Date Published 10-1-2003

Integrating Network Attached Mass Storage Systems into Educational Networks: An Initial Examination of Performance and Security Issues

Dennis Guster
St. Cloud State University

James E. Weber
St. Cloud State University

Charles Hall
St. Cloud State University

INTEGRATING NETWORK ATTACHED MASS STORAGE SYSTEMS INTO EDUCATIONAL NETWORKS: AN INITIAL EXAMINATION OF PERFORMANCE AND SECURITY ISSUES

DR. DENNIS GUSTER, DR. JAMES E. WEBER, & CHARLES HALL
ST. CLOUD STATE UNIVERSITY

ABSTRACT

As colleges and universities digitalize record storage, classroom support and administrative processes, data storage needs are increasing dramatically. Institutions of higher education are turning to mass storage systems to fill these needs. NAS (network attached storage) is often selected over alternatives for cost, scalability and ease of administration reasons. This paper reports the results of an initial study assessing security concerns, performance characteristics, ease of installation, configuration, upgradeability, and use by end users, in an academic installation of network attached storage.

I. INTRODUCTION

Computer networks have become ubiquitous in our modern world, with almost all electronic devices in business or education having access to some form of connectivity. Although this connectivity opens doors to the exchange of vast amounts of information, the infrastructure supporting the storage and distribution of this information is often taxed to the limit. In particular, the need for secondary storage is growing explosively (Pfenning, 2001; Schuchart, 2001).

In business, e-commerce and data warehousing are putting increased emphasis on database systems and storage (Ling, Yen & Chou, 2000-2001; Schuchart, 2001). In higher education, however, the need for additional secondary storage has multiple drivers. Data mining classes require data warehouses containing "tons of data" (Olsen, 1999; pg. A52) and some propose the use of data warehouses, hosted by educational institutions, to speed the flow of data on the internet (Young, 1999). A significant portion of the need for additional storage comes because an increased amount of data from everyday activities is stored indefinitely. White (2000) gives an example of a student's everyday activities being tracked and stored as they browse the web, check materials out of the library, e-mail friends and family, and visit the health

clinic. But educational administrators view distance education as having the greatest potential to explode in strategic significance and demand great expenditures of resources for storage needs (Lembke & Rudy, 2001; Young, 2001). Online courses, which can include streaming video, simulations, animation, and high-end graphics threaten to make current storage and bandwidth resources obsolete (Taylor, 2001).

Traditionally, secondary storage was increased at the server level, where the pool of additional storage was based on the operating system of that device. In other words, space attached to a Novell server would be organized in Novell format, and a user must be attached to that server to access it. This logic often requires users to have pools of space on several different servers/hosts to support all of the varied applications they might run and acquiring additional space can be cumbersome or require administrative permissions that can be difficult to obtain. However, this diversity of space might have some advantages in regard to performance. If the space is distributed across several devices, one can theoretically surmise that the contention on any one device would be lessened.

In an attempt to address some of the limitations of server/host-attached storage, vendors have come up with several competing solutions. Of the available options, Storage Area Networks (SAN) and Network Attached Storage (NAS) are the most popular (Farley, 2001). Each option has advantages and disadvantages, and businesses have had difficult choices to make in selecting the best option for their situation (Anderberg, 2002). NAS was the first of these solutions offered, while SAN was introduced to overcome shortcomings of the NAS architecture, including overload of the local area network and management issues (Garvey, 2001). SAN consists of an additional, separate high-speed network and storage facility that bypasses the existing local area network and attaches directly to servers. As such, SANs are optimized for delivering large blocks of data to servers (Fiero, 2001; Lee & Rigney, 2000). But NAS has some advantages too, offering superior scalability, simple setup and administration and low total cost of ownership. It may also offer some security advantages because of the stripped-down network operating system used (Fetters, 2000; Fiero, 2001).

Typically, a NAS is accessed through a single network entry point (i.e., a samba server) and is modular in nature. This means that a user can begin with a small single device and then upgrade painlessly to many devices offering hundreds of gigabytes of space (Lee & Rigney, 2000; Petreley, 2001). In fact, the process may only involve mounting the additional storage units on a rack,

plugging them into an Ethernet switch, and spending less than a half-hour redefining the configuration through a web accessible interface. This is in sharp contrast to adding space to a Linux or Novell server. Some suggest that NAS and SAN are starting to converge (Moore, 2000), but at the current time the two options appeal to different market segments, with NAS being distinctly less expensive (Baltazar, 2001). The cost advantage is likely to be the controlling factor in decisions made by educational institutions, which often don't have the resources that are available to for-profit businesses. From this perspective, NAS seems the likely choice for educational institutions.

Although the advantages of NAS from cost, installation, management and scalability perspectives are apparent, what about its performance characteristics? On the surface it looks like a centralized system using existing networks may cause contention problems as the number of users and the amount of potential space available to them increases. Furthermore, does distributing this space result in security concerns beyond what would be expected with dedicated units?

II. ARCHITECTURE

Servers, storage and networking have been described as the three pillars of computing (Dahl, 2001) that need to support the goal of providing the correct information to the correct place in a timely and secure fashion. Researching this simply stated goal is often easier said than done. In fact, there are often numerous trade-offs regarding network design decisions that affect performance.

Currently, there appear to be two schools of thought concerning how secondary storage should be integrated into the servers-storage-network model (Farley, 2001; Schuchart, 2002). One school advocates server-attached storage that is linked via high-speed interfaces, such as fiber channel (Lee & Rigney, 2000; Moore, 2000). This school advocates the SAN approach. The other school feels that storage should be independent of servers and their platforms and be directly attached to the network, hence the acronym, NAS (network attached storage devices). From an architectural perspective, the latter is a radical departure from the previous model in that it uses the existing network rather than a dedicated physical channel to transfer data and to perform management functions related to the storage function.

But the NAS approach raises the question of how well this new type of traffic is integrated into already taxed network environments. The manner in which the storage devices attach to the network infrastructure provides some

insight. For the sake of flexibility, redundancy and expandability, these storage systems are typically built upon the concept of independent modular units. Each unit contains about 100 gigabytes of raw storage space. A typical configuration would involve 4 units and because redundancy is built in, the yield would be somewhat less than the 400 gigabytes expected. Although each unit is an independent module, the logic built into each unit is designed so that all units work together in concert as a single network attached mass storage system. In terms of expandability, it is not unreasonable to configure up to 16 units as a single system (Tricord Systems, 2002).

How is this modularity supported by the physical network? Each unit has one 100BASETX connection (a front channel) supporting both data and management communications, which is connected to the LAN via a switch or hub (Tricord Systems, 2002). The 100BASETX connections are a relatively high-performance industry standard, and potential bandwidth increases as additional units are added to the stack. This configuration is shown in Figure 1, which depicts the physical connections between units. In this configuration, the samba server functions as an authentication mechanism for packets addressed to the NAS. Theoretically each unit could also have two channels; a back channel for communications between units for management purposes and a front channel which would then be used only for data transmission to and from the local area network (LAN). The back channels would also be connected together through a switch or hub. This configuration is illustrated in Figure 2.

Figure 1.
Physical Connectivity of Mass Storage System.
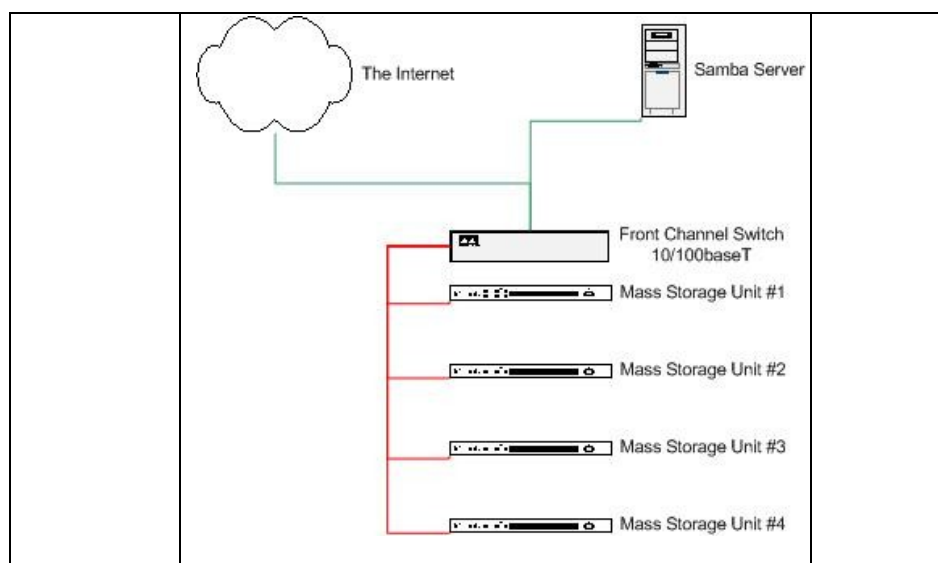
Connection of all units through the back channel would be highly desirable. In addition to data packets transferred in and out of the mass storage system, a number of strictly management packets are transmitted as well. These packets are generated by CPUs located on each storage unit and are for synchronizing the individual units into a single mass storage system. By connecting all units to a single hub or switch via the back channel, the management communication path tends to be shorter and quicker. Furthermore, this management traffic remains isolated and does not interfere or cause degradation on other parts of the network.

Figure 2.
Physical Connectivity of Mass Storage System with Back Channel
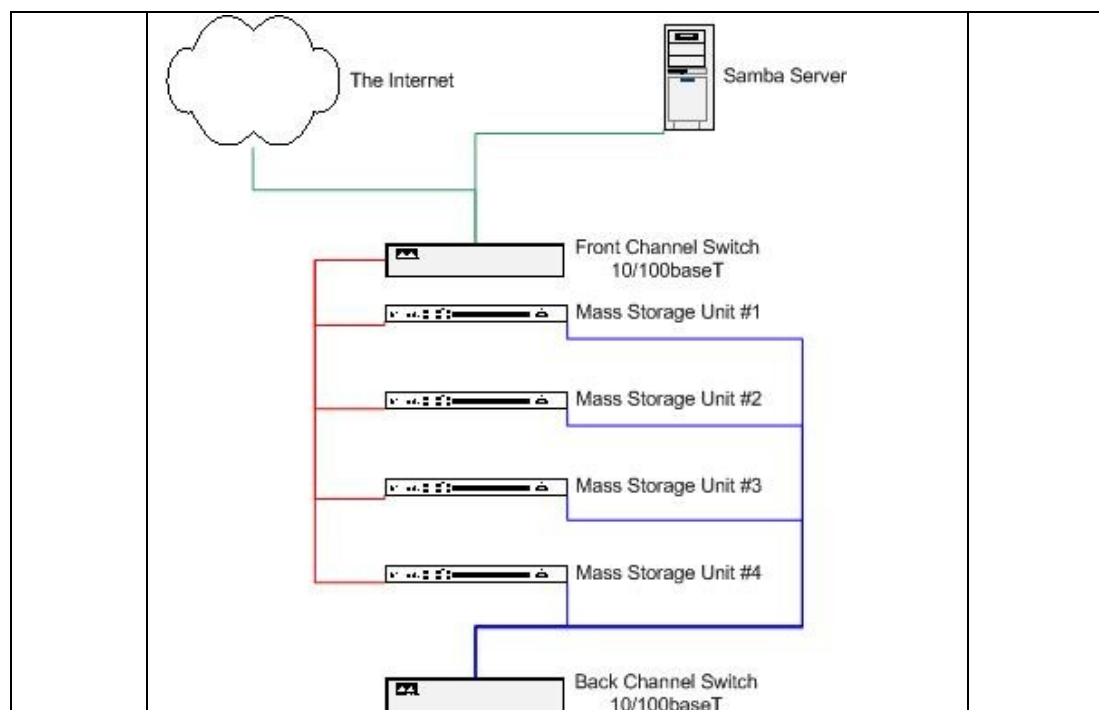
While this solution appears to be efficient for handling the management overhead, how it affects the flow of data packets to a workstation running an application supported by data stored on the mass storage system merits further discussion. To a certain extent, efficiency would be influenced by topology and architecture. In other words, the connective relationship among the hub/switch containing the mass storage system, the placement of the workstation, and the location of server to which the mass storage system is logically attached all should affect transmission efficiency.

III. THIS STUDY

A preliminary installation of a NAS system, configured as shown in Figure 1 with four individual 100 gigabyte units, was implemented. Because the testing was done with beta units, no back channel was present and all traffic, both management and data, took place on the same channel. Twenty users with workstations were allowed access to the NAS, and archived network traffic files from the LAN were also stored on the NAS as they were generated. An initial analysis of total traffic coming and going from the individual mass storage system units reveals a packet arrival on the network approximately every .014 of a second. The traffic pattern involves incoming and outgoing packets in every possible combination among the mass storage units. A large number of management packets were included in the load. The purpose of the management traffic was to maintain the integrity of the redundant array of mass-storage system units. A second category of traffic is that of user management/configuration requests via a browser to a Java applet. This traffic was of limited intensity and sporadic especially after the original configuration wa solidified. The last major category of traffic was applications requesting or writing data to/from the mass storage system. The intensity of this traffic of course is a function of the workload, the sum of all applications that will access the mass storage system, its server, and the network itself.

1. ISOLATION OF OVERHEAD TRAFFIC

Based on a sample of 10,000 packets recorded from a 143.9261 second time interval, one can conclude that the traffic pattern is very intense. In other words, about 70 packets per second, on average, arrived on the network. In obtaining this traffic, only packets addressed to the samba server, the four mass storage units (including management packets), and the requesting workstation were recorded. The ratio of data traffic to overhead traffic was quite skewed in favor of the overhead traffic. In fact, only about 200 packets containing data or a request to set up a virtual data circuit were recorded.

This ratio is not all that surprising in that only a single workstation was requesting data during the capture session. However, this overhead load needs to be further examined if intelligent network design decisions are to be made. Therefore, the question needs to be addressed – should the overhead and data traffic be combined in a single switch as shown in Figure 1 or separated into two physical channels (including a back channel) requiring two switches as shown in Figure 2?

2. COMPARISON OF COMBINED VERSUS A TWO-CHANNEL CONFIGURATION

To provide objective data about these two different configurations, a simulation was programmed in Comnet III, a network simulation tool. In one model, all traffic - both data and overhead - was transmitted together in the same physical channel as shown in Figure 1. A second model was devised in which the data traffic and the overhead traffic were separated into two physical channels as shown in Figure 2. As much as possible, values obtained from the initial investigation described above were used to program the simulation. However, two major limitations need to be stated. First, certain values needed for the simulation were not readily available, such as the delay to be expected in each mass storage unit. In such cases, the realistic default values were used, which certainly reduced the validity of the simulation. Second, there was a wide variety of the average packet interarrival rates, ranging from .014 to .0001, among the samples taken. For the sake of simplicity, the simulation was run using the worst-case scenario logic. In other words, the .0001 value was used, again raising certain validity questions.

This simulation would provide information of limited value if the goal were to understand how a specific mass storage system would perform. However, the goal for this study was to gain objective data about the relative performance of each topology under the same conditions. In that regard and that regard only, the results obtained are useful. Table 1 provides performance statistics regarding the network link(s) for each model.

The link utilization of the combined link was reduced from 57.67 percent to about 40 percent on the data link and 36 percent on the overhead link. The 57.67 percent is getting dangerously close to the 80 percent saturation level defined in basic queuing theory (Arnold, 1978). By splitting the channels, there certainly would be more tolerance for increased loads. The delay and frame size need to be analyzed together. Because the data packets in the front channel tend to be larger by nature, it takes longer to get the whole packet across the link. It is also interesting to note that the average frame size for the data channel is fairly well optimized while in the combined channel, it is reduced to about 200 bytes because of the influence of the overhead traffic.

Although it is more realistic to expect the physical channel(s) to be connected to a switch (a more efficient, newer technology), in this simulation connections were made via a hub to help illustrate potential contention

problems. Use of the hub allowed us to capture data unavailable if a switch were used, and to illustrate how traffic intensity can cause processing delays. In the three columns in Table 1, contention problems occurred at a surprisingly similar ratio. Therefore, the true interarrival rates of both the data and overhead packets need to be carefully examined in any design that implements a mass storage system. In terms of deferral, when the channel was busy, the values again appear to be a function of the average packet size. The number of frames delivered in the 15-second simulation in which identical workload definitions were applied to each model reveals that about 3,000 more packets were passed through the two-channel model. Perhaps this indicates more work done in the same time period. This would make sense if information is being delayed along the way and therefore sitting in queues longer in the combined mode simulation. To a certain extent, this is supported by the results in Tables 2 and 3.

Table 2 reports that the results of delay at the mass-storage-system node level in the mass storage system while Table 3 reports the delay in getting a message from the workstations to the mass storage system nodes. In almost all cases, the delay is markedly less in the two-channel system. Again, this would support the contention that the two-channel system is reducing delays and hence queue lengths and therefore resulting in better performance.

Table 1
Link Characteristics

	Combined	two-channel front	back
Utilization	57.67%	40.18%	36.27%
avg/delay (link)	.108 ms	.186 ms	.019 ms
avg frame size	208 bytes	1030 bytes	90 bytes
collided frames	3998	474	5017
avg deferral delay	.03 ms	.06 ms	.001 ms
delivered frames	5246	736	7513

Table 2
Average Node Delays in Mass Storage System

Node	Combined Channels Send	Receive	Two-Channel Send	Receive
1	.0001	6.917	.025	.619
2	2.946	3.825	.103	.505
3	1.133	2.369	.027	.386
4	2.236	3.251	.038	.307

Table 3
Average Delays at the Workstation Level

	Combined	Two-Channel
1	.661	.376
2	.743	.215
3	.344	.146
4	.799	.166

3. SECURITY CONCERNS

Although it has been noted that NAS systems have certain inherent security advantages over alternatives (Fetters, 2000), some concerns remain. One problematic area is related to some of the overhead traffic generated that goes beyond the stations directly involved. For example, the NTP (network time protocol) packets are broadcast to all stations on the home network. There have been documented cases of hackers devising attacks using this protocol mainly to reset the time back to an earlier time and replacing an authentication string (Bishop, 1990). Therefore, filtering this traffic beyond its intended audience may be a good idea. Also, jini-announcements (java-based network coordination commands) are multicast to address 224.0.1.84 and although they may perform some valid function (or be a byproduct) in a java-based application, the need to multicast them over the internet needs to be examined (Baker & Smith, 2001; Coffee, 2001; McElligott, 2001).

Within the rest of the packet traffic observed, the source and/or destination addresses were to either mass storage system units or servers/workstations requiring access. A good share of the management traffic appears to be proprietary in nature and its payload is not easily read with a packet sniffer. Some traffic, however, is readable, such as ping requests from the primary unit to the secondary units.

Therefore, protection of these units as much as possible would be recommended. The idea of isolating the units on their own switch would reduce the viewing domain if a packet sniffer were involved. Also, some type of packet filtering device should be placed between that switch and the outside to help protect against threats originating primarily from OSI layers 3 and 4, such as source routing or redirecting a TCP session.

IV. CONCLUSIONS

It appears that a mass storage system has a number of merits to be considered when devising a strategic plan for networked storage in higher

education, and NAS seems the most likely candidate to be implemented. The ease of installation and configuration could be big time savers for network administrators. The same is also true for its ease of expandability and flexibility of access.

In terms of performance, the data samples analyzed did indicate that the packet interarrival times can be quite dense. However, their densities do not appear to be problematic on the 100BASETX architecture. Furthermore, the simulation programmed to test performance of a one-versus two-channel topology yielded some interesting results. This simulation, which was programmed to reflect the worst-case scenario in a 200-node domain, indicated that even when data and overhead traffic were combined into a single channel, performance was still acceptable. Also, this simulation illustrated that the two-channel approach had advantages over the single-channel approach. This separation of management from data traffic may well have security advantages as well. Colleges and universities should consider using a back channel when implementing NAS.

There appear to be several security concerns associated with the installation of a mass storage system, though NAS, with its simpler installation, comes with fewer concerns than other options. There are concerns with any addition to a network. Any networked device that supports the prime function of the university needs to be carefully studied, analyzed and integrated into the comprehensive security plan of that university. These storage devices should, at a minimum, require firewall protection just like any traditional host. Configuration of the firewall to limit distribution areas for network timing and coordination announcements needs to be carefully undertaken.

In summary, this initial study has provided valuable information that institutions of higher education need to take into account in making adoption and configuration decisions. Technology is rapidly changing, and it is evident that mass storage units will soon be available with both front and back channel options. This study has discussed and examined some of the considerations involved in that decision and summarized other concerns.

ACKNOWLEDGEMENTS

REFERENCES

Anderberg, K. (2002). Tough choices. *Communications News, 39*(1), 4-7.

Arnold, A. O. (1978). *Probability, Statistics and Queuing Theory with Computer Science Applications*. Orlando: Academic Press.

Baker, M. & Smith, G. (2001). Establishing a reliable JINI infrastructure for parallel applications. *Parallel Processing Letters, 11*(2-3), 203-221.

Baltazar, H. (2001). Deciphering NAS, SAN storage wars. *eWeek, 18*(14), 26.

Bishop, M. (1990). A security analysis of the NTP protocol. *Proceedings of the 6th Annual Computer Security Conference. Tuscon, AZ, December*, 20-27.

Coffee, P. (2001). The next 20 years. *eWeek, 18*(30), 13.

Dahl, G. (2001). Server appliances as network devices: Combining the best of the three pillars. White Paper, Tricord, Inc. Downloaded from the World Wide Web on 1/31/02 at *http://www.tricord.com/appliance/aggregation?PID=Detail.html&CID=18429 1415f65460388780784 90679e03&DID=ee35fb021ddc443b804b49ec0dcf208c*

Farley, M. (2001). Three models for SAN-based data sharing. White Paper, Infostar. Downloaded from the World Wide Web on 2/1/02 at *http://is.pennnet.com/Articles/article_display.cfm?Section=Articles&ARTICLE _ID=77714&Publication_ID=23*

Fetters, D. (2000). Midrange NAS: Happy medium. *Network Computing, 11*(19), 201-205.

Fiero, D. (2001). NAS vs. SAN. *Communications News, 38*(4), 18.

Garvey, M. J. (2001). Storage under control. *Information Week*, 837, 79-81.

Lee, M. & Rigney, S. (2000). Data management & storage technology. *Network Computing, 11*(25), 147-154.

Lembke, R. & Rudy, J. (2001). Second annual EDUCAUSE survey identifies key IT issues. Downloaded from the world Wide Web on 1/31/02 at *http://www.educause.edu/issues/survey2001/report.asp*

Ling, R. R., Yen, D. C. & Chou, D. C. (2000-2001). From database to web browser: The solutions to data access. *Journal of Computer Information Systems, 41*(2), 58-63.

McElligott, T. (2001). Catalyst project uncorks JINI. *Telephony, 241*(17), 52-54.

Moore, F. (2000). Storage networks start to converge. NAS gains momentum and respect. *Computer Technology Review, 20*(4), 1.

Olsen, F. (1999). At business schools, students learn to mine data on consumer behavior. *The Chronicle of Higher Education, 45*, A52.

Petreley, N. (2001). Nothing new under the SAN. *Computerworld, 35*(42), 66.

Pfenning, A. (2001). Storage needs are growing, but budgets aren't. *InternetWeek,* 879, 40-44.

Schuchart, Jr., S. J. (2001). Data management & storage technology. *Network Computing, 12*(26), 77-82.

Schuchart, Jr., S. J. (2002). Restoring SANity. *Network Computing, 13*(1), 32-40.

Taylor, M. C. (2001). Unplanned obsolescence and the new network culture. *The Chronicle of Higher Education, 48*, B14.

Tricord Systems, Inc. (2002). FAQs. Downloaded from the World Wide Web 2-28-2002 at http://www.tricord.com/appliance/aggregation?PID=Listing.html&CID=5ef208 64517c40cb91f38338744187d8

White, L. (2000). Colleges must protect privacy in the digital age. *The Chronicle of Higher Education, 46*, B4.

Young, J. R. (1999). Speeding up the flow of scholarly data. *The Chronicle of Higher Education, 45*, A21.

Young, J. R. (2001). Survey finds concern on administrative computing. *The Chronicle of Higher Education, 47*, A33.