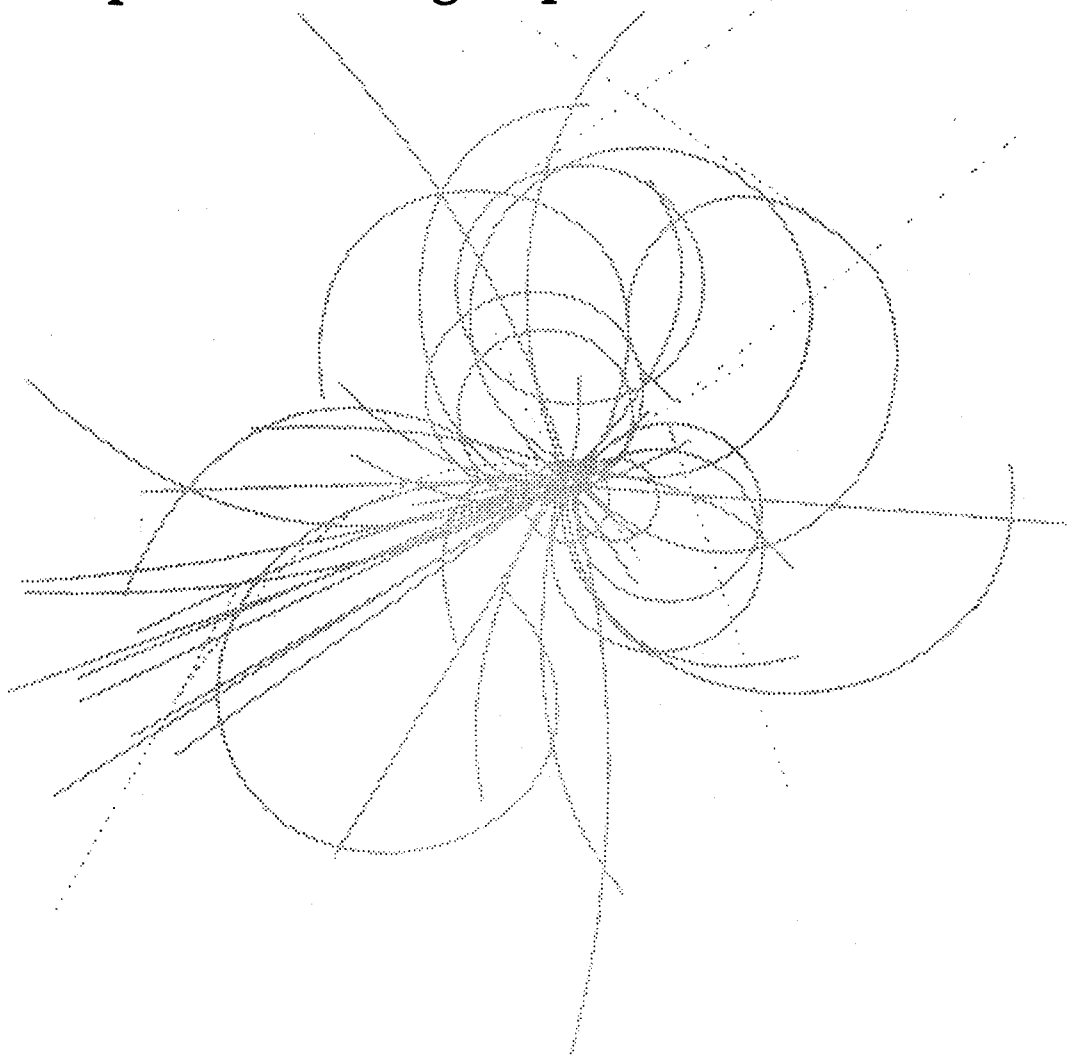
SSCL-400

Superconducting Super Collider Laboratory



High Energy Physics Computing at the SSCL

L. Cormell

March 1991

SSCL-400

High Energy Physics Computing at the SSCL*

L. R. Cormell

Physics Research Division
Superconducting Super Collider Laboratory†
2550 Beckleymeade Ave.
Dallas, TX 75237

March 1991

---

*Presented at the 9th International Conference on Computing in High Energy Physics (CHEP91), Tsukuba, Japan, March 11-15, 1991.

†Operated by the Universities Research Association, Inc., for the U.S. Department of Energy under Contract No. DE-AC02-89ER40486.

HIGH ENERGY PHYSICS COMPUTING AT THE SSCL[†]

Laird R. Cormell

Physics Research Division
Superconducting Super Collider Laboratory [§]
Dallas, TX 75237, USA

1. INTRODUCTION

It has been less than two years since the SSCL began operations in its present location in Dallas, Texas. During this time much effort has gone into beginning initial programs in all areas including magnets, accelerator, civil construction and physics research. We are now at a point where we can begin to see some concrete developments growing out of the months of planning and design. In the area of computing,for example, several rather serious computing facilities have been brought on-line in the last few months.

1.1 Computing Strategy

The SSCL has adopted a computing strategy that is intended to provide the greatest amount of low cost computing power for as many users as possible. In the 90's, to achieve the best performance at the best price requires a distributed computing environment based on networks of RISC/UNIX workstations. These workstations are connected by LAN for communication to the rest of the group, to the rest of the world, and to common file storage and compute engines. This distribution of computing resources places interactive dedicated computing on the user's desk, and at the same time provides for shared mass storage and batch computing power when needed.

By requiring open systems and conformance to industry standards the SSCL has been successful in acquiring and integrating heterogeneous networks of commercially available computers. Thus far, this strategy has worked well for several reasons.

a. First, the SSCL has been able to ride the crest of the RISC/UNIX wave by being able to select the best price/performance offers available. We are not "locked-into" a single vendor proprietary solution.

[†] Presented at the Conference on Computing in High Energy Physics (CHEP 91), Tsukuba, Japan, March, 1991.

[§] Operated by the Universities Research Association, Inc. for the U.S. Department of Energy under Contract No. DE-AC02-89ER40486.

b. Second, multi-vendor computing environments enhance competition. The SSCL, as well as other government and private institutions, has benefited from the dramatic reductions in the costs of RISC power brought on, in part, by the open competition in the market. No single vendor has held the lead in RISC price/performance for more than a few months at any time during the last few years. In order to reap the benefits of this pricing revolution, where prices are now approaching $200/MIP, the Lab is prepared to integrate multi-vendor solutions by requiring industry standard interfaces, communication, format, protocol, and the commonality of UNIX.

c. Third, the advantages gained from implementing commercially available solutions generally outweigh the advantages of "home-grown" solutions. While some advantage may be gained by developing a solution tailored to one's particular applications, such an advantage may be short-lived. More often than not, the general purpose commercially developed solution will eventually surpass the locally developed point solution. The end user may then find himself locked into the "home-grown" technology and not able to take advantage of new developements in the commercial market. Also the development of local solutions requires a considerable staff for the development, maintenance, and support efforts. This is typically beyond the means of most institutions, including the SSCL in this stage of its development.

1.2 Computing Environment

The SSCL's distributed computing environment consists of a few general purpose machines networked to several work group specific LANs. These work group LANs have been organized to optimize problem solving or special purpose applications for a particular group. For example, one Magnet Divison network consists of several HP workstations with a common file server and disks using Unigraphics to design various magnet systems.

Table 1. Computing Systems at the SSCL

SYSTEM	QTY.	MIPS	
General Purpose			
VAX 6420 (Scientific)	1	14	
VAX 6420 (MIS)	1	14	
DEC 5840 (General UNIX Server)	1	75	
		103	
Scientific, Technical, Engineering Nodes			
Workstations	186	2720	
File Servers	21	425	
		3145	
Compute Servers			
Hypercube	64 (CPUs)	1200	
PDSF Front End	30	900	
PDSF Batch Ranch	24	480	
		2580	
TOTAL		**5828**	**MIPS**
PCs and MACs		1120	

A summary of the current computers in operation at the Laboratory is given in Table 1. About one-half of the computing power at the Lab is found in desktop workstations that are used for a variety of tasks including CAD, CAE, accelerator design, and HEP code development. A large fraction of the Lab's computing power is in the 64 node Intel Hypercube. The Hypercube is used primarily by the Accelerator Division to perform orbital stability calculations. About a quarter of the Lab's computers are owned by the Physics Research Division (PRD). Most of these computers have been integrated into a single facility, the Physics and Detector Simulation Facility (PDSF).

In the Physics Research Division, one finds several application specific working group networks installed, such as the trigger and data acqusition group, the experimental facilities group, the systems and software development groups, and the PDSF. These groups share common utilities on their respective network segments, but can communicate with the remainder of the Lab over the Ethernet backbone, and to the rest of the world via T1 links to ESNET, HEPNET, etc. through SSC gate.

2.0 PDSF REQUIREMENTS AND CONCEPTS

The enormity of the computing challenge of the SSC is well known. These challenges require special solutions in almost every aspect of computing. In particular, PRD Computing must provide computing resources for on-line, off-line, and modeling and simulation needs for HEP. At this time, our immediate focus is on providing the simulation resources necessary to design the SSC detectors.

2.1 Detector Simulation Requirements

It is expected that approximately 100 full time equivalent scientists will become involved in the physics/detector simulation effort for the SSC during the next year. The work load may be distributed among 200-300 different physicists throughout the world. Initially the SSCL simulation facility will be used primarily by outside users who will log in remotely. Good network access is essential for this to be possible.

Table 2. Physics and detector simulation facility: acquisition plan.

	10/90	3/91	3/92
Computing Power	500 MIPS	1000 MIPS	4000 MIPS
On-Line Storage	50 GB	100 GB	400 GB
Tertiary Storage	.75 TB	1.5 TB	6.0 TB

(In this table a 1 MIPS processor is taken to be the approximate equivalent of a VAX-11/780 with a floating point accelerator.)

Two computer planning groups have recently reviewed the physics and detector simulation needs for the SSC.[1,2] The Price committee estimated that several thousand VAX 780 equivalents will be

[1] Report of the Task Force on Computing for the Superconducting Super Collider, SSC-N-579, M. Gilchriese, editior, Dec., 1988.

[2] Report of the SSC Computer Planning Committee, SSC-N-691, L. Price, editor, Jan., 1990.

needed for detector simulation through the proposal and early design stage of experiments for the SSC. They recommended that the Laboratory acquire computing engines in a phased acquisition plan to meet these needs. Their recommendation for the acquisition of computing power, disk storage, and tertiary storage for the next two years is summarized in Table 2.

The most cost effective way of buying CPU power at this time is in RISC-based UNIX workstations or compute servers. These now provide 20-40 MIPS per processor at a cost of a few hundred dollars per MIP. Massively-parallel machines do not seem indicated at the present time, but coupled processors are well matched to event simulation, since a) the CPU requirements are large while the memory and input/output requirements are modest, and b) only coarse-grained parallelism at the event level is envisioned at this time.

2.2 Design Philosophy

A key element in our design of a distributed computing environment for the PDSF has been the separation and distribution of the major functions. The facility has been designed to separate batch processing from interactive processing and to separate the file and tape storage functions as well. By distributing these functions it is often possible to provide higher throughput and resource availability. Similarly, our design is intended to the exploit coarse-grained (event level) parallelism in a distributed environment.[3]

2.3 Functional Model

The facility operational requirements were broken into three major subsystems by function: 1) a networked front-end for interactive usage, 2) a file server, and 3) a "ranch" of parallel batch processing compute servers. Each of the distributed subsystems is networked by a high-speed (FDDI) network.

It is intended that interactive and batch processing not be co-mingled. A separate batch ranch of compute servers is accessible through Network Queueing System (NQS) software. Cooperative Process Software (CPS) and other tools are available to assist in the porting of code developed on the front-end system to the batch ranch.

The eventual goal is that the front end/file server systems be able to access both disk and tape resources containing batch job output independently of the batch processors. Access to disk by dual/multi-ported drives and tape by multi-headed robot-based systems will accomplish this, and is therefore, reflected in the system design. However, for the first phase of development, 8-mm tape carousels will be used by the batch processors for tape storage. This is intended only as a temporary solution to the tertiary storage problem.

2.3.1 Front End Network. The front-end network provides the user the interactive computing that he needs to gain access to the facility, to retrieve and edit files, to compile and run small to medium-sized jobs, and to submit batch jobs. Each workstation, ideally one per user, provides a host unit for the users logged into the facility.

2.3.2 File Server and Tape Storage. The file server provides centralized data storage for the facility including users' source, object, input, and output files. Data produced on the front-end

---

[3] Physics and Detector Simulation Requirements, L. Cormell, et al., in AIP Conf. Proc. 209, Computing for High Luminosity and High Intensity Facilities, Santa Fe, NM, 1990.

or the batch ranch is staged on the batch ranch , or possibly the file server, disks and is then logged to 8-mm tape on the carousel.

2.3.3 Batch Processing Ranch. The major portion of the computing horsepower of the facility resides in the batch ranch. This network of symmetric multi-processors performs the large simulation jobs in parallel. The output is staged to SCSI disk and then transferred to the 8-mm tape carousels.

3. PDSF PROCUREMENT

The primary evaluation features for procuring the PDSF hardware were cost and technical factors based on several HEP codes. In addition, the SPECmark rating was used to set a minimum level of acceptable performance. The HEP specific codes included Isajet, Pythia, Jetset, a Fourier transform example, and the Geant example GEXAM1. The procurement of the equipment was openly competed and purchase orders were written separately for the batch ranch, file server, and front-end network.[4] The total purchase price including all CPUs, RAM, disks, networking, routers, tape drives, software, and maintenance was approximately $1.7 M for the 1000 VAX equivalent facility.

3.1 Benchmarks

There are many industry standard benchmarks currently available. Each of these tests some different aspect of the computer's performance. As always, the best benchmarks are those based on one's own applications. As mentioned above we used the SPECmark and our own benchmarks to evaluate CPU performance. Like the SPECmark we defined an SSC unit of processing power (SSCUP) to be the performance of a given CPU relative to a VAX-11/780. The SSCUP rating is given by the geometric mean of the performance for the Isajet, Pythia, and Jetset examples. The three jobs were run with and without some degree of optimization. A summary of our findings for several systems is given in Table 3.[5] The CERN units are based on performance for several optimized HEP codes relative to a VAX 8600. The following relationships are approximately correct: 1 SSCUP ~ 1 SPECmark ~ 0.25 CERN unit ~ 1.4 MIPS.

3.2 Hardware Configuration

The equipment purchased and installed for Phase I of the PDSF is shown schematically in Figure 1. Essentially all of the hardware was delivered and operational by the scheduled date of March 15, 1991.

3.2.1 Front-end Network

The front-end network consists of 30 Sun Sparcstation 2's labeled ws0-ws29 in the figure. These computers have 16 MB of RAM, 669 MB of disk, and are networked by individual Ethernet segments through Cisco AGS+ routers to the FDDI backbone. They do not have monitors or frame buffers, however. Access to the system is obtained by logging in to the PDSF through SSC gate on the WAN. A program running on the console concentrator cc0 assigns a Sparcstation to each user as he logs in and monitors the session via the RS-232 console output.

---

[4] Physics and Detector Simulation Facility Specifications, Computer Acquisition Working Group, SSCL-275, Jul., 1990.

[5] Benchmarks for the SSCL Physics Simulation Computing Facility, K. McFarlane, SSCL-375, Mar., 1991; Benchmark Results Summary, K. McFarlane, SSCL, in preparation, Apr., 1991.

Table 3. Benchmark Ratings

SYSTEM	SSCUPs	SSCUPs (optimized)	SPECmarks	MIPS	CERN[6]
VAX-11/780			1.0	1.0	
VAX 6410	7.0	8.4		7.5	1.9
SGI 4D/35S	15.3		23.0	33.0	
SGI 4D/310	13.5	21.6	18.5	30.0	5.3
IBM RS/520	11.9		22.0	29.5	
IBM RS/530	14.3	21.3	28.6	33.0	4.9
SUN Sparc 2	11.8	16.3	21.0	28.5	3.5
DEC Stat. 3100	6.4	10.0	10.8	14.0	2.8
DEC Stat. 5000	10.5		18.5	24.0	4.4
Apollo DN 10000	11.6	18.6	17.4	22.0	4.9
HP Apollo 9000/720*	22.7	29.4	55.5	57.0	

3.2.2 File Server

The file server, FS0 in Figure 1, is a Silicon Graphics 4D/320 with 32 MB of RAM and 30 GB of disk storage.

3.2.3 Batch Ranch

The batch ranch is comprised of 3 SGI 4D/380 symmetric multiprocessors (8 CPUs each). Each CPU has 16 MB of RAM for a total of 128 MB of shared memory per box. The batch ranch has a total of about 40 GB of SCSI disk for temporary storage of staged job output. Typically, the output of a large job which may be several hundred MB will be copied to tape after the job has completed.

3.2.4 Tape Robot

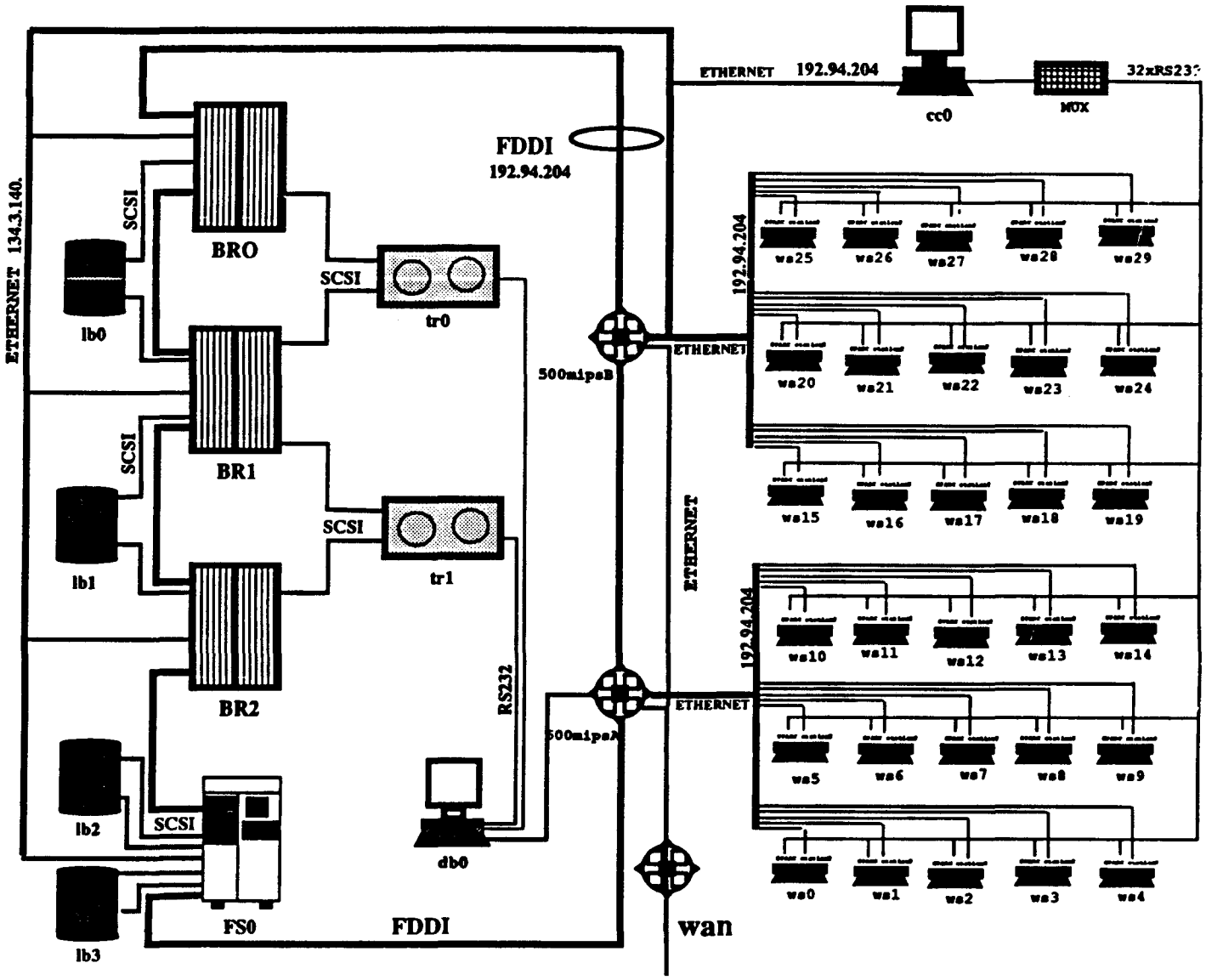
The tape robot consists of two 8-mm tape carousel storage devices from Summus, Inc. Each robot has two 500 kB/sec. drives, bar code readers, and a carousel that holds fifty-four 2.4 GB tapes for a total storage of 250 GB in the two robots. A data management system (DMS) layered on Sybase manages the data staging and storage system. Actions on the data set such as file transfer, retrieval or deletion are originated by the user, but the location (tape volume) of the files is maintained by DMS and is transparent to the user.

---

[6] E. Mcintosh, CERN, CN, private communication, Feb., 1991.

* The HP/Apollo 9000/720, just recently announced, was not evaluated as part of the PDSF procurement, but is included here for comparison.

Figure 1. PDSF Schematic Diagram



4. PDSF IMPLEMENTATION

The PDSF as shown in Figure 1 was in the final testing stage during the week of 11-15 March and is now fully operational. Those users who have attempted to use the facility have found that it fairly easy to use, not too difficult to transfer from the Sun front-end to the SGI back-end, that it is fairly accessible over the network, and that CPS parallel jobs on the batch ranch work quite well. Nevertheless, there are a number of problems that have been found and a number of tests that we must run to understand how well the system is functioning. Such tests and improvements are now in progress.

4.1 System Elements

Some of the system elements that make the PDSF a single computing system rather than a collection of individual workstations include:

- Network File System (NFS) file mounting
- Network Information System (NIS), i.e. Yellow Pages
- Network Queuing System (NQS) provides batch submission and control over a distributed network
- Cooperative Process Software (CPS) provides FORTRAN callable event parallelization on the system
- Net Central provides network monitoring daemons layered on Sybase
- Standard CERN library routines such as PAW, Geant, etc.
- Data Management System (DMS) provides a FORTRAN interface for tape storage and staging. The tape/file data base is maintained in Sybase on a Sparcstation 2 (db0).
- Robo-tape control software drives the carousel, the bar code reader, and the robot loader
- Workstation assignment system (WASH) runs on the cc0 and allocates the front-end workstation assignment to the users as they log in
- Console concentrator system (CONCH) provides the monitoring on the console output ports of the front-end workstations
- SYSPOLL and SYSMAP provide system monitoring functions and graphical operator interfaces

5. FUTURE PLANS

The immediate goals of Physics Research Computing are to improve the operation of the PDSF, to continue to build our staff, to expand our support of the detector collaborations, and in general to begin to make a larger contribution to the world-wide community of high energy physicists.

Our near term goal is the implementation of the remaining phases of the Simulation Facility. The goal is to provide up to 4000 VAX equivalents of computing power by January, 1992 and to provide several TB of tertiary storage later in the year.

In the longer term, we must address the enormous problems of data collection, storage, and the complex issues of analyzing such large volumes of data. The Lab is in the process of forming internal groups to work on these issues, but we need the help of all physicists and computer scientists who are interested in the eventual success of the SSC.