UNIX REVIEW'S

Die, Spam, Die Net Freedom

PERFORMANCE COMPUTING

THE UNIX AND WINDOWS NT ENTERPRISE MAGAZINE

The Death of the Desktop

IT Regains Control

Does Auspex's Net Server 7000 Portend A New Store-Age?

Rogue Wave's Tools.h++ Pro: An Upgrade Worth Wading For

New Column: Mary Petrosky's 'Packets & Protocols'

AUGUST 1998
US $3.95 Canada $4.95
www.performance-computing.com

A Miller Freeman Publication

Ralph Barker

AUSPEX NETSERVER 7000 MODEL 810

With much of the competition endorsing file server–specific, microkernel OSs, can a SPARC and Solaris-based system shine at the high end?

The desire for greater amounts of disk space has always been the lament of UNIX system administrators and users alike. Present-day market trends, such as Web usage, data-warehouse projects, and multimedia growth, however, have made storage an even more important issue in most organizations. Additionally, new trends in storage interfaces, such as Fibre Channel and storage area networks, have complicated the picture with additional storage options. While storage area networks, in which multiple storage systems are connected to multiple computers through a matrix of Fibre Channel links (called a "fabric" in Fibre terms), show promise for certain types of storage requirements in the future, network-attached storage will continue to provide a general-purpose solution for many applications.

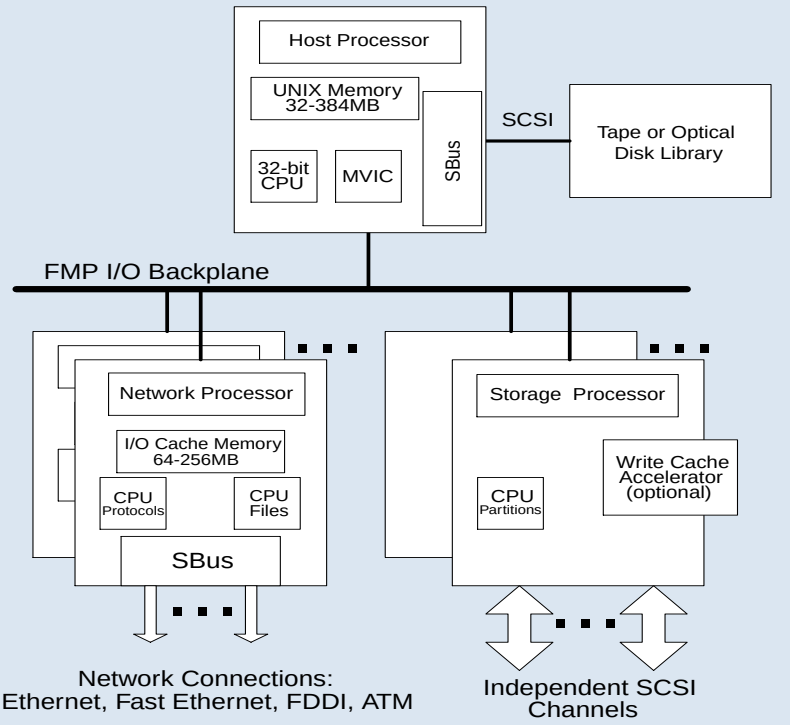
Network-attached storage is a relatively mature technology, and most frequently is based on the storage system providing NFS service to LAN-based clients. Some installations use general-purpose UNIX systems with directly attached RAID systems as NFS servers. Other installations have opted for dedicated file servers designed specifically for that application. Among the latter group, the general trend has been to use NFS-only, microkernel operating system technology similar to that originated by Network Appliance. The theory behind such designs is that the special-purpose operating system will be more efficient than a full-blown UNIX system. Vendors such as Network Appliance with its NetApp filers and Falcon Systems (recently merged into Artecon, Inc.) with its FastFile servers provided a compelling argument along that line and earned considerable marketshare.

One company that takes a different approach to network-attached storage design, however, is Auspex Systems. Its file servers, the NetServer line, use what Auspex calls "functional multiprocessing" (FMP) features to provide high-volume NFS service across the local network. Our review examines the new addition at the top end of the Auspex line, the NetServer 7000, Model 810.

OF NOTE

Although the stylish charcoal-colored cabinet and illuminated logo visually set the NS7000/810 apart from the competition, the most distinctive features of the server are inside. Like other Auspex NetServers, the NS7000/810 incorporates the company's patented functional multiprocessor server architecture. FMP separates network-protocol handling and file- and disk-I/O functions from the general-purpose OS executing those operations on dedicated I/O processors. User transparency of that functional separation and overall compatibility are maintained by running UNIX on a separate, general-purpose peer processor. Figure 1 depicts the basic elements of the FMP architecture, showing the three types of processors in the Auspex design: host, network, and storage. The network and storage processors run the functional multiprocessing kernel (FMK) and communicate via control messages across the FMP I/O backplane. The processing components also have local memory, so normal NFS request processing does not directly involve the UNIX kernel running on the host processor (HP). As such, system scalability is predictably linear—installing extra network or storage processors provides additional data paths that are independent of the host processor's functions.

Figure 1 **Auspex FMP architecture**



For example, when an NFS version 2 read request comes in from an NFS client, the request is processed by the protocol-processing CPU in the network processor module through the IP, UDP, RPC, XDR, and NFS layers. The protocol-processing CPU then constructs a control message, including a file handle and a data offset, and passes it to the file-processing CPU, also located in the network processor module. The file-processing CPU checks its hash table to determine if the data is already in cache. If so, it responds immediately with the address of the requested data. If the data is not in cache, the file-processing CPU further refines the file handle request through its locally-cached UFS metadata and constructs an I/O control message, including the I/O cache memory location reserved for the data, for the storage-processing CPU. The storage-processing CPU then translates this second control message into a physical partition request and activates the appropriate SCSI channel. The data then moves through the storage processor into the pre-designated I/O cache memory location. The storage-processing CPU tells the file-processing CPU that the request has been satisfied. The file-processing CPU, in turn, communicates the completed operation, including the memory address of the data, to the protocol-processing CPU. The protocol-processing CPU then can initiate the data transfer back to the client. Only two direct memory access (DMA) operations are involved—one when the data is moved into cache, and the second when transferred to the client. NFS version 3 requests generally are handled in a similar manner, except that the NFS version 3 protocol includes a number of performance improvements over NFS version 2.

Note that all of these operations are handled in the FMK, not UNIX. Thus, NFS operations are not subject to delays that may result from the UNIX kernel processing other system-level requests, such as cpio operations to a tape device attached to the host processor. Additionally, each processor is functionally specialized and au-

tonomous, eliminating the need for system-bus traffic associated with maintaining cache coherence in a traditional SMP design. Recognize, too, that control is achieved through interprocessor messages, not shared memory. This loose coupling of processors contributes to the linear scalability of the overall system.

The overall design of the NS7000/810 includes other notable hardware and software features:

- DriveGuard, the Auspex implementation of RAID level 5, is supported by special hardware on the storage processor to perform parity calculations as data is being transferred.
- An optional 8MB, battery-backed, nonvolatile RAM write-accelerator daughterboard attaches to the storage processor (SP), thus using no additional FMP backplane slots. The daughterboard is removable, and can be moved to another SP to recover cached data if an SP fails.
- The Auspex DataGuard software allows UNIX, running on the HP, to come to a complete halt without affecting NFS operations. The FMK running on network and storage processors continues to handle NFS requests while UNIX reboots on the HP. Thus, important applications, such as backup, can be run on the HP without the risk of an application hang interrupting file-server operations. DataGuard also permits manual reboots, so administrators can install software updates or new HP-attached SCSI devices while NFS service continues.
- The optional ServerGuard software lets two NetServers be clustered for additional high availability.

OPERATION

Installation and setup of the NS7000/810 follows industry norms for large, high-end systems. As with most systems of this scope, moderate site planning is required prior to the machine's arrival. Of particular note in this regard, the NS7000/810 requires two dedicated 30 amp, 220 volt circuits, one for the main cabinet and another for the expansion rack. A detailed hardware manual that describes electrical requirements and other hardware installation issues is available from Auspex, and should answer most pre-installation questions. It is wise to double-check such things as the NEMA codes for electrical connectors, so extra circuits that you have installed are properly equipped.

When the NS7000/810 is shipped, the complement of high-density disk assemblies (HDDA) and disk drives is packed separately from the main and expansion cabinets. This procedure provides better physical safety for the drives during shipment, and makes the cabinets substantially lighter and easier to move when they arrive. Normal installation procedure is for Auspex to dispatch a field engineer to install the drives in the proper HDDA and drive slots (the drives are appropriately tagged at the factory) and cable-up the system. A small ASCII terminal serves as the system console, and is used for basic system configuration (setting the system's hostname, IP addresses, and so on) during installation. As with any UNIX system, many maintenance operations, after initial setup, can be conducted from any workstation on the network.

Documentation for the NS7000/810 also follows large-system traditions. The printed documentation, which includes a system manager's guide, a command reference guide, and release notes, comes in large, three-ring binders. A pocket-sized system manager's quick-reference pamphlet also is included. Online documentation contains the normal man and xman references in the SunOS v. 4.1.4 operating system and a CD-ROM. Although the system software (Netserver release 1.9.2) is factory installed, a CD-ROM is packaged with the documentation set for backup. A separate CD-ROM also is included for the Auspex Premier Series Software (DriveGuard, DataGuard, FastBackup, ServerGuard, and network-related utilities).

As described earlier, the NS7000/810 incorporates the Auspex FMP design. To support that architecture, Auspex makes some changes to integrate the additional functional CPUs into the standard SunOS 4.1.4 operating system. These enhancements to the OS are well-described in the system manager's guide and the command reference guide. Most of the additional administrative commands used to manage the system have a prefix of `ax_`, making them easy to find in the documentation.

The NS7000/810 can be configured as an NIS master (or slave) and as the boot source for diskless SunOS workstations. Architecture-dependent executables for such workstations are not included with the Auspex distribution, but can be purchased from your Sun Microsystems representative or other supplier. The default configuration of the NS7000/810, however, comes with sufficient space for two sets of such executables. If more space is required, the */exports* directory can be moved to

Figure 2 **SPECnfs_A93 benchmark results comparison**

System	SPECnfs_A93 Ops/sec	Avg. Response Time (ms)	Users
Auspex NS7000/700	10,084	11.0	1,008
Network Appliance NetApp F630	4,328	9.6	433
SGI Origin 2000 with 4 250MHz R10000 CPUs	10,078	13.0	1,008
Sun Microsystems Enterprise 3002 with 6 336MHz UltraSPARC 1 CPUs	11,681	15.0	1,168

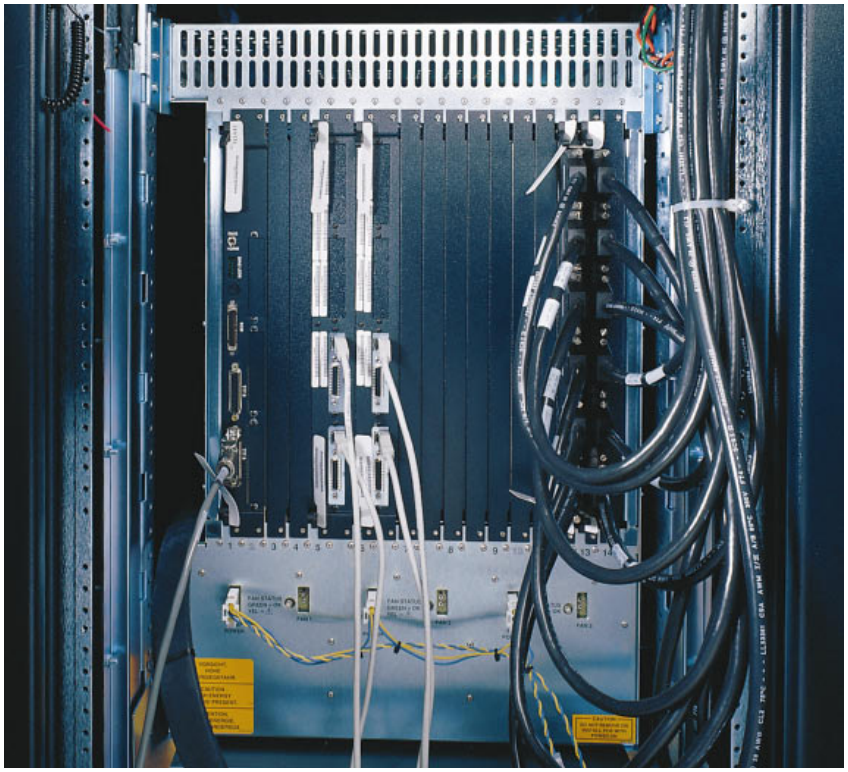
a larger partition on the system drives. Note that the boot and system drives are on a separate SCSI bus attached to the HP and are not part of the arrays managed by the storage processors. The `SetupExec` utility is for installing the appropriate executables from your media and configuring the workstation as a boot client.

The NS7000/810 implements several types of file systems depending on where the file system is located. Standard UNIX partitions on the boot disk are usually the standard UNIX 4.2 BSD file-system type, commonly known as the Fast File System (FFS), while exported file systems are what Auspex implements as local file systems designed for storage-processor handling. Local file systems default to the Tahoe File System type, often referred to as the Fat Fast File System (FFFS).

Various options exist for creating usable partitions within the file space of the NS7000/810. Physical partitions can be concatenated to create larger virtual partitions that span several drives. Virtual partitions also can be striped (RAID level 0) or mirrored (RAID level 1) using Auspex commands. The DriveGuard software adds RAID level 5 to the mix of options, but requires the optional Write Accelerator III board, SP V boards (as in the NS7000/810), and Netserve version 1.9.2 or later.

RAID 5 configuration and management on the NS7000/810 is a combination of file manipulation and maintenance commands. To create a RAID 5 array, you first must add an entry for the array in the */etc/raidtab* file in the form `ardn type [,parameters] adx1,adx2` where `n` is the array number (each SP has a pre-assigned range of numbers, with a maximum of 7 arrays per SP), `type` is either `raid 5` or `spares`, and `x` is the physical drive number. Parameters include: `spare=` (to define a specific spare), `size=` (to define the stripe size as 64K, 128K, or 256K), `pri=[hi | lo]` (to set rebuild priority), and `rebuild=[auto | manual]`. Separate `ax_raid` commands are then issued to load the array definition and initialize the array to zeros. Note that you can have both assigned spares for a particular array, and an array of floating spares for use by any array of type raid 5 supported by that particular SP. You also can partition the array after it has been initialized.

If you are using the optional ServerGuard product to cluster two NetServers, the rebuild parameter must be set to manual. An automatic rebuild will trigger a failover. The ax_raid utility and several other commands are used to manage the array over time, including the critical step of periodically verifying and correcting parity.



Main-cabinet processor modules in the Auspex 7000/810

EXPANDABILITY

In addition to the single host processor (HP VIII in the NS7000/810), up to five network processors (NP IV), and up to five storage processors (SP V) can be configured in the system. NPs can support various network interfaces including 100Base-T, ATM, and FDDI. A seven-slot drive rack is provided for the boot disk, CD-ROM drive, and other SCSI peripherals that are attached directly to the HP. Disk drives for client storage are organized into HDDAs. Each HDDA has four drawers, each of which can hold up to seven disk drives. The main cabinet can hold two HDDAs (63 drives total), and the expansion cabinet can hold up to five additional HDDAs (140 drives), for a system maximum of 203 disk drives for client storage. Supported drives include 1GB, 2GB, and 9GB.

PERFORMANCE

One advantage of the Auspex FMP architecture is that it can run performance monitoring utilities, such as ax_perfmon and ax_perfhist, without significant effect on the system's actual NFS performance for clients. These useful utilities provide statistical data that can tune the performance of the NS7000/810 further.

Although we ran our usual file server benchmark, bigdisk, on the NS7000/810, that benchmark, designed for directly attached storage devices,

has proven to be only marginally informative for NFS servers. In these tests, we connected each 100Base-T port of the NS7000/810 to a separate port on a 100Base-T Ethernet switch, and connected a mix of 100Base-T and 10Base-T clients to the 100Base-T switch and a 10Base-T switch cascaded off that switch's uplink port. Multiple copies of the benchmark were created in different directories on a typical five-disk RAID 5 array on the NS7000/810, and compiled as appropriate for each client. A base run using a single 100Base-T client, a 170MHz Sun Ultra 1, was then compared with runs done with different combinations of multiple clients. Overall throughput for the NS7000/810 was consistent with, or better than, results we have seen on other NFS servers, adjusting for the network infrastructure.

A more appropriate and informative benchmark for the NS7000/810 is the SPEC SFS, or Laddis, benchmark. Figure 2 (p. 44) shows a comparison of the published results for the NS7000/700 (essentially a single-cabinet version of the system we reviewed and should have similar performance characteristics) with other NFS servers. The SPEC SFS benchmark performs a preset mix of NFS operations against the server from multiple load generators (client systems), and expresses the results in terms of NFS operations per second with associated average response times in milliseconds. The Auspex results ranged from 998 SPECnfs_A93 Ops/sec with an average response time of 3.7ms to a high of 10,084 SPECnfs_A93 Ops/sec with an average response time of 11ms. Using the SPEC formula, the high rating translates to 1,008 SPECnfs_A93 users. The single-CPU Network Appliance NetApp F630 scored 4,328 SPECnfs_A93 Ops/sec with an average response time of 9.6ms, or 433 SPECnfs_A93 users. A four-CPU SGI Origin 2000 posted results of 10,078 SPECnfs_A93 Ops/sec with an average response time of 13ms, or 1,008 SPECnfs_A93 users—results quite similar to those of the Auspex. Meanwhile, a Sun Enterprise 3002, running six 336MHz UltraSPARC 1 CPUs, posted a score of 11,681 SPECnfs_A93 Ops/sec with an average response time of 15ms—about ten percent higher than the Auspex, but with a somewhat slower response time.

HOW IT RATES

The design of the NS7000/810 is unique among network-attached NFS servers. The Auspex FMP design uses dedicated CPUs for network and storage processes, separate from the normal UNIX processes running on the HP. The benefits of this design include greater network and storage expansion without being affected by the usual additional CPU overhead seen in single-processor NFS servers or conventional UNIX systems used for that purpose. The design of the NS7000/810 rates an excellent, or four *Performance Computing* flags.

Installation of the NS7000/810 is consistent with industry norms for a system of this stature. The factory packaging of the system (packaging the disk drives and HDDA drawers separately) makes the system relatively easy to move into place for setup. Actual setup is supported by an on-site Auspex field engineer, making the installation that much easier. Another excellent, four-flag rating for installation.

Documentation for the NS7000/810 combines printed materials, online manual pages, and a CD-ROM—a typical package. In general, the printed manuals are well-written and easy to understand. During our installation, however, we encountered an error in the hardware manual regarding the type of electrical connections. We also found the fact that setting up a RAID 5 array was not addressed in the primary manual, but rather in a separate, nonprinted DriveGuard manual to be potentially confusing. Otherwise good documentation for the NS7000/810 thus gets only a rating of average, or two *Performance Computing* flags.

Auspex's NetServer 7000/810

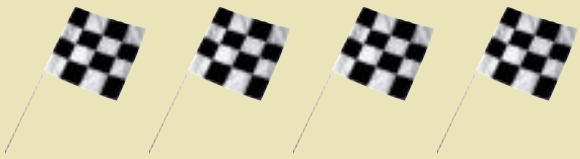
Auspex Systems Inc.
2300 Central Expy.
Santa Clara, CA 95050
408-566-2000
408-566-2020 fax
http://www.auspex.com/products/systems/800.html

TESTED CONFIGURATION: SPARC host processor 8; 256MB RAM, expandable to 384MB; Root dual SPARC Network Processor 4 with 256MB RAM, quad 100Base-T Ethernet NICs; SP5 Storage Processor with 8MB battery-backed Write Cache 3; 76-inch main cabinet with 12-slot card cage, main-cabinet drive rack with 4.29GB system disk, bootable CD-ROM drive, two 1200-watt power supplies, two 48-volt HDDA power supplies, one HDDA with four drive drawers; 76-inch expansion cabinet with power supplies, two HDDAs, each with four drive drawers; 84 9GB drives total; compact ASCII terminal for system console.

PRICE AS TESTED: $684,300

OPTIONS: Various network expansion, drive, and memory configurations. List price for base configuration with single cabinet, 128MB RAM each in host processor and network processor, one HDDA without drives, $165,000.

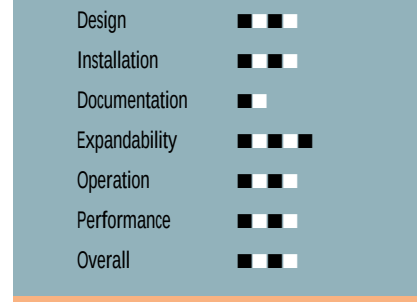
EVALUATION: Although the system is currently limited to 32-bit CPUs and lacks a graphical interface for the creation of RAID 5 arrays with the optional DriveGuard software, performance and manageability are excellent, earning an overall rating of four *Performance Computing* flags.



Expansion is a key feature of the NS7000/810. Up to five network processors and up to five storage processors can be configured. Although there are some drawbacks to the compact design of the HDDA drawers, we liked the fact that 28 drives can be installed in the space usually occupied by about ten on the average system, and 203 drives can be accommodated in the combined main and expansion cabinets. Expansion rates an outstanding on our scale, all five possible flags.

HOW IT RATES

Auspex's NetServer 7000/810



Operation of the NS7000/810 is simple and straightforward. Auspex has added extensions to the standard SunOS operating system to support FMP, and has added associated commands to make administration of the system robust. What is obviously lacking is a graphical interface to DriveGuard, the optional Auspex RAID 5 implementation. The other management and performance-tuning features of the system are sufficiently rich, however, to warrant a rating of excellent, or four flags for operation.

Performance of the NS7000/810, as reflected by the SPEC SFS (Laddis) benchmark, is on par with high-end UNIX systems used as NFS servers with respect to overall throughput (SPECnfs_A93 Ops/sec), but with a somewhat faster response time, as shown in Figure 2. We rate the performance of the NS7000/810 as excellent, or four flags.

Overall, we think the NS7000/810 is an excellent candidate system for high-end NFS server requirements where both disk and network expansion are key factors. Notwithstanding the caveat that the system's CPUs are currently 32-bit, we give the NS7000/810 an overall rating of four *Performance Computing* flags, excellent on our scale.

Ralph Barker is the senior technical editor of Performance Computing. *Write him at* rbarker@mfi.com.