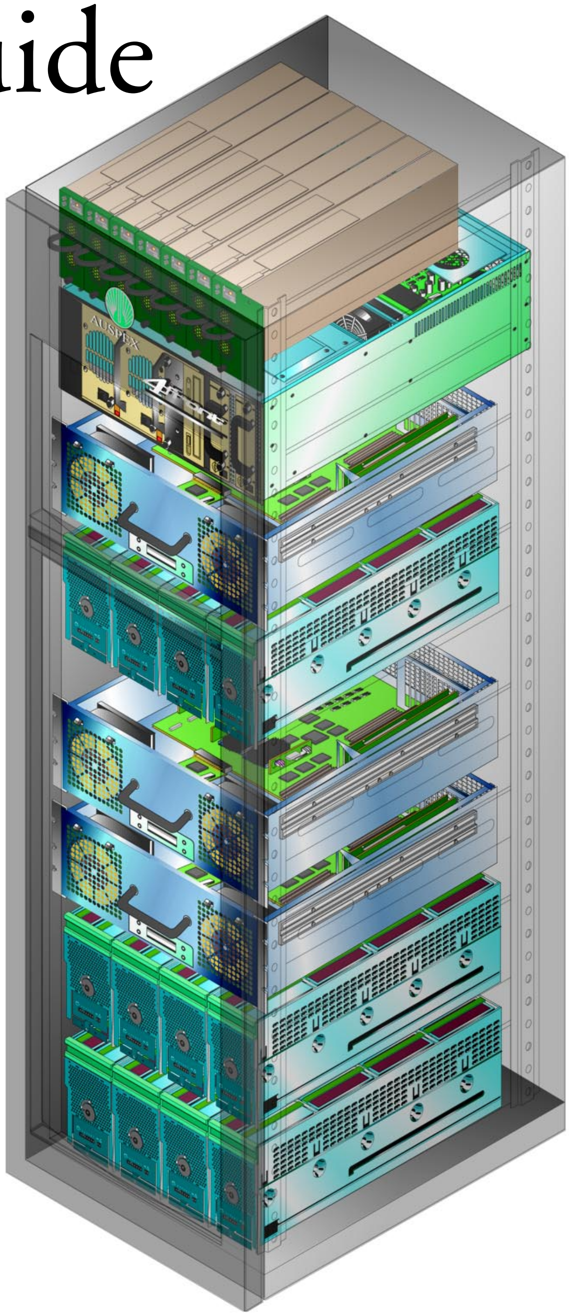
Auspex 4Front™ NS2000

Network Attached Storage for serving data to the network

Product Guide



Auspex Technical Report

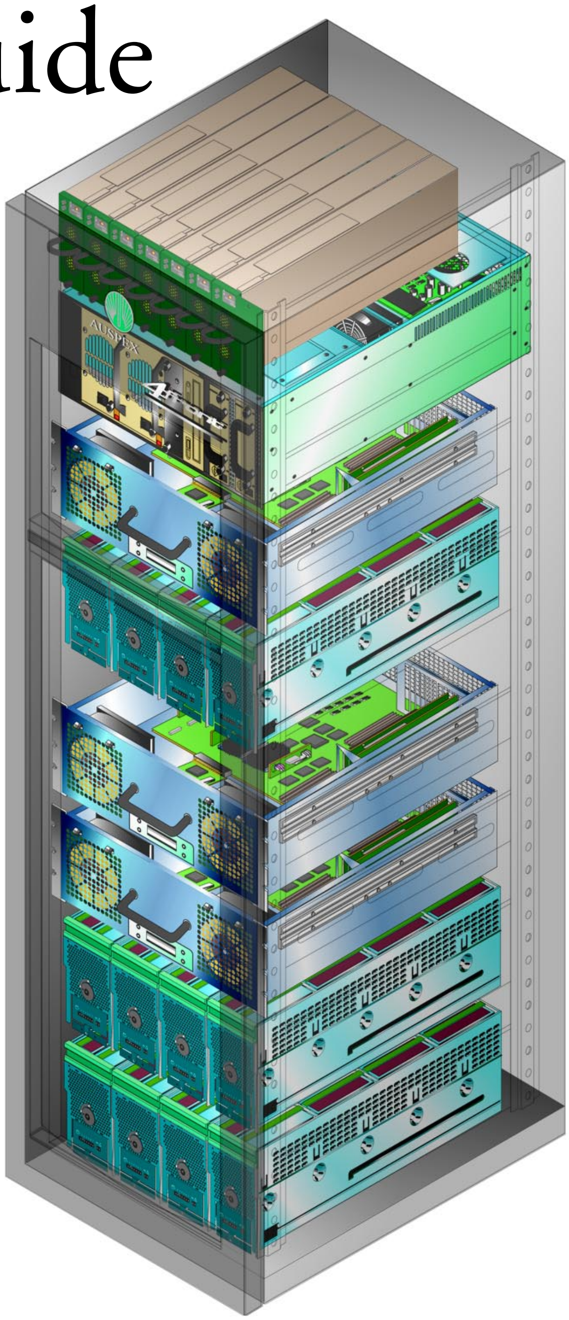
AUSPEX

Content represents the opinion of Auspex Systems Inc. and includes accurate information as far as can be determined at the time of publication. Auspex is not responsible for inaccuracies resulting from changing technologies.

Auspex 4Front™ NS2000

Product Guide

Network Attached Storage for serving data to the network



An NS2000 System Architecture Overview

Network Processors (NPs) of the NS2000 I/O Nodes

The NS2000 FastFLO Journaling File System

NeTservices: UNIX and NT file sharing on the NS2000

NS2000 storage capacities and RAID architecture

Availability purchase consider-ations

About the NS2000 Backup and Restore and Snapshot capabilities

About the NS2000 Network and System Management

Auspex Professional Services and Support

Auspex Technical Report

AUSPEX

Table of Contents

Table of Figures

Tables

Introduction

1

Network Attached Storage (NAS) and its benefits.

Network Attached Storage (NAS) evolved from the networking industry and was first pioneered by Auspex. With NAS there are strong standards for connectivity, data security and load balancing as opposed to Direct Attached Storage (DAS) and the emerging Storage Area Networks (SANs). In NAS servers such as the Auspex 4Front NetServer 2000 Series, the file system resides in the network file server with separate CPU's as opposed to DAS and SAN where the file system competes for resources with the application server's CPU. Although all three storage architectures (DAS, NAS and SAN) are appropriate for an enterprise depending on the application being supported, NAS is the best choice for UNIX and Windows NT data sharing application, consolidated file serving applications, technical and scientific applications, Internet and Intranet applications and certain types of Decision Support (DSS) applications.

Analysts predict major growth for NAS over the next five years due to its major benefits in the areas of consistent availability, performance, scalability, and manageability compared to DAS and SAN, for applications appropriate for NAS. A thorough discussion of NAS, its benefits compared to DAS and SAN, and specific decision criteria for an enterprise to use in making appropriate deployment choices for a particular application can be found in a companion Auspex report titled ***A Storage Architecture Guide***, which should be read first.

Analysts predict major growth for NAS over the next five years due to its major benefits in the areas of consistent availability, performance, scalability, and manageability

The most advanced design available in Network Attached Storage.

The Auspex NS2000 architecture is considered the most advanced design available for the specific task of serving network files with market leading performance, consistent data availability and robust security.

1. It is a ***modern parallel architecture*** introduced in 1999 and is the only choice in parallel hardware and software design among network file serving alternatives.
2. ***Very high availability*** is provided by a robust product design with customers experiencing 99.998% measured availability for systems currently in production. This results in average NetServer unscheduled down time of less than 11 minutes per year. Full environmental monitoring and hot swap capability is also provided.
3. ***Network support*** options include 10/100BaseT Ethernet, FDDI, ATM and Gigabit Ethernet.
4. It provides ***full function data sharing*** between UNIX and Windows and any other NFS or CIFS client.
5. It offers ***scalable performance*** through a multiple CPU architecture, hardware RAID, real time operating system (kernel) design, contiguous file system and superior network and system management tools.
6. ***Backup windows and disaster recovery*** are optimized with the Auspex NDMP Turbo Replicator and efficient parallelism. Data can be replicated in parallel on each of three I/O Nodes. This provides for a total backup performance of 200 GB+ per hour on a fully configured system. Data can also be replicated to remote locations for disaster prevention, using Auspex TurboCopy.
7. A ***full suite of UNIX network and system management tools*** are available in addition to Auspex custom ***Control Point*** software.
8. Auspex offers ***robust factory dial-in capability*** and a 7x24 help desk that provide world class functionality for remote diagnosis of problems when they occur.
9. It offers ***easily scalable capacity and performance*** from the smallest department to the largest enterprise as shown in Figure 1.

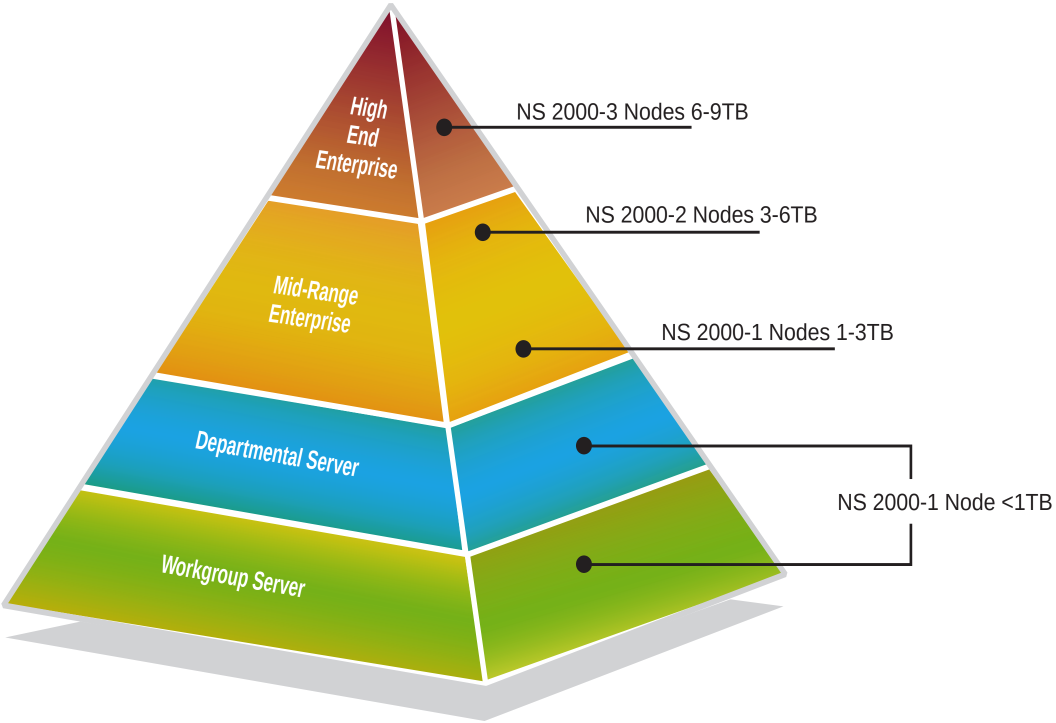


Figure 1 - The Auspex NS2000 product line scales from workgroup to data center.

The highest level of expertise available among NAS vendors.

Auspex is the originator of NAS

Being the originator of NAS, Auspex is widely considered by customers and analysts alike to be the authority in both storage and networking. Since the topic of NAS is new to many customers, Auspex is committed to provide the best public information available on optimizing the flow of accurate information and support on both a pre and post sales basis. Auspex sales and system engineering teams will often recruit additional technical support from Auspex resident specialists, who are experts in each of the areas mentioned in this report. As with any IT architecture decision, probably the most important issue is the selection of a vendor/partner with the best "total" solution. This means not only choosing a vendor who remains at the forefront of technology with the most advanced parallel architecture, but also making sure the vendor can supply the most knowledgeable professional services, consulting services and support personnel.

Other information sources available from Auspex.

For additional information on the storage market and an in-depth overview of DAS, NAS, and SAN, please see the report titled the Storage Architecture Guide that is available from Auspex. This report contains expert guidelines concerning when to use NAS and when to use SAN. It also compares different subsystem architectures available in the NAS market today. This report drills down into the important questions of how to optimally integrate NAS, DAS and SAN and why Auspex is an excellent choice as a NAS solutions provider. The Storage Architecture Guide is an important complimentary document to the Product Information Guide and should be read first. If further information is required, Auspex representatives can provide further specialized presentations, reports and expert knowledge on the topics contained in this series of reports.

An NS2000 System Architecture Overview 2

THE Auspex architecture, as shown in Figure 2 below, is the only NAS product that is based on a parallel processing architecture.

The FMP architecture is patented by Auspex

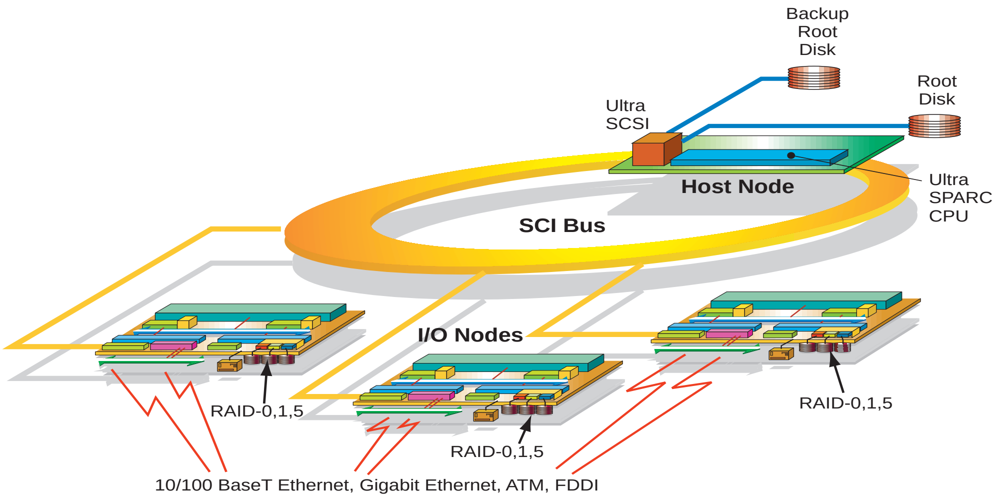


Figure 2 - System block diagram of the Auspex 4Front NS2000 FMP architecture.

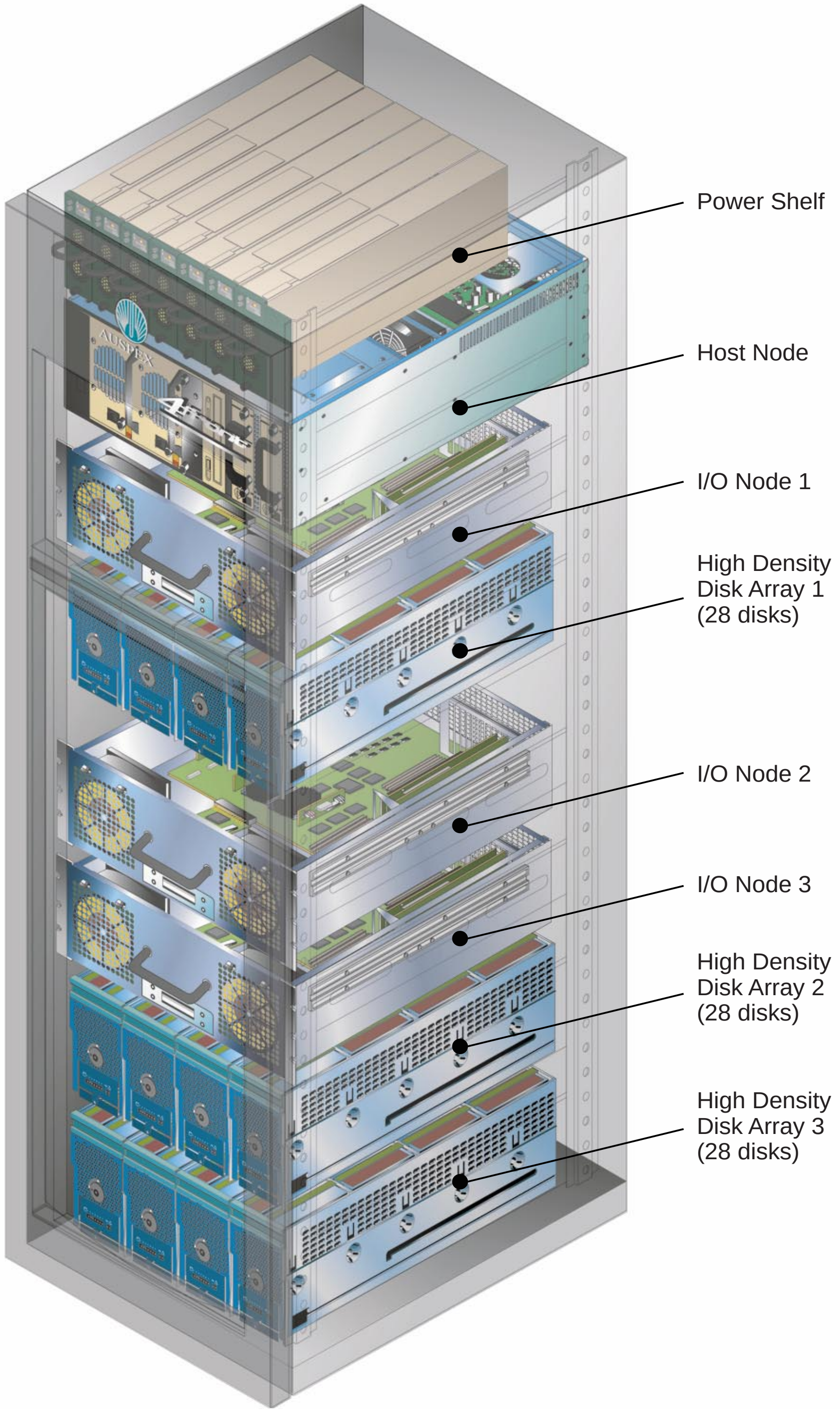
The NS2000 architecture is known as the Functional Multi Processing or FMP architecture and has been patented by Auspex. FMP is based on a "building block " concept that is highly scaleable and easy to expand. This design integrates network storage, and file system processing into a complementary architecture where tasks are assigned to specific processors thereby significantly increasing both system performance and throughput.

About NS2000 packaging.

The NS2000 is available in both one and two cabinet models. A "Host Node" serves as the control center of the NS2000. It manages the intelligent "I/O Node" modules that are connected via a high speed link based on the SCI (Scaleable Coherent Interface) standard that was designed especially for parallel processing. The system supports up to three I/O Nodes. A fully configured system is contained in two cabinets 77 inches high, 51 inches wide and 39.5 inches deep. A single cabinet system is optimized for performance and is shown in Figure 3.

The NS2000 offers one and two cabinet models

Figure 3 - A single cabinet NS2000 system.


Power Shelf
Host Node
I/O Node 1
High Density Disk Array 1 (28 disks)
I/O Node 2
I/O Node 3
High Density Disk Array 2 (28 disks)
High Density Disk Array 3 (28 disks)

This design allows for efficient centralized and consolidated storage resulting in reduced management costs, enhanced network performance, increased data availability, and lower overall total cost of ownership.

Hot-swap and N+1 system power.

Hot-swap and N+1 means non-stop file services

The NS2000 power subsystem consists of one power shelf or Power Distribution Unit (PDU) in both cabinet models. The PDU can contain from three to seven bulk 48-volt power supply modules that are hot-pluggable and hot swappable. This means they can be removed or installed during system operation. PDU's are 48v power modules that are N+1 redundant. Power supplies are added as components are added to maintain redundancy. This means that if you have an extra power supply model in the shelf (i.e., the other power supplies are capable of handling the load), then a power supply can be dynamically removed without creating a power fault. The seven power supplies of a PDU are shown in Figure 4.

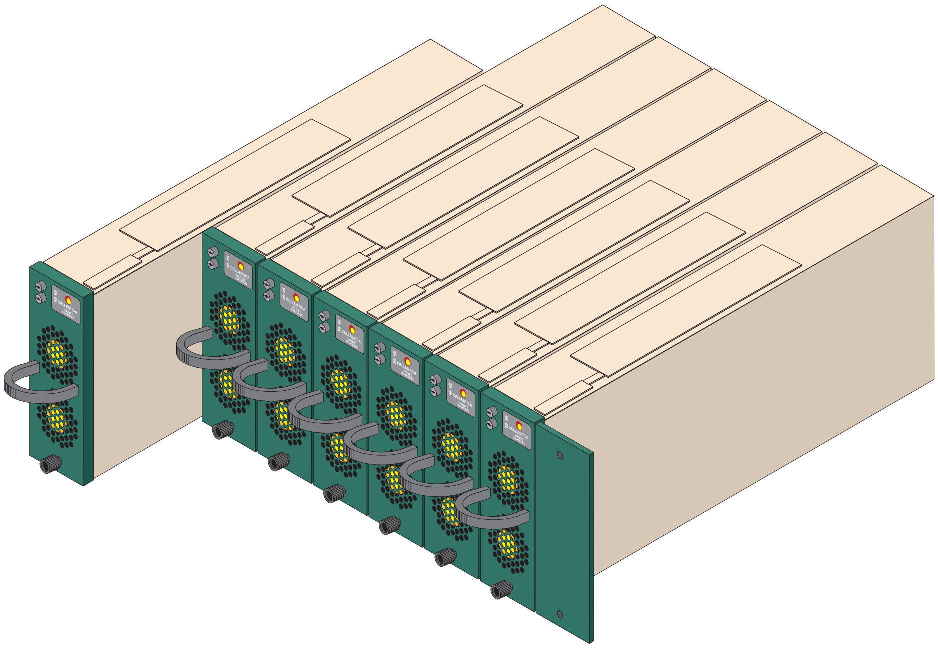
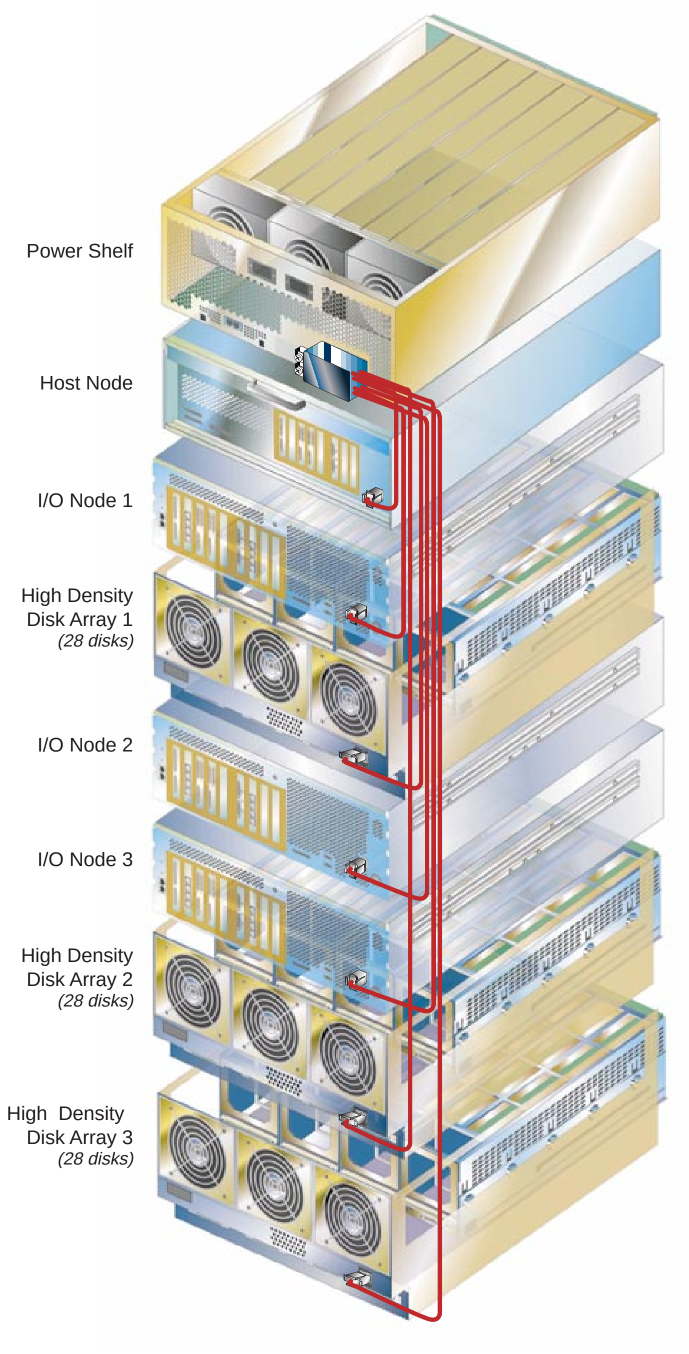


Figure 4 - The seven power supplies of a cabinet model NS2000 power shelf or Power Distribution Unit (PDU)

Cables carry 48-volt main system power from the back of the PDU to each chassis in the one or two cabinet models where the power is converted to the required voltages as shown in Figure 5.

Figure 5 - Main system (48volt) power from the PDU to each chassis in a single cabinet model.



Different AC power cord options allow all NS2000 models to connect with the different electrical outlet configurations used around the world.

The Environmental Monitoring Network (EM-Net) for system status.

The NS2000 is designed with an environmental monitoring network (EM-Net) which is cabled in a logical ring as shown in Figure 6. The EM-Net is connected to all chassis in the NS2000 system and monitors a variety of system parameters (e.g. fan speed, power, voltages, and temperature) and insures trouble free operation. EM-net not only provides for efficient predictive maintenance (replacing a component before it fails) but the failure alerts in EM-Net can expedite system repair after a component has failed. EM-Net information is passed to the Host Node. Alerts be viewed on a UNIX console attached to the network, or directly on the Host Node console.

EM-Net provides for effective preventative maintenance

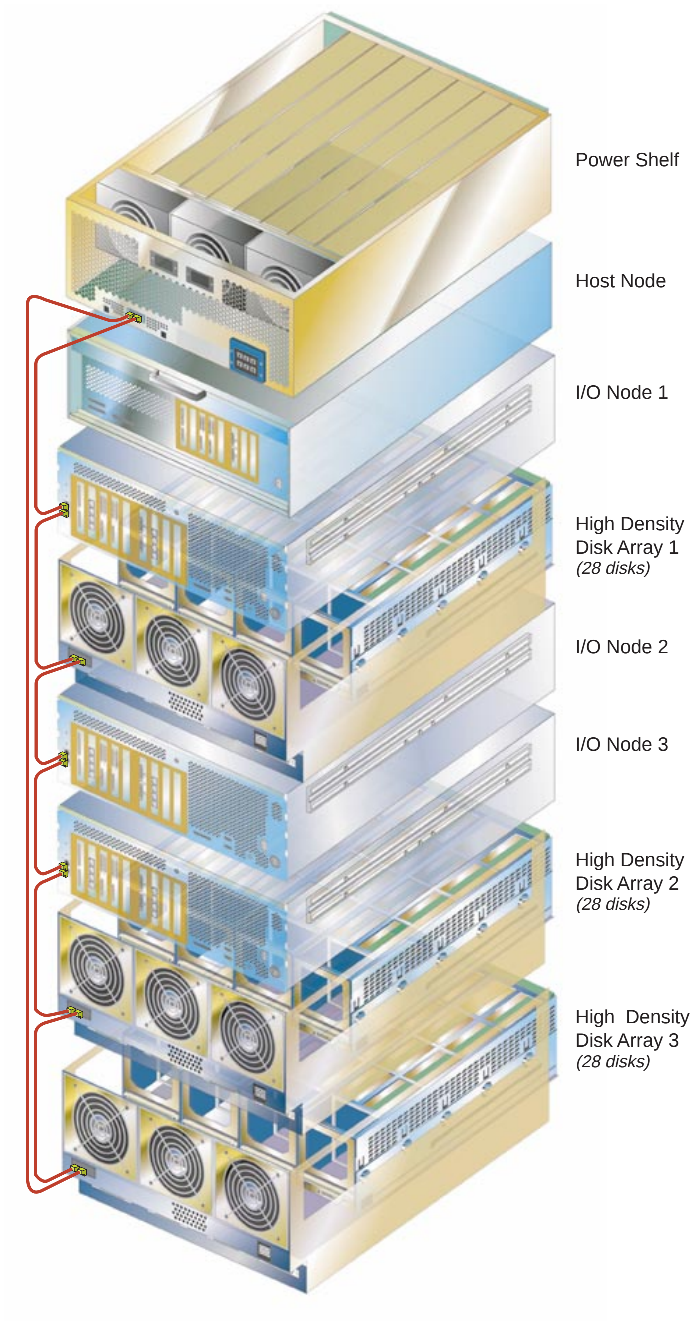


Figure 6 - Environmental monitoring cables for a single cabinet NS2000 system.

Work is divided among seven processors

The SCI bus - an efficient internal messaging network.

A highly efficient Scaleable Coherent Interface (SCI) interconnect allows a single host to manage up to three I/O Nodes. This is accomplished by the SCI bus that is a high-speed network for message exchange between computer nodes of the Auspex system. The NS2000 implements a unique and proprietary message passing protocol between I/O Nodes and the Host Node. It enables cross-mounts between I/O Nodes so that data on any I/O Node is available to any client on any network connection on the NS2000. The SCI bus is an industry standard and was specifically designed for parallel processing computer architectures.

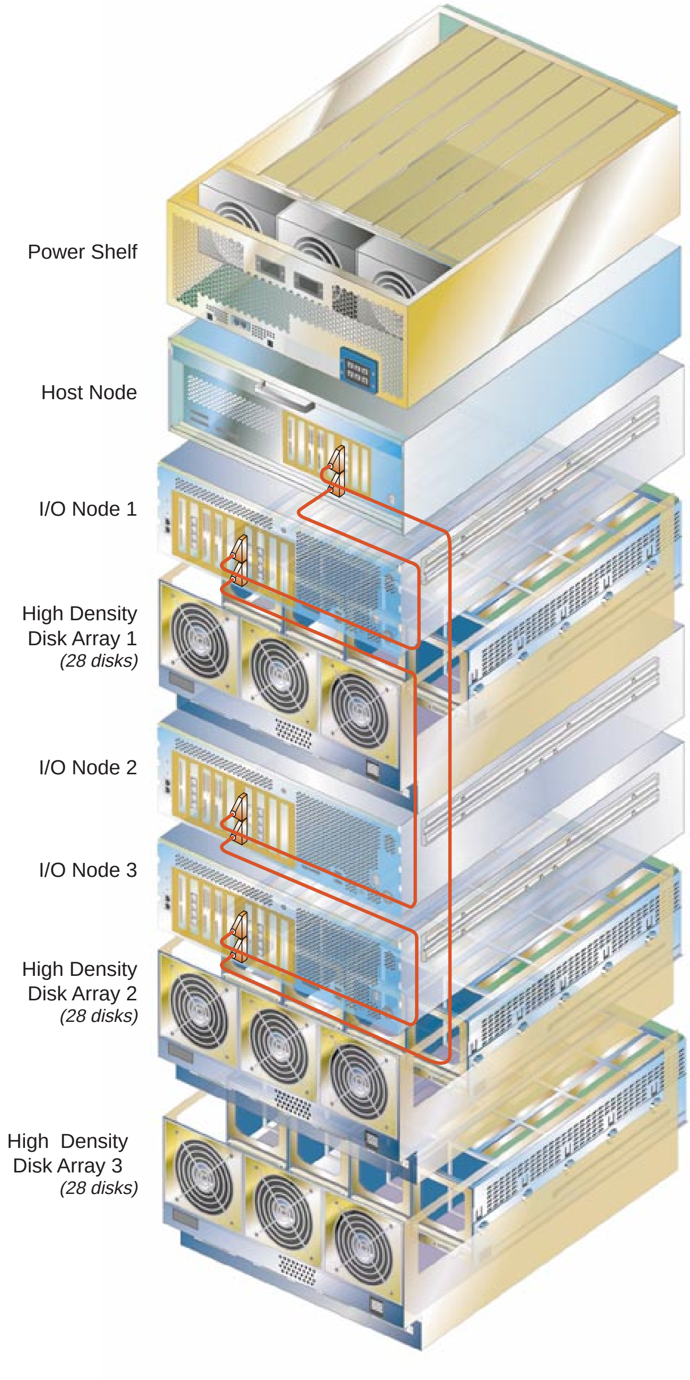

Power Shelf
Host Node
I/O Node 1
High Density Disk Array 1 (28 disks)
I/O Node 2
I/O Node 3
High Density Disk Array 2 (28 disks)
High Density Disk Array 3 (28 disks)

Figure 7 - SCI cable connections for a single cabinet NS2000 system.

Disk, tape, network connectivity and scalability.

Each I/O Node on the NS2000 has five PCI slots for disk, tape, or network interfaces. One slot is reserved for tape, one for disk drives, and one for a high speed NIC (Network Interface Card). The two remaining slots can be optionally used for Network Interface cards or SCSI disk drives. Therefore, a three-node system can have 3 SCSI disk controllers and 9 network interface cards (NICs) or 9 SCSI disk controllers and 3 network interface cards (NICs) or any configuration in between. An individual NIC can be a 4-port 10/100BaseT Ethernet card.

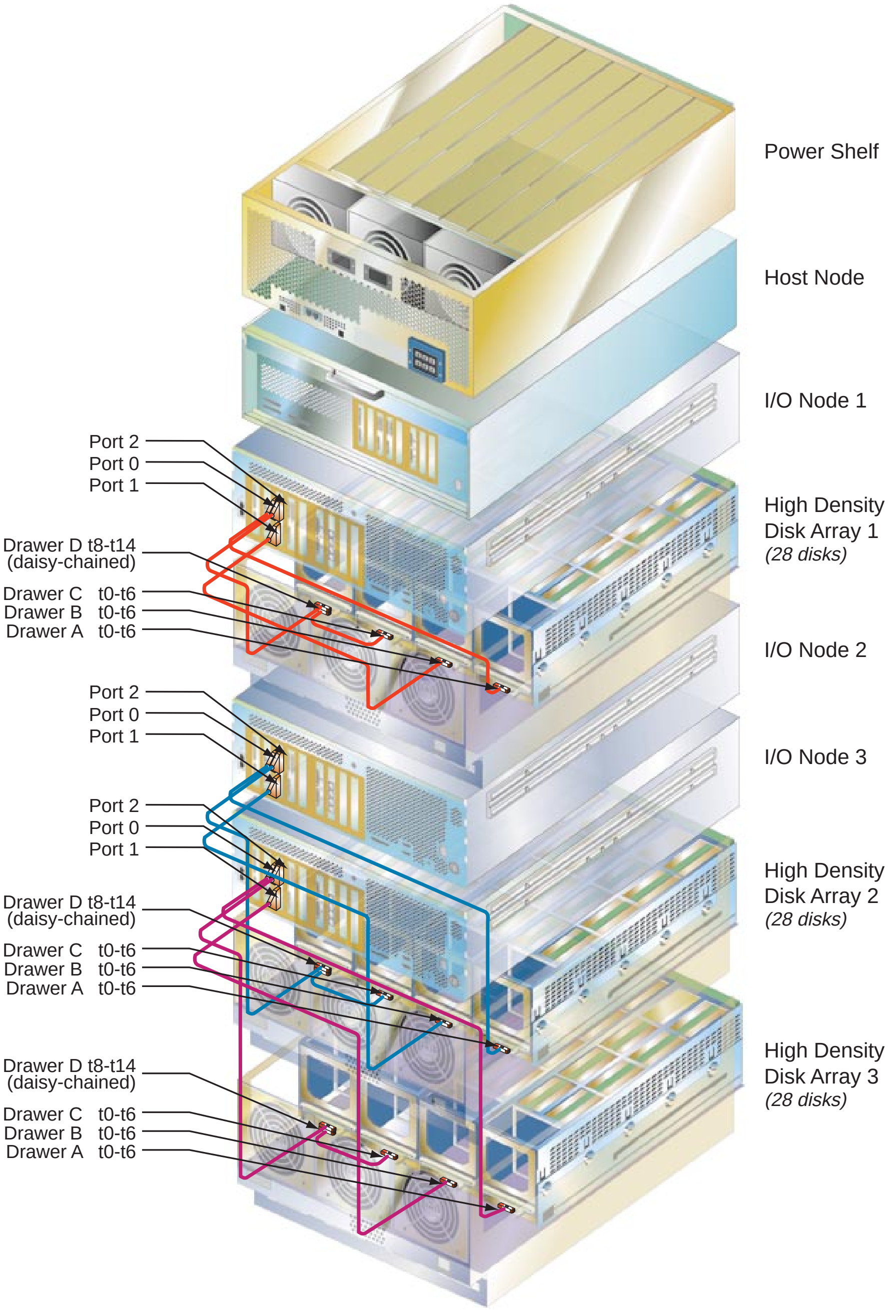
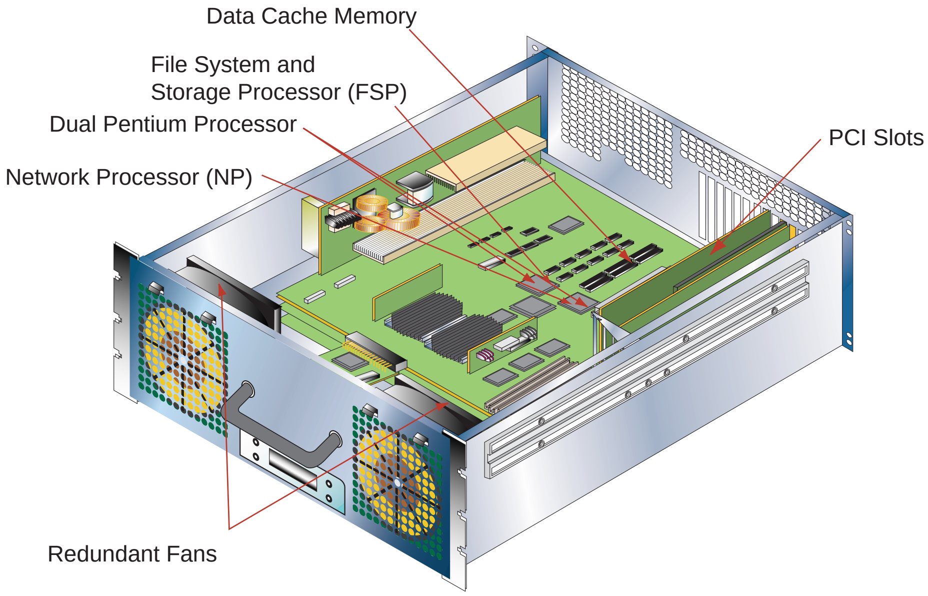


Figure 8 - SCSI cable connections from an NS2000 I/O Node to disk array shelves.

About the I/O Node building blocks

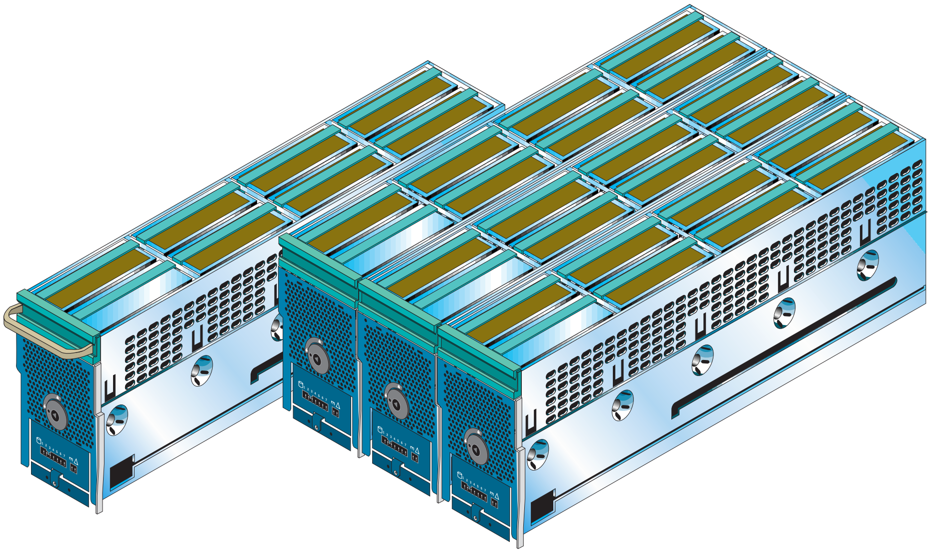
Each I/O Node, as shown in Figure 9, has two Intel Pentium processor

Figure 9 - An NS2000 I/O Node and its major hardware components.



This means a total of seven processors [six for I/O processing and one for host management] for a fully configured system. Each I/O Node can be scaled to three High-density Disk Array (HDDA) shelves with four drawers of disk drives each and a complete three I/O Node system can be scaled to nine HDDAs. Each HDDA drawer contains seven drives. An HDDA therefore contains 28 disk drives (4x7) and three HDDAs (or 3x28=84) drives are supported per IO Node. A complete three I/O Node NS2000 system supports 9 HDDAs and 252 disk drives (9x28=252). A High-density Disk Array (HDDA) shelf with 28 disk drives is shown in Figure 10.

Figure 10 - An NS2000 High-Density Disk Array (HDDA) shelf with 28 drives (4 drawers of 7 drives each).



In the case of 36GB drives this equals 3TB per node or 9TB for a fully configured system. It is convenient to think of an HDDA as 1 TB for 36 GB drives and 500 GB for 18 GB drives.

The internal hardware architecture of each I/O Node is based on dual Pentium processors running on a 64-bit GTL system bus running at 66 MHz or 528 Mbytes/sec. Both processors reside on the bus along with 1GB of DRAM memory that is used for read cache and program files. About 95% or more of the DRAM memory (950Mbytes) is available for read cache with a PCI bridge to two 32 bit 33MHz PCI busses, which act as high speed I/O busses in the Intel-based architecture. The two PCI busses are known as the primary and secondary PCI busses. These busses support the Pentium Network Processor (NP) and the Pentium File and Storage Processor (FSP). Up to three Mylex RAID controllers reside on the (FSP) processor's PCI bus. Each controller contains 64 MB of Non-Volatile Memory (NVM) or 192 MB total per node or 576 MB per three node system. The NVM write cache is battery backed up SRAM that serves as a cache for "fast writes." Writes therefore execute at memory speed for "write back" cache applications. An option exists for "write through" caching where the write goes to disk instead of cache as required by Oracle and other applications. Each node also contains a PCI add-on board with another 128MB of NVM that serves as "file systems" cache for journaling logs and fast restore. A three node system therefore contains 3 x 1 = 3 GB of memory with 95% or 2.85GB available for read cache, 3x192 = 576 MB of write cache and 3 x 128 MB = 384 MB of file system cache.

3.96GB of total cache memory

About the Auspex NS2000 Functional Multiprocessing (FMP) parallel architecture.

Although all hardware is based on the industry-standard Intel high-volume architecture, each processor is assigned a specific function instead of operating symmetrically as in a typical Intel-based SMP system. Each I/O Node consists of industry-standard dual Intel processors, dual PCI busses, associated PCI cards and ECC memory as shown in Figure 11.

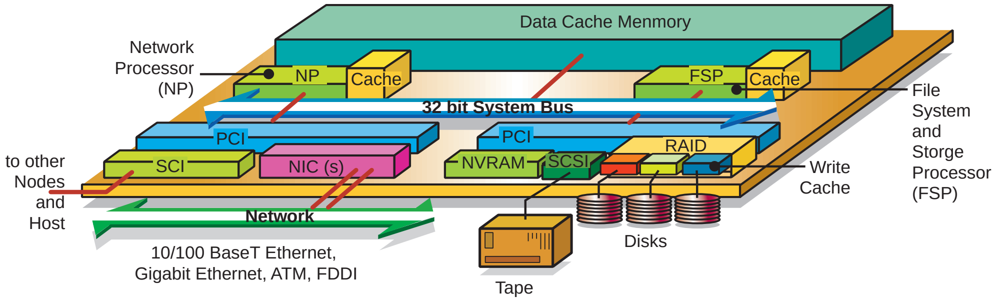


Figure 11 - I/O Node block diagram illustrating the Auspex patented Functional Multiprocessing (FMP) design.

One I/O Node processor is known as the Network Processor [NP] and manages highly reliable proprietary software that controls all network protocol and caching functions. The other I/O Node processor is known as the File and Storage Processor, or FSP, which also executes highly reliable proprietary software that handles file system processing and storage processing. The proprietary software on both processors works closely together and is known as the "DataXpress' kernel.

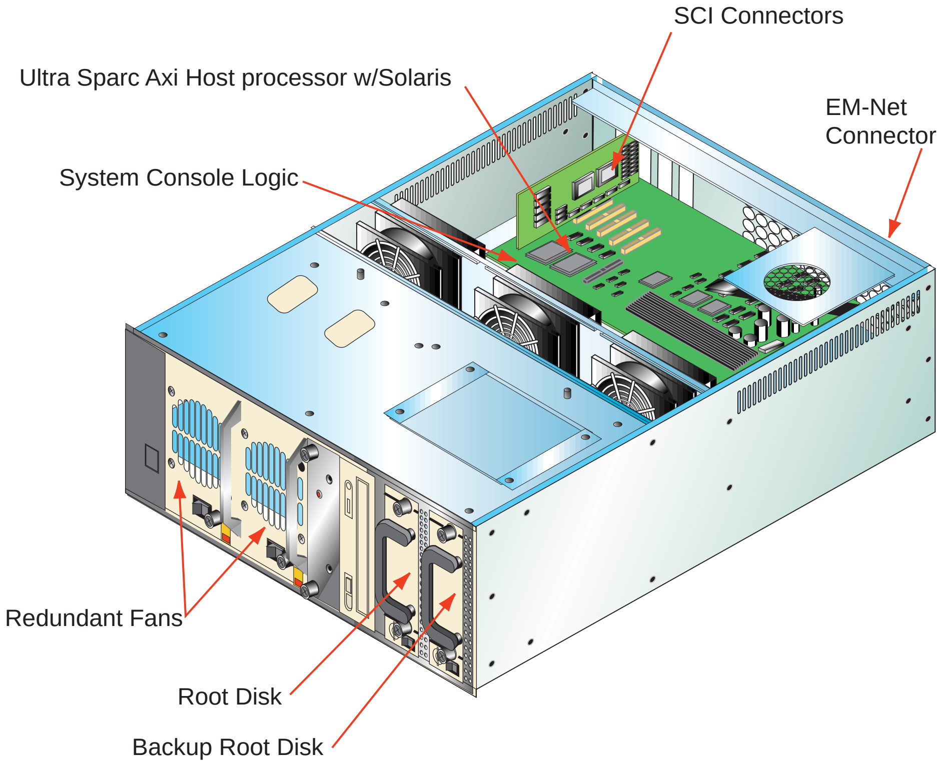
FMP provides functional specificity

FMP system software consists of a unique proprietary messaging system that enables efficient network and storage processing on the I/O Nodes and efficient communication between the I/O Nodes and the Host Node. The Host Node primarily supports system management functions. In the NS2000 architecture each node performs its assigned function in an efficient manner. The NS2000 architecture improves system availability compared to other approaches by isolating the I/O Nodes from unplanned outages of the general purpose Host OS (Solaris). This Premier Software option is called Data Guard and isolates the host from the I/O Node, permitting I/O processing to continue even in the event of a Host Node failure.

About the NS2000 Host Node and system management

The NS2000 Host I/O Node processor runs standard Solaris, which allows all management and control functions typically expected in a data center UNIX environment. This is in addition to Auspex Control Point' proprietary management software that provides NS2000 specific features such as system configuration, monitoring, backup and system control. Control Point is a Java-based GUI program that runs in standard web browsers and allows simple and effective remote control of the Auspex NS2000 from either Windows or UNIX platforms. An NS2000 Host Node is shown in Figure 12.

Figure 12 - An NS2000 Host Node and it's major hardware components.



An excellent discussion of the Auspex Control Point management control software can be referenced from the electronic version of this report (Auspex 4Front NS2000 Product Information Guide) that is available from the Auspex home page at http://www.auspex.com.

Why Auspex chose an FMP over an SMP architecture

FMP is a much better choice than SMP

Symmetric Multiprocessing (SMP) is a type of computer architecture that provides fast performance by making multiple processors available to simultaneously execute multiple software programs. SMP systems are suited for compute-intensive applications. In SMP, any processor executes any program or process. A variety of specialized operating systems are available to support SMP architectures.

In Auspex's patented FMP architecture, each processor executes a predefined set of programs or processes. In a highly predictable environment, like I/O and network processing, this architecture can provide superior performance and scalability characteristics.

Auspex NetServer 2000 links multiple I/O nodes with a host node for data and system management. This further distributes the work to many processors working in parallel. The efficient *Scaleable Coherent Interface* (SCI) interconnect allows the multiple nodes of an Auspex NetServer 2000 system to act as one. This provides a superfast network for message exchange between computer nodes of the Auspex system.

Multiprocessing systems are much more complicated than single-process operating systems because the operating system must allocate resources to competing programs, or processes, in a reasonable manner. The more processes an operating system must support, the more complex the scheduling algorithms necessary to accomplish this—and the more time the processors spend task switching and running scheduling programs to determine what to do next. Also, SMP machine performance degrades more quickly in performance than an FMP design. This is because as an SMP system gets busier, the processors must spend more time in scheduling work and less time performing work as in the FMP design of Auspex NetServer 2000. This makes an 8 processor SMP system more expensive, while yielding only marginally improved performance than a 4 processor SMP for heavy work-loads.

Not only does the NS2000 distribute different functions to the multiple CPUs in each I/O node, but additional functional specificity is achieved by distributing different functions to the host node compared to the I/O node. In addition, the NS2000 design allows multiple I/O nodes to function in parallel. This efficient parallel distribution of work illustrates the advantages of the NS2000 parallel architecture compared to SMP architecture (See **Table 1**).

This NS2000 distribution of work, both within nodes and between nodes, results in higher performance compared to SMP machines—especially at higher I/O work-loads. At the I/O node level, this is due to reduced task switching and reduced time spent running scheduling routines. By taking advantage of parallel processing, Auspex NetServer 2000 avoids the bottlenecks that result from scheduling complexities with heavy work-loads in an SMP environment (see **Table 1**). The Auspex design also provides greater predictability and consistency in file service performance compared to an SMP design. This arises from the greater predictability of the time to complete each program on each node and each processor because of the greatly reduced task switching and scheduling overhead compared to SMP.

Architecture	**NS2000 (FMP)**							**SMP**
Processor Type	**Host**	**Network Processors**			**File System and Storage Processors**			**Host**
# Processors	**One**	**NP1**	**NP2**	**NP3**	**FSP1**	**FSP2**	**FSP3**	**One**
Network Processing	No	Yes	Yes	Yes	No	No	No	Yes
File System Processing	No	No	No	No	Yes	Yes	Yes	Yes
Storage Processing	No	No	No	No	Yes	Yes	Yes	Yes
Management software	Yes	No	No	No	No	No	No	Yes
Peripheral management	Yes	No	No	No	No	No	No	Yes
Complex scheduling	No	No	No	No	No	No	No	Yes

Table 1 - The NS2000 distributes processing functions not only among processors within nodes but also among processors between nodes. SMP computers perform all functions in one node.

Network Processors (NPs) of the NS2000 I/O Nodes. 3

Types of Network Interfaces Supported

The Auspex NS2000 supports 10/100baseT Ethernet, FDDI, ATM OC12 and Gigabit Ethernet. The maximum transmission rates of these interfaces are shown in Table 2.

Table 2 - Transmission rates of supported network interfaces.

Supported Interface	Maximum Transmission Rate
10BaseT Ethernet	10 Mbits/sec
100BaseT Ethernet	100 Mbits/sec
FDDI	100 Mbits/sec
ATM OC12	25Mbits/sec - 2.46Gbits/sec
Gigabit Ethernet	1000 Mbits/sec

Although later to market than ATM, Auspex believes that Gigabit Ethernet will be the preferred network interface compared to ATM because Gigabit Ethernet will run all applications currently running on slower versions of Ethernet without a protocol translation performance penalty as is necessary with ATM. All interfaces are supported however for maximum user choice.

Number of Network Interfaces Supported

Complete network support

Table 3 shows the number of each type of network interfaces supported by the NS2000.

Table 3 - Number of network interface cards (NICs) supported Network protocols supported

Network Interface Adapters	Auspex NS2000 Support
10BaseT Ethernet PCI Adapters	Quad Port (4 ports per NIC)
Max 10BaseT ports per system	36 = (9 adapters x 4 ports)
100BaseT Ethernet = Fast Ethernet	Quad Port (4 ports per NIC)
Max 100BaseT ports per system	36 = (9 adapters x 4 ports)
Gigabit Ethernet	Yes (1 port per NIC)
Max Gigabit Ethernet ports per system	6 = 2/node
ATM OC12	Yes (1 port per NIC)
Max ATM ports per system	3 = 1/node
FDDI	Yes (1 port per NIC)
Max FDDI ports per system	6 = 2/node

Network protocols supported

The Auspex NS2000 supports the Network File System protocol or NFS version 2 (v.2) and version 3 (v.3) over UDP (Universal Data Protocol) or TCP (Transmission Control Protocol) with Internet Protocol (IP) routing. The NS2000 network protocol software architecture is shown in Figure 13

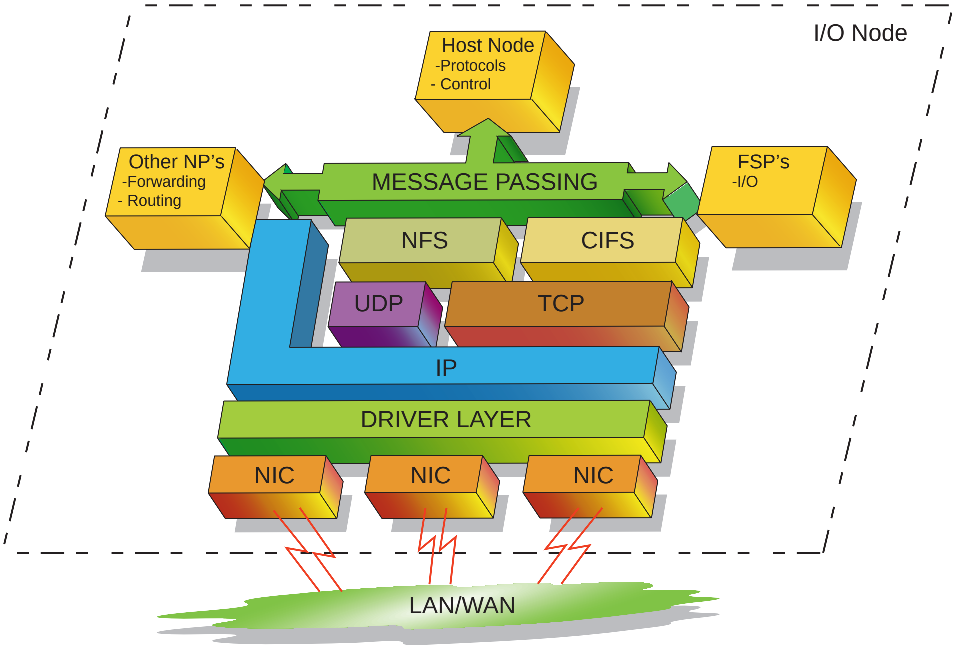


Figure 13 - NS2000 Network Processing (NP) software architecture.

NFS is the standard UNIX protocol for accessing files and printers remotely. In addition the NS2000 supports the Common Internet File System or CIFS which is Microsoft's latest dialect of the Server Message Block (SMB) protocol. CIFS is based on the SMB dialect known as Windows NT LAN Manager (NT LM) version 0.12. These protocols are fully integrated and provide for true UNIX and NT file sharing which is discussed more thoroughly in Chapter 4 on File System Processing and File sharing.

The NS2000 FastFLO Journaling File System

4

The NS2000 FastFLO file system

The FastFLO file system provides local file operations similar to NFS remote operations

At the heart of the Auspex Software Architecture is the patented FastFLO file system. The FastFLO file system is a file system type developed by Auspex and used in the NS2000 for file-system communication between the network processors and the file processor and between the host processor and the file processor. The FastFLO file system provides local file operations similar to NFS remote operations, but without the protocol processing overhead. The Auspex software architecture is shown in Figure 14.

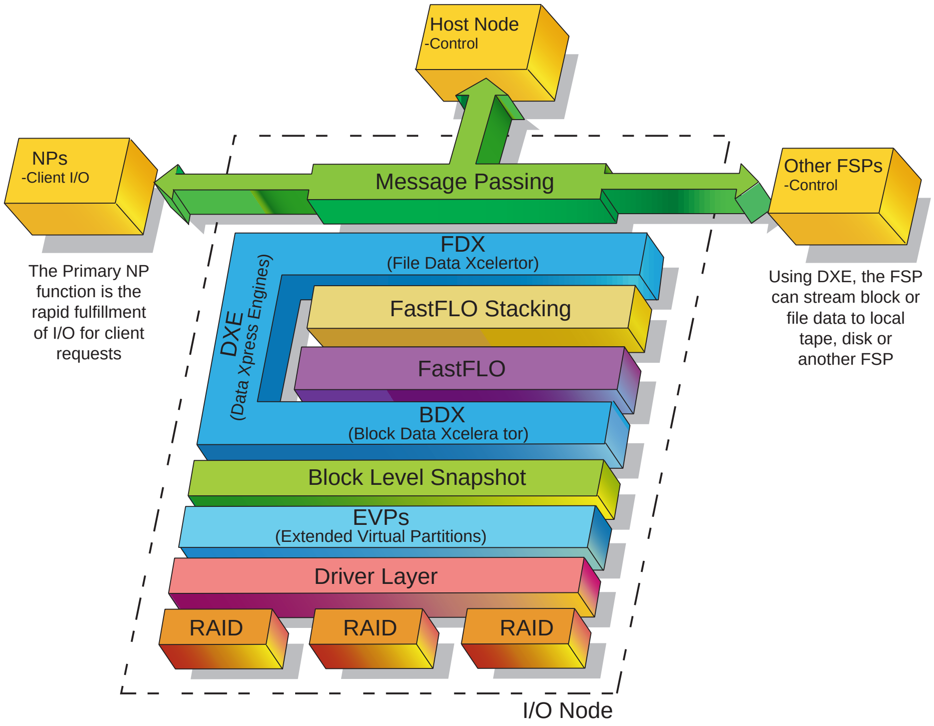


Figure 14 - NS2000 File System and Storage Processing (FSP) software architecture.

File semantics and attributes of both UNIX and NT

Auspex File System Innovations

Although it is often overlooked, the file system plays a critical role in data management. As the hunger for disk storage continues to grow, the file system must be able to scale to handle large file systems, very large files, large numbers of files, single directories with thousands of entries, and the file semantics and attributes of both the UNIX and NT environments. At the same time, the file system must be able to maintain its internal accounting and recover quickly in the face of an unexpected power outage or other failure.

The NS2000 architecture includes the innovative FastFLO File system, which is an integral part of the FSP kernel. FastFLO provides more than double the throughput of the standard UNIX File System (UFS) and includes the following features:

- Journaling for reliability and fast recovery.
- Dramatically fast system reboot performance.
- Checkpointing increases reliability.
- Aggressive write-clustering for high performance.
- Contiguous block allocation for high speed sequential access.
- Read-ahead for high performance sequential reads.
- Support for very large files and large directories.
- Full support for both UNIX and NT semantics and metadata.
- Multiple access method support.
- Online File system expansion.
- Dynamic inode allocation.
- Online file system expansion.
- Stackable architecture for future enhancements.

Journaling for reliability and fast recovery.

FastFLO uses journaling technology to ensure that file system structural integrity is guaranteed at all times. Traditional file systems always maintain a certain amount of important internal file system metadata (information about the files within the file system and about the file system internal organization) in memory to enhance performance. These data structures are periodically flushed to disk to ensure that the file system on disk is consistent. However, if a disruption occurs before the data is flushed, the file system on disk is left in an inconsistent state. This is why checking programs like fsck are required.

Unfortunately, as the administrators of most large systems have discovered, checking programs like fsck can take a prohibitively long time to run on systems with many large file systems. It is not unheard of for a system to take many hours to complete checking all file systems after a system crash. This amount of downtime is obviously unacceptable.

In addition, some of the internal data within the file system must always be written synchronously to ensure that the file system stays consistent. These synchronous operations decrease overall file system performance, since pending operations must wait while the disk (or disk volume) completes the I/O operation.

Journaling file systems like FastFLO avoid these problems by writing a record of each disk transaction to a separate log before any data is written to disk

Journaling file systems like FastFLO avoid these problems by writing a record of each disk transaction to a separate log before any data is written to disk. If a failure occurs, the file system need only examine the contents of the log and verify that transactions in progress have been completed. Incomplete transactions are either completed from information in the log or backed out to return the file system to a consistent state. This typically takes a few seconds versus the hours that checking programs may spend checking all file system data structures for consistency.   Within FastFLO, a transaction daemon processes outstanding disk transactions, groups them together to minimize the number of I/O operations, and commits them to disk. By grouping the transactions together, multiple metadata operations that affect the same disk block can be written together, thus significantly reducing the total amount of I/O that must be performed. The transaction daemon is responsible for recording the data in the log before the FastFLO transactions are committed to disk. On the NS2000, FastFLO uses non-volatile RAM to store its transaction log.This further accelerates file system performance since NVRAM can be written much more rapidly than disk.

Dramatically fast system reboot performance.

The FastFLO file system has a dramatic effect on overall system reboot performance, especially after a system crash or other unexpected outage. Journaling allows the entire system, regardless of configuration, to recover and come back online in a few minutes, just as it would from a clean system boot.It is important to note that FastFLO, like almost all other file systems whether they use journaling or not, will lose transactions that are in memory but have not yet been logged or committed to disk. For network clients using NFS this does not present a problem since NFS writes are either required to be synchronous (NFSv2), or are performed using a safe asynchronous write protocol (NFSv3). Both mechanisms ensure that data is committed to disk before the client recognizes the write as successful. CIFS, however, does not require this guarantee, and thus, some client data could be lost if a system failure occurred at the wrong time.

Checkpointing increases reliability.

FastFLO provides an optional checkpointing mechanism that increases the frequency with which user file data is flushed to disk. This mechanism aggressively flushes data to disk when there are no outstanding transactions or when demand on the file system is low. This improves the reliability of the file system without adversely affecting performance. This mechanism is primarily applicable to file system metadata and CIFS, NFSv3 and data written from the host. As mentioned, NFSv2 requires that all file writes be performed synchronously. (These writes are actually cached in NVRAM on the RAID controllers and don’t have to wait to be written to disk.)

Aggressively flushes data to disk

Aggressive write-clustering for high performance.

FastFLO uses a read and write cluster size of 128 Kilobytes; clustering for reads and writes is performed using a 128KB window. For write operations, when “dirty pages” (pages with updated data) are ready to be written to disk, the file system will cluster all the pages, construct one logical request to the disk driver and issue a write.

Contiguous block allocation for high speed sequential access.

FastFLO uses a delayed block allocation mechanism for allocating disk blocks to files. Unlike UFS where blocks are allocated early, disk blocks in FastFLO are allocated only when the file system is ready to write data to disk. The file system attempts to allocate contiguous disk blocks for all the dirty pages that make up the cluster. This helps to maximize contiguous block allocation. Contiguous block allocation is highly advantageous for applications like Mechanical Design (MCAD) where a relatively small number of large files are read and written by a few clients. Under such conditions, the I/O request stream seen by the server is often highly sequential. By storing file data contiguously read performance can be substantially improved.

Disk blocks in FastFLO are allocated only when the file system is ready to write data to disk

The block allocation mechanism is further tuned for the underlying RAID 5 implementation such that it generates full stripe writes whenever possible. Full stripe writes allow a full stripe of data plus parity information to be written to the RAID array without requiring any data to be read from the stripe. By comparison, partial stripe writes require data to be read from the stripe to generate correct parity.

Read-ahead for high performance sequential reads.

Read clustering in FastFLO is enabled using built-in heuristics to read ahead for every file. The heuristics are applied to determine if the file is being accessed sequentially. If so, then the file system reads ahead pages corresponding to that file. Read-ahead helps to ensure that when a client read request is received the requested data will already be stored in the data cache, so the request can be satisfied immediately.

Support for large files and large directories.

FastFLO allows file systems and individual files to scale up dramatically

FastFLO features full support for large file systems, large files and large directories. All file offsets are stored internally as full 64-bit numbers, allowing file systems and individual files to scale up dramatically. The standard method is a linear search similar to the method used by UFS and works well in most situations, but a different method can be selected (either at mount time or using administrative tools) to accelerate directory access for directories with a very large number of files.

Multiple access methods support.

FastFLO supports multiple access methods for data. An access method is an attribute of a file. Access methods allow an administrator to select the method of storing file data that is optimal for the particular data type. Example access methods include, standard byte-stream I/O, Windows file system compatibility, and record oriented I/O. As the need arises, new access methods can be designed to meet particular needs without necessitating a re-design of the file system. For instance, applications like those used in the Oil and Gas industry perform sequential access to very large files. When handling files that approach or even exceed the size of the data cache, the normal caching algorithms are no longer advantageous.

Dynamic inode allocation

A direct I/O access method would allow such applications to access files directly, bypassing standard caching mechanisms. FastFLO dynamically allocates inodes in the file system, as they are needed. There is no need for the system administrator to manually tune the number of inodes. This allows the file system to easily support a large number of small files such as are typically created by news feeds or a small number of large files, efficiently.

Online file system expansion

Should a file system need to increase in size, expansion can be accomplished online.

Stackable architecture for future enhancements

Perhaps the most unique feature of FastFLO is its stackable architecture. FastFLO incorporates a unique feature called Stacking, which layers on top of FastFLO. Stacking allows FastFLO to be complemented by additional plug-in modules that modify and enhance native FastFLO capabilities. An example of a stacking plug-in module would be the Auspex Snapshot feature.

NeTservices: UNIX and NT file sharing on the NS2000 5

The Importance of True File Sharing

The NS2000 supports both NFS and CIFS network file sharing protocols to allow ***true file sharing*** between UNIX and Windows NT hosts. This is important because true file sharing can often result in a business advantage for the enterprise. For this reason, about 75% of NS2000 users share data between UNIX and Windows NT. The NS2000 file sharing feature offers standardized, reliable and integrated file locking which is a major advantage compared to other approaches discussed in this chapter. For example, in the manufacturing industry a UNIX-based electronic document control system can be accessed by Windows NT clients in the purchasing department to retrieve engineering drawings for vendor quotes.

Four basic approaches for supporting mixed UNIX/NT environmentsFour basic approaches have emerged for supporting mixed UNIX and NT environments.

- Separate UNIX and NT servers accessed by different clients.
- Client-based emulation.
- Server-based emulation.
- Bilingual Network Attached Storage (NAS) such as the Auspex NS2000.

Separate servers make data sharing difficult.

Sharing data is difficult with general purpose file systems

Maintaining separate NT and UNIX servers, with separate sets of clients, is the path of least resistance. However, this approach makes it difficult for UNIX and Windows systems to share files and system management facilities such as backup. Users moving from UNIX to NT are likely to be dissatisfied with performance, availability and reliability.

The biggest disadvantage of the approach is that sharing files between different types of clients may require a difficult-to-manage backup scheme resulting in additional administration cost for administrators to learn. In addition, UNIX system management facilities like backup are available only for UNIX users and NT backup facilities are available only for NT users. If NT runs on typical PC servers, users moving from UNIX to NT may become less productive and less satisfied due to the lower availability and performance of PC servers. This approach avoids the complexities involved in evaluation and implementing the other three approaches for administrative personnel. However, the cost of fragmenting the work environment and lost user productivity is often far greater than the initial investment in a superior architecture.

Client-based emulation uses processing power inefficiently.

Client-based emulation can be an administrator's nightmare

Client-based emulation implements an NFS protocol stack on a PC client or a CIFS protocol stack on a UNIX client. For example, SunSoft's PC-NFS implements an NFS protocol stack on a PC client. This approach has the advantage of requiring no changes at the server by system administration personnel and any problems with the product affect only clients using the emulation software. Technically savvy users can often manage the solution themselves. For this situation or when users only access files on a foreign system on an occasional basis, client-based emulation can be an appropriate solution to mixed UNIX and

NT environments.

However, these products tend to be relatively slow because of the extra work that the client processor must do to emulate the foreign protocol. They can also create stability problems on the client. To allow two-way file sharing, two different products are needed: one to allow Windows clients to access UNIX systems and another to allow UNIX clients to access Windows systems. There is also a major cost associated with client-based emulation that is the administration overhead of keeping all the clients current with new releases or getting the software installed on all new clients. Furthermore, since it is quite difficult to mask fundamental differences in file systems client-emulation is usually imperfect. Perhaps the biggest drawback to client-based emulation occurs when the user cannot solve all installation, configuration and administration problems placing a burden on system administrators. Finally, if there are large numbers of clients, total costs to the organization can be high.

Restrictions of server-based emulation.

Server-based emulation suboptimizes performance

Server-based emulation implements foreign protocol conversion software on a server, for instance, CIFS on a UNIX server or NFS on an NT server. An example this is the Samba suite of freeware components that implements a CIFS protocol stack on a UNIX server. TotalNET Advanced Server (TAS) from Syntax is bundled with Sun's Netra 150 server product and is an example of a commercial server-based emulation product.

Server-based emulation is generally better than client-based emulation in terms of availability, performance and manageability since no special software is required on the client machine. Since servers are usually more powerful than clients, performance tends to be better. Availability tends to be better than client-based emulation since servers tend to be more tightly controlled, monitored and configured. In addition management difficulties that do occur are confined to servers and not spread over an entire client population. Finally, server-based emulation is likely to be better because the product's central location is more strategic.

Server-based emulation products however execute as user-level processes as opposed to running in the UNIX kernel. This is not the most efficient way to tune a protocol on a server and performance is less than kernel based software since it takes far more instructions to accomplish the same amount of work. Like client-based emulation, server-based emulation is usually imperfect due to the difficulties of emulating facilities like file locking and security on a system that has different features and is a fundamentally different machine. It is particularly difficult to support Windows users on UNIX, since Windows is more flexible and offers more options. In essence clients that use emulation are still "second-class citizens" when compared to native clients. Although server-based emulation is a step up from client-based emulation and is appropriate for more users and provides more intensive file access of the foreign system. It's performance limitations make it less than ideal for large numbers of users or even moderate numbers of users with high-intensity application.

Advantages of bilingual Network Attached Storage.

Bilingual file servers are the best choice

Bilingual file servers such as the NS2000 are typically the most appropriate solution for high-intensity mixed UNIX and NT environments. They become more attractive as the amount of data increases and as the number of clients requiring access to both UNIX and Windows files increases. Technically, the bilingual file server is greatly superior to both of the emulation-based approaches. Performance and reliability are likely to be much better, since protocol stacks are part of the kernel, not user-level add-ons. Furthermore there are no "second-class citizens" since a bilingual server treats CIFS and NFS as peers even under heavy concurrent loads on both protocols. Bilingual servers provide the best structure for integrated management of file locking and can make sure that NFS users cannot violate CIFS locks. The bilingual file server also provides "best of breed" administration and management. For example, UNIX backup tools can be used for all files while NT administrators can manage the system using standard administrative tools. Because of the scalability, manageability, and reliability of this approach, the total cost of ownership over time is less than other approaches especially for high-intensity applications.

Although acquisition costs may be higher for a bilingual file server such as the NS2000, total cost of ownership is lower over the life of the product due to:

- centralized backup
- higher productivity
- data integrity
- lower administration

NeTservices overview

NeTservices is the premier solution for deploying enterprise-level, shared file services for mixed UNIX and Windows environments. It allows Auspex NS2000 systems to provide data consolidation and file-sharing for such environments without compromising support of native file access performance, NT integration, and NT remote administration capabilities. NeTservices supports an implementation of Microsoft native CIFS file sharing protocol that leverages the proven FMP architecture to provide industry-leading performance and scalability. Furthermore, it delivers the NT 4.0 networking environment including directory services, file security, and remote administration, that is essential for deploying NT in corporate environments. Finally, NeTservices delivers a best-of-breed administrative environment that fully supports remote NT administrative tools for user/group account and file server properties management as well as enterprise-level administration tools for managing disk storage, RAID, and backup/restore. NeTservices was developed to deliver enterprise-class file services allowing data consolidation and file sharing in a mixed UNIX and Windows environment. It provides the following benefits:

NeTservices supports an implementation of Microsoft native CIFS file sharing protocol that leverages the proven FMP architecture

Enterprise-level consolidation for UNIX and Windows data.

With NeTservices, both UNIX and Windows data can be managed on the same Auspex server, reducing costs, and simplifying management. Further, customers can now obtain very high levels of data availability for both UNIX and Windows data by consolidating server attached storage to the NS2000 which offers less than 15 minutes of unscheduled downtime per year.

Secure, flexible file sharing among UNIX and Windows users.

NeTservices allows customers to manage only one physical copy of the shared data. It provides features allowing transparent, yet secure sharing of individual files by UNIX and Windows clients. Further, it provides support for mechanisms allowing data protection in situations with concurrent file access by UNIX and Windows clients.

Enterprise-level Support.

NeTservices is an Auspex-developed product. It is sold and supported by Auspex. As such, it will take advantage of Auspex's pre- and post-sales support organizations experienced in supporting enterprise-level file services deployment.

A key requirement for support of enterprise-level shared file services is to concurrently deliver NFS and CIFS file services

NeTservices Performance

In order to support the overall goal of enterprise-level consolidation and sharing of UNIX and Windows data, NeTservices delivers an optimized implementation of the CIFS file services protocol providing very high NT file services performance. A single I/O Node provides significantly greater performance than a 4-way SMP Intel-based NT file server does. The Auspex CIFS implementation delivers near-linear CIFS performance scalability as additional I/O Nodes are added to the server system. A system with multiple I/O Nodes can be expected to provide performance equivalent to multiple NT file servers, all in an easy to manage single system image.

A key requirement for support of enterprise-level shared file services is for the data consolidation and sharing platform to concurrently deliver NFS and CIFS file services with undiminished performance and scalability. The Auspex CIFS implementation allows each I/O Node to simultaneously deliver NFS and CIFS protocols with sustained performance and scalability.

Windows NT Domain Security

The NT domain security model provides single log-on capability for NT networks. An NT Domain is defined as a group of servers running Windows NT Server that share common security policies and user group account databases. Therefore, the Windows NT Domain is the basic unit of security and centralized administration for Windows NT clients and servers in the domain, which in some ways, can be viewed as a single system.

One server running Windows NT Server acts as the Primary Domain Controller (PDC), that maintains the centralized security databases for the domain. Other computers running Windows NT Server in the domain function as Backup Domain Controllers (BDC) and can authenticate logon requests. The PDC or BDC authenticates users of a Windows NT Domain. Changes in security policies are implemented on the PDC and transparently replicated to BDCs. An alternative configuration is to setup an NT server as a stand-alone server that can participate in a domain and share its resources with other nodes on the network. This is typically referred to as a Member Server.

Another key concept in Windows NT Domains is the Trust Relationship. A Trust Relationship is a link between two domains that enables a user with an account in one domain to have access to resources, such as files and directories, in another domain.

NeTservices provides full support for the NT Domain security model

NeTservices provides full support for the NT Domain security model including support for PDC, BDC, and Member Server, mode of operation. NeTservices support of the NT Domain security model also includes the capability to respond to validation requests from users/groups in trusted domains and support of authentication of local and global groups.

Windows NT security

Windows NT uses a set of standard Access Control Lists (ACLs) for granting access to shares, directories, and files. The ACLs offer useful combinations of specific types of access, which are called individual permissions. Individual permissions are somewhat analogous to UNIX permissions. They consist of read (R), Write (W), Execute (X), Delete (D), Change Permissions (P), and Take Ownership (O). UNIX supports three sets of file and directory permissions: owner, group, and world. This is the familiar –rwx-rwx-rwx (read-write-execute) that shows up in the output from

With Windows NT, permissions can be granted to either individual users or to groups. The major difference between Windows NT and UNIX is that in Windows NT each user or group can be granted its own set of permissions for each file and/or directory. This allows a finer degree of access control and therefore greater flexibility. In UNIX assignment of access control is effectively limited to three entities, the owner of the file/directory, the primary group, and the rest of the world.

NeTservices includes full support for both share-level and file/directory-level permissions

NeTservices includes full support for both share-level and file/directory-level permissions. Authorized administrators using Windows Explorer, Server Manager or File Manager GUI tools can accomplish management of such permissions.

NeTservices Administration

NeTservices supports “best-of-breed” administration environment for NT services and data. NT administrators can autonomously administer all NT services, such as file sharing, user account, and file security, running on NS2000 systems. In addition, enterprise-class solutions running on Auspex servers that deliver high-performance and robust storage management for UNIX data can be leveraged for Windows data.

Windows NT administration is based on a client/server model utilizing Windows NT Remote Procedure Call (NT RPC) technology. This is a decentralized model, which dramatically simplifies many of the tasks usually associated with system administration. The Windows NT Server 4.0 administrative environment, which is fully integrated with its Windows-based, graphical user interface (GUI), includes the following tools:

- ***User Manager*** for Domains provides the same function as the UNIX method of manually editing the /etc/passwd and /etc/groups files such as adding, modifying, renaming users/groups, and managing security policies.
- With auditing enabled, ***Event Viewer*** can be used to monitor system events such as such as when a particular user last logged on to the domain.
- ***Server Manager*** is the NT GUI tool for monitoring and managing server properties, such as who is connected to a server, how long they have been connected, and what resources they have open. It can also be used for management functions such as closing open resources, and disconnecting users connected to a share.

NeTservices provides full support for NT 4.0 tools for remote administration, including ***User Manager, Event Viewer,*** and ***Server Manager.*** This support is enabled by the inclusion of a full implementation of NT RPC in NeTservices. Finally, centralized UNIX-based data management, for tasks such as managing disk storage, RAID, and backup, is fully supported for Windows data.

NeTservices provides full support for NT 4.0 tools for remote administration

File Sharing Using NeTservices

Support of transparent sharing of individual files among UNIX and Windows users is a key goal behind the delivery of the NeTservices product. Each file system on a NetServer running NeTservices supports file systems that can be used by NFS, CIFS or both without special configurations. A number of value-added facilities in the locking and file-sharing areas are provided that enable secure and robust high-performance sharing of files among UNIX and Windows users.

File Locking

NeTservices includes support for PC-style mandatory file/record locking that is fully compatible with CIFS file/record locking. The goal behind such mandatory locking functionality is to allow robust access by multiple authorized Windows clients to the same file or record. CIFS locking is the locking mechanism used by Windows clients when accessing file systems. NFS Lock Manager (lockd) will continue to be supported as the advisory locking mechanism used by clients when accessing file systems.

NeTservices provides robust access by multiple authorized Windows clients to the same file or record

Coordinated Locking

The need for sharing individual files by UNIX and Windows users is primarily driven by the emergence of UNIX and Windows versions of specific applications supporting a common, interchangeable file format. File systems that can be shared across UNIX and Windows clients enable file sharing in a robust fashion. NeTservices has been designed to support two alternative means for coordinated, “safe” (that is, file data is fully protected against corruption) access by UNIX and Windows clients to individual files.

1. NeTservices provides a facility that allows CIFS locking to be enforced with respect to NFS access so that Windows clients are fully protected against concurrent NFS accesses to the same file.

NeTservices is fully compatible with high-level application-based locking

2. NeTservices is fully compatible with high-level application-based locking. This refers to applications using Relational Database Management Systems (RDBMS) or shadow files to support locking information. Examples include MCAD applications such as ProEngineer, EDS/Unigraphics and Catia that use Product Data Managers (PDMs), and Frame's use of shadow files to manage locked file access across UNIX and Windows Clients. NeTservices is designed to allow interoperation with such applications.

NeTservices also supports server-based coordination between NFS and CIFS locking to allow coordinated access across UNIX and Windows clients. This is to support safe, coordinated access to data by multi-platform applications using platform-specific locking, such as NFS Lock Manager and CIFS file/record locking (versus applications using built-in locking functionality to provide coordinated access as in alternative 2 above).

File Access Control

UNIX file permissions automatically work with NTFS permissions

Automatic coordination between the UNIX and Windows file access control mechanisms is critical if file sharing across UNIX and Windows clients is to be supported in a flexible and secure fashion.

NeTservices provides mechanisms that allow UNIX file permissions to automatically work with NTFS permissions to allow secure file sharing among UNIX and Windows clients with a minimum of administrative overhead. Specifically, NT ACLs (which are set by using NT Server Manager, File Manager, or Explorer tools) function as the first level of file security for access by Windows users to files. UNIX file/directory permissions are used as the second level of file security for file accesses by Windows users. This is accomplished through a mechanism that allows mapping of NT domain-based user/group accounts to UNIX user accounts. (Note that logon validation for NT users/groups continues to be provided by NT PDC/BDC servers). UNIX permissions continue to be the means for managing file security for UNIX users.

NS2000 storage capacities and RAID architecture.

6

Storage Capacity of the NS2000

Up to 9TB of storage capacity

Each of the three I/O Nodes of the NS2000 manages three shelves of four drawers; each with seven drives per drawer or 3x4x7=84 drives per I/O Node. In the case of 36GB drives this equals 3TB per node or 9TB for a fully configured system. Table 4 shows system capacities and number of disk drives for various NS2000 system configurations.

System Capacities and Number of disk drives for various configurations.

System Configuration	Capacity with 18 GB drives	Capacity with 36 GB drives
One Node	1.5 TB – 84 disks	3TB – 84 disks
Two Nodes	3TB – 168 disks	6TB – 168 disks
Three Nodes	4.5TB – 252 disks	9TB – 252 disks

Table 4 - NS2000 system capacities and number of disk drives at various configurations.

Advantages of the NS2000 RAID Hardware

Hardware RAID is a significant advantage compared to software RAID

Auspex selects best-of-breed intelligent PCI RAID controllers to interface to its disk subsystems. These intelligent controllers provide disk interface and RAID management, offloading these tasks from the FSP CPU. This is a significant advantage compared to software based RAID subsystems that use the central CPU to manage RAID. Each NS2000 controller supports RAID 0 (striping), RAID 1 (mirroring), and RAID 5 (parity RAID). All disks in the system are defined as part of a RAID array of one type or another. Non-volatile RAM on each controller accelerates RAID functions, particularly disk writes.

RAID 0 – Striping.

RAID 0 arrays consist of from 2 to 8 disks. RAID 0 accelerates disk access by striping data across the disk array, thereby spreading I/O evenly across multiple disk spindles. RAID 0 provides the best overall performance but provides no resilience to disk failures.

RAID 1 – Mirroring.

RAID 1 arrays consist of pairs of disks in which each disk is maintained as an exact replica or mirror of the other. This method provides for data redundancy and resilience. It provides two paths to the data and does not incur the write penalty. It is also the most expensive RAID alternative since it requires disk space equivalent to twice the usable capacity.

RAID 5 – Distributed Parity.

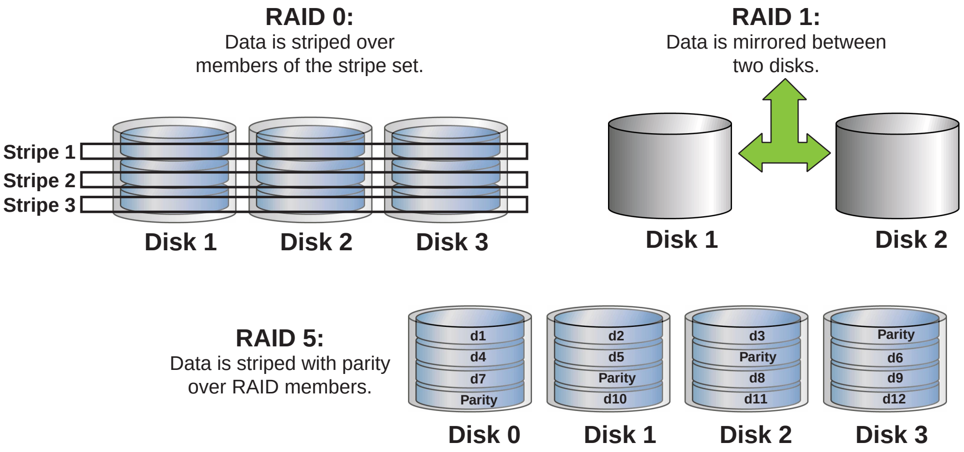
RAID 5 arrays consist of from 3 to 8 disks. Data is striped across RAID 5 arrays in a fashion similar to RAID 0, but RAID 5 provides fault resilience by creating and maintaining parity information on each stripe of data. If a failure occurs, the contents of that block can be recreated by reading back the other blocks in the stripe along with the parity. Parity information is distributed throughout the array to minimize potential bottlenecks. The storage overhead of RAID 5 is equivalent to one disk drive regardless of the size of the array.

RAID Rebuild Capability

No CPU cycles are used to rebuild the array in the event of a drive failure

In the event of a disk failure, RAID 1 or 5 arrays can be rapidly and automatically rebuilt using available “hot-spare” drives. The caching policy may also be specified for each array as either write-back or write-through. With write-back caching, once data is committed to Non-volatile RAM, it is considered complete and written to disk at a later time. With the NS2000 “write-through” feature enabled, data is written to disk while a copy is preserved in controller NVRAM. The RAID controllers allow RAID arrays to be expanded online. Configurations for RAID 0, RAID 1, and RAID 5 are shown in Figure 15. In that the NS2000 uses hardware RAID controllers, no CPU cycles are used to rebuild the array in the event of a drive failure. A drive failure in a software supported system will impact throughput during rebuild. This is not the case with the Auspex NS2000 design.

Figure 15 - RAID 0, RAID 1, and RAID 5 arrays.



High-density storage arrays

The Auspex RAID controllers connect directly to the High-density Disk Array (HDDA) with system capacities as shown in Table 4 above. The HDDA provides a very large amount of storage in a package that has the same form factor as the I/O Node itself. All drives in the array are hot pluggable to ensure rapid, online replacement in the event of a problem.

Extended Virtual Partitions

The RAID arrays of the NS2000 are supported by the Auspex RAID controllers in each I/O Node. The ability to define larger virtual partitions made up of multiple RAID arrays is known as the Extended Virtual Partition Feature or EVPs. EVPs can span multiple RAID controllers within an I/O node to create very large file systems. All EVPs take full advantage of the underlying redundancy of the individual RAID arrays that make up the EVP. Three types of EVPs are supported as shown in Figure 16.

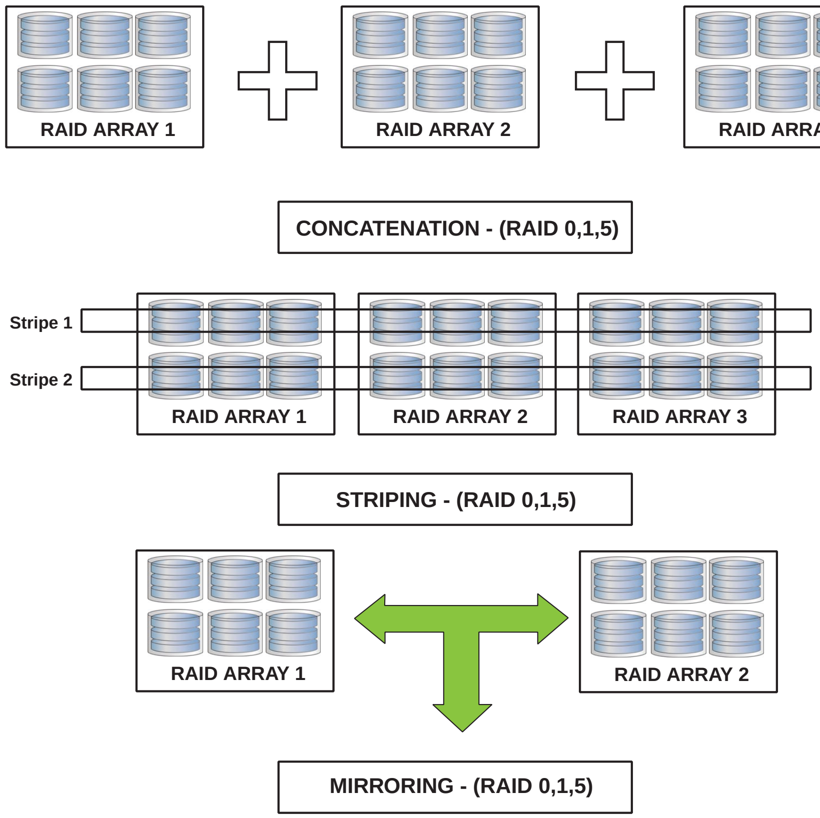


Figure 16 -Three types of Extended Virtual Partitions (EVPs) - Concatenation, Striping and Mirroring

Concatenations

Concatenations join multiple RAID arrays (RAID 0, 1, or 5) together to create larger storage volumes.

Stripes

Stripes are similar to concatenations, but with stripes, the data is spread more evenly across the disks of the multiple RAID arrays, which can be RAID 0, 1, or 5.

Mirrors

Mirrors allow a duplicate copy of all data on one RAID to be duplicated on a second RAID array. When two RAID 0 arrays are mirrored it is known as RAID 0+1. This configuration of striping with mirroring creates outstanding performance and reliability.

Availability purchase considerations. 7

The NS2000 offers a full range of high availability options

From the network to the disk drive, the Auspex NS2000 offers a full range of high availability options across the availability pyramid. This is illustrated in Figure 17.

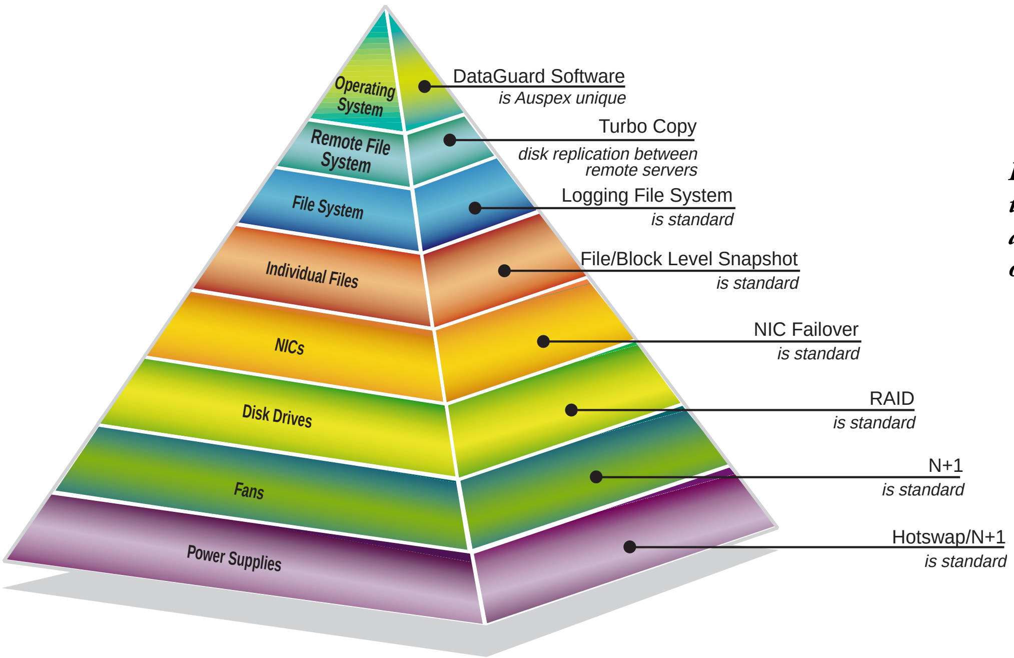


Figure 17 - From the disk to the network, the NS2000 offers a full range of availability options.

Superior availability compared to general-purpose servers.

The availability features of the NS2000 results in much less unplanned downtime per year than for general-purpose computers that are used for network file serving. This difference is shown in Table 5 below.

The NS2000 results in much less unplanned downtime per year than for general-purpose computers

Table 5 - Availability of the NS2000 compared to general purpose computers used as file servers.

Measurement	General Purpose Computers used as File Servers	Auspex NS2000
Annual Availability of Data	99.86%	99.998%
Unplanned downtime per year	12 hours = 720 minutes	<11 minutes

This dramatic difference in part relates to the fact that one of the major causes of system outages when general-purpose computers are used as file servers is the failure of their UNIX or NT operating system. This is opposed to the specifically designed NS2000 that insulates the function of file service from failures in a general-purpose operating system. The NS2000 NetOS operating system (known as a "real time" operating system kernel) involves approximately 10MB of code and is designed to manage only those functions necessary for network file service. General-purpose file servers use the UNIX or NT operating system. These are known as "fat O/S" and can often exceed 3 GB of code. These large operating system software structures fail more often due to their complexity and the impossibility to test for all possible failure modes. The NS2000 kernel is approximately 10MB and has been more thoroughly tested for all possible failure modes. This results in significant advantages for the Auspex NS2000 in terms of performance compared to general-purpose file servers. This advantage is shown in Table 4 above.

Standard Hot Swapped and N+1 Power Supplies

The NS2000 provides full system power in the event of a power supply failure

As discussed in Chapter 2, the NS2000 provides an N+1 design for power supplies and fans to insure uninterrupted power and cooling to all chassis in the system. Each power distribution Unit (PDU) is independently redundant based on its N+1 power supply design. An N+1 power shelf design provides for one redundant power supply on the shelf to provide full system power in the event of a power supply failure. In the event of such a failure the power supply that has failed can be "hot swapped" from the shelf without disruption to the power subsystem.

Standard N+1 Fans for cooling

Each chassis in the NS2000 cabinet or stack models are equipped with standard N+1 cooling fans. The speed and operation of these fans are monitored by the EM-Net as discussed in Chapter 2. The fans utilize an N+1 design, which provides for one redundant fan in the event of a fan failure. This guarantees that a fan can fail without causing a cooling problem for the chassis. A failed cooling fan can therefore be quickly replaced with the remaining fans being sufficient to operate the NS2000 within factory cooling specifications.

Unmatched competitive advantage in RAID choices.

The NS2000 offers the unique capability to intermix RAID levels on a file system by file system basis

The flexibility in choice for NS2000 RAID protection for disk drives is unmatched. It is important to have choices in RAID protection since workloads have different read/write characteristics which influence performance throughput. RAID 1 (mirroring) offers the highest performance (read from either disk in the mirror pair) during normal operation. RAID 1 also offers the highest performance in the event of a failed disk since it is not necessary to read parity (and then data) in the event of a disk failure in other RAID configurations.

The NS2000 offers RAID 0, RAID 1, and RAID 5 and the ***unique capability to intermix RAID levels by file system*** within a controller.

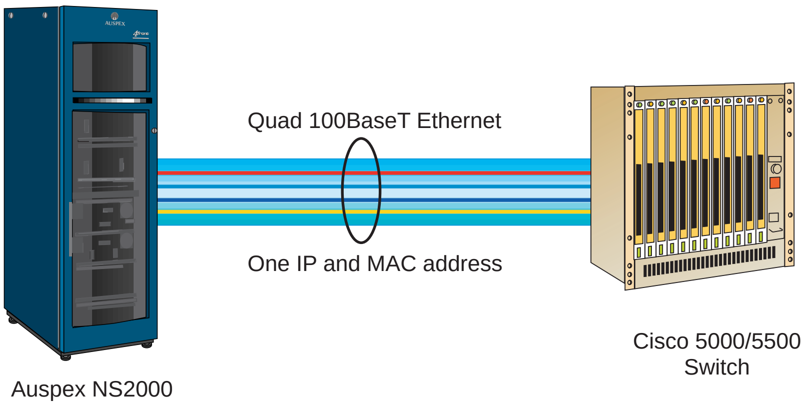


Figure 18 - EtherBand supports both load balancing and NIC failover.

Optional Network Interface Card (NIC) failover

Failover in computer systems is the ability for one component to seamlessly take over the workload of a failed component of a similar type. The NS2000 EtherBand™ feature allows for multiple redundant network connections to provide protection in the event of a failed network adapter card. This feature requires support for Cisco Fast EtherChannel, which not only provides port failover but also balances port workloads to the switch.

Without the EtherBand feature, four separate IP and MAC addresses are needed to balance the load and provide failover capability in the event of a NIC outage. Figure 18 illustrates that with the EtherBand feature, only one IP and MAC address is required for load balancing and NIC failover. Up to 4 EtherChannels per NS2000 I/O Node are supported

NS2000 logical volume level snapshot capability is standard

Snapshots are typically used to provide a consistent unchanging view of the file system for backup. Regularly performed, snapshots can also be used to provide a rudimentary form of file versioning, and also allow for rapid easy recovery of data that is accidentally overwritten or deleted due to user error. In this sense, snapshots are an important data availability characteristic of storage systems.

Device level snapshots occur at the level of the physical disk volume to create a snapshot, all cached data is first flushed to disk and the volume is made momentarily quiescent to ensure that it is stable. Once the snapshot occurs, any time a data block is changed, a copy of the original is made and saved in a separate designated snapshot partition. These snapshots can be exported and mounted by network clients, thereby providing a view of the file system exactly as it was at the time the snapshot was taken. Snapshots can also be backed up using a variety of methods, which will be discussed shortly. Up to 16 simultaneous snapshots are supported for any given volume.

Device level snapshots occur at the level of the physical disk volume to create a snapshot

Figure 19 illustrates that after a snapshot has been taken, and as new data blocks are written to the source volume, the original data blocks are copied and saved to the snapshot volume. To maintain a static view of the snapshot data, the snapshot view accesses unmodified data blocks from the source volume and any modified data blocks from the snapshot volume. Two maps (bitmap and remap) are maintained in FSP memory and on the snapshot volume to provide the necessary bookkeeping information.

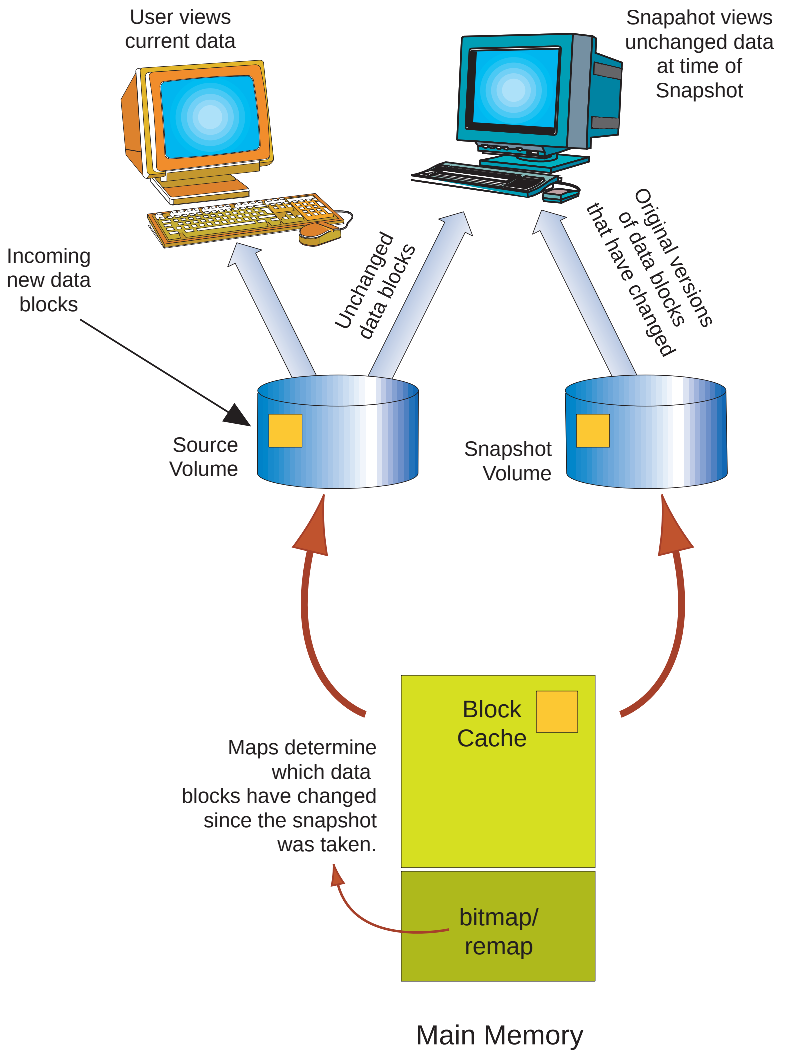


Figure 19 – Maintaining an unchanged snapshot view in the device level snapshot process of the NS2000.

Logging file system is standard

As discussed in Chapter 4 the NS2000 FastFLO file system is a logging file system, which is also very important to overall system availability. A logging file system keeps track of all changes that are made to data from a particular point in time and is therefore important to consistent high data availability. To restore a file system in the event of a system failure, the FastFLO file system restores only the data that has been changed instead of requiring the entire system to be restored. This feature saves time in rebuild performance and improves overall system availability.

Disaster tolerance using optional TurboCopy software.

File systems can be replicated to remote NS2000 servers for disaster recovery planning. The optional NS2000 TurboCopy Feature provides for bi-directional transfers between NS2000 across TCP/IP LAN and WAN network connections. This software runs on top of optimized NDMP compliant Turbo Replicator software and is managed by a Solaris or NS2000 host Node. Remote replication of file and metadata to remote disks is illustrated in Figure 20.

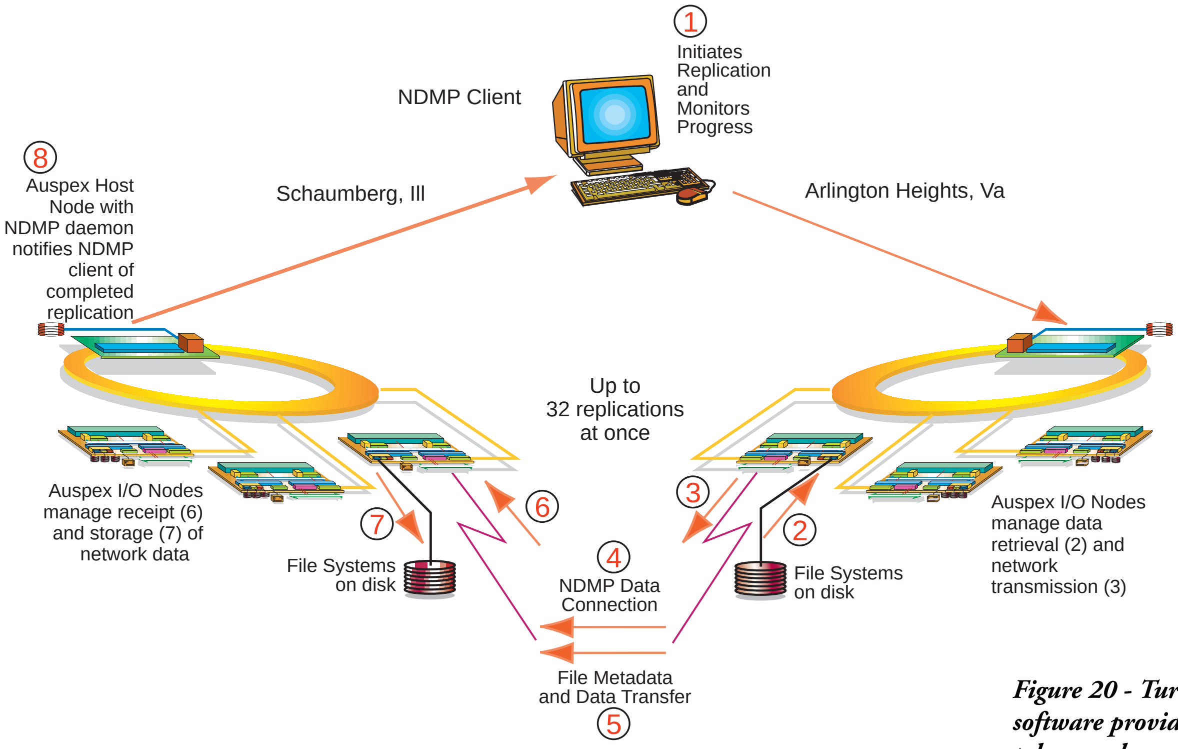

1
Initiates Replication and Monitors Progress
NDMP Client
8
Auspex Host Node with NDMP daemon notifies NDMP client of completed replication
Schaumberg, Ill
Arlington Heights, Va
Up to 32 replications at once
Auspex I/O Nodes manage receipt (6) and storage (7) of network data
7
6
File Systems on disk
3
2
4
NDMP Data Connection
File Systems on disk
Auspex I/O Nodes manage data retrieval (2) and network transmission (3)
File Metadata and Data Transfer
5

Figure 20 - TurboCopy software provides for disaster tolerance by remote file system replication.

Optimized data replication windows using optional TurboCopy software.

TurboCopy can also be used to optimize data replication windows. Data can be replicated in parallel on each of three I/O Nodes for a total data replication performance of 200 GB+ per hour when the feature is enabled.

Data Guard provides higher availability.

Data Guard provides a firewall between the NS2000 I/O Nodes and the Host Node. For this reason a Host Node failure does not interrupt I/O Node operations. Since Solaris runs on the NS2000 Host Node, all UNIX data management and control tools are available in this unique design. The NS2000 therefore provides the best of both worlds i.e., the use of all Solaris management tools and the highest availability for network file services, since the I/O Nodes continue file service in the event of a Solaris failure. Data Guard is the reason that the NS2000 exhibits higher availability and less unplanned downtime than general-purpose computers used as file servers. This is discussed at the beginning of this chapter, and the availability advantage is shown in Table 4.

The NS2000 therefore provides the best of both worlds

About the NS2000 Backup, Restore and Snapshot capabilities.

8

Advantages of the NS2000 parallel backup

Parallel backup provides major advantages in reducing backup windows

Because of its modern parallel architecture, the Auspex NS2000 provides users with simplified scale-up of storage capacity, processors, network connections and performance. In addition, this design provides major advantages in reducing backup windows due to parallel backup of data on each I/O Node. The importance of backup windows is illustrated by the fact that this was the single most important storage concern in a recent survey of 80 enterprises by ITcentrix.

Parallel backup performance of the NS2000

Unsurpassed backup rates of 195 Gbytes/hr for BTE and 186 Gbytes/hr for FTE

In evaluating backup performance in computer systems, the performance bottleneck is always the transfer rate of the tape drives themselves. Auspex has a major advantage over competitive designs such as Network Appliance and EMC in this respect since tape drives can be attached to each node. This allows backup to occur in parallel for an entire NS2000 system. With the NS2000, both block (BTE) and File (FTE) backup performance for a 9TB system can be accomplished at the rate of 195 Gbytes/hr for BTE and 186 Gbytes/hr for FTE with the maximum number of tape drives configured locally to each I/O Node for parallel operation. Although backup speed depends on file size, compression, and the type of RAID applied to the file system, a 2:1 compression ratio is typical. Assuming this compression ratio a maximum configured NS2000 system can be backed up completely in 11.8 hours for RAID 1 and 20 hours for RAID 5 at the block level. Similarly, a complete system can be backed up completely in 12.4 hours for RAID 1 and 21 hours for RAID 5 at the file level. A more normal scenario is to do incremental file backup and only backup changed files. Assuming a 25% hit ratio for damaged data a maximum configured three I/O Node NS2000 system can be incrementally backup up in 3.1 hours for RAID 1 and 5.2 hours for RAID 5. This is shown in Table 6.

Configuration for Backup	RAID 1 4.5TB useable	RAID 5 7.7TB useable
Backup BTE archive = 195 GB/hr	11.8 hrs	20.0 hrs
Level 0 (FTE) = 186 GB/hr	12.4 hrs	21.0 hrs
Incremental FTE assuming 25% hit ratio	3.1 hrs	5.2 hrs

Table 6 - Backup performance to tape drives directly attached to I/O Nodes

Third Party UNIX-based backup tools.

The Auspex NS2000 system supports a range of third party, UNIX-based backup tools, including products from Legato and Veritas. NeTservices allows Windows data to be backed up using these UNIX-based, enterprise-level backup products.

NDMP backup with device and file level acceleration.

Auspex NS2000 system software provides a Network Data Management Protocol (NDMP) server with two data moving engines and interfaces that are designed to dramatically accelerate system backup. NDMP is a standard protocol that can be implemented on any server or backup device. NDMP hides the unique interfaces from third party backup software which allows this software to execute on any NDMP compliant system on the network (such as the NS2000 I/O Node) and control backups on the NS2000 using standard commands.

BDF streams blocks of data from disk to tape

The NS2000 uses snapshot along with the ***Block DataXceleration Engine (BDX)*** which provides the capability to stream blocks of data from disk to tape, and provides the foundation for extremely rapid image backup. NDMP can also use the ***File DataXceleration Engine (FDX)*** which provides a similar interface that acts at the file system level, and provides the basis for rapid file-by-file or directory backup. Data passes directly from disk to tape if the tape is attached to the same I/O Node or from disk on one FSP across the SCI network to tape on another FSP if the tape is attached to another I/O node. These engines in conjunction with NDMP are robust data management tools, which enable high-performance backup and restore of file systems, directories, and individual files along with UNIX and Windows security information, including NT ACLs, SID, etc.

The NS2000 supports NDMP compliant products from Veritas, Legato and other companies

NDMP compliant products from Veritas, Legato and other companies, can be used for high-performance backups of data on Auspex servers. Auspex NS2000 system software also provides UNIX commands that are capable of backup and restore of UNIX and Windows data along with associated security information.

Device level and file level snapshots are supported.

To create a device level snapshot all data is first flushed to disk and the volume is made momentarily quiescent to ensure it is stable. Once a snapshot occurs, any time a data block is changed a copy of the original is made and saved in a separate designated partition. Both types of snapshots preserve both UNIX and Windows data along with associated security information. Users can specify the interval at which these snapshot copies are made. A list of currently certified products is available from Auspex upon request.

About the NS2000 Network and System Management 9

ALL data processing managers realize the importance of configuration, maintenance, and monitoring of all the equipment under their control. Initial configuration, as well as any required re-configuration and maintenance of the NS2000 is accomplished through Auspex Control Point software. This Java applet allows any Web browser, either on the server console or a remote network-attached client, to configure, maintain, and monitor single or multiple NS2000 servers once authenticated through root privleges. In addition to configuration and reconfiguration capacities, Control Point's Performance Monitor can display server performance in real-time as well as capture server performance metrics on an on-going basis, giving system managers the historical information they need to calibrate baseline server performance and plan for server growth.

Monitoring the NS2000 is also accomplished through Control Point's Event Manager. Critical events, such as high levels of network traffic on particular interfaces, which could be a sign of changing network requirements, can cause an alert to be set which will send an E-mail or other notification to appropriate system administrators.

The NS2000 also includes SNMP MIBs for integration with Enterprise Management Systems. This allows for monitoring of the NS2000 to be included with monitoring of other network devices on platforms such as HP OpenView, CA Unicenter, and Tivoli TME.

The Enterprise Management Platform vendors are signing cooperative agreements with systems vendors to provide greater levels of integration such as custom icons, expanded component views, and integration of traps and alerts with their more traditional 'changed color' icon system views.

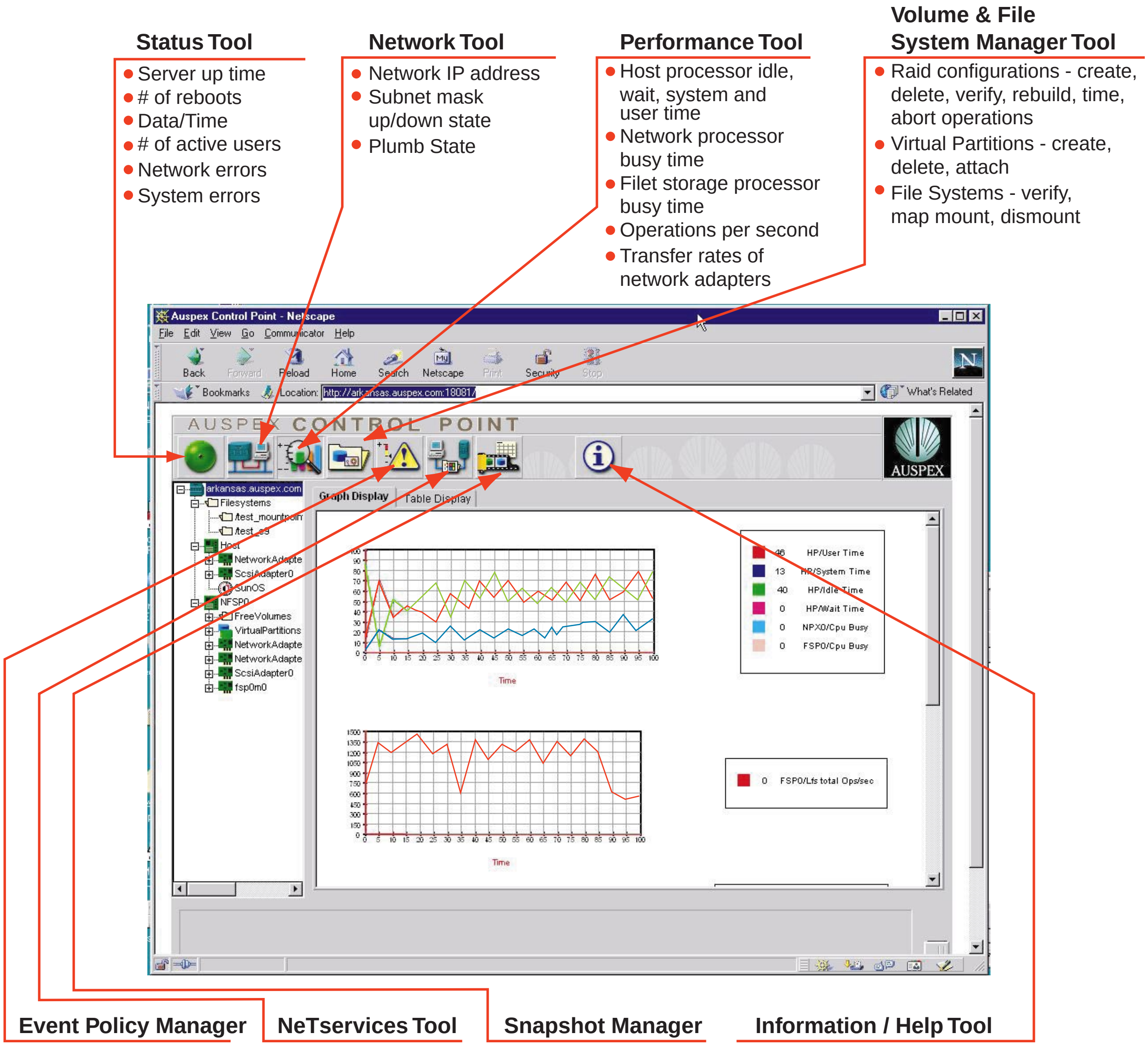
System Status

This selection provides a summary of NS2000 system status and system configuration options. Information includes server uptime, number of reboots, system date and time, number of active users, and information about system and network errors. In addition to this information, system configuration information is also available, detailing specific information about the system's inventory, including information about all the network interfaces and storage options contained in the server, down to serial number and firmware revision level.

Network Interface(s)

This selection reports on the specific configuration information about all the network interfaces contained in the system. Network interface type, IP address, subnet mask, speed, and EtherBand configuration information is all reported. In addition to being the reporting interface for this detailed information, this selection is also used to configure the server's network interfaces. Configuration can be done in real time, such that changes which are made to server network interface configuration are immediately reflected in server operation.

Figure 21 - Auspex Control Point provides a view of logical and physical server components (left panel) as well as rich configuration, maintenance, and monitoring functions.



Status Tool

- Server up time
- # of reboots
- Data/Time
- # of active users
- Network errors
- System errors

Network Tool

- Network IP address
- Subnet mask up/down state
- Plumb State

Performance Tool

- Host processor idle, wait, system and user time
- Network processor busy time
- Filet storage processor busy time
- Operations per second
- Transfer rates of network adapters

Volume & File System Manager Tool

- Raid configurations - create, delete, verify, rebuild, time, abort operations
- Virtual Partitions - create, delete, attach
- File Systems - verify, map mount, dismount

Event Policy Manager

-Emails
-Pop up windows
-Scripts
-Pager

NeTservices Tool

- User name mapping
- Creation of NT shares
- Mount points

Snapshot Manager

- Hourly, daily, weekly
- File system specific
- Repeat integral

Information / Help Tool

- Context sensitive
- Server monitoring
- Server management
- Tool definitions
- Contacting Auspex
- Glossary

Performance Monitor

Control Point's Performance Monitor features two valuable functions for the system administrator. All server performance metrics, such as network interface utilization, host processor utilization, operations per second of file systems, and overall system cache utilization can be displayed in real time in either tabular or graphical format. The Performance Monitor also features a data logging utility which allows a system manager to select metrics to be monitored, frequency of that monitoring, and time to start and stop metric monitoring. These collected statistics are saved on the server in .CSV (Comma-Separated Variable) files for easy import into standard data analysis packages. This ability allows system administrators to gain additional visibility and perspective into the server's operation over time and use this information to plan for future requirements and server upgrades.

Control Point offers robust systems management tools.

Volume and File System Manager

The Control Point Volume and File System Manager is where all details of the server's storage configuration is managed. All RAID configuration options, such as creation, deletion, verification, rebuild, and tuning are supported, as are all Virtual Partition configuration options, which include creation, deletion, and member attachment. File System operations, including mount, dismount, map, and verification are also supported through the Volume and File System Manager.

Event Policy Manager

One of the major requirements of any system management package is alerting the system administrator when certain critical event thresholds are reached. The Control Point Event Policy Manager allows for multiple thresholds (2 low thresholds and 2 high thresholds) to be set for all server performance metrics. When thresholds are crossed multiple event notification actions can be triggered. These actions include an E-mail or a pop-up browser notification, logging of the event to a file, or execution of a Unix script.

NeTservices Manager

The Control Point NeTservices Manager allows for configuration of the Auspex NeTservices optional software. NeTservices allows an Auspex NetServer to serve Windows NT files as well as Unix files and enables true bilingual file sharing in a centralized storage environment. The NeTservices Manager allows a system administrator to map Unix user account names to NT user account names as well as creating NT file share names and mountpoints. For further information about NeTservices, see Chapter 5

Snapshot Manager

The Control Point Snapshot Manager provides for the time scheduling and administration of file system snapshots. Snapshots can be scheduled for any time of the day and can be automatically repeated hourly, daily, or weekly on a per file system basis. For further information about Snapshot, see Chapter 8

Information/Online help

Control Point features context sensitive on-line help. For example, if a system administrator were configuring a network interface and wasn't comfortable with some of the terminology used, clicking the Information button will lauch the administrator into the specific portion of on-line help associated with configuration of network interfaces.

Auspex Professional Services and Support 10

The Auspex commitment to *Solutions.*

Service is often the key difference between a product and a solution. A company's ability to provide the best possible information and support from pre-sale through post-sale creates true customer partnerships and long term relationships. Auspex is committed to delivering the most responsive, comprehensive and cost-effective support and value-added services available. The Auspex service team will assist prospects and customers in realizing the highest return possible on their IT investments and maximizing the level of availability and productivity from NS2000 equipment. Being the primary innovator of NAS, Auspex is widely considered by customers and industry analysts alike to have the highest level of expertise in both storage and networking technology.

Being the primary innovator of NAS, Auspex is widely considered by customers and industry analysts alike to have the highest level of expertise in both storage and networking

Auspex Consulting Services

The Auspex Data Solutions Consulting Group is a knowledge-based consultative practice that identifies, and delivers optimized application infrastructure consulting for continuous data accessibility. The consulting team is comprised of experts in infrastructure planning and simulation.

The team uses proven consulting methodology and field-tested analytical tools including predictive modeling techniques. Working in constant interaction with the client, Auspex consultants look carefully at how data flows through the network and what can be done to enhance that flow. Simulation models test network re-configurations or technology upgrades to determine whether performance benefits will be achievable. This information helps clients directly evaluate productivity-related business benefits such as capacity scaling and time to market cycles, In addition the client is provided with a predictive road map that allows clients to achieve the maximum possible return from infrastructure asset investments.

Auspex offers three levels of Consulting Services a user can choose from depending on the level of service desired. Additionally, a client's return on investment from a typical Auspex consulting engagement is typically measured in months, not years.

The team uses proven consulting methodology and field-tested analytical tools

Auspex Professional Services

Auspex offers a full range of professional services designed to install, bring up, relocate, re-certify, recover from, and avoid any problems involved with a customer's Auspex technology investment. The Auspex professional services offering also includes a comprehensive Business Protection Services (BPS) that evaluates and provides a variety of client-specific environment services such as power conditioning and cabling for UPS.

Additionally the company can provide both technical and administrative staff with expertise in a variety of IT operational areas. Auspex can allow this to be accomplished on an outsourced basis. Such an outsourcing program can allow clients to save significant time and money compared to continually developing and training full time staff.

The Auspex professional services offering also includes a comprehensive Business Protection Services (BPS)

Basic NS2000 Warranty

The Basic NS2000 Warranty includes 24x7 telephone NetOS support for 90 days and telephone, parts and labor support for hardware from 8AM to 5PM Monday through Friday, with a next-business-day, on-site response for 1 year. In addition to a solid basic warranty, Auspex offers additional levels of warranty and support services.

Premier NS2000 Warranty

The Premier NS2000 Warranty extends NetOS telephone support to 1 year and enhances hardware support to 24x7 and shortens the on-site response time to 4 hours from time of problem diagnosis.

Auspex Premier Software Option

The Auspex Premier Software Option delivers trusted large scale support 24 hours a day, seven days a week, in mission-critical environments that require continuous access to crucial data. It is available as a 90-day warranty, a 9 month warranty or a 1 year warranty.

The Auspex Competitive Advantage

Robust on-board diagnostics combines the UNIX command set and extensive NetOS capabilities

Auspex offers robust factory dial-in capability that provides world class ability to experts to diagnose problems if they occur. This significantly shortens the time for problem diagnosis and resolution. In addition the NS2000 comes with robust on-board diagnostics. This combined suite of tools use the UNIX command set and extensive NetOS capabilities. Remote capability allows the SE to run and view the performance monitor "perfmon," assess and adjust network configurations, and assess and adjust file system configurations. This provides a "Virtual SE" with each machine, with strict customer controlled access.

Most significantly Auspex has many more years of experience in solving network problems. Auspex Customer Engineers (ACE) handle both pre and post sales work thereby providing for continuity in the installation, training and management of NS2000 systems.

Glossary of Terms

A

10BaseT

Ethernet with a data transfer rate of 10 Mbits/sec.

100BaseT

Also known as Fast Ethernet with a data transfer rate of 100 Mbits/sec.

ACL

Access Control List. Windows NT uses a set of standard Access Control Lists (ACLs) for granting access to shares directories, and files. The ACLs offer useful combinations of specific types of access, which are called individual permissions. Individual permissions are somewhat analogous to UNIX permissions. They consist of read (R), Write (W), Execute (X), Delete (D), Change Permissions (P), and Take Ownership (O). UNIX supports three sets of file and directory permissions: owner, group, and world. This is the familiar -rwx-rwx-rwx that shows up in the output from the UNIX ls-al command.

API

An Application Programmer's Interface or API is a standardized set of software commands (calls) that can be used to access a particular software program in a consistent and reliable way.

ATM

Asynchronous Transfer Mode. A suite of network protocols providing low-level services spanning local and wide-area networks. ATM is intended to provide the switching and multiplexing services necessary to carry voice, data and video and multimedia traffic using fixed 53-byte cells. Standards are defined to allow ATM to emulate traditional LANs (LANE).

b

Abbreviation for "bit" where 8 "bits" comprise a byte.

B

Abbreviation for byte or the equivalent of one character in text.

BDC

One server running Windows NT Server acts as the Primary Domain Controller (PDC), that maintains the centralized security databases for the domain. Other computers running Windows NT Server in the domain function as Backup Domain Controllers (BDC) and can authenticate logon requests. The PDC or BDC authenticates users of a Windows NT Domain. See also PDC.

BDX

The NS2000 Block DataXceleration Engine provides the capability to stream blocks of data from disk to tape, or node to node, and provides the foundation for extremely rapid backup.

CIFS

Common Internet File System. A connection-oriented, network file-sharing protocol developed by IBM and Microsoft as part of LAN Manager. CIFS is the native file sharing protocol for systems running Windows for Workgroups, Windows95 and Windows NT. Sometimes referred to as SMB. Control Point—Auspex's proprietary management control software.

CPU

Central Processing Unit. Can refer to either a processor chip such as Sun's SPARC or Intel's Pentium, or to a processor chip or chips and support circuitry on a CPU board.

DataXpress

Communication among the NS2000's multiple hardware processors and software processes are handled by DataXpress, a low-overhead message-passing kernel executing on each processor.

ECAD

Electrical Computer Aided Design

EM-Net

The NS2000 Environmental Monitoring Network that connects to all chassis in an NS2000 system and reports a variety of control information to the Host Node.

Ethernet

A Local Area Network (LAN) protocol developed by Xerox in cooperation with Digital Equipment and Intel in 1976. Ethernet supports a star or bus topology and supports a data transfer rate of 10 megabits per second or 10 Mbps. The Ethernet specification formed the basis of the IEEE 802.3 standard, which specifies the physical and lower software layers. Ethernet uses the CSMA/CD access method for handling simultaneous demands and is one of the most widely implemented LAN standards.

Fast Ethernet

Fast Ethernet or 100BaseT, defined by the IEEE 802.3 committee, provides a 100 Mbps standard that is compatible with existing 10BaseT installations, preserving the CSMA/CD media access control (MAC) protocol.

FastFLO

The NS2000 proprietary file system that is optimized for providing high performance and consistent file services.

FC

An acronym for Fibre Channel

FCP

FCP is an acronym for Fibre Channel Protocol, an ANSI standard covering Fibre Channel protocol for SCSI.

FDDI

Fiber Distributed Data Interface. A standard for local area networks that typically uses fiber-optic media capable of data rates up to 100 megabits/second over distances up to 100 km. An FDDI network is a token-based logical ring, and is often constructed as a pair of counter-rotating redundant rings (called dual-attachment mode) for reliability. Ethernet, in contrast, is a bus-based, non-token-based, 10-megabits/second network standard.

FDX

The NS2000 File DataXceleration Engine (FDN) provides an interface that acts at the file system level to provide the basis for rapid file-by-file or directory backup.

Fibre Channel

ANSI standard designed to provide high-speed data transfers between workstations, servers, desktop computers and peripherals. Fibre channel makes use of a circuit packet switched topology capable of providing multiple simultaneous point-to-point connections between devices. The technology has gained interest as a channel for the attachment of storage devices, but has limited popularity as high-speed networks interconnect. Fibre channel can be deployed in point-to-point, arbitrated loop (FC-AL), or switched topologies. Fibre channel nodes log in with each other and the switch to exchange operating information on node attributes and characteristics. This information includes port names and port IDs and is used to establish interoperability parameters.

Fibre Channel Protocol (FCP)

FCP is an ANSI standard covering Fibre Channel protocol for SCSI.

FMP

Functional Multiprocessing (FMP) is the term Auspex uses for its unique distributed parallel processing NS2000 architecture. Each NS2000 I/O Node is based on an Asymmetric Multiprocessing design with two processors and a unique real time OS called the DataXpress™ kernel. Each processor simultaneously and efficiently executes different functions in the network file serving process. One processor handles network processing and the other processor handles File and Storage Processing. A Host Node is based on the traditional general-purpose single CPU computer running the general purpose Solaris OS, and is used primarily for system management activity. Up to three I/O Nodes and one Host Node are connected by a Scalable Coherent Interface (SCI). System software consists of a unique custom messaging system that enables efficient network and storage processing on the I/O Nodes and efficient system and data management on the Host Node. The FMP architecture improves system availability compared to other approaches by isolating the I/O Nodes from unplanned outages of the general purpose OS (Solaris). I/O processing can therefore continue in the event that the Host Node is down. See also SMP, parallel processing, SCI.

FSP

The File System and Storage Processor refers to one of the two Intel Pentium processors on an I/O Node of an NS2000 system. This processor runs highly optimized microcode that manages all file system and storage processing of the I/O Node and communicates with other I/O Nodes and the Host Node. See also Network Processor (NP).

Gigabit Ethernet

Standard of the IEEE 802.3 committee that provides a mechanism for conveying Ethernet format packets at GB/s speeds. The goals of the gigabit Ethernet effort include: preserve the CSMA/CD access method with support for 1 repeater, use the 802.3 frame format, provide simple forwarding between Ethernet, fast Ethernet and gigabit Ethernet, support both fiber and copper, and accommodate the proposed standard for flow control.

Gigabyte

1024 Megabytes.

GSL

The GSL bus is the main system bus in standard Intel CPU board design.

GUI

An acronym referring to a Graphical User Interface that is the screen presented to a user in any computer application.

HDDA

A term that refers to a High-Density Disk Array shelf in an NS2000 containing 28 drives arranged in four drawers of 7 drives each. A maximum three I/O Node NS2000 system contains nine HDDAs or 9x28=252 disk drives.

Inode

In UNIX, an inode is an index to files.

IP

The IP (Internet Protocol) is the underlying protocol for routing packets on the Internet and other TCP/IP-based networks. IP is an internetwork protocol that provides a communication standard that works across different types of linked networks for example Ethernet, FDDI or ATM. In an internetwork, the individual networks that are joined are called subnetworks or subnets. IP provides a universal way of packaging information for delivery across heterogeneous subnet boundaries. See also TCP Transmission Control Protocol.

Java

Developed by Sun Microsystems, Java is now a standard software language for developing plug-in applications.

Journaling

A journaling file system keeps track of all changes to files as transactions occur in real time. In the event of unexpected system problems, the file system can be restored to a consistent state by updating a prior copy of the file system for the changes made from the point in time that the copy was made.

LADDIS

An acronym formed by names of the group (Legato, Auspex, Data General, Digital Equipment Corporation, Interphase, and Sun) that developed and popularized SPEC's vendor-neutral NFS server benchmark of the same name. See also SPEC.

LAN

Local area networks or LANs are networks of computers that are geographically close together; this usually means within the same building.

MAC

Media Access Controls or MACs are the rules defined within a specific network type that determines how each station accesses the network cable. Using a token-passing method, a carrier sensing and collision detection method or a demand priority method prevents simultaneous access to the cable. The MAC used for 100BaseT Ethernet as implemented in the Auspex EtherBand™ feature is based on the "demand priority" access method in which the central hub scans all its ports in a round-robin fashion to detect stations that want to transmit a frame. Higher priorities can be requested by ports to transmit real-time information like video or audio.

MCAD

Mechanical Computer Aided Design

MIB

Management Information Base is a set of standards for detailed system information that is reported to a control console for SNMP compliance. Its intent is to provide common metrics for heterogeneous computer systems.

MTBF

Mean Time Between Failure. A key component of the availability equation, AVAILABILITY = (MTBF – MTTR) ÷ MTBF. Example: A server that on average fails once every 5,000 hours and on average takes 2 hours to diagnose, replace faulty components and reboot would have an availability rating of (5,000 – 2) ÷ 5,000 = 99.96%.

MTTR

Mean Time To Repair. Includes the time taken to diagnose the failure, replace or repair faulty component(s) and reboot the system. See MTBF.

N+1

An N+1 power supply design provides for one redundant power supply on the shelf to provide full system power in the event of a power supply failure.

NAS

Network Attached Storage.

NDMP

NDMP is a standard protocol for network-based backup of network-attached storage. NDMP hides the unique interfaces from third party backup software which allows this software to execute on any NDMP compliant system on the network (such as the NS2000 Host Node, and control backups on the NS2000 using standard commands.

NeTservices

The Auspex software product that provides for consistent high performance UNIX and NT file services and makes the NS2000 a true bilingual file server.

NIC

Network Interface Cards (or NICs) in the NS2000 support 10/100BaseT Ethernet, Gigabit Ethernet, FDDI or ATM. There are from one to three on each I/O Node. See Table 3.

NIS/NIS+

Network Information Service. This is ONC's general name-binding and name-resolution protocol and service.

NFS

Network File System. NFS is an ONC application-layer protocol for peer-to-peer, distributed, file system communication. NFS allows a remote file system (often located on a file server) to be mounted transparently by client workstations. The client cannot perceive any functional difference in service between remote and local file systems (with trivial exceptions). NFS is the most popular ONC service, has been licensed to over 300 computer system vendors, and runs on an estimated 10 million nodes. It is a de facto UNIX standard. See also VFS, ONC, and NFSv3.

NFSv3

NFS version 3. References to NFS generally imply NFS version 2 protocol. NFS version 3 is an update to the NFS protocol. Significant among the many changes made for NFSv3 are the adoption of a safe asynchronous write protocol and the use of block sizes up to 64 KB. Other protocol changes are intended to improve the overall network and client efficiency and provide improved support for client-side caching.NFS ops/s NFS operations per second. Typical NFS operations include: lookup, read, write, getattr, readlink, readdir, create, remove, setattr, and statfs.

Node

See Fibre Channel.

NP

The Network Processor (NP) refers to one of the two Intel Pentium processors on an I/O Node of an NS2000 system. This processor runs highly optimized microcode that manages all network processing of the NS2000 I/O Node and communicates with other I/O Nodes and the Host Node. See also File System and Storage Processor (FSP).

NTFS

A term that refers to Windows NT file system.

NVM

Non volatile memory is a term used to refer to battery backed up DRAM so that data will not be lost in the event of power failure.

NVRAM

Non-volatile random access memory such as static RAM will not lose data in the event that power is lost to the chip.

ONC

Open Network Computing. The trade name for the suite of standard IP-based network services—including RPC, XDR and NFS—promulgated by Sun Microsystems.

Operating System

The operating system is the most important software program that runs on a computer. The Operating System (OS) performs basic tasks such as recognizing input from a keyboard, sending output to the display screen, keeping track of files and directories on the disk and controlling peripheral devices such as disk drive and printers or a mouse. The OS acts as a traffic cop and schedules the various programs that the computer executes. The OS is also responsible for security, ensuring that unauthorized users do not access the system. Operating systems can be classified as follows:
1) Multi-user – allows two or more users to run programs at the same time. 2) Multiprocessing – supports running a program on more than one CPU. 3) Multi-tasking – allows more than one program to run concurrently. 4) Multi-threading – allows different parts of a single program to run concurrently. 5) Real Time – Usually a stripped down OS that responds to input instantly.

Parallel processing

Parallel processing refers to when a single computer simultaneously uses more than one CPU to execute a program. Ideally parallel processing makes a program run faster because there are more CPUs running it. In practice, it is often difficult to divide a program so that separate CPUs can execute different portions without interfering with each other. Among NAS vendors, only the Auspex NS2000 effectively overcomes this problem by designing each I/O Node with two processors each performing separate portions of the network file-serving task. In addition the NS2000 links multiple I/O Nodes together by a highly efficient Scaleable Coherent Interface (SCI) interconnect that allows the multiple nodes to act as one system. See also Functional Multiprocessing (FMP).

PCI

The Peripheral Channel Interconnect is an ANSI standard for an I/O bus used predominantly in PC design.

PDC

One server running Windows NT Server acts as the Primary Domain Controller (PDC), that maintains the centralized security databases for the domain. Other computers running Windows NT Server in the domain function as Backup Domain Controllers (BDC) and can authenticate logon requests. The PDC or BDC authenticates users of a Windows NT Domain.

PDU

Power Distribution Unit or Power Shelf in the NS2000. A cabinet model contains from three to seven power supplies and is N+1 redundant. See also N+1.

Port / Port ID

See Fibre Channel.

RPC

Remote Procedure Call. An RPC is an (almost) transparent subroutine call between two computers in a distributed system. ONC RPC is a Sun-defined session-layer protocol for peer-to-peer RPC communication between ONC hosts.
ONC RPC underlies NFS.

RAID

Redundant Array of Independent Disks. RAID is used to increase the reliability of disk arrays by providing redundancy either through complete duplication of the data (RAID 1, i.e., mirroring) or through construction of parity data for each data stripe in the array (RAID 3, 4, 5). RAID 5, which distributes parity information across all disks in an array, is among the most popular means of providing parity RAID since it avoids the bottlenecks of a single parity disk.

RAID Controllers

The NS2000 RAID controllers provide a highly optimized scheme for securely managing RAID configurations on NS2000 systems. The Auspex RAID controllers allow RAID arrays to be expanded online, and support conversion of an array from one RAID level to another.

SCI

Scalable Coherent Interface is an ANSI standard (#1596-1992) that is the modern equivalent of a processor-memory-I/O bus and a Local Area Network combined and made parallel to support distributed multiprocessing. The SCI interconnect has very high bandwidth, very low latency and a scaleable architecture. This allows building large high performance systems and is used by Convex/HP supercomputers, Sun Clusters, Sequent, Auspex and others. Network latency has been measured at 150 times less than previous network connections for efficient and fast communication between computer nodes.

SCSI

Small Computer System Interface. An intelligent bus-level interface that defines a standard I/O bus and a set of high-level I/O commands. The SCSI busses in the NS2000 are used to connect multiple peripheral devices such as disk drives tape drives. Each SCSI device has an intelligent SCSI controller built into it. There are currently many flavors of SCSI defined by different bus widths and clock speeds. The seven major variations of SCSI are SCSI 1, SCSI 2 (Fast / Narrow), SCSI 2 (Fast / Wide), Ultra SCSI (Fast / Narrow), Ultra SCSI (Fast / Wide) – also called SCSI 3, Ultra 2 SCSI (Narrow), Ultra 2 SCSI Wide. Single ended SCSI is used when the peripheral device is close to the point of attachment as in the NS2000 method of attaching disk drives. Differential SCSI provides for reliable operation over greater distances and is used in the NS2000 for tape drive connections.

SE

System Engineer(s) perform a variety of technical pre and post sales services for customers and prospects.

SID

An NT term meaning System Identification (SID).

SMB
Server Message Block protocol. See CIFS.

Snapshot
A term that refers to a copy of a file system at a certain point in time. Snapshots are used for backup and recovery.

SMP
Symmetric Multi-Processing. A computer architecture in which processing tasks are executed in parallel on multiple, identical, general-purpose CPUs that share a common memory. SMP computer systems usually have modified operating systems that can themselves execute concurrently. The SMP architecture offers high computational throughput, but not necessarily high I/O throughput. See FMP.

SNMP
Simple Network Management Protocol. SNMP is a protocol used for communication between simple, server-resident SNMP agents that respond to network administration requests from simple-to-sophisticated SNMP manager tools running on remote workstations.

Solaris 2.x
Sun's UNIX operating system.

SPARC
Scalable Processor Architecture. SPARC International's specification for the Reduced-Instruction-Set-Computer (RISC) CPUs found in systems sold by Sun Microsystems, Auspex, etc.

SPEC
Standard Performance Evaluation Corporation. A nonprofit corporation of vendors' technical representatives that develops and certifies accurate, vendor-neutral, computer-system benchmarks. As an example, popular SPEC CPU benchmark metrics include SPECint, SPECfp and the now obsolete SPECmarks. See LADDIS.

SPECnfs
SPECnfs measures ops/s as a measure of NFS performance standardized by SPEC. This unit of measure is often used interchangeably with SPECNFS-A93 ops/s. The A93 suffix indicates the first of what may evolve into a series of workloads, each corresponding to different LADDIS variations simulating the loads and traffic patterns of application environments like ECAD, MCAD, imaging, etc. The current version is SFS97 and incorporates NFSv3 testing.

Stack
In software protocols the CPU processes one instruction after another in a serial fashion. The exact sequence of these instructions is referred to as a stack.

Stripe
In RAID terminology, a stripe is when data is read or written in parallel to or from multiple disks instead of reading or writing all data to one disk. Striping provides much higher performance through its parallel design.

TB
A Terabyte (TB) equals 1024 Gigabytes.

TCP

Transmission Control Protocol or TCP is a transport layer component of the Internet's TCP/IP protocol suite. It sits above IP in the protocol stack and provides reliable data delivery services over connection-oriented links. TCP uses IP to deliver information across a network and makes up for the deficiency of IP providing a guarantee of reliable delivery services that IP does not. TCP messages and data are encapsulated into IP datagrams and IP delivers them across the network.

UFS

UNIX File System. UFS is the standard file system type in the BSD 4.3 kernel.

WAN

Wide Area Networks or WANs are networks of computers that are geographically dispersed and connected by radio waves, telephone lines or satellites.

Zoning

In a SAN environment this is a workaround to security problems with the Fibre Channel specification whereby data pools are assigned to a specific server. This defeats the basic premise of SAN whereby "any application" can have access to "any data."

Corporate Headquarters
2300 Central Expressway
Santa Clara, CA 95050
USA
408.566.2000 **phone**
800.735.3177 **fax**

http://www.auspex.com **web**

Japan

Auspex K.K.
ATT Shinkan, 6K
2-11-7 Akasaka
Minato-ku
Tokyo 107-0052
Japan

+81 3 3586 1751 **phone**
+81 3 3586 1778 **fax**

United Kingdom

Auspex Systems, Ltd
670 Eskdale Road
Winnersh Triangle
WOKINGHAM
Berks RG41 ST5
United Kingdom

+44 1189690900 **phone**
+44 1189692266 **fax**

France

Auspex Systems, SA
40, Boulevard Henri Cedex
France

+33 1 42 04 9150 **phone**
+33 1 42 04 9151 **fax**

Germany

Auspex Systems, GmbH
Ingolstaedter Stasse 22
80807 Muenchen
Germany

+39 89 350151-0 **phone**
+49 89 350151-50 **fax**

Auspex Worldwide

www.auspex.com

AUSPEX