An Evaluation of Storage Systems based on Network-attached Disks *

Gang Ma
Dept. of Computer Science,

A. L. Narasimha Reddy
Dept. of Electrical Engineering,

Texas A & M University
College Station, TX 77843-3128

Abstract: *The emergence of network-attached disks provides the possibility of transferring data between the storage system and the client directly. This offers new possibilities in building a distributed storage system. In this paper, we examine different storage organizations based on network-attached disks and compare the performance of these systems to a traditional system. Trace-driven simulations are used to measure the average response times of the client requests in two different workloads. The results indicate that it is possible to reduce the workload on the file manager and improve performance in some workloads. However, in other workloads, reduced cache hit ratios may offset the gains of distributing the network processing workload.*

1 Introduction

With the increasing processor speeds and increasing use of I/O intensive applications such as video retrieval, distributed and parallel storage organizations have received considerable attention recently. However, it has been observed that even as parallelism is being exploited in data retrieval, the file manager managing the data remains a bottleneck. In a traditional file system, file server has to be involved intensively in order to satisfy a client's request. With the rapid growth of the number of network users and the number of requests, demands on the file server continue to increase. It has become imperative to find ways to reduce workload on the file servers(for example through client caching) and to increase the throughput capacity of the servers(for example, through parallel and distributed systems).

Earlier work has been done in AFS [1] to reduce the file server load. AFS clients use the local disk as a cache to store the files that have been accessed before and will probably be needed in the future. While client caching can significantly reduce the requests going to the file server, the load on the file server could still be too high due to a large number of misses in the client cache caused by file sharing and increasing file sizes.

NetServer of Auspex exploits a unique functional multiprocessor architecture by coupling storage devices to network as directly as possible [2] to improve file system scalability. Server level striping is shown to be effective in improving data throughput in ZEBRA [3] and in SWIFT [4]. NCSA HTTP server [5] employs an AFS based organization to distribute the workload at the file server to multiple HTTP servers. Cooperative caching scheme [6]in xFS file system utilizes the idle client's cache to obtain higher global hit ratio.

With the introduction of network-attached disks, the workload of file server could be reduced by transferring data directly between clients and the storage system. Fibre Channel attached disks [7] and Servernet-attached disks [8] are examples of such an approach. The file server may need to set up a session between a client's request and the disk controller. After that data can be transferred directly between the client and the disk without going through the file server. This can reduce the server load and thus can potentially improve the overall performance. Several organizations for a file system based on network-attached disks are studied in [9]. The authors measured the workload in the file server under the the proposed NetSCSI and NASD architecture and showed that the server load could be reduced by a factor of up to ten in NFS and five in AFS. However, the average response time of the client requests, an important factor of overall performance was not measured.

In this paper, we study several issues in organizing a storage system based on network-attached disks. In a regular file system, file server manages the file name space, allocates physical disk space and supports user's requests for file service. A file server typically uses memory caching to improve response time for user requests. Without extensive modifications of the file system, only a few of these functions can be delegated to other components in the system in order to reduce the load at the file server. Network-attached disks, by

*This work is supported in part by an NSF Career grant and by a grant from Texas Higher Education Board

MICROSOFT CORP.
EXHIBIT 1010

Authorized licensed use limited to: Klarquist Sparkman LLP. Downloaded on February 07,2023 at 18:05:21 UTC from IEEE Xplore. Restrictions apply.

their ability to supply data directly to the user, can reduce the network precessing overhead on the server. This is one of the main perceived advantages of a storage system based on network-attached disks. Then how much impact does this make on user's response time? If the network precessing work is to be done at the disks, how much processing capacity do the disks require to offload this work from the server system? Typically disks have a small cache on the disk read-write arm. The size of this cache is in the range of 512k-2MB. Is this cache sufficient for providing better response time than a traditional system that employs server caching? These are some of the issues that will be studied in this paper.

In order to answer the above questions, we use trace-driven simulations to compare storage systems based on network-attached disks with a traditional file system? We use the average response time of client requests as the primary measure of evaluation.

The rest of this paper is organized as follows. Section 2 describes the system organizations we will study in this paper. In section 3, we discuss the simulations performed. In section 4, we present and analyze the simulation results. Section 5 describes related work in this area. Section 6 provides a summary of the results and directions for future work.

2 Storage System Organizations

In the following subsections, we discuss two organizations for a storage system.

2.1 Server-disk: Regular storage system with server-attached disks

This storage organization is used by the current distributed file systems. The disks are attached to the file server via a private bus(typically SCSI bus). Clients access the file server through a public network(typically a LAN). All the client requests are sent to the file server first. The file server is responsible for handling and interpreting the requests. File servers typically employing caching to reduce the load on the disks at the server and to improve response time for client requests. If a miss occurs in the server cache, file server will issue a request to the disk, which in turn accesses the data and sends it back to the file server. Finally file server will satisfy the client request by sending the data to the client. In this situation, file server is intensively involved in handling every request from the clients. Fig. 1 shows the sequence of operations that take place to serve a user's request in this organization.

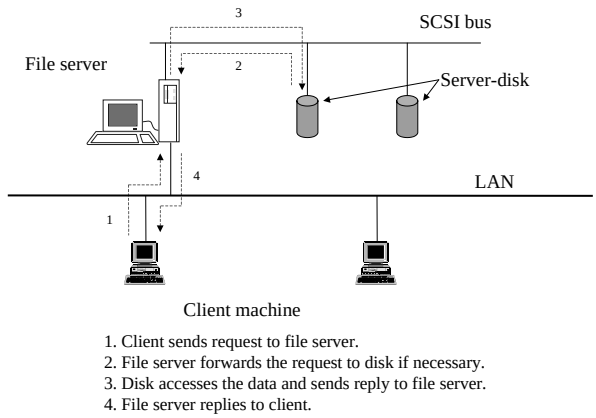

Figure 1: Server-disk: Regular server-attached disk.

2.2 Net-disk: Storage System based on Network-attached disks

In this system, the disks are attached to the file server via a private network. Disks are also directly connected to the LAN which connects clients and the file server together. Clients still send requests to the file server first. The file server processes the requests and forwards them to the disks via the private network if necessary. Rather than sending the data back to the file server again, the disks send the desired data directly to the client via the LAN. Since the file server is no longer involved in block data transfer, the load on it will possibly be less than in a regular server-disk system. On the other hand, to offload network processing work to disks, the file server does not employ data caching. Then how does the performance of this organization compare to regular server-disk organization? If we allow the server to continue to cache and reply data to the client, the network processing can not be offloaded to the network-attached disks on cache hits. Then the network-attached disks can reply data to clients only on cache misses at the server. Fig. 2 describes the sequence of operations that take place to serve a user's request in this organization.

Net-disk and server-disk organizations employ caching at different places in the system. A single large cache space in the file manager in server-disk enables efficient utilization of cache space across all the data sets in the system. The smaller caches at each disk in net-disk organizations can lead to inefficiencies if disks are accessed non-uniformly. Our assumption of system-wide disk striping [10] however reduces this disadvantage. We assume that the file manager caches metadata in both the organizations and replies directly

Authorized licensed use limited to: Klarquist Sparkman LLP. Downloaded on February 07,2023 at 18:05:21 UTC from IEEE Xplore. Restrictions apply.

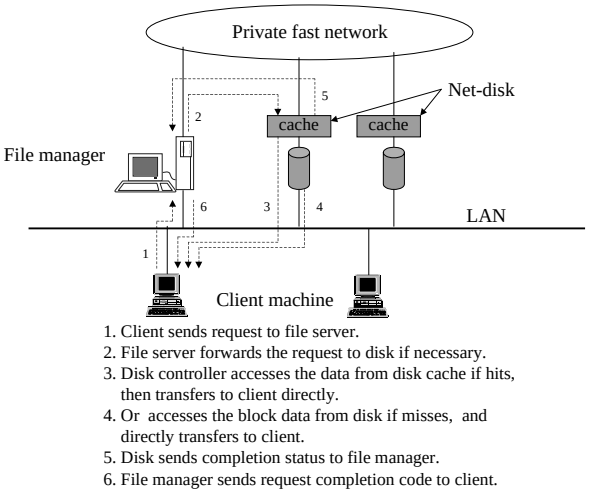

Figure 2: Net-disk:Storage system based on Network-attached disks.

to clients on metadata related requests. However, the file data caching is performed differently in different organizations. Server-disk caches file data in the file manager cache while net-disk organization caches file data at the disk.

We are also interested in other issues in organizing the storage system. For example, how powerful should the disk processor be in a net-disk system? How large a cache should disks have in order to maintain comparably high hit ratios? How will the overall performance change if we vary the number of disk nodes? We also want to identify which component among file server, network and disks is the bottleneck that impacts the system performance in these different approaches.

3 Simulations

In order to measure the overall system performance under the different storage system organizations, we performed a set of trace-driven simulations.

In each organization, the simulated system consists of a file manager and a set of disks. The file manager and the disks are interconnected differently in the different organizations. In the server-disk, the disks are attached to the file manager through a SCSI bus and in the net-disk case, the disks are attached to the file manager through a private fast network. In both the cases, cost of control messages to talk to disk is assumed to be 50us.

In both organizations, we assume that the file manager caches the metadata of the file system. In the server-disk case, the file manager also caches the block data. The size of the file manager cache is varied from 128MB to 2GB. We vary the size of the disk cache from 4MB to 128MB in the network-attached disk. We assume LRU is used as the replacement policy for the caches both at the file managers and at the disks. We assume that the caches are organized on a 4KB block basis.

It is assumed that data is striped across the disks in order to distribute the load evenly among the disks. The data will be stored in all available disks in a stripe size of 16 blocks(64KB). In our system, the first chunk (64KB) of each file is stored on a random starting disk, s. Subsequent chunks of that file are consecutively stored on disks $(s+1) mod\ n$, $(s+2) mod\ n$, ..., where n is the number of disks. Equal number of disks are employed in both the organizations.

The simulations are written in CSIM. All the requests are sent to the file server first via regular network(LAN), and then processed in separate ways depending on the organization as shown in figures 1 and 2. The file manager processor is assumed to have a processing capacity of 50MIPS. We vary the disk processor speed from 10 MIPS to 100 MIPS to study the impact of disk processor power on the system performance.

Table 1 lists the costs of different operations that we will use as parameters in the simulations. We assume that the disk access latency is 15.0 ms. The cost for cache access(hit/miss)in a file manager and the network processing load are based on measurements done on IBM's OS/2 operating system [11]. We assume that the cache search time in the network-attached disk is half of that in the file manager cache because the file manager needs to do extra work(translating file name into disk block address, checking access permissions, etc.). Costs for disk access and network transfer are based on the size of the request. The costs per 4KB block and 64KB chunk reflect the efficiency of transferring data in larger size blocks (from the disk and over the network). We assume that the disk has a transfer rate of 8MB per second. Processor dependent costs are given in number of instructions such that the impact of changing the processor MIPS can be easily studied.

In our experiments, we use request response time as the performance measure. Previous studies [9] have used the load on the server as a measure of evaluating the network-attached disk organizations. The goal of our study is to investigate the impact of the new organizations on the response time of user's requests. The request response time is measured as the time interval to service all the blocks in a given request.

We use two different traces in our simulations. A

Authorized licensed use limited to: Klarquist Sparkman LLP. Downloaded on February 07,2023 at 18:05:21 UTC from IEEE Xplore. Restrictions apply.

Table 1: Costs for different operations.

Operations	Per-block (4KB) cost	Per-chunk (64KB) cost
Getattr & similar	750 ins.	N/A
Cache hit in file manager	3,000 ins.	N/A
Cache miss in file manager	7,000 ins.	N/A
Cache hit in network-attached disk cache	1,500 ins.	N/A
Cache miss in network-attached disk cache	3,500 ins.	N/A
Disk access time	15 ms	22.5 ms
Private network transfer time	0.1 ms	1.6 ms
Private network latency (for control messages)	50 us	50 us
Public network(LAN) transfer time	10,000 ins.	108,000 ins.

trace based on NFS workload and a trace based on web accesses are used in this study. The NFS trace data is obtained from University of California at Berkeley [6]. This trace consists of network requests from 237 clients during one week period that are serviced by an Auspex file server. The trace was taken by snooping the network, so it only contains the post-client-cache request data. In other words, these requests are actually the local misses occurred at client caches. The original NFS trace includes a large amount of backup activity over weekends and at night. We only used the daytime activity (between 8 AM to 5 PM) on the server as input to our simulations.

The other trace we used in our simulations is the workload trace of the ClarkNet WWW server [12]. ClarkNet is a busy Internet service provider for the Metro Baltimore-Washington DC area. This trace contains the requests from 161,140 clients during a two week period. The original trace consists of successful accesses as well as unsuccessful accesses. We only consider the successful accesses since unsuccessful accesses do not incur any data transfer. Table 2 lists a brief description of these two trace files.

The NFS trace file consists of three major types of activities: getattr & setattr, block read & write, directory read & write. In all the organizations we study here, file manager has to deal with the accesses to file metadata and the translation of client requests into disk commands. Therefore the getattr & setattr and directory read & write operations are handled by the file manager in all the systems. In the net-disk organizations, block read & write will occur between disk(or disk cache) and client without going through the file manager. The Internet trace file simply consists of file access operations, mostly file read operations.

Table 2: Description of the two trace files.

	NFS trace	Web trace
# of client machines	231	161,140
Total # of requests	6,302,418	2,955,038
Requested bytes (MB)	10,100	27,810
Trace Duration (hours)	40	236

4 Results

4.1 NFS trace

Fig. 3 shows the average request response time as a function of the disk processor MIPS. We used one file manager and 16 disk nodes in this experiment. The file manager has a processor capacity of 50 MIPS and a 128MB cache. The disk cache in the net-disk organizations is varied from 4MB to 32 MB. Since requests are processed entirely at the file manager in the server-disk organization, changes in disk processing capacity do not affect its performance. Higher disk processor capacity improves performance considerably in net-disk organizations since block data transfer and network processing are performed by the disk processor. However, we see that the response time is not always better than in the server-disk organization. With 4MB disk cache, net-disk organization has worse performance than the server-disk organization. Only with larger disk caches, net-disk organizations have better response times than server-disk organization. This shows that the benefit of distribution of data transfer brought by network-attached disks can be offset by the costs of extra disk accesses introduced due to insufficient caching. With 32 MB of disk cache and 25 disk MIPS, we see an improvement in overall performance-response time by up to 33% compared to server-disk organization. Even at a low disk processing speed of 10MIPS, the processing capacity at the disks(total of 160MIPS) outweighs the capacity at the file manager. This is part of the reason why a net-disk organization outperforms the server-disk organization. Net-disk organization also has extra amount of cache at the disk nodes. It is not clear if the higher MIPS in the system or the larger amount of cache space or the distribution of workload contributes to the improvement in response time in the net-disk organization. We will explore these issues further in

Authorized licensed use limited to: Klarquist Sparkman LLP. Downloaded on February 07,2023 at 18:05:21 UTC from IEEE Xplore. Restrictions apply.

later experiments.

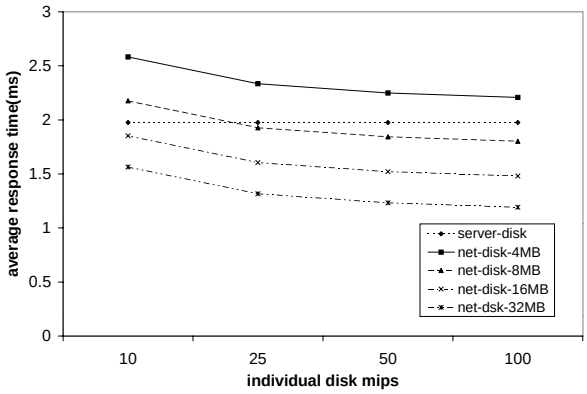

Figure 3: Response time vs. disk processor MIPS.

Fig. 4 shows the average request response time as a function of the number of disk nodes. We assumed that the system has a file manager with 50 MIPS and 128 MB of cache. With a larger number of disks, the disk accesses can be distributed over more disks to reduce the disk waiting times in all the organizations. Net-disk organizations see even more improvement because the data transfer load can be distributed to more disk nodes. However, we notice that with 4MB-8MB of disk cache, the net-disk organization has a higher response time than the server-disk organization. From figures 3 and 4, it seems that a disk cache of 16MB-32MB is needed for this NFS workload for net-disk organizations to outperform the server-disk organization.

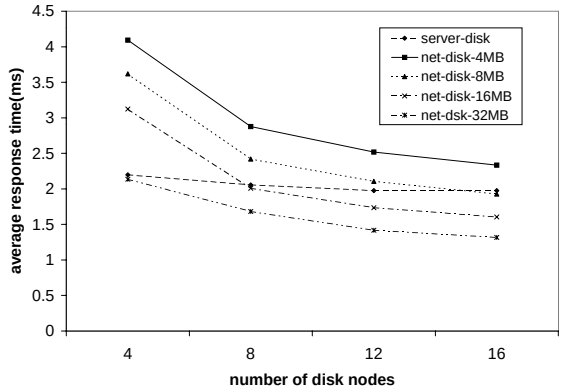

Figure 4: Response time vs. number of disks.

So far, we have considered varying the disk parameters (number, MIPS and cache sizes) while keeping the configuration of the file manager constant in all the organizations. The server-disk cannot effectively utilize the enhancements at the disk (increased MIPS or increased caches). Given the same processing capacity and cache memory, the different organizations can utilize them more effectively in different locations in the system. How do the systems compare if we put the extra memory and the extra processing power at the file manager in the server-disk organization? To answer this question, we consider systems with equal processing capacity and equal amounts of cache memory.

Fig. 5 shows the average request response time as a function of total memory in the system. In server-disk organization, all the memory is located at the file manager. We assume that the file manager in the net-disk has 128 MB and the rest of the memory is distributed equally across the 16 disks in the system. For example, with a total memory of 192 MB, each disk has (192-128)/16 = 4MB in the net-disk organization. We assume that file managers have equal processing capacity of 50 MIPS in all the organizations. The disks are assumed to have a capacity of 25 MIPS. We notice that the response time of the server-disk organization is better than the net-disk organization in all the cases. However, as the amount of memory on each disk increases, the differences in response times reduce. Beyond 32MB per disk, the difference in response times doesn't reduce significantly further.

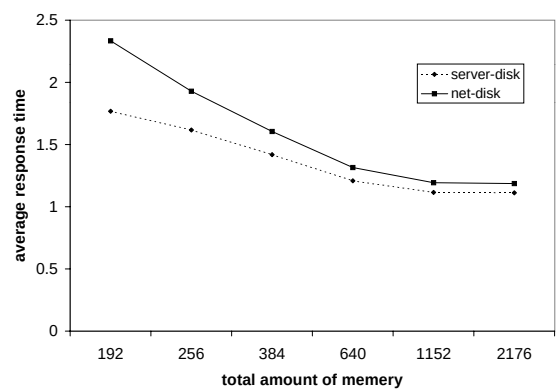

Figure 5: Response time vs. total amount of memory.

Next we considered file managers with different processing capacities. In net-disk organizations, the processing capacity and the load is distributed between the file managers and the disks. In server-disk organization, the processing capacity at the disks is unutilized and hence seems wasteful to locate any MIPS at the disk. For example, a net-disk organization with 50 MIPS at the file manager and 25 MIPS at each of the

Authorized licensed use limited to: Klarquist Sparkman LLP. Downloaded on February 07,2023 at 18:05:21 UTC from IEEE Xplore. Restrictions apply.

16 disks in the system has a total processing capacity of 50+16*25 = 450 MIPS. How would such a system compare to a server-disk organization where all the 450 MIPS are located at the file manager? Fig. 6 shows the results from such experiments. We consider two processing capacities of 450 MIPS (as explained above) and 210 MIPS (with 10 MIPS at 16 disks and a 50 MIPS file manager in the net-disk organization). We also kept the total amount of the memory the same (640MB) in both the organizations. The server-disk organization has significantly better response times for both the processing capacities. With increased amount of memory, the performance differences are reduced. It is also noticed that the differences in response times have increased from the earlier results in figure 5. It could be possible to divide the total MIPS more optimally in the net-disk organization to reduce the differences in performance. But, Fig. 5 indicates that when the total memory in the two systems is the same, with NFS workload, the net-disk organization has worse performance even with extra processing capacity.

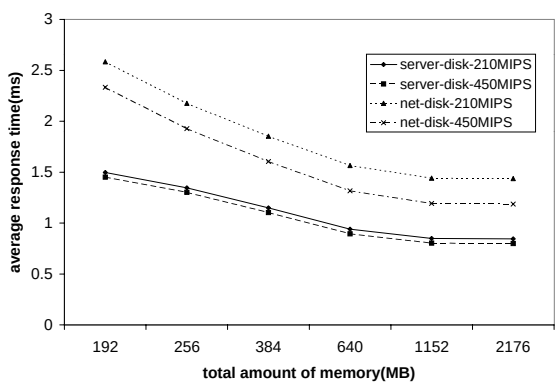

Figure 6: Response time with equal system MIPS.

Metadata requests constitute roughly 75% of the NFS workload requests and about 54% of the bytes accessed. These requests are processed at the file manager in both server-disk and net-disk organizations. With lower processor MIPS at the file manager in the net-disk organization, these requests will experience worse response times than in the server-disk organization. What about the response times for data requests? We observed that the data requests experience lower hit ratios in the net-disk organization as shown in figure 7. This reduced hit ratio offsets the gains due to distributing the network processing load. Hence, the overall response time is higher in the net-disk organization.

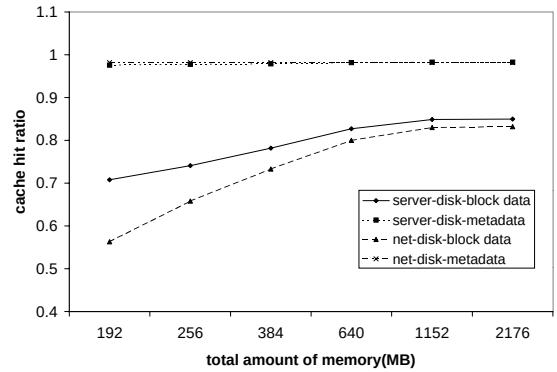

Figure 7: Hit ratios as a function of memory.

Fig. 8 shows the different components contributing to the response times in server-disk (on the left) and the net-disk (on the right) organizations. Fig. 8 shows that the disk access times are significantly higher in the net-disk organization. As shown earlier, this is mainly due to lower hit ratios for block data at the disks in the net-disk organizations. This lower hit ratio, which results in higher average disk service times, dominates the other possible advantages of the net-disk organization for the NFS workload.

4.2 Web trace

We performed the same experiments using Clarknet trace file. In Web access, popular files are accessed frequently and other files are accessed rarely. It is shown in [12] that 10% of the distinct files were responsible for 85% of all the requests received by the server. Fig. 9 shows the performance of different organizations as a function of total memory in the system with a file manager of 50 MIPS and 128MB of cache memory and 16 disks (each with 25 MIPS). In all our experiments, cache hit ratio is already so high that increasing the total memory in the system does not show much performance improvement beyond 384MB. We observe that net-disk organization has 25% lower response time than the server-disk organization. This is in contrast to the earlier results observed in the NFS workload as shown in Fig. 5. This is mainly due to the high hit ratios (95-98%) observed even with small caches at the disks in the web workload.

Fig. 10 shows the contribution of different components to response times in both the organizations configured as explained above. Service times and waiting times at the file manager are the main contributers to the response time in the server-disk organization.

Authorized licensed use limited to: Klarquist Sparkman LLP. Downloaded on February 07,2023 at 18:05:21 UTC from IEEE Xplore. Restrictions apply.

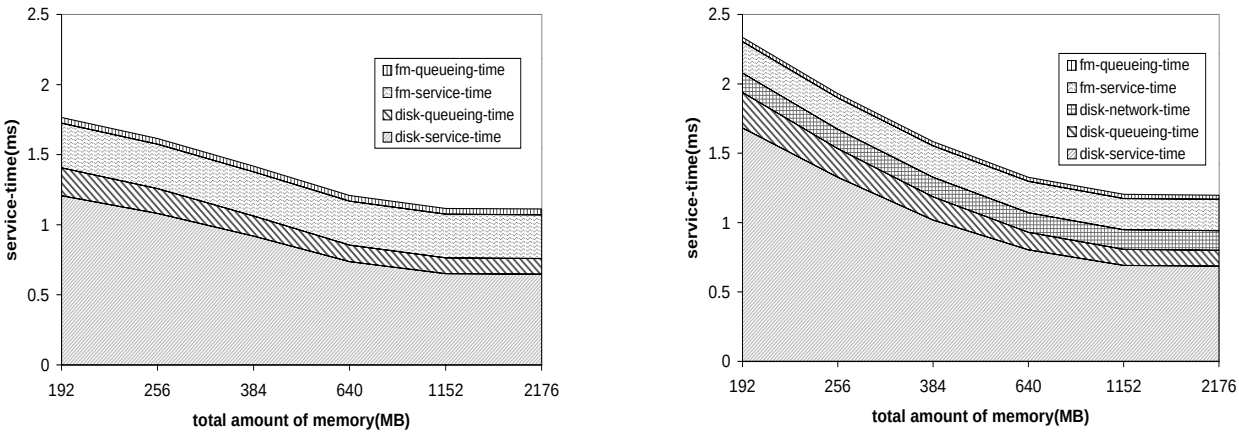

Figure 8: Components of request response time – NFS trace.

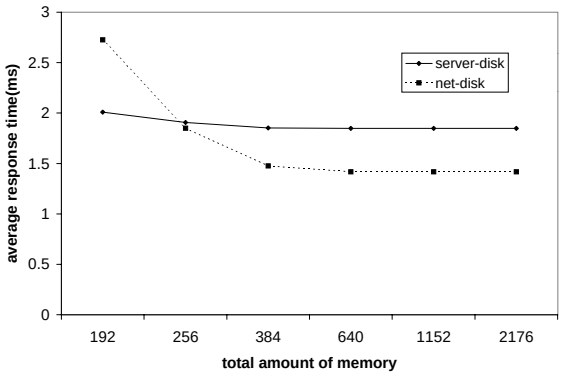

Figure 9: Response time vs. total amount of memory–Web trace.

The disk access time is not a major factor. It is observed that the distribution of network processing cost to the disks in the net-disk organization resulted in smaller service and waiting times at the file manager. This results in improved performance in the net-disk organization. At smaller amounts of disk cache, the disk access times become significant (because of lower hit ratios) in the net-disk organization.

5 Conclusions and future work

In this paper, we studied a number of issues in organizing a storage system based on network-attached disks. Through trace-driven simulations, we found that although using network-attached disks can reduce workload at the file server, if sufficient caching is not employed at the disks, the overall system performance will be worse than the traditional system.

The two traces we have studied exhibited different behavior. Caching made a bigger impact on the NFS workload and distribution of network processing load made a bigger impact on the Clarknet Web workload. With the NFS workload, distributed caching at multiple disk caches performed worse than a single centralized pool at the file manager offsetting any possible gains due to distributing network processing load. However in the Clarknet Web workload, where a small cache is enough to make a big contribution to hit ratio, network processing played a bigger role in determining the performance. We conducted a number of other experiments to study the impact of varying the cost of a network reply, the cost of an average disk access, the fraction of metadata related requests in the workload and the amount of client caching. These results can be found in [13].

When systems with equal memory and equal processing capacity are compared, the traditional server-disk organization performed better than any net-disk organization in both the workloads. If we relax the constraint of equal processing capacity, the net-disk organizations could provide better performance than a server-disk organization in the web workload. If the server has to process the data, rather than just supplying it to the client, the server has to store this data in memory. In this case, caching at the disks in net-disk will not be useful for processing the data at the server. Database and transaction processing systems (which process retrieved data) may not be able

Authorized licensed use limited to: Klarquist Sparkman LLP. Downloaded on February 07,2023 at 18:05:21 UTC from IEEE Xplore. Restrictions apply.

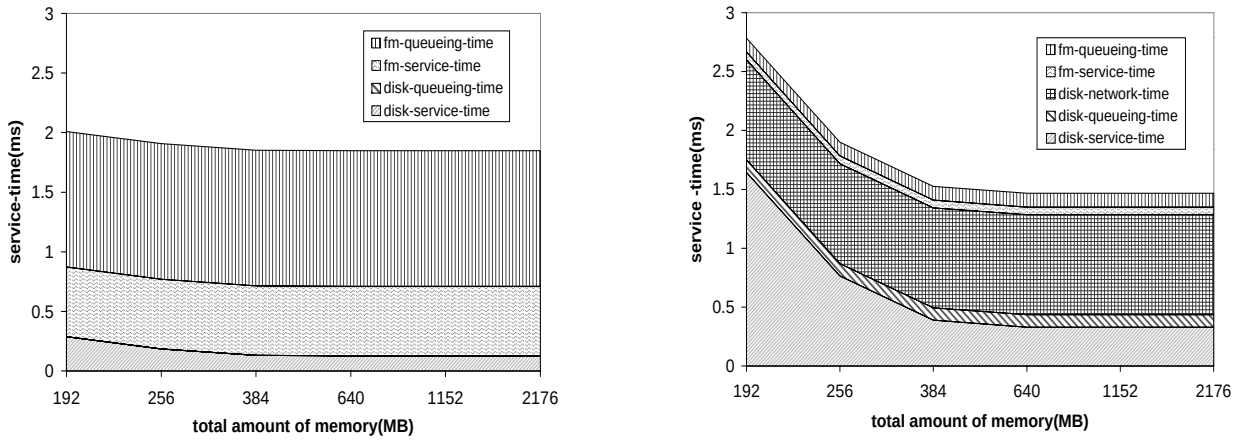

Figure 10: Components of request response time–Web trace.

to exploit the advantages of network-attached disks as well as a file system.

A number of issues require further study. Can a net-disk organization support higher loads due to the reduced load on the file manager? How does the workload impact the minimum required size of the disk cache in the net-disk organization? What workload characteristics determine if net-disk organization can provide better performance than a server-disk organization?

References

[1] J. H. Howard, M. L. Kazar, S. G. Menees, D.A. Nichols, M. Satyanarayanan, R. N. Sidebotham, and M. J. West. Scale and performance in a distributed file system. *ACM Transactions on Computer Systems*, 6(1):51–81, Feb 1988.

[2] P. Trautman, B. Nelson, and Auspex Engineering. An overview of NFS server using functional multiprocessing. Technical Report 10, Auspex Corporation, April 1997.

[3] J. H. Hartman and J. K. Ousterhout. The Zebra striped network file system. In *Proc. of the 14th SOSP*, pages 29–43, Dec. 1993.

[4] D. D. E. Long, B. R. Montague, and L. Cabrera. Swift/RAID: A distributed RAID system. *Computing Systems*, 7(3):333–359, 1994.

[5] E. D. Katz, M. Butler, and R. McGrath. A scalable HTTP server: The NCSA prototype. In *Proceedings of the First International WWW Conference*, May 1994.

[6] M. D. Dahlin, R. Y. Wang, Anderson T. E., and D. A. Patterson. Cooperative caching: Using remote client memory to improve file system performance. In *Proceedings of the First Symposium on Operating System Design and Implementation*, pages 267–280, November 1994.

[7] Seagate Corporation. Fibre channel: The digital highway made practical. Technical report, Seagate Corporation, 1994. `http://www.seagate.com/support/disc/papers/fibp.shtml`.

[8] Robert W. Horst. TNet: A reliable system area network. *IEEE Micro*, 15(1):37–45, Feb. 1995.

[9] Garth A. Gibson and et al. File server scaling with network-attached secure disks. In *Proceedings of the Sigmetrics Conference on Measurement and Modeling of Computer Systems*, Seattle, June 1997.

[10] D. A. Patterson, G. Gibson, and R. H. Katz. A case for redundant arrays of inexpensive disks(RAID). In *Proc. of SIGMOD*, pages 109–116, June 1988.

[11] B. Dimpsey. Private communication. *IBM Austin*, 1996.

[12] M. F. Arlitt and C. L. Williamson. Web server workload characterization: The search for invariants. In *Proceedings of the Sigmetrics Conference on Measurement and Modeling of Computer Systems*, pages 126–137, May 1996.

[13] Gang Ma. Evaluation of storage systems based on network-attached disks. *M.S. thesis, Texas A & M University*, May 1998.

Authorized licensed use limited to: Klarquist Sparkman LLP. Downloaded on February 07,2023 at 18:05:21 UTC from IEEE Xplore. Restrictions apply.