial link). Figure 5 shows such a board.

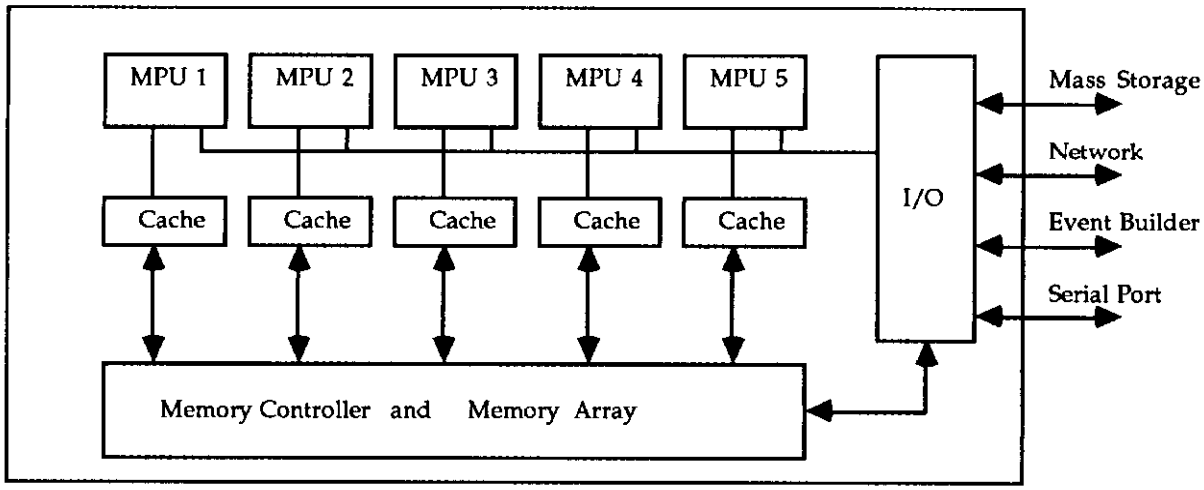

Figure 5.

A discussion of the event builder and the protocol of data transfer is beyond the scope of this paper. For the sake of completeness, a farm consisting of these boards is shown in figure 6.

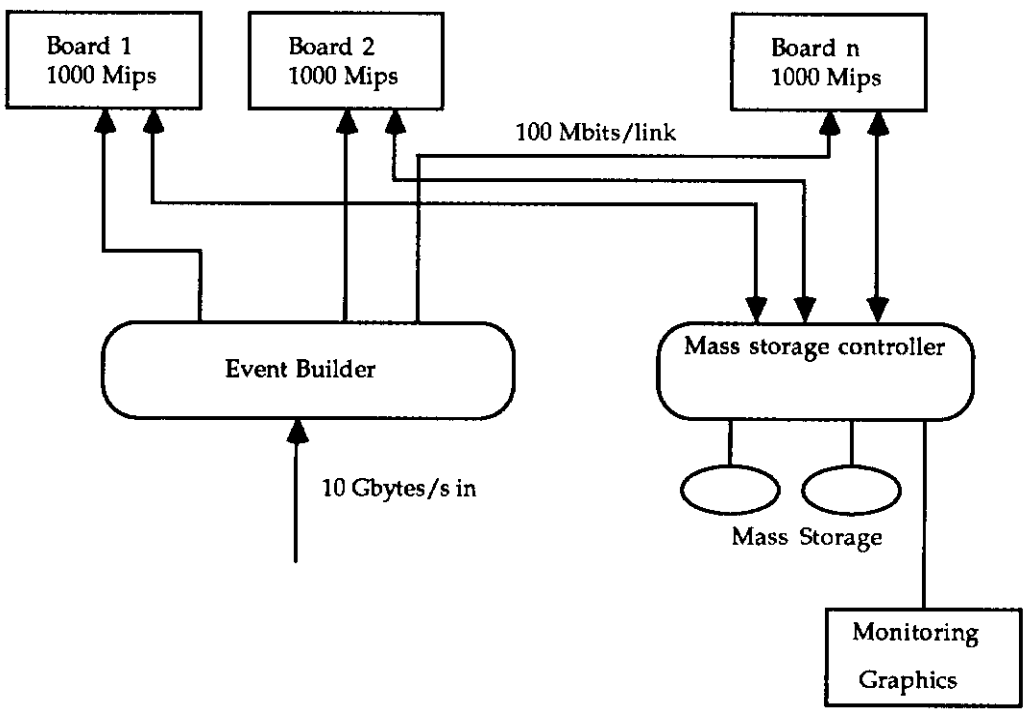

Figure 6.

The mass storage controller is similar in architecture to the event builder, connecting any node to the first available storage unit. It would be beneficial, if the data transfer protocol at the farm end of the mass storage controller is the same as that for the event builder. The mass storage controller is expected to be a very sophisticated device requiring careful study and will be discussed in a later paper. An offline farm looks exactly like an online farm with the event builder excluded.

Monitoring and graphics could be accomplished with the same boards as used for event reconstruction. In this case, the I/O interface will require graphics capability. It may also be desirable to implement a terminal server that can connect any terminal to any desired node. The terminal server could be essentially as simple as a patch panel or could be nearly as sophisticated as the mass storage controller.

Conclusion

The technology trends indicate that it is possible to obtain a million mips of computing power in the near future. The floor area required to implement such a farm, can be reduced by housing multiple nodes per board. Data transfers are expected to be through high speed serial links and no bus system (VME, Fastbus) is expected to play a part in the interconnecting schemes.

Acknowledgments

The author wishes to thank his colleagues in the Advanced Computer Program for many valuable discussions over the years and Richard Hance and Umesh Joshi for many helpful suggestions.

References

1. ECL/CAMAC Trigger Processor System. Fermilab TM821, 1982.
2. Level 3 System at CDF. Nucl. Instr. Meth A263, 1988.
3. Self-Routing Parallel Event Builder for SSC. Ed Barsotti, Private Communication.-