The Effects of Disk Organization on Lotus Notes Performance

Executive Summary

Most IT professionals understand that disk array performance improves as drives are added to an array. The performance gain results from distributing the same amount of disk I/O activity across a greater number of drives. So a general rule-of-thumb is that larger arrays deliver better performance than do smaller arrays. However, to obtain increased performance, the software must scale I/O demand by increasing the I/O request rate to the array. If the software does not increase the disk load as drives are added to an array, performance will not scale as the number of drives is increased.

As the number of files increases, disk storage requirements also increase. Coping with the increasing amounts of data could include organizing files into logical volumes, which could satisfy data management requirements (e.g., easing the tasks of performing backup and restore of files). However, in a client/server environment, disk partitioning can introduce disk-access latency resulting from disk-head movement across partitions. Task switching results in file accesses alternating between the different disk volumes, which can increase average disk seek time, adversely affecting disk subsystem performance.

Alternatively, a large array could be divided into two arrays, forming two independent logical volumes. This solution avoids the disk-head movement due to task-switching between disk partitions. The downside of this approach is that using two smaller arrays may slow performance because one of the arrays could become a hot-spot or bottleneck.

What effect does disk drive organization have on performance in a Lotus Notes e-mail environment?

This paper seeks to answer the question by presenting the results and analysis of performance measurements conducted using Lotus Domino Mail server, the server-side companion to Lotus Notes Mail. Understand that the results obtained with Lotus Notes are not likely to be the same for other applications. One should not assume that these techniques will apply to other application environments.

The goal of the first study was to determine the efficiency of RAID-1 arrays of various sizes. The goal of the second study was to determine the effect of different organizations of a constant number of disk drives at various realistic user loads.

June 2000

Introduction

Lotus Notes lends itself to distributing user mail files evenly across multiple logical volumes so that disk I/O activity is evenly distributed across them.

The Domino Mail Server accesses various types of files for which usage is very different. Performance can be improved by organizing these files in different physical volumes based on file-access characteristics. Instead of putting all the code and data in one physical volume or disk array, we recommend organizing the files by type as follows:

- Type 1: Operating system code and main data director, which holds the frequently accessed files such as the directory (formerly known as the Name and Address Book), one or more in-boxes for all incoming mail to the server, and the server log
- Type 2: Domino Server Transaction log, for which disk activities are primarily short sequential writes to the log file to commit the transactions.
- Type 3: Dedicated storage to hold all the users' mail files, which can be distributed to multiple physical volumes (i.e., arrays) with the use of directory or file links

Type 3 storage generates the majority of all disk I/O activity; therefore, this analysis will compare performance resulting from changes only to the Type 3 storage while keeping the other types of file storage constant.

RAID-5 is often used because of its lower cost compared to RAID-1 and because of increased reliability compared to RAID-0. However, RAID-5 requires overhead to manage the checksum necessary for data protection. This reduces performance, especially for environments that perform frequent write operations. For applications such as Lotus Notes that perform frequent write operations, RAID-1 should be used when optimal performance is desired. We used RAID-1E[1] for these performance studies.

[1] RAID-1E is an IBM-exclusive version of RAID-1. RAID-1E ensures that each block of data is mirrored onto another physical drive within an array. This is somewhat different from RAID-1 where the entire drive is mirrored, and RAID-10 where an entire group of drives is mirrored. RAID-1E allows users to configure odd numbers of drives into an array. RAID-1 and RAID-10 can only be configured with pairs of drives.

Study One: Array efficiency vs. size

Methodology

To determine array efficiency, we measured the maximum number of mail users that can be supported by RAID-1E arrays of different sizes (number of drives in the array). The criterion used to determine efficiency was the number of users supported per disk drive. Each user in the workload performs a prescribed set of activities at a prescribed frequency (see Appendix A for details).

The criteria for supporting a user include an acceptable service level, defined as a 5-second user response time. When the average user response time exceeded 5 seconds, we discarded the results of that run. Although this criterion results in a user capacity higher than is recommended for a production environment, it is routinely used to allow capture of the ”knee” of the curve showing User Response vs. User Count performance. The knee is the point at which the increase in user response time becomes disproportionate to the increase in the number of users (see Figure 3). Beyond this point, it is very difficult to ensure satisfactory user response time in a production environment. Although the Netfinity ServeRAID-3H controller used for the measurements supported array sizes up to 16 drives, we limited the study to array sizes of two to ten drives, which is an economical range given the current capacity (maximum of 10 drives) of external SCSI expansion used.

Measurement results and analysis

A Netfinity 5500 M20 with four 500MHz[2] Pentium® III Xeon™ processors with 1MB L2 cache and 2GB memory was used for collecting data in this study. An IBM EXP15 Storage Expansion Enclosure in a split back-plane configuration was used to hold ten 9.1GB[3] 10K rpm Wide Ultra SCSI drives. The expansion unit was connected to two SCSI channels on the Netfinity ServeRAID-3H controller. A two-disk RAID-1 array in the server held the system boot drive, the Domino Server code and its main data directory. The Notes mail files were stored on the RAID-1E array in the external expansion unit. With this memory and disk configuration, we were able to support approximately 3,500 users.

At the highest user level analyzed, maximum CPU utilization was less than 21 percent. Minimum Available Memory was about 720MB. These levels of CPU and memory usage ensured that the only performance bottleneck at any given time was due to the disk subsystem.

The workload was run using several different drive configurations in which only the RAID-1E array holding the mail files was varied. Measurements were made at array sizes of 2, 4, 5, 7 and 9 drives to approximate an even interval. The user-capacity measured for each configuration was divided by the number of disk drives in the array(s) in order to produce an efficiency matrix (Number of Users Supported per Drive). The message size used for these runs was 10 Kbytes, which is the same size used for Domino R5 Notes Mail benchmark reports.

[2] MHz only measures microprocessor internal clock speed, not application performance. Many factors affect application performance.

[3] When referring to hard drive capacity, GB stands for one thousand million bytes. Total user-accessible capacity may vary depending on operating environments.

Table 1. The Effect on Performance from Increasing the Size of a RAID-1E Array

Number of Drives in the RAID-1E Array	Maximum Number of Users Supported	Number of Users Supported per Drive	Percentage Increase of Users per Additional Drive
2	1,600	800	NA
4	2,300	575	22%
5	2,500	500	19%
7	3,000	429	18%
9	3,350	372	16%

These results show that the percentage gain in users supported is significantly smaller than what you might expect from the addition of more drives. In addition, there is a slight drop in this performance gain as the size of the RAID-1 arrays grows larger. These results beg the question: Could four drives organized as two RAID-1 arrays with two drives each support 3,200 users as against a single four-drive RAID-1 array supporting 2,300 users? The second study provides some insight into the answer.

While this part of the study was planned as a effort to produce sizing-guide information for RAID-1 storage arrays, the result of the analysis produced insight into the efficiency of an array as a function of its size for Lotus Notes application environments

Study Two: Evaluation of various data distribution

Methodology

In the second study, we selected three disk configurations, each consisting of 10 disk drives. All three configurations consisted of one or more RAID-1E arrays of identical sizes. The three disk configurations used included one 10-drive RAID-1E array, two RAID-1E arrays each consisting of five drives, and five RAID-1 arrays each consisting of two drives. We selected user levels that would tolerate high load variability typical of a production environment. After some of the measurements were made, we settled on user levels of 2,000; 3,000; and 4,000.

Based on statistics collected from IBM’s own use of Notes Mail and customers’ feedback, Lotus Development determined that the average message size of Notes Mail is in the range of 50KB to 100KB for year 2000 and beyond. For this reason, we selected a 50KB message size. A larger message size was not used for this analysis due to consideration of file storage required for the study. When using a 50KB message for runs made with the 10-drive RAID-1 array, the mail-file growth left free space in the array of around 30 percent. When runs were made on multiple-array configurations, unevenness in file access further reduced this buffer on one or more arrays in the configuration.

Because we are interested in whether the performance difference is significant at user levels normally associated with a production environment instead of at peak capacity, we measured and compared such metrics as Average Disk Queue Length and Average User Response Time

Measurement results and analysis

The measurements for this second study were made using a Netfinity 8500R server populated with eight 550MHz Pentium III Xeon processors with 2MB L2 cache and 4GB of memory. A two-drive RAID-1 array in the server held Windows NT, Domino server code and the main data directory. A Netfinity EXP200 Storage Expansion Enclosure in a split back-plane configuration held ten 18.2GB 10K rpm Wide Ultra2 SCSI drives on which the mail files were stored. The significant performance difference between this platform and that used in Study One does not affect our analysis because there is no need to correlate the data from the two studies. The difference resulted in larger user capacities seen in the results of Study Two. Appendix B discusses the factors contributing to the performance difference and also presents a comparison of the two study environments for the interested reader.

At 4,000 users, Average CPU Utilization was under 24 percent; Minimum Available Memory was above 2GB; and Average Network Utilization was below 14 percent. These levels ensured that the only performance bottleneck at any given time was due to the disk subsystem.

A few shorthand notations are introduced here to facilitate presentation of the information from this point on.

Notation used:

5x2disk: Five RAID-1 arrays, two drives each

2x5disk: Two RAID-1E arrays, five drives each

1x10disk: One RAID-1E array of 10 drives

To approximate even distribution of disk I/O as mail-file storage organization was changed, we distributed user mail files evenly across all arrays that form the mail-file storage. We depended on the workload's randomization algorithm to spread mail deliveries evenly across all users, resulting in even writes to all mail-files. We checked I/O access distribution by inspecting the:

- Average disk transfers per second
- Total mail-file growth for each data array

Average Disk Transfers per Second for each array in the 5x2disk and 2x5disk configurations varied by no more than 1 percent between the highest and lowest. File-growth analysis showed considerable variability between different arrays in a particular configuration. For the 5x2disk configuration, the difference in file growth per array was 24 percent between the highest and the lowest at 3,000 users. At 4,000 users, this ratio was 29 percent. For the 2x5disk configuration, the same metric was 3 percent at 3,000 users and 6 percent at 4,000 users, respectively.

For the 5x2disk and the 2x5disk configurations, the individual arrays were nearly 60 percent populated at the beginning of the workload run. At the conclusion of the run, the data filled up to about 70 to 80 percent of the array, depending on the user level.

Figure 1 on the next page shows Average Disk Queue Length per disk for the mail-file storage at different user levels for all three configurations. Average Disk Queue Length for an array in a configuration divided by the number of drives per array produces this metric. The lower average queue length (Figure 1) associated with the configuration with more physical arrays contributed to the lower user response time (Figure 2). Figure 2 shows Average User Response Time calculated for every user in the workload.

Figure 1. Effect of User Load for Different Disk Configurations

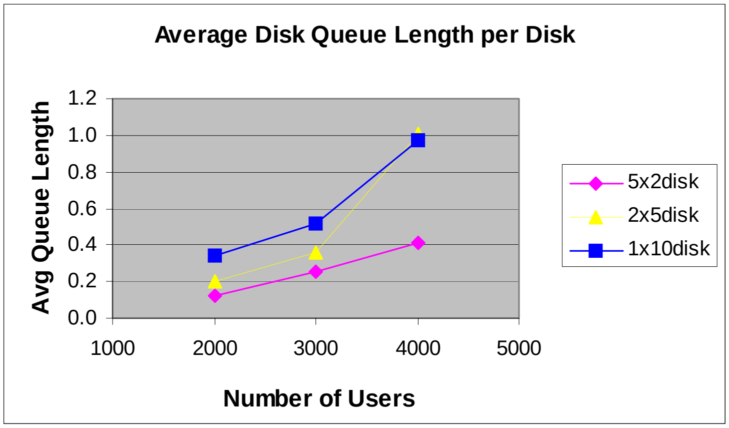


Figure 2. Response Time of Different Disk Configurations

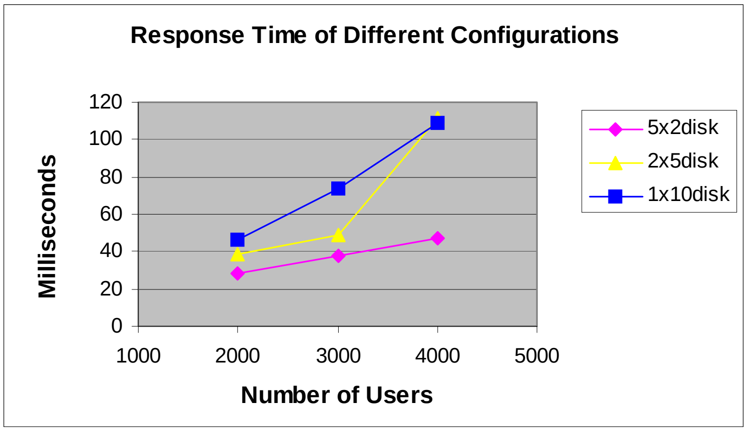


Table 2: Response Time Reduction Compared to 1x10d Configuration

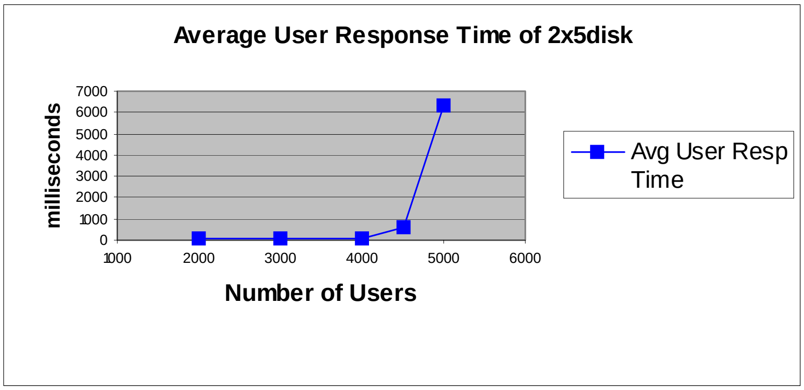
Number of Users	2x5disk	5x2disk
2,000	15%	39%
3,000	33%	49%
4,000	-2%	57%

Figure 2 and Table 2 show that the performance (Average User Response Time) for a fixed number of drives in a RAID-1E array improves (decreases) as the array is reorganized into smaller arrays. There is less certainty in realizing this benefit as the user load approaches 4,000 users. This is most likely caused by our choice of the total number of drives. With our 5x2 drive configuration, users are randomly selected by the workload as mail recipients, subjecting their mail files to write access as mail is delivered to their mail folder. The unevenness of the random algorithm compounded by the large amount of mail generated at 4,000 users contributed to significantly greater disk-head travel in one or more arrays in this configuration. This in turn caused a disproportionately high user response time for users whose mail file was in the more fully populated arrays, and skewed the average user response time.

Validating the optimal user range

Plotting the Average User Response Time for runs made on the 2x5disk configuration shows a knee between 4,000 and 4,500 users, above which response time increased dramatically and non-linearly, making such user loads on this configuration inadvisable for a production environment. The runs made on the 5x2disk and 1x10disk configurations show a similar knee above 4,000 users. Data gathered in Study One showed that a user increase of 3 to 5 percent above the level at which this knee occurs often takes the response time to an unacceptable level. This makes the user level at which the knee occurs a valid upper end for this system configuration in a production environment. We chose 2,000 users as the low end because it is 50 percent of 4,000, approximately the maximum indicated by Lotus' general sizing guideline for the disk configuration used here.

Figure 3. User Response Time for 2x5disk

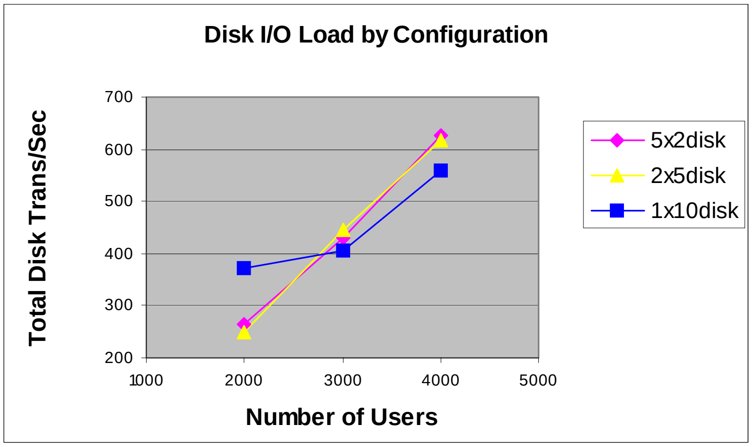


Difference between finding and current Knowledge

Given that the same workload was run for all three configurations, one could expect that the disk I/O load would be similar. Applying the same disk I/O load to all three configurations should produce similar Average User Response Times, unless software introduced a bottleneck. To help confirm this suspicion, we summed the Disk I/Os per second measured for all arrays in a configuration to arrive at a "disk I/O load" for each configuration.

Figure 4 shows that the I/O activity for each of the three configurations is different. Current knowledge of the internals of Domino's interaction with Windows NT does not explain the difference we observed in the disk I/O activity for each of the three configurations.

Figure 4. Mail-File Storage Disk I/O Load



Validating the choice of message size

The processor utilization, network utilization and Average User Response Time calculated for the workload were analyzed for the three user levels with the workload running with three different message sizes. Figures 5, 6 and 7 summarize the effects. As expected, processor utilization increased when user load increased and was not significantly affected by a change in message size. Network utilization increased nearly linearly with user load. At each user load, it also increased as the message size increased. The increase in network utilization when message size increased from 30KB to 50KB was less than when the message size increased from 10KB to 30KB.

Figure 7 shows that Average User Response Time increased slightly as message size increased from 10KB to 30KB, then to 50KB at the 2,000- and 3,000-user levels. This trend was expected due to increasing user load. At 4,000 users, the performance was much more sensitive to the message size change from 30KB to 50KB. This made the 50KB message a better environment to illustrate the performance difference of various mail-file storage configurations.

Validating the cause of performance differences

Figures 5 and 6 show that processor or network utilization contributes to only a small fraction of the server performance differences measured for different user loads. The April 2000 NotesBench Report for the Netfinity 7600 demonstrated that Lotus Domino R5 Server can support more than 10,000 mail users under Windows NT in a single-partition environment. So the application is not the bottleneck for any of the runs made in this study. This leaves the disk subsystem as the main component driving the performance difference in our results.

Figure 5. Effect of Average Message Size on CPU Utilization

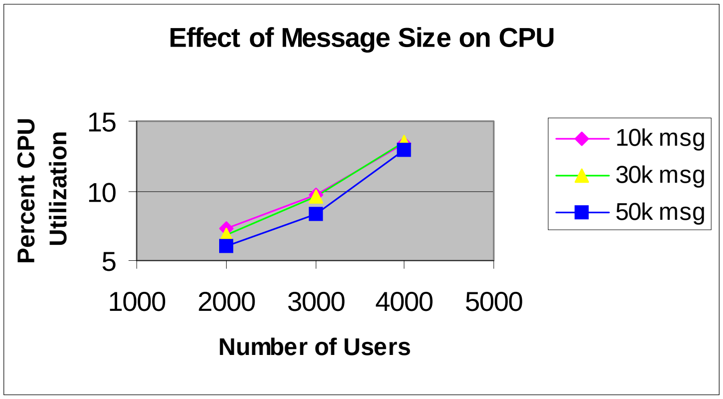


Figure 6. Effect of Average Message Size on Network Utilization

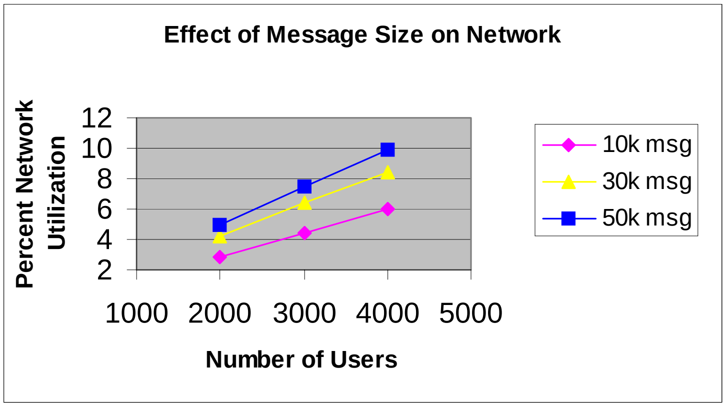
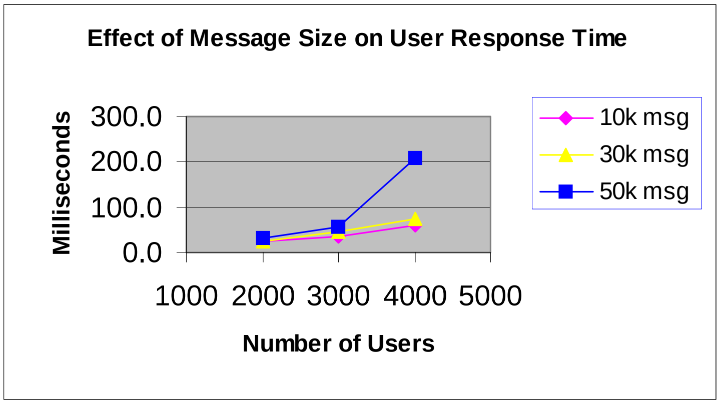


Figure 7. Effect of Average Message Size on User Response Time



Conclusion

Provided that the network and client performance are similar, the administrator will find it easier to ensure that all users enjoy the same level of service if a single volume is used to hold all the mail files.

As the storage requirements for the mail files grow and as drives are added to the data volume, there will be practical and design constraints that limit the number of drives that can be added to the array used as the data volume. The design of many RAID controllers usually limits the number of drives in an array. The external SCSI storage expansion physically limits the number of drives that can be attached to a RAID controller. Faced with these limits, as well as performance concerns when mail-file storage is to be partitioned to facilitate backup, an administrator must consider distributing the data over multiple arrays. This report showed that, for RAID-1E, Average User Response Time improved noticeably as the 10 drives holding the mail files were organized into more arrays. Because of difficulty in achieving even usage among the different physical arrays that constitute the mail-file storage, and the significant difference in disk I/O activity seen in this controlled lab environment, the author does not recommend that anyone should partition existing data volumes for performance gain alone. The gain can be illusive. However, if partitioning is necessary for other reasons and even distribution of disk activity can be arranged, especially for a new install, the effect on server performance is compelling. Finally, in the case where a data volume needs to grow, splitting a large volume can be an attractive alternative to the small incremental performance gain obtained by adding another drive to an existing array.

Note that we have only demonstrated the effect with a RAID-1E array. While there is no reason to believe that the effect should not be seen with a RAID-5 array, the extent of the effect is unknown and the options are not as clear-cut.

The reason for the difference in measured Disk I/O Load (Figure 4) is likely to be buried deep within design of the Domino Server and may not be explained any time soon. With the random algorithm used to select users as mail recipients in the workload, some of the difference may also be due to the run-to-run variability. While our benchmarking experience has shown this variability to be high as we approached the limits of a particular system configuration, we have seen results to be quite repeatable at user loads below 80 percent of the maximum capacity. For this reason, based on this study, we expect that users can see this effect when the user load is between 50 and 70 percent of peak capacity. Until we have done further study and analysis with different applications and workloads, we have no reason to believe that what we see here will occur with a different application.

Any question about the methodology, measurements and analysis, or the conclusion presented here should be directed to the author, Clement Moy, at cmoy@us.ibm.com.

Appendix A: Workload Description

The Mail workload script used in these studies models an active user on a client reading and sending mail, scheduling an appointment, and sending a Calendar and Scheduling invitation. An average user will execute this script four times per hour. For each iteration of the script, there are five documents read, two documents updated, two documents deleted, one view scrolling operation, one database opened and closed, and one view opened and closed. For every sixth iteration of the script, three messages comprised of one memo to three recipients, three lookups against the Domino Server Directory (formerly the Name and Address Book), one appointment, and one invitation are sent to three recipients randomly selected from the directory on the Domino Server. The mail file of all users listed in the directory will receive mail and growth during this workload, regardless of whether the user is connected to the Domino server.

Appendix B: Comparison of Environments Used in the Studies

The measurements for Study Two were made using a Netfinity 8500R server populated with eight 550MHz Pentium III Xeon processors with 2MB L2 cache and 4GB of memory. Since there was no need to correlate the measurement results from the two studies, the file storage structure used in Study Two was different from that used in Study One. Windows NT, Domino Server code and the main data directory were placed in the boot volume consisting of a 4GB partition of a two-disk RAID-1 array. The second partition of this RAID-1 array was not used during execution of the workload. The mail files were off-loaded from the main data directory to ten 18.2GB 10K rpm Wide Ultra2 SCSI drives through the use of file links.

The Netfinity 8500R configuration used two cache coherency filters, one for each of the two processor buses. Besides helping SMP performance due to general instruction-execution, the coherency filters also improved performance of instructions that read disk data into memory. For a disk I/O-intensive workload like Notes Mail, it increased the number of users supported, using the same criterion as for Study One, on the same number of disk drives.

In Study One, we used 9.1GB 10K rpm Wide Ultra SCSI drives. For Study Two, we used 18.2GB 10K rpm Wide Ultra2 SCSI drives. While there is an insignificant performance difference between these two drives, the same amount of data occupies far less disk cylinders on an 18.2GB drive than on a 9.1GB drive. This produces a comparatively shorter average head travel during a disk seek for the same number of files, resulting in better average disk I/O performance. As a result, we were able to support 4,500 users with an Average User Response Time of less than one second for the 10-disk RAID-1 configuration, vs. somewhere not much larger than 3,500 users in Study One.

In case the reader noticed the difference in the upper end of the user range employed in Study One and Study Two, this explains why we seem to be getting better performance throughout Study Two. Since no comparison is made between results from the two different studies, we were free to use whatever system was available.

Additional Information

Information on IBM Netfinity direction, products and services is available at http://www.pc.ibm.com/netfinity.

NotesBench Reports

The Notesbench Consortium Web site at http://www.notesbench.org
IBM Netfinity Server Benchmark Web site at http://www.pc.ibm.com/ww/netfinity/benchmarks

Other Technical Resources Related to Netfinity Servers

Visit http://www.pc.ibm.com/us/netfinity/tech_library.html.

IBM Personal Computer Company
Department LO6A
3039 Cornwallis Road

Research Triangle Park
NC 27709
Printed in the United States of America

6-00

THE INFORMATION CONTAINED IN THIS DOCUMENT IS DISTRIBUTED ON AN AS IS BASIS WITHOUT ANY WARRANTY EITHER EXPRESSED OR IMPLIED. The use of this information or the implementation of any of these techniques is the customer's responsibility and depends on the customer's ability to evaluate and integrate them into the customer's operational environment. While each item has been reviewed by IBM for accuracy in a specific situation, there is no guarantee that the same or similar results will be obtained elsewhere. Customers attempting to adapt these techniques to their own environment do so at their own risk

For terms and conditions or copies of IBM's limited warranty, call 1 800 772-2227 in the U.S. Limited warranty includes International Warranty Service in those countries where this product is sold by IBM or IBM Business Partners (registration required).

References in this publication to IBM products or services do not imply that IBM intends to make them available in all countries in which IBM operates. IBM reserves the right to change specifications or other product information without notice.

IBM Netfinity systems are assembled in the U.S., Great Britain, Japan, Australia and Brazil and are comprised of U.S. and non-U.S. components.