Scalable Enterprise Storage Solutions

For Systems Based on the Intel® Pentium® II Xeon™ Processor

Revision 1.0

December 14, 1998

Enterprise Server Group

Revision History

Date	Rev	Modifications
12/16/98	1.0	Initial release.

The AD450NX MP Server System and the AC450NX MP Server System may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Conventions and Terminology

This document uses the following terms and abbreviations:

Term	Definition
AKA	Also Known As
CPU	Central Processing Unit
DMI	Desktop Management Interface
Failover	Transfer functionality to another system
FC-AL	Fibre Channel Allocated Loop
FRU	Field Replaceable Unit
HBA	Host Bus Adapter
Hot-plug	Plugging the adapters while system power is ON
IA	Intel® Architecture
ISC	Intel® Server Control
JBOD	Just a Bunch Of Disks (disks that are directly accessed as opposed to RAID)
LAN	Local Area Network
NIC	Network Interface Card
OS	Operating System
PCI	Peripheral Component Interconnect
PHP	PCI Hot-plug
RAID	Redundant Array of Inexpensive Disks
SAN	Storage Area Network
SCSI	Small Computer System Interface
Striping	Distributing data across several drives for redundancy
TB	Terabytes (approximately 1000 Billion=10^{12} bytes)
UPS	Un-interruptable Power Supply
WWW	World Wide Web
XOR	Exclusive OR

References

Refer to the following Web sites for additional information:

Fibre Channel*: http://www.fibrechannel.com
Clariion* http://www.clariion.com
Symbios*: http://www.symbios.com
Emulex*: http://www.emulex.com
QLogic*: http://www.qlc.com
Adaptec*: http://www.adaptec.com
Boxhill*: http://www.boxhill.com
DPT Corp.*: http://www.dpt.com
MAXstrat*: http://www.maxstrat.com

Table of Contents

List of Figures

List of Tables

1. Abstract

This document describes Scalable Server Storage Solutions developed on Intel® Pentium® II Xeon™ processor based systems. Currently these are the AD450NX and AC450NX systems, both of which are based on the Intel® 450NX chip set. A brief description of the different server models is given in this paper, along with the actual solutions developed and equipment used.

Until recently a scalable server storage solution on Intel® Architecture (IA) was very limited and did not provide the price/performance available on other platforms. With the advent of Fibre Channel (FC) storage and the availability of high performance four-way Pentium II Xeon processor based servers, it is now possible to configure and build enterprise level solutions with management functions that facilitate remote administration while providing a very flexible and highly scalable platform. Since 126 (125+1 host) devices can be attached to a single FC loop with 10 available PCI slots, the theoretical upper limit storage is 10x125x9G = 11,250 GB or 11 TB. Using 18G drives, which are now becoming available, this limit doubles to 22 TB attached to a **single** server. This kind of scalability was simply not possible using the conventional SCSI interface.

All of the server models developed here are based on Microsoft* Windows NT* 4.0 Enterprise Edition. Other operating systems (OSs) can be adapted easily to these configurations as well. The primary storage interface used in these examples is Fibre Channel Arbitrated Loop (FC-AL). The advantages of FC over SCSI interface are numerous and need not be listed here. If needed, the reader is encouraged to consult the reference section for further information on Fibre Channel.

2. Server Models

The simplest approach to a server solution is to provide a stand-alone server interfaced to the storage devices. The storage devices can be either Just a Bunch of Disks (JBOD) or a Redundant Array of Inexpensive Disks (RAID). There are primarily two ways to attach a storage device to the server in an enterprise environment:

- Fibre Channel
- SCSI

Fibre Channel is a serial interface and provides greater flexibility for scaling, whereas SCSI works as a parallel interface and has a limited number of devices per host adapter (15 in the case of Wide SCSI). FC can attach 126 devices per host adapter. This fact greatly enhances the scalability of systems based on the FC interface. Furthermore, FC hubs can be used in FC loops to share storage devices thus allowing multiple servers to be clustered, providing greater availability.

Even though the solutions developed here are based on RAID systems, scalable JBOD solutions have also been used and can be configured easily with Clariion* disk arrays. Two vendors have provided the storage systems used in this study. Clariion FC5500 RAID systems use a dual controller, dual loop RAID solution with Fibre connection to the drives. Symbios* RAID solution provides an FC interface to the host and has dual controllers with one FC loop per controller. The drives on the Symbios systems are differential ultra SCSI.

The FC host adapters used in this development effort are QLogic*, Emulex*, and Adaptec*. We used QLogic to attach the Clariion storage, Emulex for the Symbios storage systems, and Adaptec for Boxhill* systems.

2.1 Single Server

The most common and simple storage solution is based on a single server attached to a storage subsystem. The storage can be either JBOD- or RAID-based. There is usually a network interface card (NIC) which connects the server to the rest of the network. HP OpenView* software is installed on the client workstation to allow remote management, fault isolation, and security-related functions. The typical single server/single storage model is as follows.

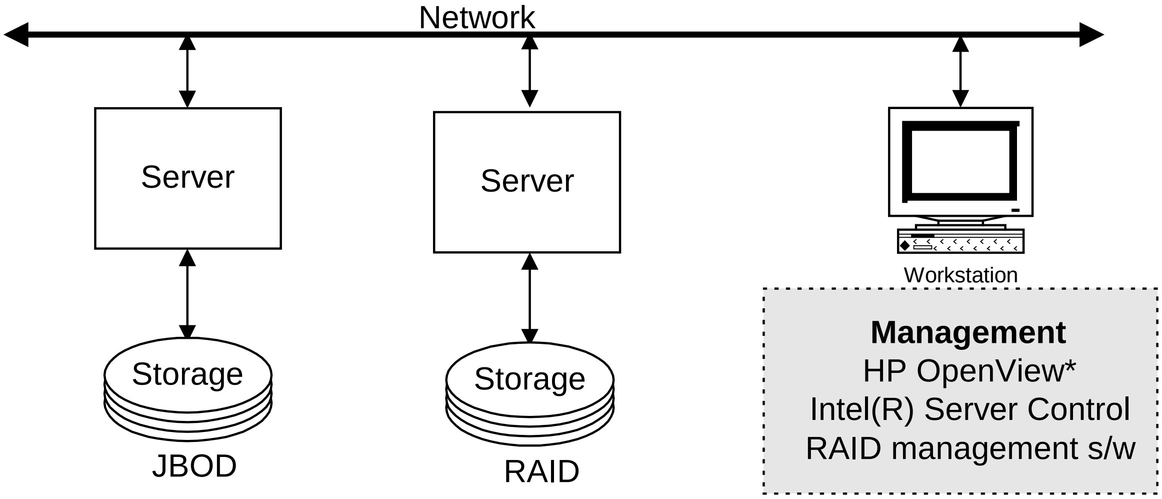

Figure 1: Single Server/Dedicated Storage Model

The two cases that were used for this model are detailed below.

2.1.1 Clariion* System

Clariion provides RAID or JBOD solutions in the FC5000 model Disk Array Enclosures and the FC5500 Disk Processor Enclosures respectively. These subsystems offer a dual loop, dual controller FC host interface to increase availability and reliability. All components of the Clariion systems can be hot-plugged providing a highly available configuration. The Clariion systems were used in stand-alone and clustered modes. The configuration used for the stand-alone mode is described below. With 9 GB Seagate* Barracuda* drives, the storage capacity of this configuration is in the order of 1 TB. When 18 GB drives are available, the storage capacity of this system can be augmented to 2 TB.

Note that the Clariion system has dual controller and dual Fibre link cards for the drive interfaces. These redundant components greatly enhance the availability of the system by providing alternate paths from the storage to the host system via a dual loop/dual adapter configuration.

For enterprise-level management, HP OpenView is installed on the client system. Intel® Server Control software provides component level fault isolation of the server and the Navisphere Manager* provides the RAID management functions right from the workstation. Navisphere also provides status monitoring and fault isolation for the RAID hardware modules such as controllers, fans, FC link cards, power supplies, and drives.

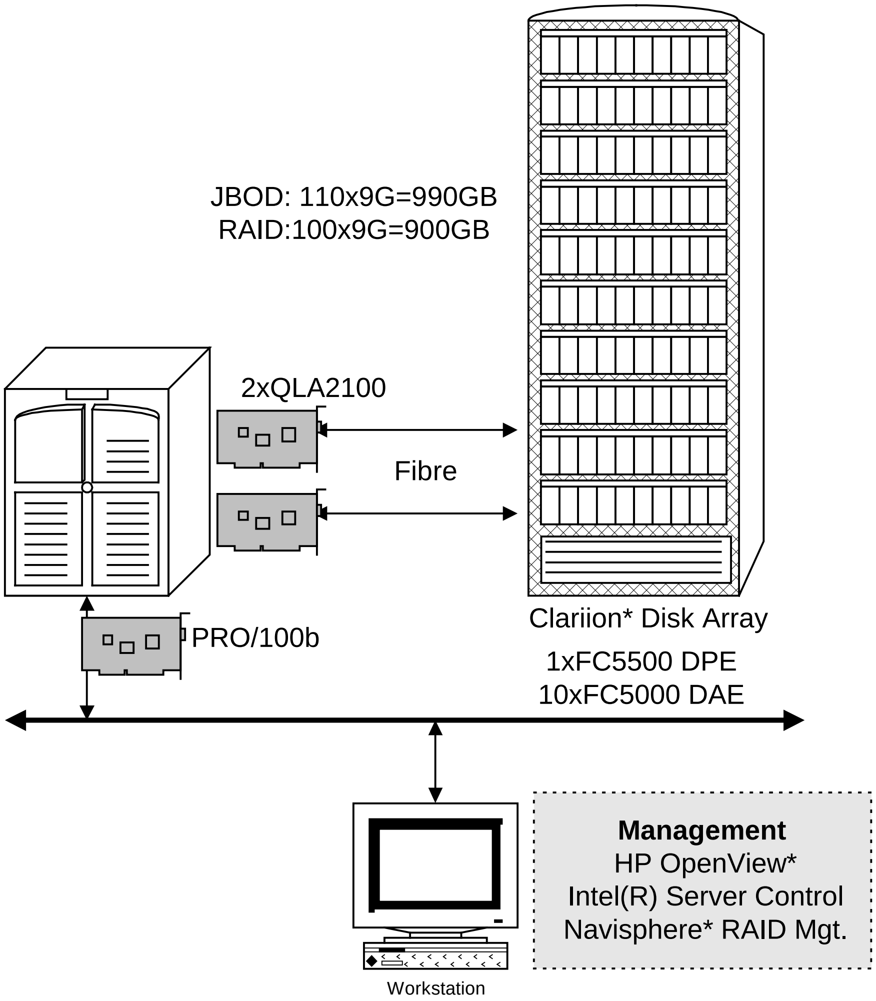

Figure 2: Clariion Stand-alone Server Mode*

The fact that Clariion disk packs have native Fibre interface gives the user a very flexible approach to scaling. Each disk pack (10 drives) can be used either as a JBOD or as part of a RAID system and can be daisy-chained as necessary. Furthermore, the AC450NX servers support PCI Hot-plug (PHP) slots which makes replacing defective host adapters possible without shutting off the system. The current PHP solution supports “like for like” replacement which means a PCI card can be replaced with an identical version and continue using the same drivers. The driver support for hot-plugging will be available in the near future.

System Summary

The following table summarizes the various components used to build the system. Note that the network hubs and the client systems are not included in this list.

Table 1: Single Server System Summary

Item	Model	Qty.	Notes
Server: Intel Intel® AC450NX/AD450NX System, four 400-MHz Pentium® II Xeon™ processors, 128 MB RAM	AC450NX/AD450NX	1	System uses onboard SCSI drives to boot to Windows NT* 4.0 Enterprise Edition. On AC450NX: Hot-plug PCI slots for “like for like“ replacement.
Storage: Clariion* DPE RAID controller DAE disk subsystems	 FC5500 FC5000	 1 10	Includes 2 x storage processors + 10 drives 10x9G-ST39173FC Seagate* FC drives (total ~1 TB storage).
Disk: Seagate*	ST19171FC	100	Barracuda* 9G.
Bus adapters: QLogic* Corp.	QLA2100	2	64-bit PCI HBA for FC-AL.
Network: Intel	Pro 100B	1	Running at 100Mbs .
Software: Operating system RAID management Enterprise management Server management	 Windows NT* 4.0 Navisphere* HP OpenView* Intel® Server Control	 1 1 1 1	 Microsoft* Clariion Hewlett-Packard* Intel

2.1.2 Symbios* System

The Symbios RAID subsystem uses differential SCSI drives to interface to the control modules. The host interface however is based on FC. The RAID controller SYM1000 has a 10Mbs Ethernet Interface for remote management. SYMplicity* Storage Manager runs on a client workstation and provides the RAID management interfaces. A direct null modem connection to the SYM1000 controller also provides low-level management functions. Because Symbios currently supports a 10Mbs Net interface, a network switch must be used to connect to a 100Mbs backbone. For the purposes of this demonstration, a 10Mbs hub was used.

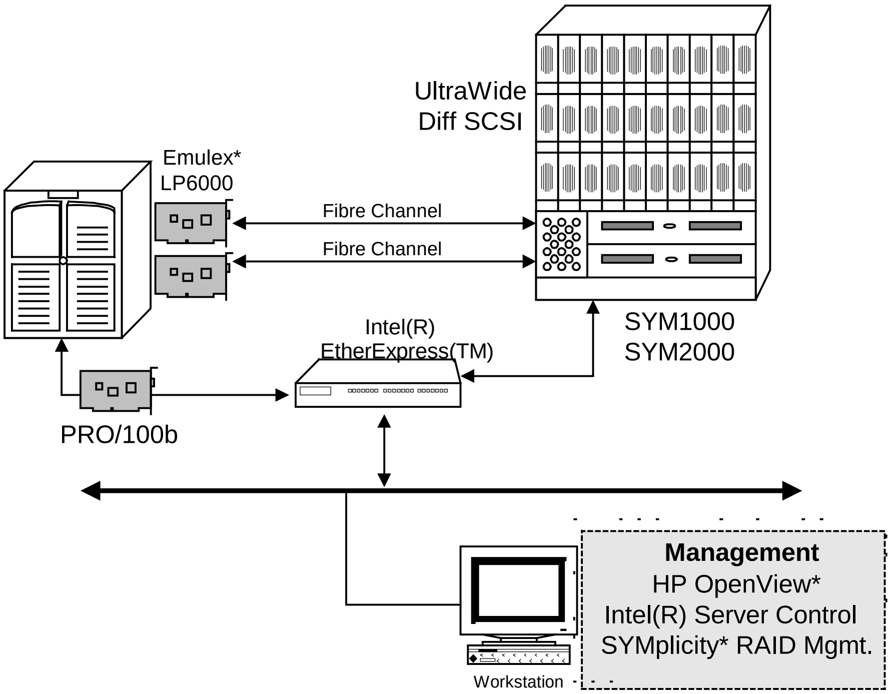

Figure 3: Symbios Stand-alone Server Mode*

Table 2: Symbios System Summary*

Item	Model	Qty.	Notes
Server: Intel Intel® AC450NX/AD450NX System, four 400-MHz Pentium® II Xeon™ processors, 128 MB RAM	AC450NX/AD450NX	1	System uses onboard SCSI drives to boot to Windows NT* 4.0 Enterprise Edition. On AC450NX: Hot-plug PCI slots for "like for like" replacement.
Storage: Symbios* RAID controller Storage	SYM1000 SYM2000	1 3	Fibre Channel host interface, Ultra Wide Diff. SCSI interface for drives. 10 drives/unit.
Disk: Seagate*	ST19171FC	30	Barracuda* 9G.
Bus adapters: Emulex*	LP6000	2	32-bit PCI HBA for FC-AL.
Network: Intel® EtherExpress™ adapter	PRO100b	1-2	May need a second one for public net.
Software: Operating system RAID management Enterprise management Server management	 Windows NT* 4.0 SYMplicity* HP OpenView* Intel® Server Control	 1 1 1 1	 Microsoft* Symbios* Hewlett Packard* Intel

2.2 High Availability Cluster Model

The clustered server model has the advantage of a "failover" mechanism to prevent loss of server functions when one server malfunctions or crashes. One clustered server solution that is available for Windows NT systems is provided in the Windows NT 4.0 Enterprise Edition. The solution that was developed for Clariion and Symbios systems was used with Microsoft Cluster Server*.

The following diagram shows the model of this configuration.

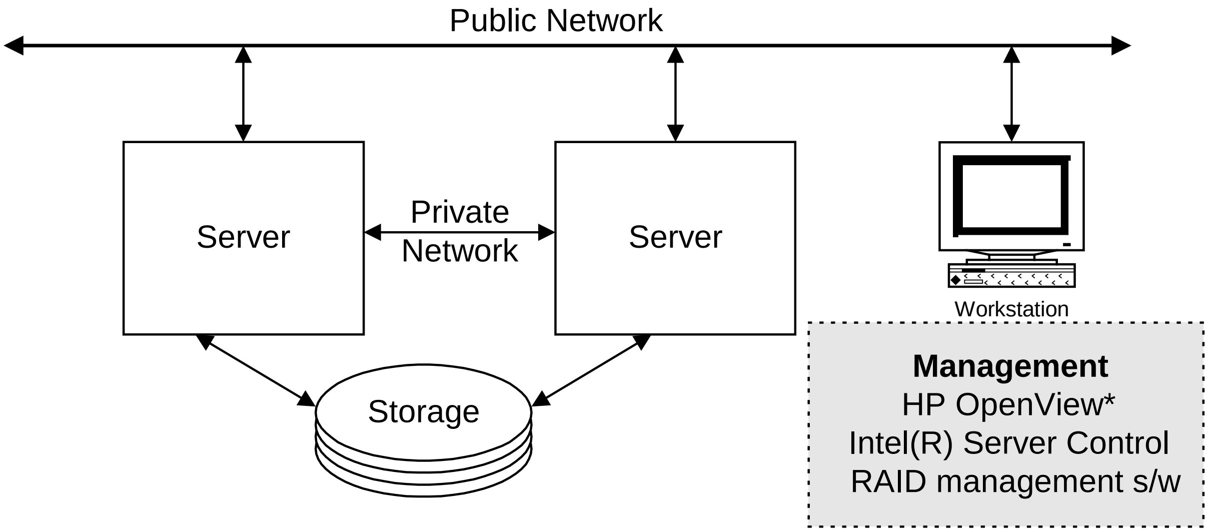

Figure 4: High Availability Cluster Model

In this mode, a portion of the common shared storage is used to store "keep alive" information. The servers can communicate via a private network as well as a public network. The private network requires a second set of NICs and a hub. The management workstation will manage both servers and the cluster using HP OpenView and Intel Server Control (ISC) software. The following configurations were developed with Microsoft Cluster Server.

2.2.1 Clariion*

Essentially the configuration is similar to the single server model, except this configuration has an additional server to share the storage and failover redundancy.

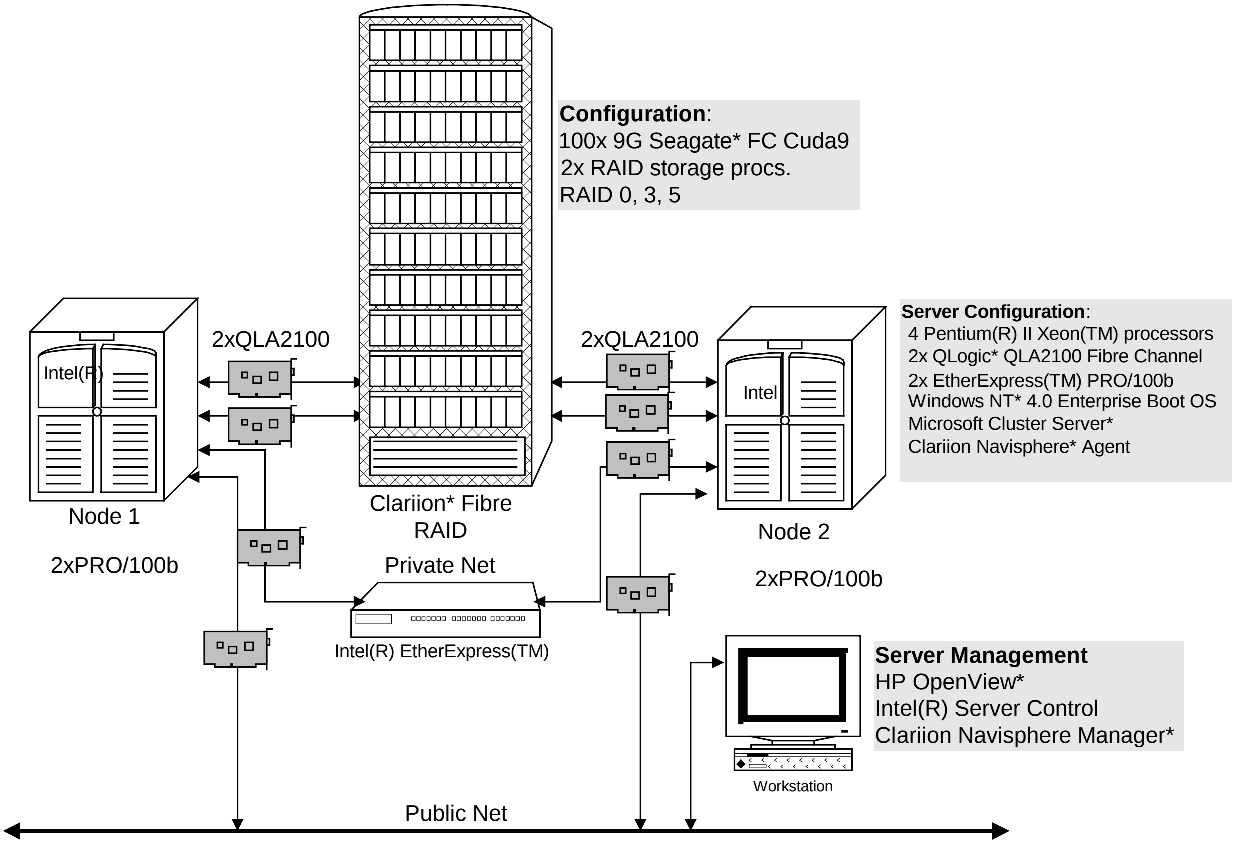
Configuration:
100x 9G Seagate* FC Cuda9
2x RAID storage procs.
RAID 0, 3, 5
Intel(R)
Node 1
2xQLA2100
2xPRO/100b
Clariion* Fibre RAID
Private Net
Intel(R) EtherExpress(TM)
2xQLA2100
Intel
Node 2
2xPRO/100b
Server Configuration:
4 Pentium(R) II Xeon(TM) processors
2x QLogic* QLA2100 Fibre Channel
2x EtherExpress(TM) PRO/100b
Windows NT* 4.0 Enterprise Boot OS
Microsoft Cluster Server*
Clariion Navisphere* Agent
Server Management
HP OpenView*
Intel(R) Server Control
Clariion Navisphere Manager*
Workstation
Public Net

Figure 5: Clariion, Microsoft Cluster Server* Cluster*

Windows NT 4.0 Enterprise Edition provides all the components needed to run a clustered server solution. In this mode, both servers provide a pool of services and either server can take over the host functions should the other server fail to continue reliable operation. The dual adapter/dual loop further enhances the redundancy nature of the FC configuration. The disk packs used in the Clariion system are modular and thus provide a highly scalable storage solution that can grow as the need for storage grows.

Again, enterprise-level management functions are provided by the HP OpenView, Intel Server Control and Clariion Navisphere software components that help with the monitoring, managing, and fault isolation tasks from any remote client attached to the network.

The following table summarizes the various hardware and software components used to help build this configuration.

Table 3: Clustering System Summary

Item	Model	Qty.	Notes
Server: Intel® AC450NX/AD450NX System, four 400-MHz Pentium® II Xeon™ processors, 128 MB RAM	AC450NX/AD450NX	2	System uses onboard SCSI drives to boot to Windows NT* 4.0 Enterprise Edition. On AC450NX: Hot-pug PCI slots for “like for like“ replacement.
Storage: Clariion*			
DPE RAID controller	FC5500	1	Includes 2 x storage processors + 10 drives
DAE disk subsystems	FC5000	10	10x9G-ST39173FC Seagate* FC drives (total 1 TB storage).
Disk: Seagate*	ST19171FC	100	Barracuda* 9G
Bus adapters: QLogic* Corp.	QLA2100	4	64-it PCI HBA for FC-AL
Network: Intel® EtherExpress™ adapter	PRO100B	4	Running at 100Mbs, public, private nets.
Software:			
Operating System	Windows NT* 4.0	1	Microsoft*
RAID management	Navisphere*	1	Clariion
Enterprise management	HP OpenView*	1	Hewlett-Packard
Server management	Intel® Server Control	1	Intel

2.2.2 Symbios*

The clustered configuration used for Symbios has the same scalable storage components as the single server model with the exception of the second server that forms the Microsoft Cluster Server cluster. Additionally, a second network card and an FC hub is used to augment the FC loop. The final tested configuration is shown in the following diagram. Essentially this setup is similar to the Clariion system, except an FC hub is added to the FC loop to provide access to both controllers by either one of the nodes. This is done to enhance redundancy and provide uninterrupted storage interface functions in case one of the RAID controllers faults. The PCI slots support hot-plug features so the server can stay online while a defective host bus adapter is replaced. The following is a summary of the configuration.

Table 4: Symbios System Summary*

Item	Model	Qty.	Notes
Server: Intel® AC450NX/AD450NX System, four 400-MHz Pentium® II Xeon™ processors, 128 MB RAM	AC450NX/AD450NX	2	System uses onboard SCSI drives to boot to Windows NT* 4.0 Enterprise Edition. On AC450NX: Hot-plug PCI slots for “like for like“ replacement.
Storage: Symbios*			Fibre Channel host interface, Ultra Wide SCSI
RAID controller	SYM1000	1	interface for drives.
RAID storage	SYM2000	3	10 drives/unit.
Disk: Seagate*	ST19171WC	30	Barracuda* 9G.
Bus adapters: Emulex	LP6000	2	32-bit PCI HBA for FC-AL.
Network: Intel® EtherExpress™ adapter	PRO100B	2	Private net, public net interfaces.
Hub: Gadzoox* FC-AL	FCL1063TW	1	Provides redundant loops for high availability.
Software:			
Operating system	Windows NT* 4.0	1	Enterprise, Microsoft Cluster Server*
RAID management	SYMplicity*	1	Symbios*
Enterprise management	HP OpenView*	1	Hewlett-Packard
Server management	Intel® Server Control	1	Intel

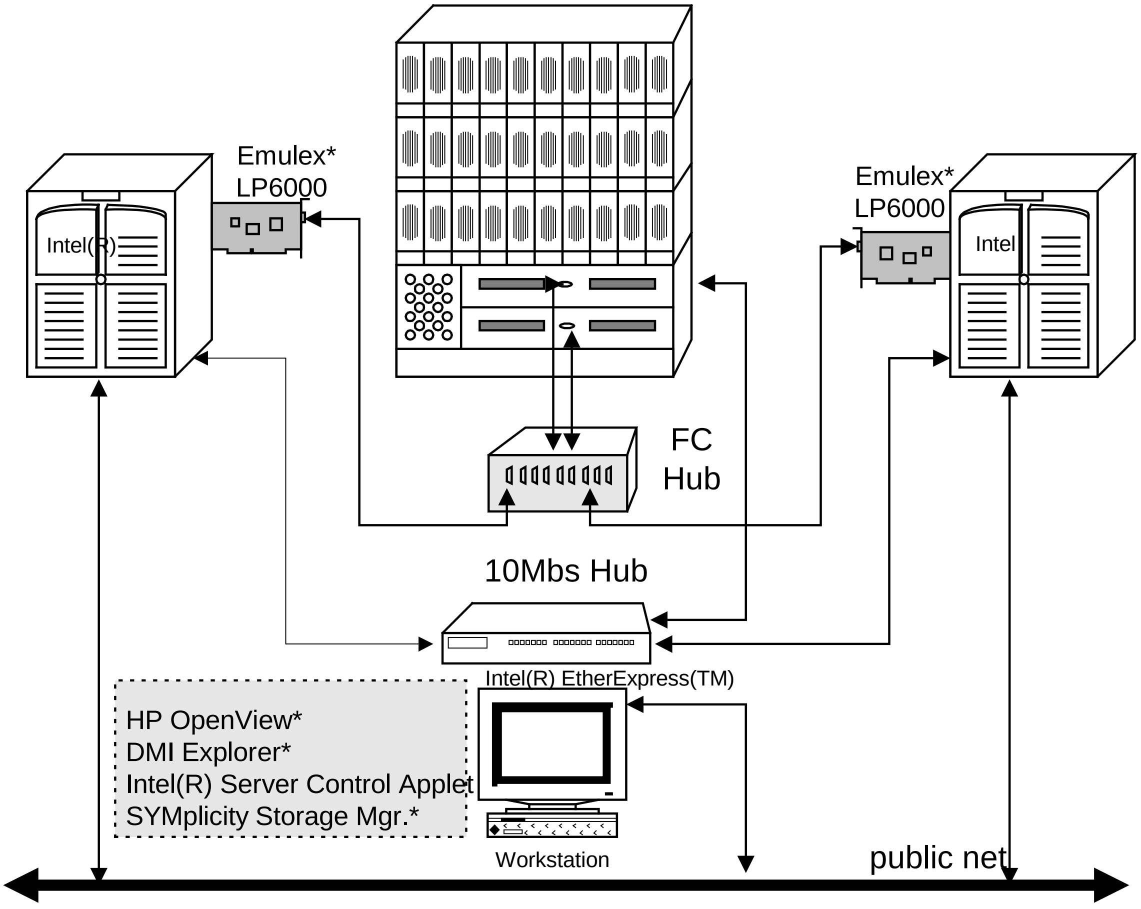

Figure 6: Symbios, Microsoft Cluster Server* Cluster*

2.3 Enterprise Management Model

The next step of integration involves multiple server models coexisting on the same network and managed from a central location and workstation. Both Clariion and Symbios systems with Microsoft Cluster Server were attached to the same public network and managed from a single workstation seamlessly. This model will be the most common configuration, considering there are many types of servers and storage systems present at any given time in the enterprise.

The following diagram illustrates this model. The server management which is running on the remote workstation will be able to identify failed components via Desktop Management Interface (DMI) and ISC down to the basic field replaceable units (FRUs). Since all servers used will have PCI hot-plug functionality, it will be possible to replace defective HBAs without taking any server offline.

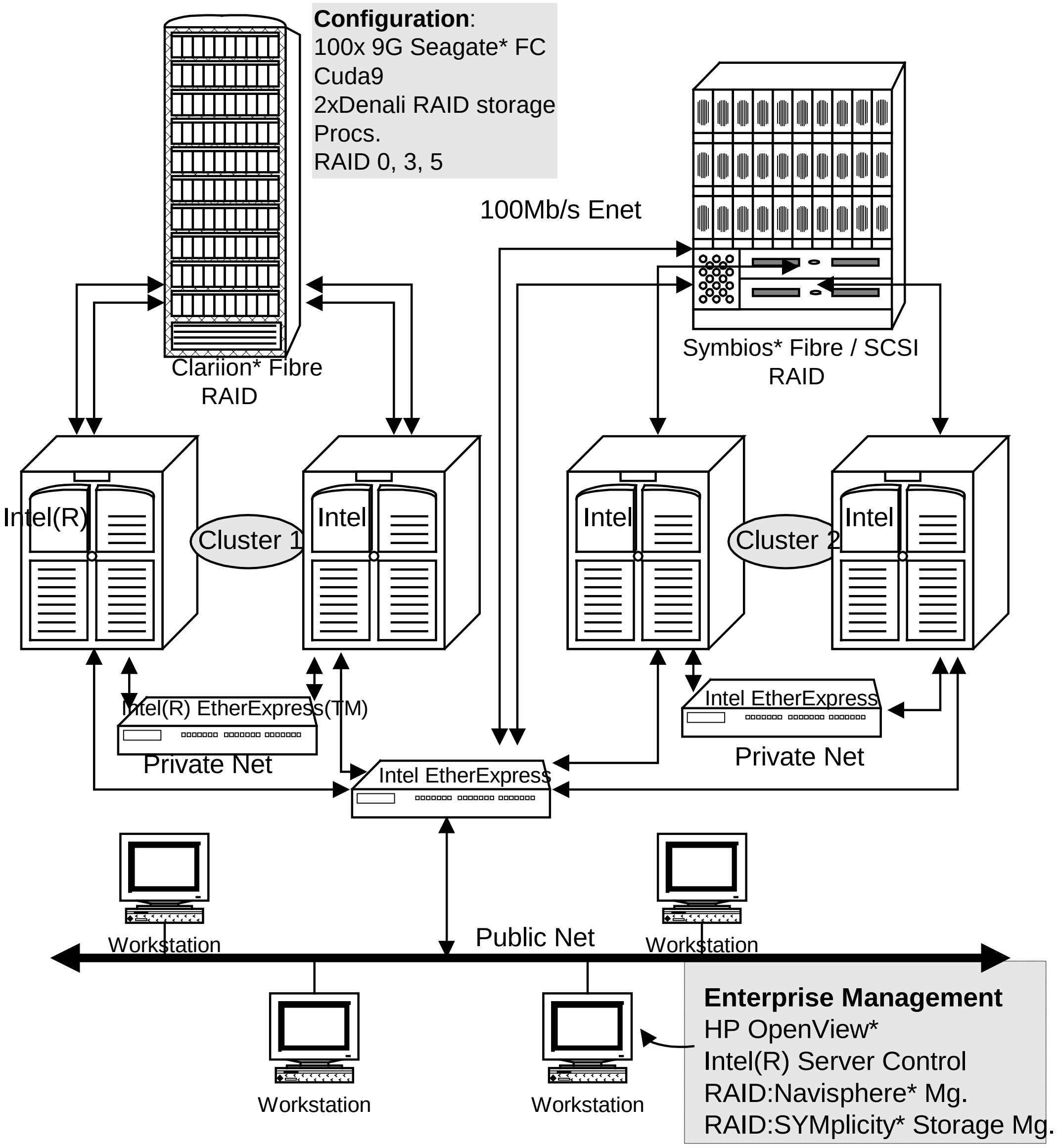

Figure 7: Enterprise Management Solution

2.4 XOR RAID

In SCSI type storage solutions, the only way to create a RAID is to either have the OS handle the striping or have specific RAID hardware that takes control of the drives and provide the RAID functionality. In FC, a third option has emerged and it is called XOR RAID. This is a built-in feature in FC drives that provide the RAID functionality at the device driver and HBA firmware level. The RAID operations are executed by the drives and the adapter BIOS and the OS is not burdened by the complex and CPU-intensive task of providing redundancy for data storage. In our tests using the FibreBox* from Boxhill Systems for FC RAID storage, we found this to be a very reliable and good performing solution. Our test configuration was created using one server and one FibreBox but it is possible and very easy to daisy-chain the storage to scale to multi-terabyte levels. The following table shows the configuration tested. For more information, please refer to http://www.boxhill.com.

Table 5: XOR RAID System Summary

Item	Model	Qty.	Notes
Server: Intel® AC450NX/AD450NX System, four 400-MHz Pentium® II Xeon™ processors, 128 MB RAM	AC450NX/AD450NX	1	System uses onboard SCSI drives to boot to Windows NT 4.0 Enterprise Edition. On AC450NX: Hot-plug PCI slots for "like for like" replacement.
Storage: Boxhill Systems*	FibreBox	1	Built-in XOR RAID support.
Disk: Seagate*	ST19171FC	10	Barracuda* 9G.
Bus adapters: Adaptec*	AHA950	1	64-bit PCI HBA for FC-AL.
Software: OS RAID Management	Windows NT* 4.0 ArrayExplorer*	1 1	Enterprise Boxhill Systems

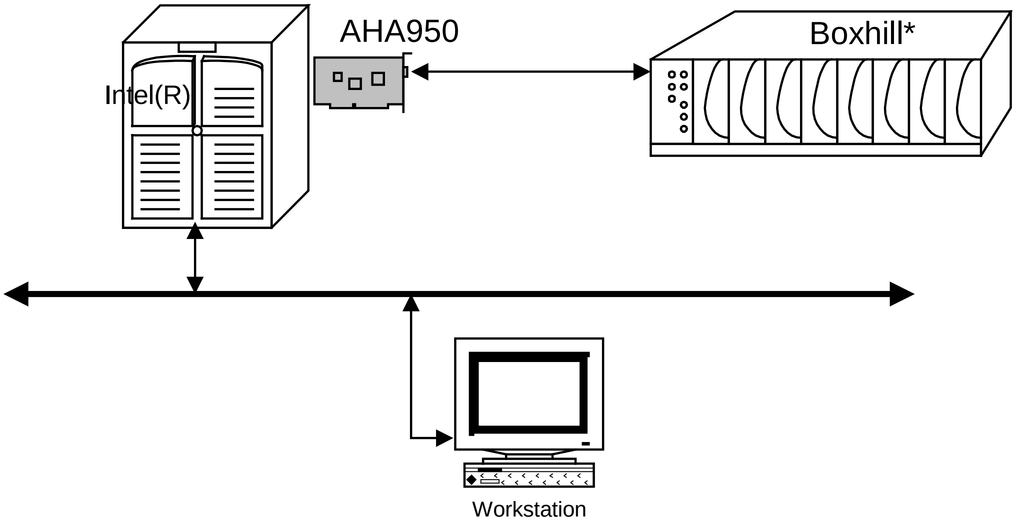

Figure 8: XOR RAID Boxhill Systems*

2.5 MAXstrat* Fibre Storage Solution

Another FC based solution used was the MAXstrat Noble System. This system supports RAID and has UPS built into the storage box.

Table 6: MAXstrat Fibre Storage System Summary*

Item	Model	Qty.	Notes
Server: Intel® AC450NX/AD450NX System, four 400-MHz Pentium® II Xeon™ processors, 128 MB RAM	AC450NX/AD450NX	1	System uses onboard SCSI drives to boot to Windows NT 4.0 Enterprise Edition. On AC450NX: Hot-plug PCI slots for "like for like" replacement.
Storage: MAXstrat Corp.*	Noble	1	Built-in RAID support, UPS, redundant cooling and power.

Item	Model	Qty.	Notes
Disk: Seagate*	ST19171FC	9	Barracuda* 9G.
Bus adapters: Emulex*	LP6000	1	32-bit PCI HBA for FC-AL.
Software: OS	Windows NT* 4.0	1 1	Enterprise.

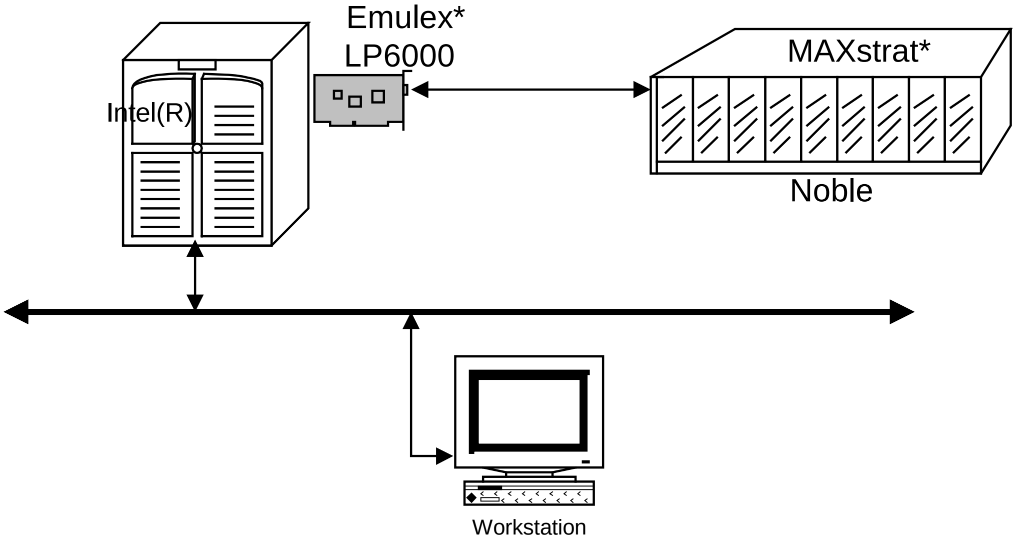

Figure 9: MAXstrat Noble Fibre RAID*

2.6 Industry Solution Examples

There are several industry solution examples that use some of the technology presented in this paper. Among these are:

- IBM Netfinity* 7000 Server. Uses Symbios-Emulex. Refer to http://www.pc.ibm.com/us/netfinity/7000m10.html.
- Dell* Poweredge 6300. Uses Clariion, QLogic. See http://www.dell.com.
- NCR* Worldmark 4400. Uses Clariion, QLogic. See http://www.ncr.com. For a report from http://www.tpc.org see http://www.tpc.org/execsum_TPCC.html.
- Microsoft Terraserver*. This application uses 2x AP450GX (four Intel® Pentium® Pro processors), Clariion, and Emulex to do the highly storage-intensive tasks to create the satellite imagery data base shown at http://terraserver.microsoft.com.

2.7 Future Work and Trends

This section describes a PCI-based RAID for FC and the Storage Area Network (SAN).

2.7.1 PCI RAID

PCI RAID is a familiar concept with respect to SCSI but not a common solution when it comes to a FC interface. All of the solutions developed and referenced in this paper provide the RAID functionality at the storage level (Clariion, Symbios, and Boxhill). The next level of RAID in FC storage is the PCI-based RAID solution which is very similar in concept to SCSI RAID HBAs.

2.7.2 Storage Area Network

Scalability of storage solves one of the problems facing the enterprise level data storage problem. Even though more storage can be added to a server or a cluster of servers, this does not provide a flexible solution, should some storage needs to be diverted from one server to another. Therefore the next step for a scalable storage solution is to create a SAN. The basic idea behind SAN architecture is to share a common storage pool among many servers and dynamically allocate the storage resources among different servers without the need for physical configuration changes, thus avoiding down time. The most important enabling factor for this type of architecture is the connectivity advantage of the FC interface vs. SCSI. The following diagram illustrates the concept behind the SAN topology.

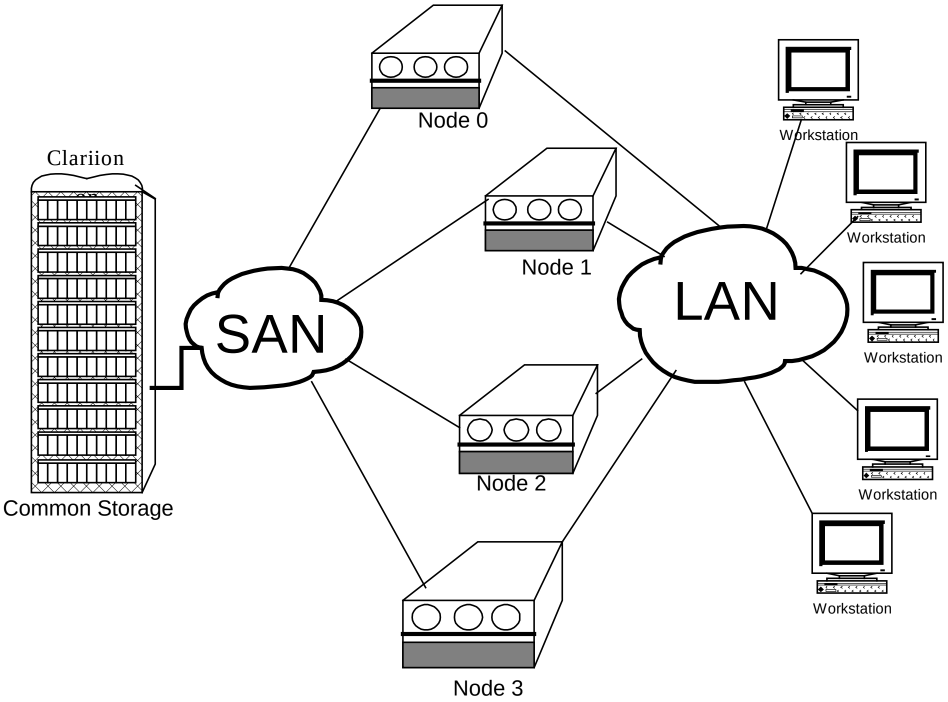

Figure 10: Storage Area Network

The ultimate advantage to this type of solution will be to dynamically add/remove servers or storage components without affecting the applications running on the clients.

3. Conclusion

This short overview demonstrates that highly scalable and available storage-server solutions are possible using Intel Pentium II Xeon processor based servers, and that there are already a number of companies providing such solutions. These solutions are expected to grow through the year 2000 and beyond.