Next Generation Farms at Fermilab

Ron Cudzewicz, Lisa Giacchetti, Mark Leininger, Tanya Levshina
Raymond Pasetes, Marilyn Schweitzer, Stephen Wolbers

Fermi National Accelerator Laboratory, Batavia, IL

The current generation of UNIX farms at Fermilab are rapidly approaching the end of their useful life. The workstations were purchased during the years 1991-1992 and represented the most cost-effective computing available at that time. Acquisition of new workstations is being made to upgrade the UNIX farms for the purpose of providing large amounts of computing for reconstruction of data being collected at the 1996-1997 fixed-target run, as well as to provide simulation computing for CMS, the Auger project, accelerator calculations and other projects that require massive amounts of CPU.

Key words: Computing Farms, Large Systems, Parallel Processing

1 Introduction

UNIX farms at Fermilab have had a long history of providing massive amounts of CPU power required for the event reconstruction of large datasets taken by experiments at Fermilab[1]. This paper will describe the upgrades to the farm that are being implemented to provide sufficient computing for event processing of current experiments and for simulation needs of magnet, accelerator, and experiment design.

2 History and Needs for Expansion

The 300+ workstation farm at Fermilab has been used for many years to provide computing for the 1990-91 fixed-target experiments and for the CDF/D0 collider run (RUN 1) that ended in early 1996. The farms worked very well and were able to provide sufficient computing for CDF and D0 to reconstruct their data in 'quasi-real time', i.e., at the same rate as the data was being collected but with a few days to a couple of weeks delay needed for final calibration constants.

The CPU capacity of the farm is approximately 8000 'MIPS', where a 'MIP' is defined by the performance of a small simulation and reconstruction code (named TINY). At least three factors contributed to the decision to expand the farms. First, the needs of the 1996-1997 fixed-target run were estimated to be at least 15,000 MIP-years. This estimate was based on assumptions about code performance and data volumes which were the best known at the time (early 1996). Our experience has been that these numbers are underestimates – both the CPU time per event and the data volumes are typically larger than estimated. When one factors that in as well as the efficiency of using the farms it is quickly seen that 8000 MIPS is not sufficient for a timely reconstruction of data from the 1996-1997 run.

Second, there are other large CPU needs at Fermilab which the farms can satisfy. Simulations for the CMS experiment, the Auger project, the Recycler Ring being designed for the next collider run, and superconducting magnets all require large amounts of CPU power. Some theory calculations are also best done on the farms because they require large CPU resources. These programs require large integrated CPU but also larger real memory and processor speeds compared to the 300+ nodes. Finally, the old nodes are becoming more difficult to maintain and reconfigure. The old SGI nodes, being R3000 based, cannot run IRIX 6.x and the IBM nodes would require additional disk space to allow them to run AIX 4.x. The old farms are divided by routers into Ethernet segments and are logically divided into many NIS domains with multiple NFS servers. Upgrading the OS or memory or modifying the node allocation is difficult on the old farms. We are addressing all of these issues with the design of the new farm.

3 Farm Expansion

3.1 Capacity and bids

A project to expand the farms by 15,000 or more MIPS was undertaken in the spring of 1996. To avoid a major porting and testing program it was decided that only the 4 already supported UNIX operating systems would be allowed to be used on the new farms. They are Digital UNIX, IBM AIX, SUN Solaris and SGI IRIX. A bid package was prepared for the four vendors. The expansion was spread over two fiscal years with 7,500 MIPS to be purchased in each year. We purchased 5,000 MIPS from the winner and 2,500 MIPS from the second place bidder in 1996. In 1997, are purchasing an additional

5,000 MIPS from the winner and 2,500 MIPS from the second place bidder. There was no rebid between the first and second purchases. However, each vendor had the opportunity to modify the computing offered based on the availability of newer technology, as long as the cost per MIP was less. There was also a purchase of the necessary peripherals and network equipment to build a complete farm.

3.2 Farms Expansion Details–I

The details of the first farm expansion are found in Table 1.

Table 1
Details of first half of farm expansion

CPU	Number of CPU's	MIPS/CPU	Total MIPS
SGI R5000 Challenge S	45	114	5130
IBM RS6000/43P (133)	22	115	2530
SGI Challenge DM	2	105	210
IBM J40	2	97	194

The SGI Challenge S and IBM 43P computers comprise the worker nodes and the SGI Challenge DM and the IBM J40 the I/O nodes in our language. Each worker node has 64 MB of real memory and 2 GB of local disk. The I/O nodes each have 128 MB of real memory. In addition to the CPU, peripherals and a switch fabric was purchased to complete the farm. A total of 150 GB of disk and 15 8mm tapedrives were purchased and were connected to the two I/O nodes. The entire farm is connected to a Catalyst 5000 switch. The switch has been configured with 72 Ethernet ports, 2 fast-Ethernet ports and 1 FDDI port. The performance of the switch is more than sufficient for the types of computing that will be typically performed on the farm. This can be shown as follows. If all of the jobs which run on the farms are event reconstruction then we expect that the CPU/IO ratio is 1000-5000 instructions/byte, based on previous experience. Each worker node therefore requires at least $115 \times 10^6/1000 = 115$ KB/s. This is easily achievable as each worker node has a full Ethernet of bandwidth. The aggregate bandwidth out of the two I/O nodes must be $7,500 \times 10^6/1000 = 7.5$ MB/s. FDDI should be able to provide this bandwidth. Because FDDI is shared among the I/O nodes we are moving the I/O nodes to fast Ethernet, which will expand the total I/O capacity.

A drawing of the new farm is shown in Figure 1. There is much more flexibility built into the farm than was possible in the old farm. All worker and I/O nodes communicate with each other through the switch. This allows rather arbitrary configurations for individual users and jobs. The entire farm is a single NIS

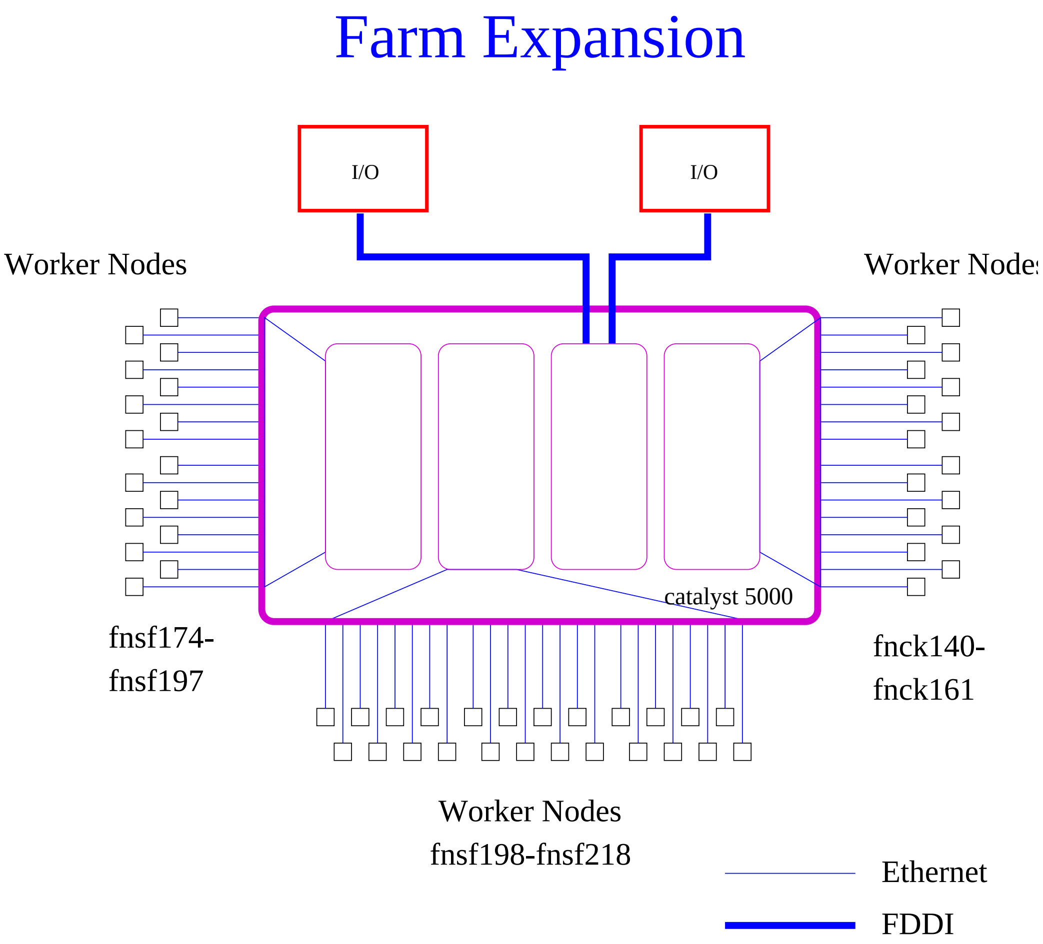

Fig. 1. Configuration of the first half of the Fermilab UNIX Farm Expansion

domain and file systems are NFS mounted across the farm (AFS is also a possibility). This configuration should allow a much better utilization of the resources of the farm than was possible before.

3.3 Farms Expansion Details – II

The second half of the farm expansion is just being completed. The hardware purchased is detailed in Table 2. At this time we are unsure whether the IBM 43P's will be 133 MHz or 200 MHz models.

SGI offered the new Origin 200 series for this half of the expansion. We were interested in the new architecture as well as the faster R10000 processors and decided to purchase 6 systems with 4 processors each as the worker nodes and one system with 2 processors as the I/O node. The peripherals purchased are the same type and numbers as the first expansion. The Catalyst 5000 will be replaced by a Catalyst 5500 which has a much larger capacity for ports than the Catalyst 5000. The original ports have been retained and Ethernet ports

Table 2
Details of second half of farm expansion

CPU	Number of CPU's	MIPS/CPU	Total MIPS
SGI Origin 200	24	208	4992
IBM RS6000/43P (133)	22	115	2530
SGI Origin 200	2	208	416
IBM J40	2	97	194

have been added for the new IBM worker nodes and fast Ethernet ports have been added for the Origin 200 systems.

4 Software and First Experiences

The software that runs on the farms includes CPS(Cooperative Processes Software)[2], CPS BATCH[3] and OCS(Operator Console Software)[4]. CPS is a toolkit that is used to modify a program and allow it to run across many computers in parallel. The parallelization is normally at the level of an event or collection of events. CPS BATCH is a simple queuing system for jobs. OCS is a program that provides tapedrive and tape-mount access at the Fermilab Computing Center. CPS has been upgraded to a new version (3.0) which supports 64 bit operating systems. The functionality of the program is left unchanged. CPS BATCH was rewritten to use the same underlying database system as that used by OCS. OCS has been ported to the new systems.

The first half of the expansion farm was available for users in early December of 1996. The first users include E781, a fixed target experiment, and a large theory calculation of W/Z p_t at the collider. E831 has been added recently and their use of the farm is increasing. No significant problems were encountered although some problems associated with the modifications to CPS and CPS BATCH have been encountered and fixed. The Auger collaboration has started to use the farm to simulate very high energy cosmic ray showers. Magnet calculations are also being considered for the farm.

5 Future and Summary

During the coming year the second half of the expansion will be integrated into the farms. We expect the fixed-target experiments to finalize their reconstruction code and to start to use the farms heavily. Other other large users of CPU will start to use the farm to solve their problems. The challenge for this coming year is to provide a stable system that is used efficiently. It is felt that the faster computers, larger memories, larger disk space, uniform file system, and the switch fabric will all make this easier than in the past.

The farm expansion will increase the total CPU power in the farms to over 20,000 MIPS. This, along with the other facilities at Fermilab, should be sufficient for the near future computing needs. We are starting to think about the computing that will be needed for CDF and D0 during the next collider run, scheduled to begin in 1999. The amount of computing required will be much larger than 20,000 MIPS, meaning that we will be looking at a major increase in computing power near the end of the century. Possible candidates include large SMP's, more UNIX farms, and PC farms.

References

[1] Frank Rinaldo, Stephen Wolbers, Comput.Phys.7:184-190,1993

[2] Matt Fausey, CPS and the Fermilab Farms, Fermilab-Conf-92/163, 1992.

[3] M. Fausey, et al., CPS and CPS Batch Reference Guide, Version 3.0, April 26, 1996.

[4] M. Fausey, et al., OCS Reference Guide, Version 2.4, July 24, 1995.