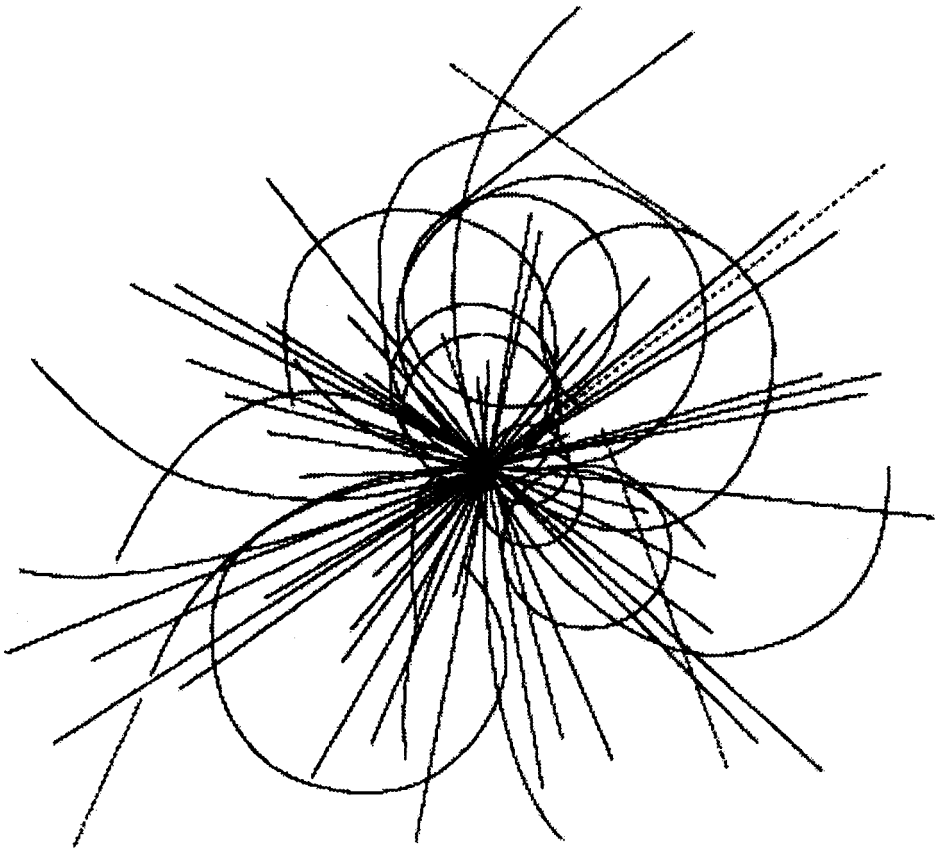
SSCL-Preprint-324

SSCL-Preprint-324
May 1993
Distribution Category: 400

B. Scipioni

Physics Detector Simulation Facility (PDSF) Architecture/Utilization



Superconducting Super Collider Laboratory

To be published in *Supercollider 5* SSCL-Preprint-324

Physics Detector Simulation Facility (PDSF) Architecture/Utilization*

B. Scipioni

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

May 1993

*Presented at the Fifth Annual International Symposium on the Super Collider, May 6–8, 1993 San Francisco, CA.
†Operated by the Universities Research Association, Inc., for the U.S. Department of Energy under Contract No. DE-AC35-89ER40486.

PHYSICS DETECTOR SIMULATION FACILITY (PDSF) ARCHITECTURE/UTILIZATION

B. Scipioni

Physics Computing Department
Superconducting Super Collider Laboratory*
2550 Beckleymeade Avenue
Dallas, TX 75237

ABSTRACT

The current systems architecture for the SSCL's Physics Detector Simulation Facility (PDSF) is presented. Systems analysis data is presented and discussed. In particular, these data disclose the effectiveness of utilization of the facility for meeting the needs of physics computing, especially as concerns parallel architecture and processing. Detailed design plans for the highly networked, symmetric, parallel, UNIX workstation-based facility are given and discussed in light of the design philosophy. Included are network, CPU, disk, router, concentrator, tape, user and job capacities and throughput.

INTRODUCTION

The SSCL Physics Detector Simulation Facility (PDSF) is currently in Phase II of a multi-phase growth plan. The initial plans called for 500 MIPS of computing in Phase I, 1000 MIPS for Phase II and 4000 MIPS for Phase III. However, due to technology advances, Phase II has 3000 MIPS (or 2000 VAX 11/780 equivalents). The original requirements[1, 2, 3] have recently been re-evaluated to determine whether or how the facility will continue to evolve. The original requirements led to system specification,[4] which was implemented in March of 1991. This initial, Phase I, implementation was a prototype. Lessons learned resulted in Phase II which was implemented in March of 1992. This paper is concerned with the architecture and performance of Phase II.

ARCHITECTURE

After having studied the performance characteristics of Phase I PDSF several conclusions were reached. For parallel processing of GEANT detector simulations it was clear that there was no advantage (from a price/performance standpoint) to the parallel SMP computers (SGI 4D/380) over the single CPU workstation (SPARCstation 2). This was no great surprise based upon the loosely coupled nature of the parallelism and the less than 1/1 speedup of the SMP machines and their premium price. (This situation has changed since and will be discussed below). However, data throughput, especially parallel I/O achieved through disk striping and multithreaded raw disk I/O, showed the SMPs to be the better data machines. In addition, threading the Remote Procedure Call (RPC) mechanism turns a (free) client/server technology into a simple-to-use parallel programming paradigm for networked computers. This led to the selection of the SMP computer as both data server machine and launchpads

*Operated by the Universities Research Association, Inc., for the U.S. Department of Energy under Contract No. DE-AC35-89ER40486.

for parallel programs. The ability to have two FDDI interfaces on the SGIs allows for multiple FDDI networks tied together to extend the parallelism to network I/O. This combination of parallel processing, parallel disk I/O and parallel networking has given the PDSF the capability of true overall supercomputer throughput at a fraction of the cost. Another lesson learned might be called spontaneous symmetry breaking. Lack of symmetry in the distributed computing system seems to cause bottlenecks as exemplified by uneven router loads due to upstream/downstream flow of data on an FDDI ring. Sociological factors seem to cause this as well (i.e., familiar computers). Also the multiplexing of Ethernets onto FDDI greatly reduced router packet performance due to excessive fragmentation between dissimilar MTUs. The incorporation of these considerations along with additional acquisitions produced the highly symmetric, multi-networked configfiguration show in Figure 1. It consists of the following:

- 2 clusters of 15 each headless SPARCstation 2s, each with 32 MB RAM and 1 GB SCSI disk
- 2 clusters of 16 each of headless HP 9000/720s, each with 32 MB RAM and 1 GB SCSI disk
- 4 SGI 4D/360s, each with 128 MB RAM and 24 GB SCSI disk
- 2 Summus 8 mm tape robots, each with two drives and 125 GB storage
- 6 SPARCstation 2s, (two for systems database, 4 for multiplexing of consoles)
- Network equipment: 1 HP 9000/720 network management station, 3 Cisco AGS+ routers, 6 FDDI networks with 4 ODS concentrators, and 6 Ethernet networks with 5 ODS concentrators all linking the multiple networks.

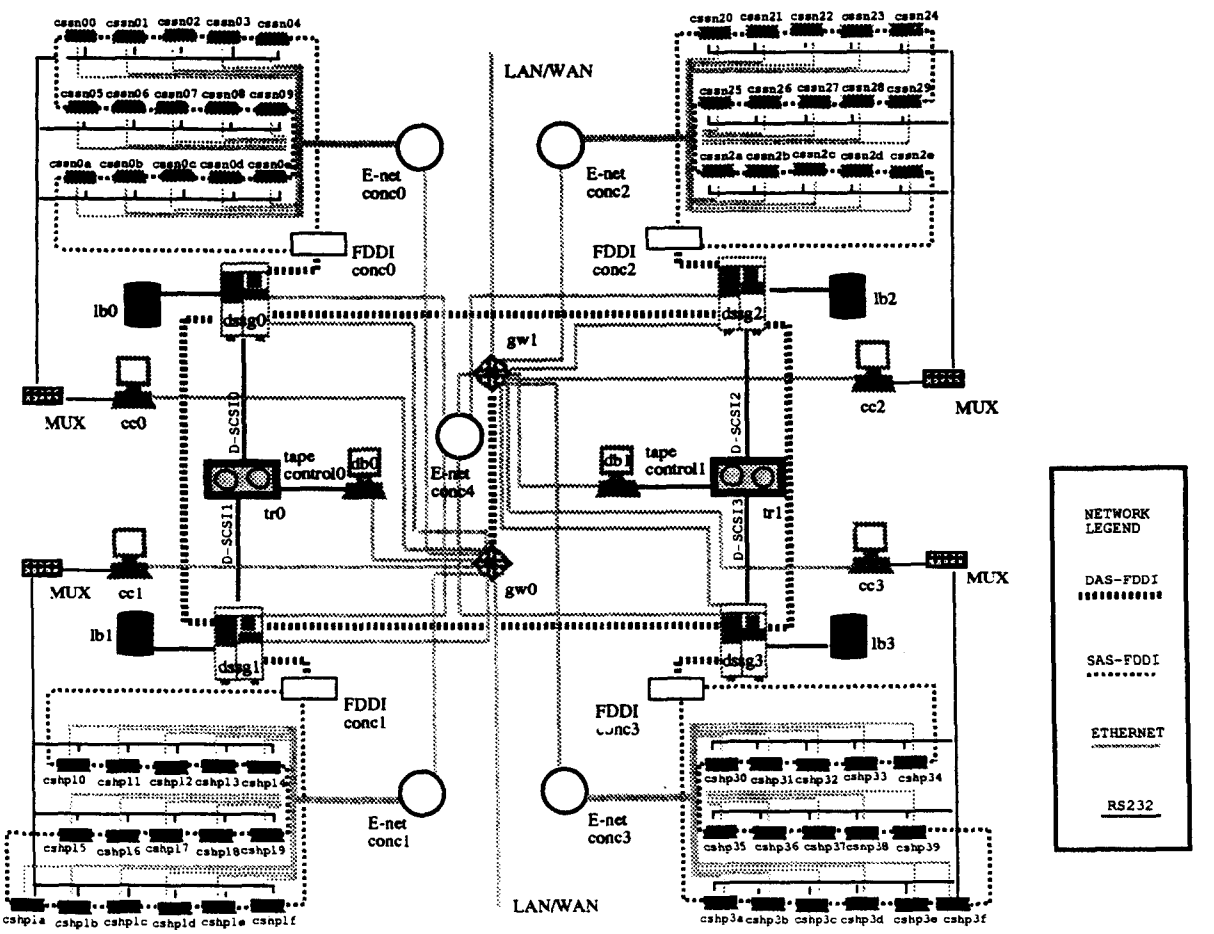


Figure 1. Physics Detector Simulation Facility (PDSF) Architecture.

UTILIZATION

The data contained in Figures 2–4 will be summarized here. There appears to be little true synchronous parallel processing taking place on the PDSF with the exception of the systems analysis software (SISSY).[5] However, there is an event-parallel GEANT which has been run by some SDC members and at least one other application using RPC with the client running outside the PDSF. This, of course, can technically be done over Internet. The parallelism of the architecture, is however, well exploited by the asynchronous (multi-job, multi-user) computing performed on the facility. Some of the more salient points of interest follow:

- **Utilization history**. Over the 1 year Phase II has been in operation certain trends are clear. Figure 2 shows some overall utilization numbers for the last 3 months. The 24 hour CPU utilization of both the SS-2s and the HP-9000/720s has increased from about 25% to approximately 90% during a typical week.The SGIs have risen from 20% to 60%. Higher utilization for SGIs has not been encouraged since they are serving dual roles as both file servers and compute servers. This situation will be remedied in the next Phase of PDSF as discussed below. These numbers fluctuate weekly and are sometimes correlated with collaboration meetings. The two main user groups and three architectures tend to load balance automatically to some extent. The maximum number of concurrent users is now fairly stable at about 80–90. Data server disks has gone from 45% capacity to 80% and has been steady for the last month or two. Among that 100 GB, some file systems remain at capacity most of the time while others fluctuate or remain lower. At the 80% level it is difficult to find large, single pieces of a filesystem for storage of single large files. While there are not many major users of the DMS,[6] there are about 180 GB backed up to 8 mm at this time as seen in Figure 3.
- **Weekly trends**. CPU utilization typically varies widely daily, sometimes reaching close to 100% for entire clusters. This means that weekly averages could only increase by balancing daily loads. However, this in undesirable due to both the low availability and the difficulty of interactive use of the system at higher utilization levels. The current weekly averages are thus considered to be near optimal.
- **Networks**. Hourly snapshots of both the maximum (5 minute average) and average (1 hour average) KB/s are plotted for the week for each computer type. It is readily seen that at times during the week the network traffic exceeds Ethernet maximum bandwidth and at many more times during the week exceeds the practical maximum. This means Ethernets used here would have been a serious bottleneck shutting down the network for many minutes at a time. However, since FFDI is used here there is still good residual capacity on the networks. More tightly coupled parallel applications could easily be run without causing problems.
- **User trends**. Most (>70%) users are from outside the lab (Figure 5). Analysis shows user groups tend to prefer certain architectures. However, during weeks where one of the collaborations is meeting, the freed CPUs tend to be picked up by other groups, although this is not always the case.

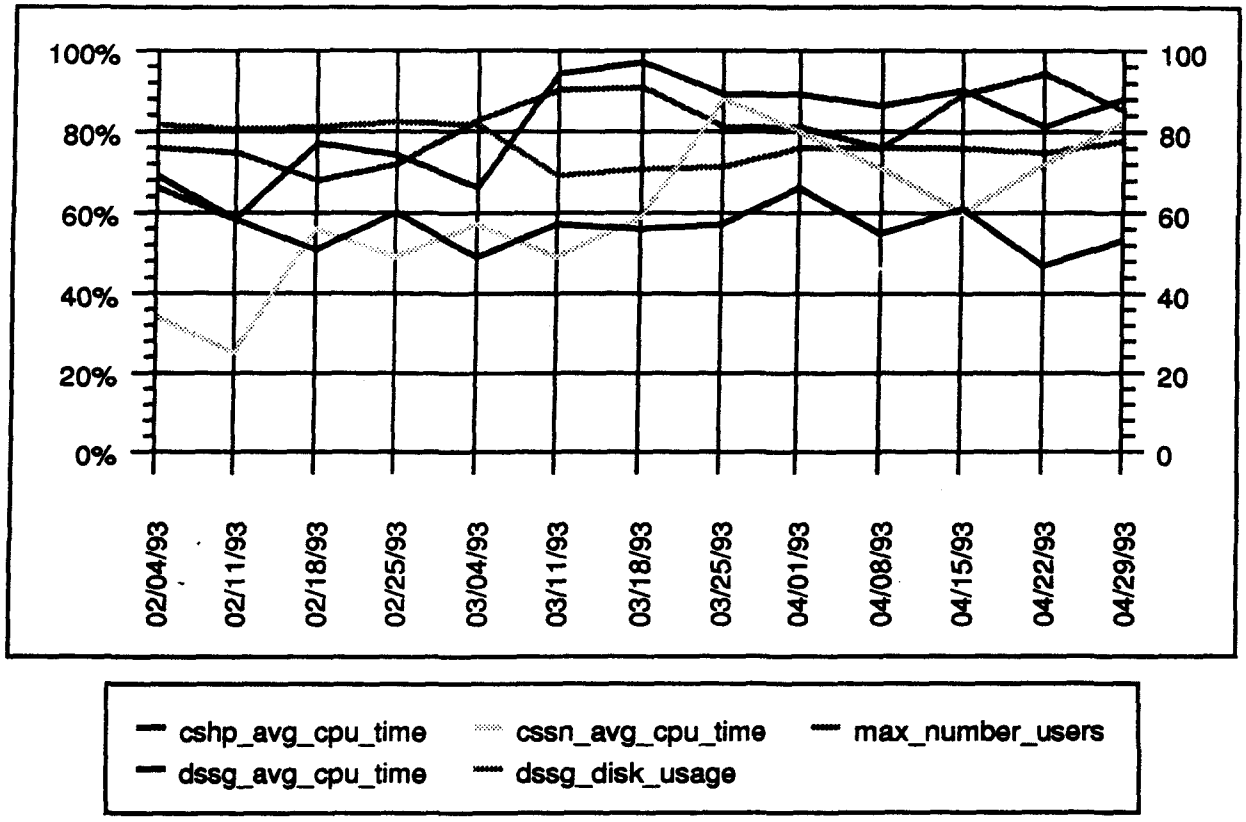


Figure 2. PDSF Utilization History.

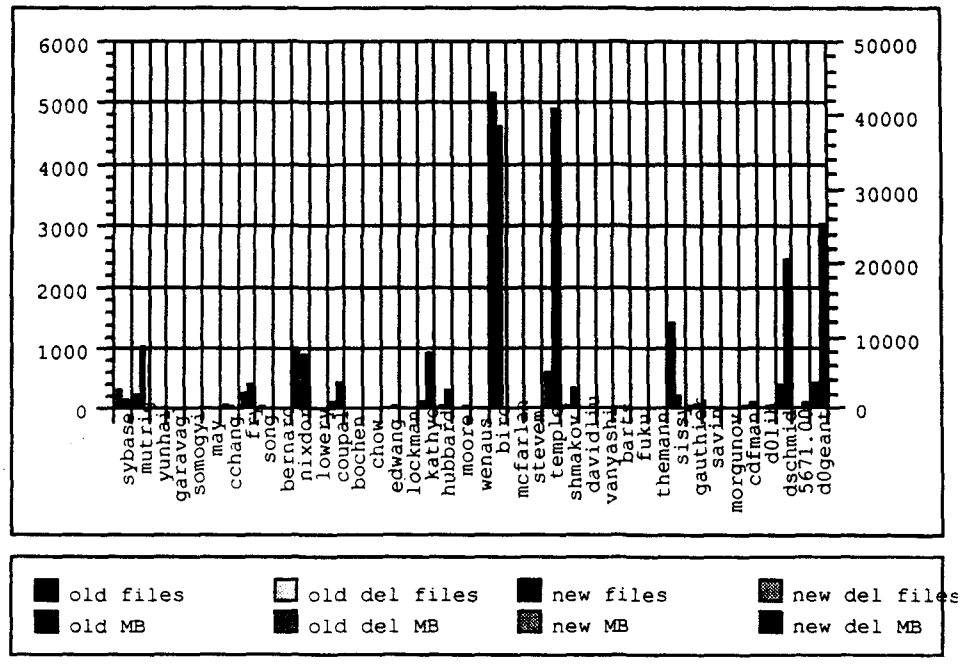


Figure 3. DMS Storage Utilization.

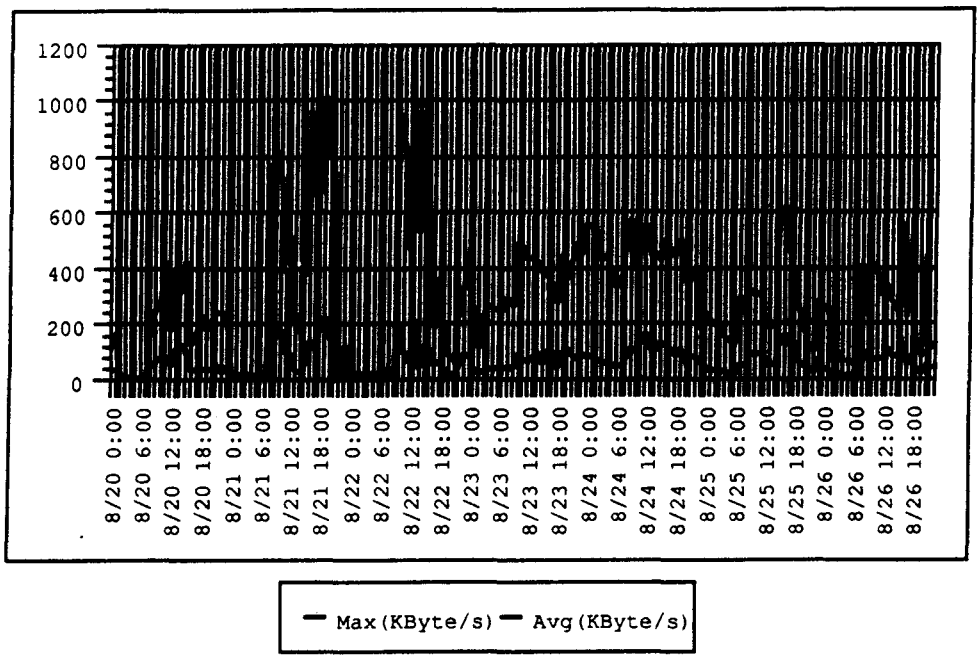


Figure 4. Network Utilization Data.

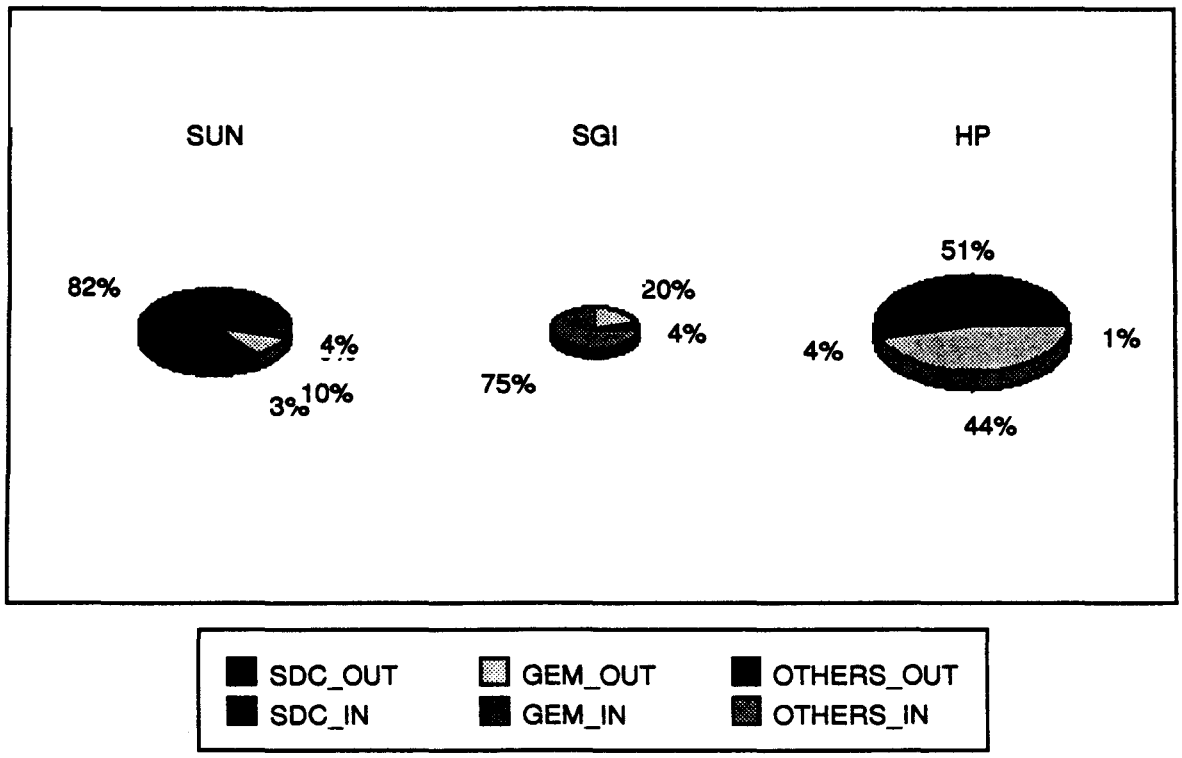


Figure 5. PDSF User Profile.

PDSF III

Based upon our experience with PDSF II and changes in the computer market several changes in strategy have adopted for the next phase of PDSF. Two of the more notable changes in philosophy of its use, and in the price/performance of SMP computers. The next phase will concentrate primarily on providing batch services, with most interactive and general purpose use being on workgroup clusters apart from but attached to the PDSF. Also the price/performance of parallel machines is now superior to single CPU workstations. Computers like the Sun SPARCstation 10 and the SGI Challenge L will be introduced in the next phase.

CONCLUSION

The overall utilization of the PDSF in terms of resource consumption is close to ideal. This indicates that the concurrency inherent in the facility is both an efficient and effective architecture. There is however a comfortable amount of network bandwidth which can be used for synchronous, network-based parallel processing. Since this is more technique-oriented it may take more time and/or encouragement to maximize its utilization as a synchronous parallel processing facility.

REFERENCES

1. M. G. D. Gilchriese, editor, "Report of the Task Force on Computing for the Superconducting Super Collider," SSC-N-579, December (1988).
2. L. Price, editor, "Report of the SSC Computer Planning Committee," SSC-N-691, January (1990).
3. L. Cormell, editor, "Physics and Detector Simulation Requirements," SSCL-259, March (1990).
4. G. Chartrand, L. Cormell, R. Hahn, D. Jacobson, H. Johnstad, P. Leibold, M. Marquez, B. Ramsey, L. Roberts, B. Scipioni, N. Shivapuja, and G. Yost, "Physics and Detector Simulation Facility Specifications," SSCL-275, Attachment A, July (1990).
5. B. L. Scipioni, D. Liu, and T. Song, "SISSY: A Multi-threaded, Networked, Object-Oriented Databased Example," IISSC 5 (these proceedings).
6. J. Allen, et. al., "SSCL - PDSF Phase II Software Support Description," IISSC 5 (these proceedings).