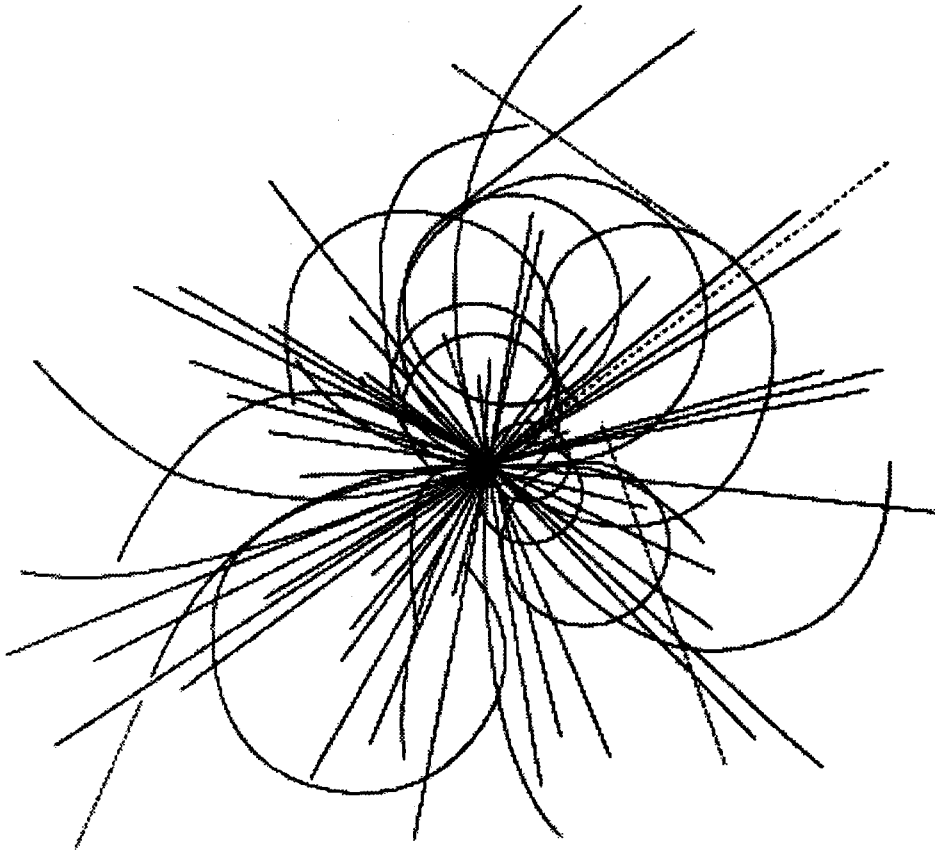
SSCL-Preprint-246

SSCL-Preprint-246
April 1993
Distribution Category: 400

M. Jaffe

FDDI Experience at the SSCL



Superconducting Super Collider Laboratory

To be published in *Supercollider 5*

SSCL-Preprint-246

FDDI Experience at the SSCL*

M. Jaffe

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

April 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.

FDDI EXPERIENCE AT THE SSCL

Mike Jaffe

Physics Research Division
Computing Department, MS-2003
Superconducting Super Collider Laboratory*
2550 Beckleymeade Ave.
Dallas, TX 75237-3997
jaffe@sscvx1.ssc.gov

ABSTRACT

The Physics Detector Simulation Facility (PDSF) is an assemblage of UNIX/RISC workstations and servers which use LAN networking components and standards in a unique way. The PDSF is configured using FDDI much like an internal system bus to a computer system and thus serves as the foundation for the entire PDSF system. This paper will describe the utilization of FDDI in the system, system monitoring, and the interfacing of the PDSF to the local site LAN and WAN environment.

INTRODUCTION

The Physics Detector Simulation Facility (PDSF) is a distributed loosely coupled parallel computing environment made up of groups of workstations and servers functioning together as a single large system. Figure 1 shows the physical and logical layout of the PDSF components.

The compute systems are RISC based architectures running the UNIX operating system and each is equipped with multiple network interfaces.They are interconnected using off-the-shelf LAN technologies in a unique way which utilizes the network as a sort of external/extended system bus or channel. All operations between machines occur via 6 internal FDDI networks which provide the primary communications pathways between systems.

COMPUTING COMPONENTS

There are two types of systems in the PDSF which are referred to as data servers and compute servers. Data servers consist of 4 Silicone Graphics Inc. (SGI) 4D/360 systems each equipped with 6 processors. The compute servers consist of 62 systems arranged in 4

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

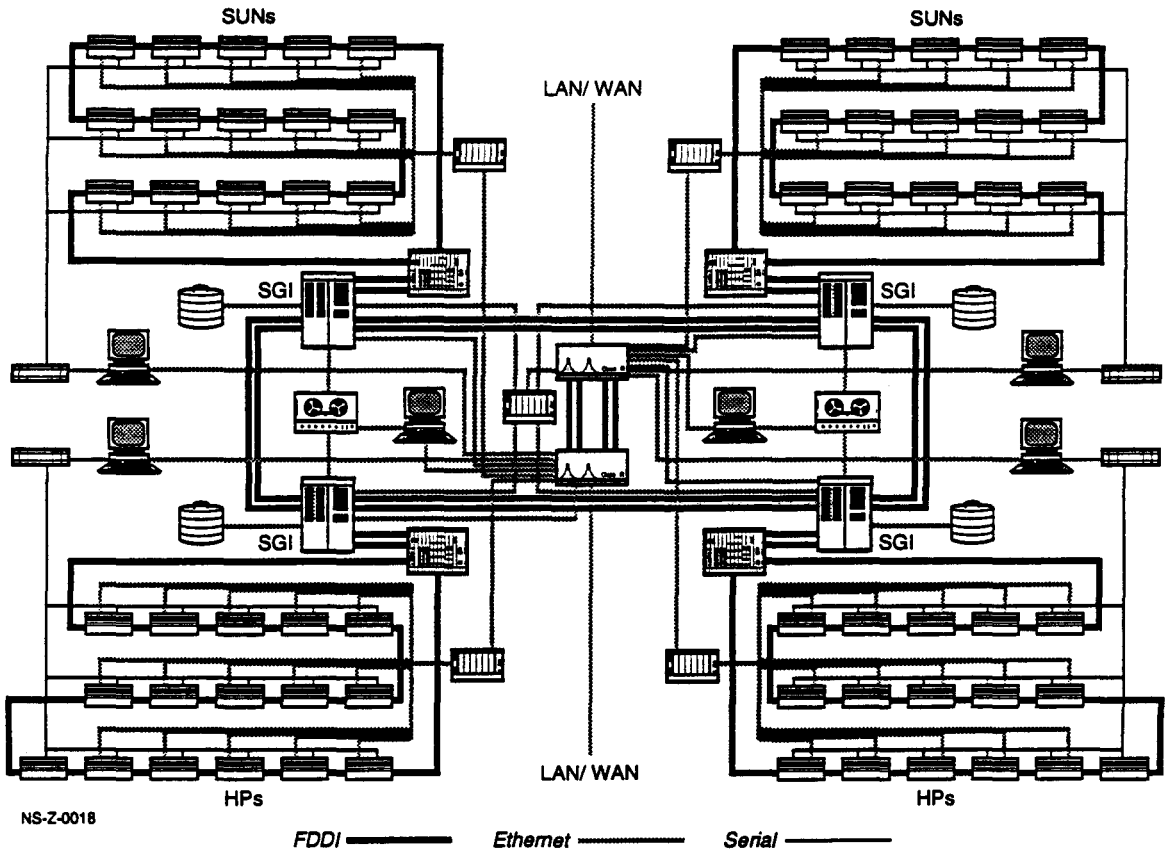


Figure 1. The FDSF Layout

groups, one group per data server, and are made up of two groups of Sun Sparc 2 and two groups of Hewlett Packard (HP) 9000/720 keyboardless and monitorless workstations. There are a total of 30 Sun and 32 HP compute servers with an additional 6 Suns and 1 HP designated for control systems. The total capacity of the PDSF is approximately 3000 MIPs or 2000 VAX equivalents. Secondary storage for the system consists of 160 Gbytes of hard disks and 250 Gbytes of 8mm tape storage arranged in two tape robot systems.

NETWORK COMPONENTS

The strength of the PDSF is its network architecture which allows communication between the many systems comprising the environment. Each compute server is equipped with 1 ethernet and 1 FDDI Single Attach Station (SAS) network interface while each data server is equipped with 2 ethernet and 2 FDDI Dual Attach Station (DAS) network interfaces. The ethernets are connected together into segments using 5 Optical Data Systems (ODS) ethernet concentrators. The FDDI systems are slightly more complex in arrangement and consist of 6 separate FDDI rings as follows: one ring connects the 4 data servers, one ring connects the two routers, four separate rings connect each of the 4 groups of compute servers to their respective data server using 4 ODS FDDI Dual Attach Concentrators (DAC). The entire system is tied together and to the outside world using 2 Cisco Systems AGS+ FDDI/ethernet routers.

FDDI'S ROLE

The foundation of the PDSF is the FDDI which connects its data servers to data servers, data servers to compute servers, compute servers to compute servers, and finally router to router. These FDDI systems make up the backbone of the system and provide a

high speed pathway similar to that of a computer system bus or channel over which the activities of the PDSF can occur. These range from resource sharing activities such as Network File System (NFS) and Yellow Pages, to operational utilities used to manage, monitor, and maintain the environment, to user's software utilizing UNIX sockets, RPC, and other network oriented access facilities. All activities internal to the PDSF occur using the FDDI rings and in fact the default names of the systems are the FDDI interfaces which insures that the majority of the network load occurs using these high speed pathways. All of these activities have the ability to access any of the individual network interfaces, and thus pathways, by specifying the IP subnet associated with the interface desired.

IMPLEMENTING FDDI

FDDI has played an instrumental role in the PDSF from its inception and it continues to challenge products, their manufacturers, and the standard itself. Implementing it was no small chore and a number of interesting problems had to be overcome in making it all work. The first difficulty encountered related to Station Management (SMT) and had to do with differences in how the product manufacturers had interpreted "proper" responses to informational request frames from stations on the rings. This didn't pose an operational problem really but did make diagnosing other problems very difficult since it wasn't clear to all stations on the ring who actually existed and who didn't. These problems were worked out with the involved vendors who have since modified their individual implementations of SMT.

Another problem experienced had to do with high signal attenuation (> 6 db) in some of the optical bypass switches which were included in the first phase of the PDSF. This attenuation was causing the systems on the ring to intermittently wrap in response to what appeared to them to be a failed fiber or system. This was the result of the 11 db FDDI defined link budget being exceeded. These particular switches were found to be defective and replacing them corrected the problem. Having experienced this problem, the whole issue of the usefulness of these devices was brought up and the jury is still out as to whether they should be required in an FDDI implementation. Although they can be useful in maintaining and diagnosing problems on a ring by allowing stations to be removed more gracefully (manually because most vendors don't yet provide software bypass control). They can also be troublesome as in the case of more than one station in a row going into bypass mode in which case the ring will probably wrap due to link budgets being exceeded. Their worth is probably best stated as being "system dependent."

Two related issues occurred and centered on the difference between packet sizes allowed on ethernet (1500 byte) and FDDI (4500 byte). The first involved packets transitioning between the FDDI rings and the ethernet segments. In order for the larger FDDI packets to fit onto the ethernet segments either the FDDI hosts or the routers had to be able to fragment them into appropriately sized units. Two solutions were available one of which restricted the FDDI hosts to using the smaller ethernet packets in which case the router passed its packets onto the ethernet segments unmodified. The other solution, and the one that proved most efficient, was to set the FDDI hosts to utilize full size FDDI packets and let the router fragment those needing to transition to an ethernet segment. The related issue dealt with the Maximum Transmission Unit (MTU) size on the FDDI interfaces for the three types of systems in the PDSF (SGI, Sun, and HP). These were found to be set differently by default but fortunately each system allowed this value to be reset to the maximum 4500 byte value.

The problems encountered in implementing FDDI in the PDSF were more of a nuisance than a show stopper. They were for the most part the result of a lack of experience and immaturity in the products, the market, and the standard itself. Additionally, problem resolution was somewhat compounded by a lack of available diagnostic tools specific to

FDDI. It was evident by the time the second phase of the system was implemented that FDDI marketplace was maturing.

ACCESSING THE SYSTEM

Of course the true power of the PDSF can only be realized through its availability to users both local and around the world. This availability is facilitated by its connectivity to the SSCL LAN/WAN which is connected to the Internet through ESNET and BITNET connections. Both of these networks are well developed providing potential T1/DS1 connection speeds to most of the remote users depending on their location and site. A typical access to the PDSF either local or remote is through an X-windows session requiring a significant amount of network bandwidth over a less typical terminal access session. Based on user feedback, response to the system is good which is significant since over 70% of the PDSF users are remote to the SSCL.

MANAGING THE NETWORK

Although the system functions like a single computer system, the PDSF incorporates as much networking as many small computing sites. Thus monitoring and managing its network is crucial to its continued operation and success. A combination of tools are used for this chore including several home-grown UNIX based utilities, Cisco System's NetCentral Simple Network Management Protocol (SNMP) software which includes the Sybase SQL database, and HP's OpenView Network Node Manager SNMP software in addition to a little creativity. These systems run continuously and are used to monitor and collect both instantaneous and trend data for all of the networks in the environment along with assisting in the isolation, diagnosis, and repair of network problems when they occur.

Most PDSF network monitoring and management is performed using the two SNMP tools mentioned but it's important to note that SNMP itself is not without its limitations. It's functionality and usefulness centers on Management Information Bases (MIBs) which amount to lists of all possible values that can be examined on a device. For example MIB entries might be defined to examine total traffic in and out of an interface on a system. In the case of FDDI specifically there are currently no MIB values available to examine its SMT components. A new MIB addressing this limitation has been proposed (RFC1285) and hopefully will be adopted and implemented quickly by the vendor community.

SUMMARY

The PDSF has been in operation in its current configuration for approximately a year now and its network design has proven sound. It just plain works and has required very little modification other than the ongoing occasional tuning of system activities and applications which is performed by its administrators. It is clear from the initial phase of the PDSF and from data gathered in its second phase that the multiple FDDI's are both key and crucial to its operation. The strength of the system is the availability of all of its networks in parallel allowing both high speed and simultaneous access for the systems and user level tasks being performed. Remarkably all of the systems in the PDSF deal very well with the multiple network interfaces and other than a few initial setup issues relating to what used what pathway and who would handle which routing etc. things work very well. All of the networks are being utilized to at least a moderate level and future expansions of the PDSF will certainly add even more FDDI to this unique computing/network environment.