Is The Bang Worth the Buck? A RAID Performance Study

Susan E. Hauser, Lewis E. Berman, George R. Thoma
Lister Hill National Center for Biomedical Communications
National Library of Medicine
Bethesda, Maryland 20894
hauser@nlm.nih.gov
Tel: 301-435-3209
Fax: 301-402-0341

Abstract

Expecting a high data delivery rate as well as data protection, the Lister Hill National Center for Biomedical Communications procured a RAID system to house image files for image delivery applications. A study was undertaken to determine the configuration of the RAID system that would provide for the fastest retrieval of image files. Average retrieval times with single and with concurrent users were measured for several stripe widths and several numbers of disks for RAID levels 0, 0+1 and 5. These are compared to each other and to average retrieval times for non-RAID configurations of the same hardware. Although the study in ongoing, a few conclusions have emerged regarding the tradeoffs among the different configurations with respect to file retrieval speed and cost.

Rationale and goals

The Lister Hill National Center for Biomedical Communications, a research and development division of the National Library of Medicine, procured a Sun SPARCstorage Array (SSA), model 101, to house image files for prototype image delivery applications. The SSA model 101 is configured with eighteen Seagate ST31200W 1.05 GB disks connected to six internal fast wide SCSI busses. The SSA is connected to a Sun SPARCstation 20 via a Fiber Channel port. SPARCstorage Volume Manager software supports use of the SSA as independent volumes or as:

RAID 0: Data is split into equal sized blocks, or stripes, and distributed among the disks in the RAID volume.
RAID 1: All data in a volume are duplicated on the mirror volume.
RAID 0+1: Both the original volume and the mirror volume are striped.
RAID 5: In addition to data blocks, RAID Level 5 includes parity blocks, which are distributed among the disks in the RAID volume [1,2].

The specifications of the Seagate disks [3] in the SSA cite a data transfer rate of 3.3 to 5.9 MB/sec. The fast wide SCSI interface has a data transfer rate of 20 MB/sec, and the Fiber Channel connector has a data transfer rate of 25 to 50 MB/sec. Those specifications the following statements from a technical white paper led us to expect very high data retrieval rates in addition to the data security available from RAID.

> "Each of the disks in a stripe are generally assumed to be on their own independent data channel, allowing the transfer rate of a RAID 0 implementation to approach the sum of the transfer rates of each of the drives." [4]
>
> " Both SPARCstorage Array models ... are capable of over 2000 two-KB input-output operations per second, and sustained transfer rates exceeding 15 MB/second." [4]

One goal of the study was to determine the optimum configuration and stripe width for fast retrieval of a variety of file sizes. Documentation from Sun [5] and other sources [1] mention the importance of "tuning" the RAID to the data and application through choices in RAID level and stripe width. Yet the guidelines for selecting these, especially for selecting stripe width, are general. One suggestion is to set the stripe width to be the length of a disk track. However, although the specifications of the Seagate drives in the SSA state that the average is 84 sectors per track, one can deduce from those specifications that the track length varies from about 72 sectors per track to about 127 sectors per track. Another suggestion is to select the stripe width such that the stripe width times the number of disks exactly matches the size of the I/O requests at the application layer. However, the SSA is intended for use with applications that read entire files of a variety of sizes into memory at once.

Another goal of the study was to determine the optimum configuration of the SSA for rapid retrieval of files by the Medical Information Retrieval System (MIRS) server program. MIRS is a client/server application that provides Internet access to biomedical databases, including X-ray medical images [6]. The SSA stores lower resolution gif format versions of high resolution digitized X-ray images. One goal is to quickly display several of the lower resolution images that match a patron's search criteria. The application reads appropriate images into server memory where they are concatenated and transmitted to the client as one file. The distribution of MIRS image file sizes is shown in Figure 1.

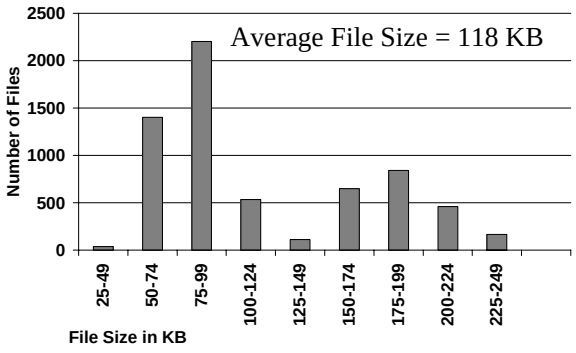

Figure 1. MIRS File Size Distribution

Study conditions

Six of the eighteen disks in the SSA, each attached to a separate SCSI bus, were used for the study. We measured performance of the RAID subsystem alone by removing such factors as reading from cache or swap space, and heavy system loads. The study measured the average time to read files from the SSA into system memory. The study concentrated

on measuring retrieval times for a single process reading files sequentially, and retrieval times for multiple processes reading files concurrently, varied by differing RAID configurations and stripe widths.

Preliminary study

In the first study we assumed that retrieval time for sequential reads was based on two performance components, Average Positioning Time and Data Transfer Rate [7], and attempted to determine these two performance indicators for the various configurations. This was done by measuring the average retrieval time for various file sizes and performing linear regression of average retrieval time as a function of file size. Two results of the linear regression are intercept, which translates to Average Positioning Time, and slope. The units of slope are seconds per byte, so the inverse of slope is the Data Transfer Rate in bytes per second.

To measure retrieval time for concurrent reads, the average retrieval time for a mix of file sizes was measured while none to several other processes were also retrieving files of the same mix of file sizes. In this case, linear regression was performed on average retrieval time as a function of the number of concurrent processes. The calculated slope of the linear relationship is the Average Additional retrieval Time per File per additional concurrent Process. Average Additional Time per File per Process is our performance indicator for concurrent processes.

The three performance indicators were measured for several configurations of the six disks in the SSA for two ranges of files sizes. The smaller range, from 50 KB to 275 KB was similar to the range of files used by MIRS. The larger range, from 1 MB to 12.5 MB, was used to determine if the optimum configuration depended on file size.

For most RAID configurations that were measured, narrower stripes yielded larger data transfer rates for sequential reads for both small and large files. Wider stripes resulted in lower data transfer rates for sequential reads, but also less additional retrieval time per file per concurrent process. The generalization holds for the case where the six drives are configured as independent non-RAID, or “simple”, volumes. A simple volume can be considered as a volume with one very wide stripe. As simple volumes, the six drives had the lowest data transfer rate and the lowest additional time per file per concurrent process. The results suggest there is a tradeoff between optimizing for sequential reads and optimizing for concurrent reads.

We also found that stripe widths less than 16 KB gave results similar to 16 KB, and stripe widths greater than 160 KB gave results similar to 160 KB. Between these two widths the changes in data transfer rate and average additional time per file per process appeared to be a monotone decreasing function of stripe width.

The maximum system throughput measured was 8.2 MB per second, which occurred with 8 processes concurrently retrieving unique files with an average size of 6.42 MB. When 6

processes retrieved files with an average size of 123 KB, the system throughput was 3.9 MB per second, the maximum measured for the smaller file sizes. These were both achieved by distributing files between two 3-disk RAID 0 volumes with a 160 KB stripe.

There were large differences in Average Positioning Time and Data Transfer Rate between the data from the small and large files sizes for a given RAID configuration. For a 6-disk RAID 0 volume and the larger files, the calculated Data Transfer Rate ranged from 5.6 MB per second to 8.3 MB per second. For the smaller files, the calculated Data Transfer Rate ranged from 3.1 MB per second to 4.8 MB per second. For both sizes, Data Transfer Rate decreased with increasing stripe width. Average Positioning Time also varied by several hundred percent, but did not appear to be a function of configuration or stripe width. We concluded that the combined effect of zone bit recording [8,9] and data striping disallowed a simple linear relationship between file size and retrieval time.

Procedures for successive studies

With knowledge gained from the initial study, we modified our performance indicators and proceeded to study the SSA performance for file sizes in the range of the MIRS data, knowing that conclusions would include a caveat about file size. The performance indicators became the Average Retrieval Time and, again, the Average Additional Time per File per Process. Average Retrieval Time is the average time to read a file into memory as measured from a single process sequentially reading files of all sizes. Average Additional Time per File per Process is the same as for the preliminary study.

A typical test set consisted of the following steps:

1. Create a volume or volumes in the configuration to be measured.

2. Fill the volume(s) with files in ten sizes from 50 KB to 275 KB. Use an equal number of files of each size, for an average file size of 162.5 KB. To minimize the effect of zone bit recording, distribute files of each size over all portions of the volume.

3. Create one randomized list of all files on the volume(s). Create twelve randomized lists, each containing approximately one twelfth of the files on the volume(s) and an equal number of each file size.

4. To determine Average Retrieval Time, a program sequentially reads every file in the one large randomized list into memory, measuring the time required to open the file and read in into memory. When all of the files are read, the program calculates the mean retrieval time, standard deviation, maximum and minimum. The sample size, calculated statistics and time of day are recorded in an output file. The program is run several times for a total sample size of several thousand.

5. To determine Average Additional Time per File per Process, a program reads all of the files from the first small randomized list onto memory, measuring the time to read each file. Then two programs run concurrently, each reading files from a different small randomized list. Then three programs run concurrently, each reading a different list of files, and so forth up to twelve programs. The same statistics described above are recorded by each program in an output file. The series is run several times for a total sample size of several hundred for each case.

Average Retrieval Time is the grand average of all the runs using the one large list of files. Average Additional Time per File per Process is determined by first calculating the grand average retrieval time for each case of concurrence, then performing a linear regression of average retrieval time as a function of the number of concurrent processes. The slope of the line returned by the regression is the Average Additional Time per File per Process.

Results

Using the procedures outlined above, we obtained the two performance indicators for the following configurations:

Three simple volumes
Six simple volumes
RAID 0 volumes with 4, 5 and 6 disks
Two 3-disk RAID 0 volumes
RAID 5 volume with 6 drives
RAID 0+1 volume with 6 drives (3 drives, mirrored)

Because of the information obtained in the preliminary study, we used only two stripe widths for the RAID configurations tested: 16 KB and 160 KB.

Figures 2 and 3 compare the results from RAID 0 volumes with 4, 5 or 6 disks. Average Retrieval Time is smaller for the narrow stripe width and also for fewer disks in the volume. Conversely, Average Additional Time per File per Process is smaller for the wider stripe width and for more disks in the volume. Again we see a potential tradeoff between optimizing for a single process and optimizing for concurrent processes.

Figures 4 and 5 compare the results from three configurations using six disks. Although two of these are RAID, none offer fault tolerance. The tradeoff between narrow and wide stripe width is still evident. Although either of the RAID configurations is faster for single processes, configuring the six disks as simple volumes is better for concurrent processes.

Figures 6 and 7 compare the results for the two configurations of six disks that offer fault tolerance to the results for six disks as simple volumes. The mirrored, striped volume (RAID 0+1) offers speed comparable to simple volumes plus the security of data redundancy, at the cost of requiring twice as much media.

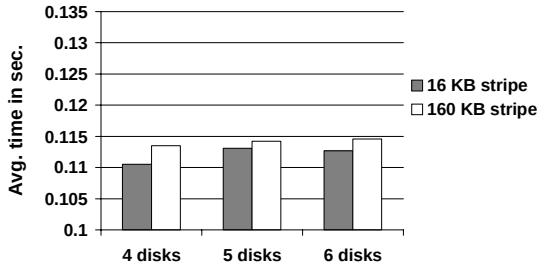

Figure 2. Average Retrieval Time, RAID 0, 3 volume sizes

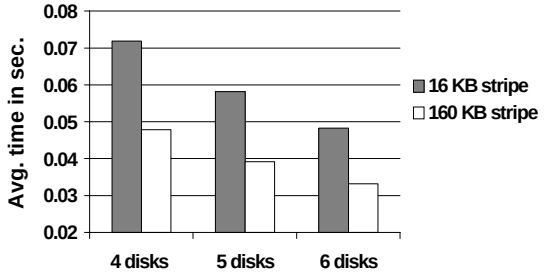

Figure 3. Additional Time per File per Process, RAID 0, 3 volume sizes

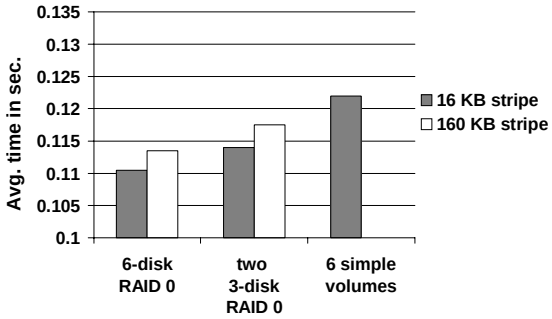

Figure 4. Average Retrieval Time, 6 disks in 3 configurations

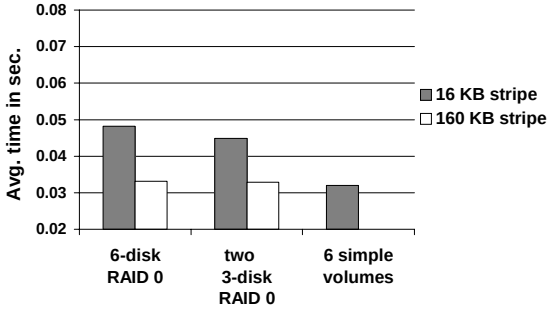

Figure 5. Additional Time per File per Process, 6 disks in 3 configurations

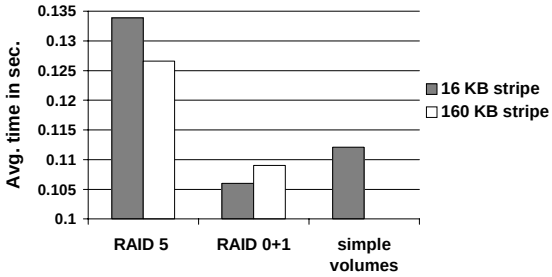

Figure 6. Average Retrieval Time, 6 disks in redundant configurations

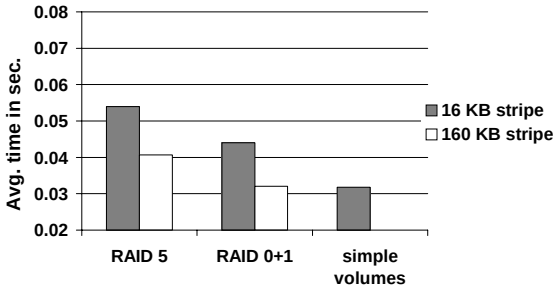

Figure 7. Additional Time per File per Process, 6 disks in redundant configurations

The maximum system throughput achieved during these tests was 4.3 MB per second, which is less than the specified maximum transfer rate of a single drive. That occurred for the RAID 0+1 configuration with 160 KB stripes, with 12 concurrent processes retrieving files with an average file size of 162.5 KB. Evidently, for files of this size the combined latencies of disk drive, SCSI and Fiber Channel interfaces and operating system overhead are great enough to counterbalance increased data transmission rates.

Configuration selection

Even if it is anticipated that access to the dataset will always be sequential reads by a single process, the choice of configuration may not be trivial. If fast retrieval is needed at

any cost, RAID 0+1 provides the fastest sequential retrieval times, excellent concurrent performance and the security of mirrored data. If cost is a consideration and some fault tolerance is required, RAID 5 is the only choice, even though it is not among the best performers for either sequential or concurrent retrieval. If cost is a consideration and fault tolerance is not, distributing the dataset across 4-disk RAID 0 volumes is a good choice.

It is more difficult to determine the optimum configuration if both sequential and concurrent access to the dataset is anticipated. In this case, cost, fault tolerance and system management issues may be more important that retrieval time, especially for small files.

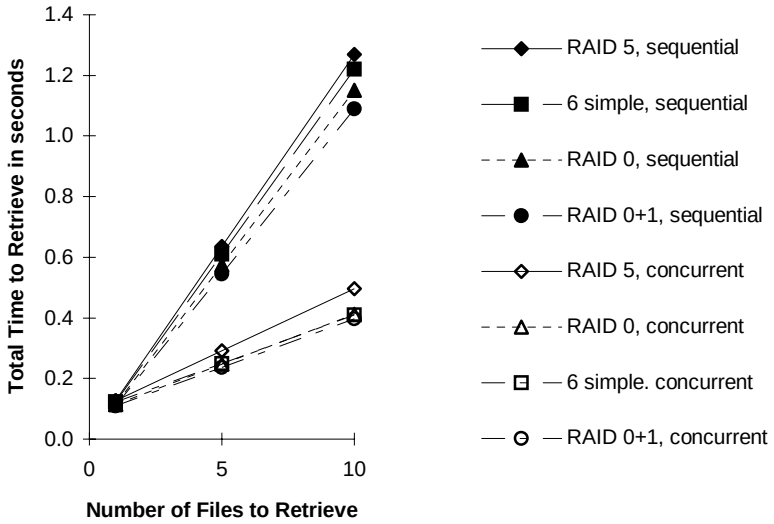

Figure 8. Total Retrieval Time of Average Size Files

Figure 8 plots total retrieval time as a function of the number of 162.5 KB files retrieved for four six-disk configurations. It shows the total time for sequential reads and the total time for concurrent reads. The lines on the graph are in the same order as the descriptions in the legend. For the occasional retrieval of a single file, any configuration of the array yields about the same retrieval time. For applications that retrieve several files at a time, concurrent access improves total retrieval time more than any RAID configuration. This is illustrated by comparing the slowest concurrent access case, RAID 5, to the fastest sequential access case, RAID 0+1. However, the difference between the best and worst configuration for either kind of access is less than 1.2 seconds for up to ten files, which may be inconsequential for many applications.

What about "Bang" and "Bucks"?

We define:

Bang = Average file size /
(Average Retrieval Time + Average Additional Time per File per Process)

Bang increases with improvement in either of the performance indicators, and gives equal weight to each. The average file size in the numerator balances the larger performance indicators that would result from larger files. This definition of Bang is only useful for

quantifying file retrieval speed. It provides no information regarding write speed, or security or system reliability, which are less easily quantified. Table I shows Bang calculated for the configurations in this study, in order of decreasing value of Bang. The average file size for the study was 162.5 KB.

Table I: Bang for several configurations

Configuration	Stripe width	Average Retrieval Time (sec)	Additional Time per File per Process (sec)	BANG
3-disk RAID 0+1 (mirrored)	160 KB	.109	.032	1.152
6-disk RAID 0	160 KB	.115	.033	1.098
3-disk RAID 0+1 (mirrored)	16 KB	.106	.044	1.083
two 3-disk RAID 0s	160 KB	.118	.033	1.076
5-disk RAID 0	160 KB	.114	.039	1.062
6 simple volumes	NA	.122	.032	1.055
two 3-disk RAID 0s	16 KB	.114	.045	1.022
6-disk RAID 0	16 KB	.113	.048	1.009
4-disk RAID 0	160 KB	.114	.048	1.003
6-disk RAID 5	160 KB	.127	.041	0.967
5-disk RAID 0	16 KB	.113	.058	0.950
3 simple volumes	NA	.116	.057	0.939
4-disk RAID 0	16 KB	.110	.072	0.893
6-disk RAID 5	16 KB	.134	.054	0.864

The price for the SSA model 101 with 18 1.05 Gbyte disks, and SBUS to Fiber Channel host adapter was $26,733. The same hardware capabilities without the RAID management features would have been approximately $20,000. Each disk provides approximately 863 MB of space for user data, whether formatted as a simple volume or as part of a RAID volume. Thus 18 disks offer a total of 15.534 GBytes of data storage when configured as simple volumes or as RAID 0, 12.945 GBytes when configured in 6-disk RAID 5 volumes, or 7.767 GBytes when configured in 3-disk RAID 0+1 volumes. Table II shows the calculated Bang per Buck per Gbyte of data storage, for four configurations of the SSA or of the equivalent hardware.

Table II: Bang per Buck per Gigabyte

	Bang	Thousands of $ per Gbyte of data storage	Bang / K$ / GB
3-disk RAID 0+1	1.152	3.442	0.335
6-disk RAID 5	0.967	2.065	0.468
6-disk RAID 0	1.098	1.721	0.638
simple volumes	1.055	1.545	0.683

Summary and conclusions

For retrieval of files of a few hundred KB or less, Bang alone is not worth the Bucks. RAID offers many other attractive features, such as fault tolerance, ease of storage management, and, in many cases, a compact, well designed peripheral. If the subsystem is just one component of a large system, the extra cost of RAID may be worth these conveniences alone.

For the MIRS application, where a set of files between 50 KB to 275 KB must reside in fault tolerant storage that maintains the retrieval speeds that are available from hardware alone, there is no choice but RAID 0+1, even though it is expensive. RAID 0+1 is also the choice if fast retrieval is of primary importance and cost is not. Although narrow stripes produce slightly faster sequential retrieval times and wide stripes produce slightly faster concurrent retrieval times, the performance difference between wide and narrow stripes for this range of file sizes is so small that any choice would be acceptable. We recommend that the database of lower resolution MIRS images reside on a RAID 0+1 volume with a 16 KB stripe width. Because the MIRS application software reads images sequentially, the narrow stripe should give slightly faster retrieval times.

For applications where fault tolerance is required, and retrieval speeds can be slower than those available from hardware alone, RAID 5 is the best choice. For RAID 5, wider stripes appear to improve both sequential access speed and concurrent access speed for files in this range.

If either RAID 0+1 or RAID 5 is selected, retrieval times may be faster for more or fewer than six disks per volume. We plan to measure the performance several configurations in the next phase of the study.

For applications where fault tolerance is not important and funds are limited, balancing the load across several volumes without the benefit of RAID management can yield fast retrieval speeds at significant cost savings.

We find no reasons for choosing RAID 0 for applications involving small files. The slight performance advantage for sequential file retrieval is offset by the cost of the RAID management capabilities and the reliability risk incurred by distributing each file across several disks.

References

1. National Peripherals, Inc., “An Introduction to Disk Arrays and RAID Levels.”

2. Procom Technology, Inc., http://www.procom.com/homepage/raid_def.html, RAID definitions, 1996.

3. Seagate Technology, Inc., http://www.seagate.com/tech/techtop.shtml, Technical specifications for drive model ST31200W, 1995.

4. Sun Microsystems, Inc., “The SPARCstorage Array Architecture, Technical White Paper,” February, 1995.

5. Sun Microsystems, Inc., “SPARCstorage Array Performance Brief, Technical White Paper,” Revision 2, July, 1994.

6. Long LR, et.al., “A Prototype Client/Server Application for Biomedical Text/Image Retrieval on the Internet,” Proceedings of SPIE Storage and Retrieval for Still Image and Video Databases IV, San Jose, CA, February 1-2, 1996, pp.362-372.

7. Chen PM and Lee KL, “Striping in a RAID Level 5 Disk Array,” Technical Report CSE-TR-181-93, University of Michigan, November 1993.

8. Sun Microsystems, Inc., Technical Product Marketing, “Configuration Planning for Sun Servers,” Third Edition, January, 1994.

9. Van Meter R, et.al., comp.arch.storage FAQ, Subject [6.1], June, 1996.