RISC Architecture Microprocessor Farm for offline analysis

C. de Barros, M. Mendes, M. Miranda, A. Nigri,
E. Paiva, A. Santoro, B. Schulze, C. Silva.
LAFEX Centro Brasileiro de Pesquisas Físicas
Rua Dr. Xavier Sigaud 150
Urca - Rio de Janeiro
22290 - BRAZIL

H. Areti, J. Biel, A. Cook, J. Deppe, M. Edel, M. Fischler, I. Gaines,
M. Gao, B. Haynes,D. Husby,M. Isely, T. Nash, T. Zmuda.
Fermi National Laboratory
Batavia, Il.
60510 - USA

Microprocessor farms have been successfully employed in high energy physics for both offline analysis and online triggers. The amount of data collected by the experiments are steadily growing over the past years requiring huge computing needs. Development of computer farms became necessary to perform the new computing needs.
By the other hand, the advent of RISC architecture processors with excellent price/performance ratio has been dominated the market over the last years.
To take advantage of the low cost and to fulfill the computing needs of Fermilab's E791 experiment, a RISC computer Farm has been developed at Fermilab using parallel computing technology.

Introduction

Computer "farms" have been successfully employed in High Energy Physics for both offline analysis and online triggers. The amount of data collected by the experiments are steadily growing over the past years, for instance, in Fermilab we have had: E691 generated 100 milions of events; E769 generated 400 milions of events; E791 Experiment we have had 20 billions of events, each one with 4Kbytes of data, or 80 Terabytes to be reconstructed and analysed, requiring fast and powerful computers. The size of the reconstruction codes also have been increasing fast along the years requiring computers with big memory configuration, today those codes are in the tens of megabytes range.

The reconstruction of data, require huge computing but carries an intrinsic parallelism because the interaction events are independent of each other. Digitized raw data from different events remain independent and can be reconstructed into physical parameters (individual track momenta,vertices,etc.) and can be analysed in an event per event basis.

Development of computer farms, have taken advantage of the intrinsic parallelism for the reconstruction, each event is sent to individual processors in the farm for reconstruction, the output is written to Data Summary Tapes (DSTs) as soon as the processor

is done and each farm "node" is read to reconstruct another event. This technique can be also be successfully employed for simulation, another event parallel problem.

The ACP Project

The LAFEX/FERMILAB collaboration started in 1985, during the ACPI SYSTEM development. This project intended to provide the computer power necessary for HEP at a low cost.

The ACP came into being a fortuitous time, just as 32 bit microprocessors were about to be introduced into the market. Equipped with Fortran compilers which actually could compile, these were the first chips capable of running real sizeable scientific programs. The ACP designed VME modules with these microporcessors and their floating point coprocessors(Motorola 68020/6881) and up to 8Mbytes of DRAM. For a VAX 11/780 price a 100 VAX CPU power system could be built.

Typical parallel farms of the computer "nodes" were configured with VME crates each holding 16-20 processors. A host micro VAX computer reads individual event data in the raw form written by an experiment. The host passes each event to a processing node to reconstruct into physics information.

The ACPI SYSTEM concept was revolutionary ,the research win the R-100 award from Research and Development magazine.

THE CBPF ACPI SYSTEM have also being used by others brazilian Reasearch Institutes, producing more then 10 teses, and innumerus papers.

The ACP II

The advent of high performance Reduced Instruction Set Computer (RISC) architecture, with an order of magnitude better performance than that provided by the 68020, made attractive to upgrade the processors used in the ACP I systems.

The Second Generation of the ACP system, also called ACP II, is a high performance system composed of ACP/R3000 boards which are VME modules with MIPS RISC R3000 CPU, R3010 Floating Point coprocessor, 128 Kbytes of instruction cache, 128 Kbytes of data cache, 8 to 32 Mbytes of RAM memory, 256 Kbytes o EPROM and a performance of 20 VUPs (1 VUP = 1 VAX 11/780 equivalent).

A typical ACP II system consists of a set of ACP/R3000 boards connected in one or more VME crates with disk, tape and Ethernet controllers. VME crates can be connected through a proprietary bus, the Branch Bus. The Branch Bus requires two modules in each crate: A VBBC (slave in VME and master in BB) and a BVI (master in VME and slave in BB).

Each ACP/R3000 modulee runs a modified version of MIPS risc/os operating system. It is an AT&T UNIX System V release 3 version with BSD and POSIX compatibility.

The port of riscos to the ACP/R3000 included modifications for a diskless version of UNIX. Some CPUs control the disks and are disk server for the other bords (diskless nodes) that don't have a disk. For a diskless CPU to boot, the server loads the operating system in the clients memory and tell it to start running the operating system.

The swapping device in the diskless CPU was modified to redirect the swapping operations to the server node, which has a swapping server process.

A new physical layer was added to the TCP/IP software. The CPUs are connect through VME/Branch Bus instead of Ethernet. One ACP/R3000 can be connected to Ethernet as well as VME/BB and become a router.

Besides the above described hardware and software, the ACP project comprises a set of tools for parallel programing initially designed to be used with the ACP/R3000 multiprocessor system. Today, this software known as CPS (Cooperative Processes Software), runs in a wide variety of machines. These include IBM R6000, Silicon Graphics, Sun and others.

The CPS is a package of software tools that makes it easy to split a computational task among a set of processes distributed over one or more computers. The primary tools comprise a Job Manager program and a set of subroutines. These tools, based on message passing, support a wide range of parallel programming models.

Within a single process, the user has the normal mechanisms provided by a high level language (such as FORTRAN or C) and the computer operating system to work with. The CPS routines provide the necessary enhancements to deal with a multiprocess job. Other functions necessary to the smooth running of a multiprocessor application are handled by the Job Manager. The application programs need not pay attention to these function.

CBPF is now taking the full responsibility to support the ACPII system. CBPF software engineers are porting the new version of the UNIX operational system (RISCos 4.52 based on SYTEM V REASE 3). CBPF hardware engineers designed a 32 MBYTES upgrade for the ACPR3000/25Mhz and a ACPR3000/33Mhz daughter board, and is now working in a fully ACPR3000/33Mhz redesign.

Following the ACPI tradition, CBPF ACPII SYSTEM is being used by other reasearch institutes, running parallel applications from different fields of reasearch.

Other Farms

The same functionality of ACP II farms can be achieved using commercial RISC based computers connected through Ethernet. This approach has the advantage of using commercialy supported machines and operating systems. The disavantage is having a smaller communication bandwidth and a lower cost/performance ratio. This is speacilly true in the brazilian market.

In Fermilab, some called "production farms" are currently used to accomplish the reconstruction computing needs. An IBM RS6000 farm with 1700 VUP, a Silicon Graphics farm with 300 VUPs and an ACP farm with 1000 VUPs are available, all running the ACP multiprocessor software developed in Fermilab called CPS (Cooperative Processes Software).

The IBM farm configuration is:

I/O nodes: 2 IBM RS6000/530 and 2 RS6000/320 with 11 8mm tape drives, 23 GBytes of disk;

Worker nodes: 64 IBM6000/320.

The Silicon Graphics farm consists of:
I/O nodes: 3 SGI 4D/25 with 9 tape drives and 4 GBytes of disk;
Worker nodes: 22 SGI 4D/25.

The ACP Farm consists of:
I/O nodes: 3 ACPR3000/25 with 6 tape drives and 2.1GB of disk;
Worker nodes: 48 ACPR3000/25

Two more farms are being installed at this time:
An IBM RS6000 with:
I/O nodes: 2 IBM RS6000/530
worker nodes: 36 IBM RS6000/320H

and a Silicon Graphics farm with:
I/O nodes: 3 SGI 310s
worker nodes: 78 SGI 4D/35s

There are currently six experiments doing raw data reconstruction or Monte Carlo on these Unix based farms.

Conclusion

The use of Computer Farms, exploiting the powerfull RISC processors and the intrinsic parallelism of High Energy Physics problems has proven to be a solution for HEP community. LAFEX has been involved in the development of such systems for already 8 years and the results of this work is that we have been capable of doing high quality physics in Brazil which would be untinkable without a system like the ACP.

We can expect that RISC processors power will still continue to increase in the next future. A problem that we have to be concerned is with obtaining a higher communication bandwith in order to keep up with the faster processors.