Fermi National Accelerator Laboratory

FERMILAB-Conf-94/111

Production Farms at Fermilab

Mark Fischler, Frank Rinaldo and Stephen Wolbers

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

May 1994

Presented at *the Computing in High Energy Physics* 94 *Conference,* San Francisco, California, April 21-27, 1994

Disclaimer

PRODUCTION FARMS AT FERMILAB *

Mark Fischler, Frank Rinaldo, Stephen Wolbers
Fermi National Accelerator Laboratory
P.O. Box 500
Batavia IL 60510 USA

Abstract

UNIX Farms at Fermilab have been used for more than three years to solve the problem of providing massive amounts of CPU processing power for event reconstruction. System configurations, parallel processing software, administration and allocation issues, production issues and other experiences and plans are discussed.

INTRODUCTION

The UNIX Farms Systems at Fermilab are a large and growing example of loosely-coupled parallel computing. There currently are over 300 UNIX workstations (SGI and IBM) in the Fermilab Farms in full production, with a rated computing power of over 10000 MIPS. In addition to the CPU's there are quite a few peripherals (disk and tape) attached to the systems to allow access to the large datasets that are to be processed. The Farms are utilized for reconstruction (pattern-finding and fitting) of the raw data coming from high-energy physics experiments at Fermilab and for CPU-intensive Monte Carlo simulations. Many experiments have used the Farms to successfully do event reconstruction for very large datasets.

*This work is supported by the U.S. Department of Energy under Contract No. DE-AC02-76CH03000.

The software for parallel computing (cps) and for batch scheduling (**cps-batch)** is maturing and improving in capability and robustness. Additional tools for debugging and performance as well as for computer operations (tape mounts) are being developed and delivered in order to make maximum use of these computing cycles. This paper discusses experiences in making large systems work, including efforts put into improving those tools.

The systems are currently being used to complete the reconstruction of the multi-Terabyte datasets from the last fixed-target runs at Fermilab and for the essentially real-time reconstruction of data from the 1994 Collider run.

FARM HARDWARE

The farms consist of UNIX workstations and servers configured logically in two different sets. The vast majority of the workstations (over 300) are "worker

nodes": Each is a UNIX workstation with 16-24 MB of memory, a local system disk, and ethernet and power connections. These workstations are a mix of SGI 4D/25, 4D/35, and R3000 Indigo's, as well as IBM RS6000/320, 320H and 220's. The combined CPU power is approximately 10,000 MIPS. The other set of workstations and servers are designated as I/O nodes; these provide connectivity to the datasets (via SCSI-connected 8mm tape) as well as additional disk space and memory to handle other job functions. There are 17 I/O nodes: SGI 4D/420's and Challenge XL, and IBM RS6000/580, 590, 530, 530H, 320 and 320H workstations.

Configuration

The balance of I/O and worker nodes must be tuned to obtain maximum utilization of the available compute power. The average ratio of worker nodes to an I/O node is approximately 16-the ratio used for a given application is determined by experience and by a knowledge of its CPU/IO needs. To avoid saturating ethernet segments, the worker nodes are divided into subnets attached to routers, with each subnet consisting of between 8 and 20 workstations. The users of the farms are assigned production systems which contain an I/O node, tape drives on the I/O node, and multiple worker nodes. Each production system is typically assigned to only one experiment and each experiment is assigned multiple production systems depending on priority and need.

FARM SOFTWARE

The success of the UNIX Farms depends to a large extent on the software provided to utilize the available CPU power.

cps

cps (Cooperative Processes Software)[1] was written to allow an arbitrary program to share data and computing across many processors. The package allows many different types of parallel computing but typically in event reconstruction the structure used is to have an input and output process and many reconstruction processes. Events are read from tape or disk, passed to a process on a worker node, reconstructed there, passed to an output process and written onto disk or tape. The changes to a typical reconstruction package to allow it to run with cps are fairly modest and do not require fundamental restructuring of the program. Performance tuning of the components of **cps** has been done as differing needs were identified and bottlenecks were found and removed. **cps** relies on a single job manager process, and thus is not indefinitely scalable. But in our experience the relevant size limit is imposed by I/O requirements, not by cps-induced bottlenecks.

cps-batch

Due to the nature of farm computing at Fermilab (hundreds or thousands of tapes to be processed) a batch system is necessary to allow queuing of jobs. There is no ubiquitous batch management tool in the UNIX environment. **cps-batch** is a simple queueing mechanism which allows users to submit multiple jobs to be executed one at a time on each production system. Each experiment is assigned mul-

tiple production systems to use as part of their overall farms allocation. Enforcement of resource control is primitive: A given system is assigned to only one experiment .

OCS

As farm usage began to grow, it became clear that tape-mounting by operators was an extremely important part of the overall processing problem. Reliable and fast tape mounts are needed to handle the hundreds of tapes that are processed each day on the many I/O nodes of the systems. OCS (Operator Communications Software)[2] allows robust tapedrive allocation and tape-mounting in a distributed UNIX environment such as the farms. It handles the bookkeeping needed to tell operators what tapes to mount and where, and to cope with operator replies indicating anomalous situations. This system is invaluable in providing a successful production environment.

Utilities

Many utilities have been written to manage, debug, operate, and tune the farms. The **jobview** log file tool allows post-mortem analysis of the flow of data and control, to help tune the system. A distributed debugger, **jmdb,** lets the user debug the parallel aspects of his code. Activity on entire farms can be monitored coherently using the graphical **cps_xpsmon** and **xcpssysmon** tools, and **xfalive** watches for hardware failures by probing the worker nodes in various ways.

CONCLUSIONS

The UNIX Farms at Fermilab have been extremely effective in solving the needs of event reconstruction of high-energy physics data. The Farms have allowed both collider detectors to reconstruct their data in real-time and to handle special reconstruction needs. The success of the farms has depended on large efforts from many people to produce a production system which handles many hundreds of tapes each and every day. Many issues which were difficult enough to handle in the context of dedicated systems would present insurmountable problems in the context of scavenging time on diverse under-utilized workstations.

The current farms are being modified to handle the increased data-taking of the current collider run, and to expand the range of HEP data processing activities to which they can be applied. Further improvements and enhancements will be made as technology allows and as demands grow.

ACKNOWLEDGEMENTS

The development and maintenance of the hardware integration and software products make the farms a valuable resource. Credit for the success of this effort is due the Fermilab Farms Group and UNIX Systems Support Group.

REFERENCES

1. M. Fausey, et al., "CPS User's Guide, CPS Version 2.9", Fermilab Computing Division Library GA009, June 24, 1993.
2. M. Fausey, M. Schweitzer, "Operator Communications Software Reference Guide V2.2", Fermilab Computing Division Library GA0012, April 1, 1994.