CRI RAID

Robert A. Albers, Jr., Cray Research, Inc., Eagan, Minnesota

ABSTRACT: *Cray Research has several RAID offerings. This includes RAID implemented at both the software and hardware levels. As part of our hardware RAID offerings, the DCA-3 Disk Arrays and the Network Disk Arrays (HIPPI Disks) are covered, including their features and performance. In addition, Network Disk Array performance, space utilization, and system configuration recommendations are also given. For a more in depth discussion of Network Disk Array issues, see SN-2185, the "Network Disk Array (HIPPI Disk) Configuration Options And Performance Technical Note".*

1 INTRODUCTION

Cray Research has several RAID offerings. These are presented in the context of the various RAID levels that exist. These offerings include both software and hardware implementations of RAID. For RAID implemented at the hardware level, the DA-60, DA-301, and ND-12 are covered. Preliminary information for the ND-40 is also covered. As will be seen from this information, the DCA-3 Disk Arrays (DA series disks) are easy to use whereas the Network Disk Arrays (ND series disks) are complex devices and thus harder to use. For a further discussion of Network Disk Arrays, see Technical Note SN-2185, "Network Disk Array (HIPPI Disk) Configuration Options And Performance Technical Note".

2 RAID LEVELS

The following RAID (Redundant Array of Independent Disks) levels exist.

- RAID-0 - striped data (without parity)
- RAID-1 - mirrored data
- RAID-2 - bit striped data with SECDED (Hamming code ECC)
- RAID-3 - bit or byte striped data with a dedicated parity disk
- RAID-4 - block striped data with a dedicated parity disk; has an associated write penalty
- RAID-5 - block striped data with distributed parity; has an associated write penalty

3 SOFTWARE RAID

The following UNICOS features support the given RAID level.

- Striped disk device (sdd) - RAID-0
- Mirrored disk device (mdd) - RAID-1

For added resiliency, the alternate path feature also exists. This is where a disk is accessible through another I/O path.

4 HARDWARE RAID

The following hardware supports RAID at the stated level.

4.1 Early Disk Arrays

- DC-40/41/42 disk controllers - RAID-0

  Supported the following disks:
  DD-40
  DD-41
  DD-42

4.2 DCA-3 Disk Arrays

- DCA-3 disk channel adapter - 4 + P RAID-3

  Supports the following disks:
  DA-62
  DA-60
  DA-301
  DA-302 (April 1995)

The DA series disk arrays are DCA-3 disk channel adapter based disks. They are general purpose devices and are easy to use. They have high reliability. There are no significant configuration issues with them. They have high performance across a range of transfer sizes. They are directly attached to the IOS-E.

4.3 Network Disk Arrays

- Network disk arrays (HIPPI disk device (hdd))

  ND-12/14
  ND-40 (July 1995)

  Support the following RAID levels:
  RAID-1
  RAID-5
  RAID-3 (ND-40)

As an added resiliency feature, there are dual HIPPI ports on the ND-12. The ND-40 supports up to four HIPPI ports. The disk can be reached from any of the HIPPI channels connected to it.

The ND series network disk arrays are special purpose disks. They are complex devices - a person needs to understand them to use them. They have high reliability due to RAID. There are significant configuration issues with them due to the many configuration options available for them. They have low performance for small transfers and high performance for large transfers. They can be directly attached to an IOS-E, Y-MP EL, or J90, or exist on a network attached through a HIPPI switch.

4.3.1 ND-12

The ND-12 consists of data modules. There are 10 data modules in a bank in a ND-12. There are 2 disk spindles per data module which are connected through a device module controller (DMC) in the ND-12. These data modules can be grouped together to form a facility. In a 8 + P + S RAID-5 configuration, 8 data modules are used for user data, one data module is used for parity, and one data module is used as the standby (hot spare) data module. Note, however, that parity is spread across all 9 (8 + P) data modules being used. Parity accounts for only one data module worth of space, however. For a picture of this see Figure 1.

The following two items are of importance in RAID-5 mode. For more information on these issues, refer to SN-2185, the "Network Disk Array (HIPPI Disk) Configuration Options And Performance Technical Note".

4.3.1.1 Parity Group

In RAID-5 mode, a parity group is a set of disk spindles that share common parity. This is equal to the number of data modules in a facility. In the case of an 8 + P configuration, this is the number of data modules used for user data which is 8. Thus, a parity group consists of 8 spindles in this case. The amount of usable space associated with this is 8 times the logical sector size. A parity group is half a stripe width. A parity group consists of the gray disks (parity included) in Figure 1.

4.3.1.2 Stripe Width

In RAID-5 mode, a stripe width is a set of sequential logical sectors distributed across all the disk spindles in a facility. This is two times the number of data modules in a facility since there are two spindles per data module. In the case of an 8 + P configuration, this is equal to the number of data modules used for user data times two which is 16. Thus, a stripe width consists of 16 spindles (8 times 2) in this case. The amount of usable space associated with this is 16 times the logical sector size. A stripe width consists of two parity groups. A stripe width consists of the gray and black disks (parity included) in Figure 1.

4.3.1.3 Configuration Options

The following ND-12 configuration options exist.

- 8 + P (9) + S
- 9 + P (10)
- Dual 4 + P
- User defined
- RAID partitioning
- 64 KByte sector
- 32 KByte sector
- RAID-1
- RAID-5
- Read Parity Check On

4.3.2 ND-40

A ND-40 bank consists of 12 data modules. There are 2 disk spindles per data module which are connected through a device module controller (DMC) in the ND-40.

5 TEST ENVIRONMENT

The test environment used consisted of the following.

- CRAY Y-MP/2E
- 2 CPUs
- 16 MWords of memory
- 128 MWord Solid-state Storage Device (SSD) with 1 Very High Speed (VHISP) channel
- IOS-E with 1 IOC

  1 HISP channel from the IOC to memory

  1 HISP channel from the IOC to SSD
- The DA series disks were connected to a DCA-3 channel adapter
- The ND series disks were connected to a NSC PS-32 HIPPI switch which was connected to a HCA-3/4 channel adapter; they were also connected directly to a HCA-3/4 channel adapter
- The disks were used in the secondary allocation area of a primary/secondary allocation area file system with SSD as the primary allocation area
- UNICOS 7.C
- IOS-E 8.0

The testing period ran from March 1994 to the present.

6 TEST DESCRIPTIONS

Two tests were used for performance measurement and evaluation. These tests were designed to measure sustained peak transfer rates and aggregate I/O performance. Unless noted otherwise, they started timing at the first request processed and ended timing at the last request processed. They were written in C. File space was preallocated using setf.

The disk tested was used in the secondary allocation area of a primary/secondary allocation area file system. As a result, file system (metadata) I/O was not measured. Also as a result, for

the ND series disks, the file was aligned on a stripe width boundary.

One test did sequential raw asynchronous I/O to a file using writea/reada. Either one buffer per file or four buffers per file were used. Transfer sizes ranged from one sector to the number of sectors it took to equal one MByte. In several cases, transfers beyond one MByte were done. 100 transfers were done in each direction. Flaw free contiguous space was used on the DA series disks for this test.

The other one did random raw asynchronous I/O to a file using listio. The random requests were uniformly distributed across the entire disk. The number of outstanding requests ranged from one to 64. The transfer size was one sector. 400 or 10,000 transfers were done in each direction.

7 TEST RESULTS

7.1 DA-60

Only sequential performance was evaluated for the DA-60. 100 transfers of 1 through 16 sectors (64 KBytes through 1 MByte) were done in each direction. This I/O was primed, meaning the first request in each direction was discarded so as to measure sustained peak performance. Sustained peak performance along with single sector results are in the following table.

Table 1.

DA-60		
Sequential I/O	**Write**	**Read**
Sustained Peak MBytes/Sec.	80.4	80.5
Single Sector MBytes/Sec.	65.5	80.5
Sector I/Os/Sec.	1002	1231

For this test, performance was pretty much level across the whole range of transfer sizes. Single sector write performance started out somewhat slowly due to IOS-E channel buffer utilization constraints. For a graph of this data see Figure 2.

7.2 DA-301

Sequential and random performance were evaluated for the DA-301. For sequential I/O, 100 transfers of 1 through 32 sectors (16 KBytes through 512 KBytes) were done in each direction. This I/O was primed meaning the first request in each direction was discarded so as to measure sustained peak performance. Sustained peak performance along with single sector results are in the following table.

Table 2.

DA-301		
SEQUENTIAL I/O	**WRITE**	**READ**
Sustained Peak MBytes/Sec.	34.6	32.9
Single Sector MBytes/Sec.	30.3	32.9
Sector I/Os/Sec.	1858	2016

For this test, performance was level across the whole range of transfer sizes. For a graph of this data see Figure 3.

For random I/O, 400 transfers of one sector were done in each direction. One all the way up to 64 outstanding requests were measured so as to measure sustained peak performance. Sustained peak performance and one outstanding request results are in the following table.

Table 3.

DA-301		
RANDOM I/O	**WRITE**	**READ**
Sustained Peak I/Os/Sec.	58	57
1 Outstanding Request I/Os/Sec.	58	**52**

As can be seen, performance was pretty much equivalent for one versus many outstanding random I/O requests.

7.3 ND-12

For results and analysis of ND-12 performance for all configuration options, refer to SN-2185, "Network Disk Array (HIPPI Disk) Configuration Options And Performance Technical Note".

Sequential and random performance were evaluated for the ND-12. For sequential I/O, 100 transfers of 1 through 32 sectors (32 KBytes through 1 MByte) were done in each direction. Both single buffered and quadruple buffered tests to measure sustained peak performance were run. This was done for both RAID-5 and RAID-1. RAID-5 sustained peak performance and single sector single buffered performance are in the following table.

For the graphs of this data see Figures 4 and 5. For this test, performance increased gradually as the transfer size increased. The spikes on RAID-5 writes are for writes that occur on a parity group boundary and thus don't incur the RAID-5 write penalty. For single buffered I/O, RAID-5 versus RAID-1 performance is pretty much equal except for the RAID-5 parity group writes which are better. For quadruple buffered I/O (4 outstanding requests), RAID-1 write performance for smaller transfers is better, with RAID-5 performance being better in the rest of the cases.

A sector size of 32 KBytes was also compared to a sector size of 64 KBytes. This data is presented in a graph in Figure 6. As can be seen from this graph, for transfers through one MByte, a sector size of 32 KBytes delivers better performance. This is because transfers are striped across multiple disk spindles sooner with a sector size of 32 KBytes. This sector size also has more RAID-5 parity group write opportunities.

Table 4.

ND-12 8 + P 32 KBYTE SECTOR RAID-5		
SEQUENTIAL I/O	**WRITE**	**READ**
Sustained Peak MBytes/Sec.	44.6	50.3
Single Sector MBytes/Sec.	.8	1.6
Sector I/Os/Sec.	23	50

ND-12 performance was also compared to DA-301 performance in the graph in Figure 7. The different performance curves of these devices is clearly seen here. The DA-301 performance curve is essentially flat whereas the ND-12 performance curve gradually increases with transfer size.

For random I/O, either 400 transfers or 10,000 transfers of one sector were done in each direction. One all the way up to 64 outstanding requests were measured so as to measure sustained peak performance. Sustained peak performance and one outstanding request results are in the following table.

Table 5.

ND-12 8 + P 32 KBYTE SECTOR RAID-5		
RANDOM I/O	**WRITE**	**READ**
Sustained Peak I/Os/Sec.	130	434
1 Outstanding Request I/Os/Sec.	19	35

As can be seen, performance improved greatly from one outstanding random request to 64 outstanding random requests. This demonstrates the parallelism in the disk which is taken advantage of with many outstanding requests. The reason write performance wasn't as good as read performance is due to the RAID-5 write penalty for single sector requests.

7.4 ND-40 (Preliminary)

Sequential and random performance were evaluated for an early version of the ND-40. For sequential I/O, 100 transfers of 1 through 32 sectors (32 KBytes through 1 MByte) were done in each direction. Both single buffered and quadruple buffered tests to measure sustained peak performance were run. RAID-5 sustained peak performance and single sector single buffered performance are in the following table.

Table 6.

ND-40 10 + P 32 KBYTE SECTOR RAID-5		
SEQUENTIAL I/O	**WRITE**	**READ**
Sustained Peak MBytes/Sec.	83.7	88.5
Single Sector MBytes/Sec.	1.0	2.9
Sector I/Os/Sec.	29	88

ND-40 performance was compared to ND-12 performance in the graph in Figure 8. Again, with the ND-40, performance increased gradually as the transfer size increased. As shown in this graph, the ND-40 has overall higher performance than the ND-12.

For random I/O, either 400 transfers or 10,000 transfers of one sector were done in each direction. One all the way up to 59 outstanding requests were measured so as to measure sustained peak performance. Sustained peak performance and one outstanding request results are in the following table.

Table 7.

ND-40 10 + P 32 KBYTE SECTOR RAID-5		
RANDOM I/O	**WRITE**	**READ**
Sustained Peak I/Os/Sec.	170	569
1 Outstanding Request I/Os/Sec.	24	43

As can be seen, performance improved greatly from one outstanding random request to 59 outstanding random requests. This demonstrates the parallelism in the disk which is taken advantage of with many outstanding requests. The reason write performance wasn't as good as read performance is due to the RAID-5 write penalty for single sector requests.

8 RECOMMENDATIONS FOR NETWORK DISK ARRAYS

In general, the best choices from an overall performance standpoint and space utilization standpoint for a Network Disk (ND) are:

- 32 KByte sector size
- RAID-5

8.1 Performance

- A sector size of 32 KBytes has the best overall performance for small transfers and sustained peak performance
- RAID-5 has better overall performance
- RAID-1 has better small transfer write performance

8.2 Space Utilization

- A 32 KByte sector results in a 6% reduction in usable space versus a 64 KByte sector
- A 32 KByte sector provides 89% more sectors versus a 64 KByte sector which is beneficial if there are a lot of small files
- RAID-5 provides 44% more usable space than RAID-1

8.3 System Configuration

- If possible, put the ND in the secondary allocation area of a primary/secondary allocation area file system to shield it from file system (metadata) I/O
- Have **mkfs -S** set to a parity group for RAID-5 secondary allocation area partitions to force alignment of files on a parity group boundary
- **ldcache** the ND if possible
- In multiple partition file systems where **ldcache** is used, it is imperative that all partitions be a multiple of a stripe width in size so as not to get misaligned into the ND
- The **ldcache** unit size should be a multiple of a parity group so as not to incur the RAID-5 write penalty

9 CONCLUSIONS

As can be seen, Cray Research has many RAID offerings. Of these, the DA series disks deliver good performance across a range of transfer sizes. Based on these results, once the I/O is started up, the disk should normally be able to run at full speed. The results presented here also show that the ND series disks performance ramps up gradually as the transfer size increases. The main advantage of the ND series disks is that they can exist on a network and not be directly dependent on any one system.

RAID-5

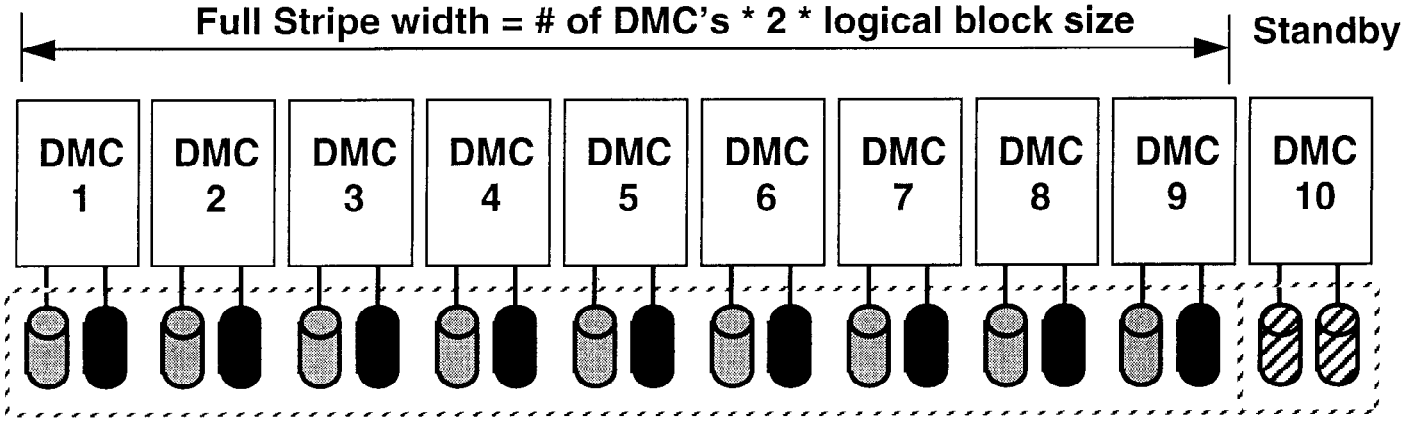

Figure 1

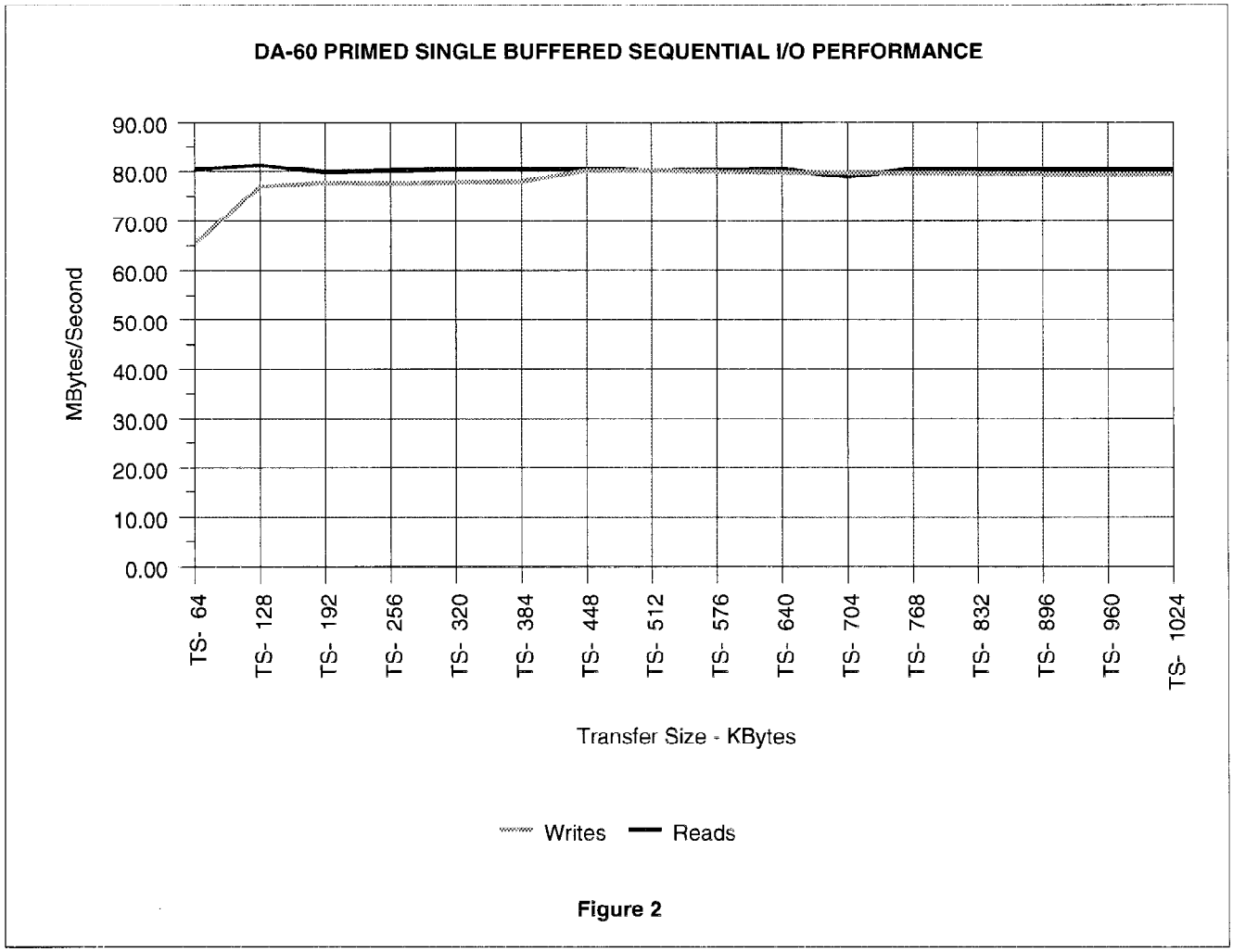

Figure 2

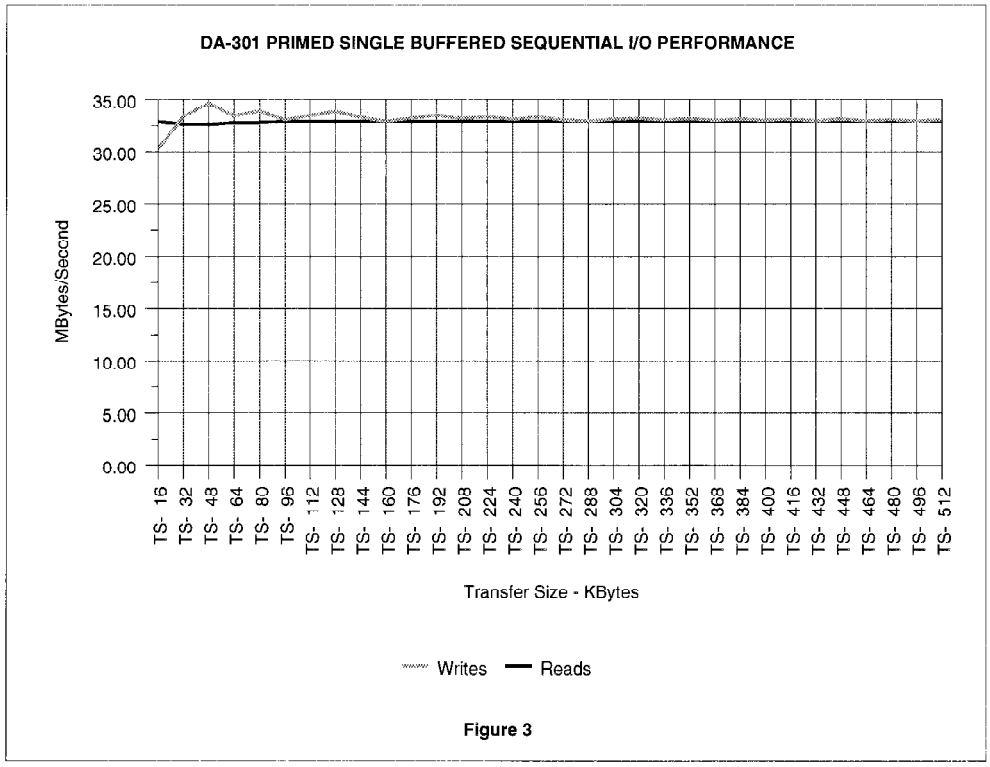

Figure 3

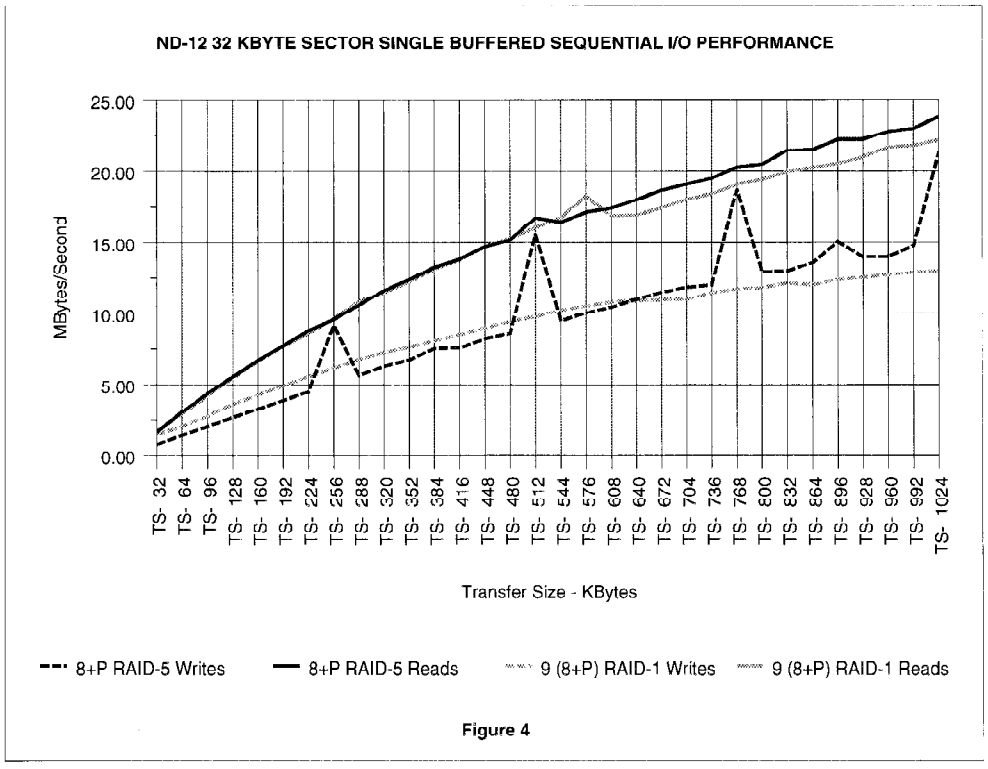

Figure 4

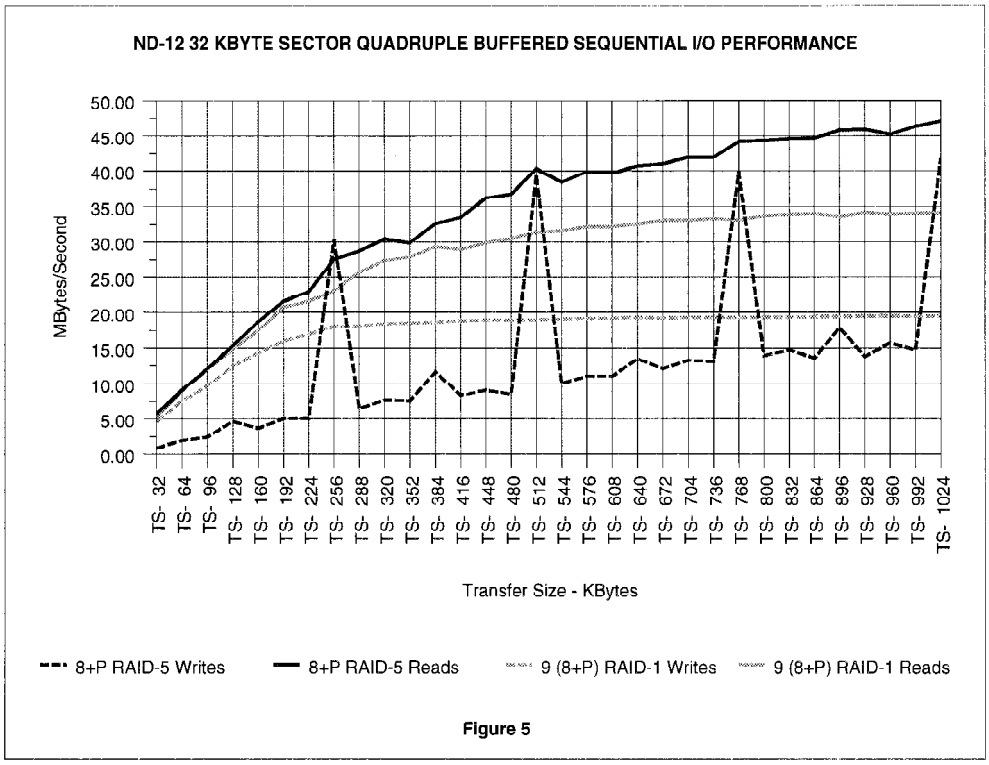

Figure 5

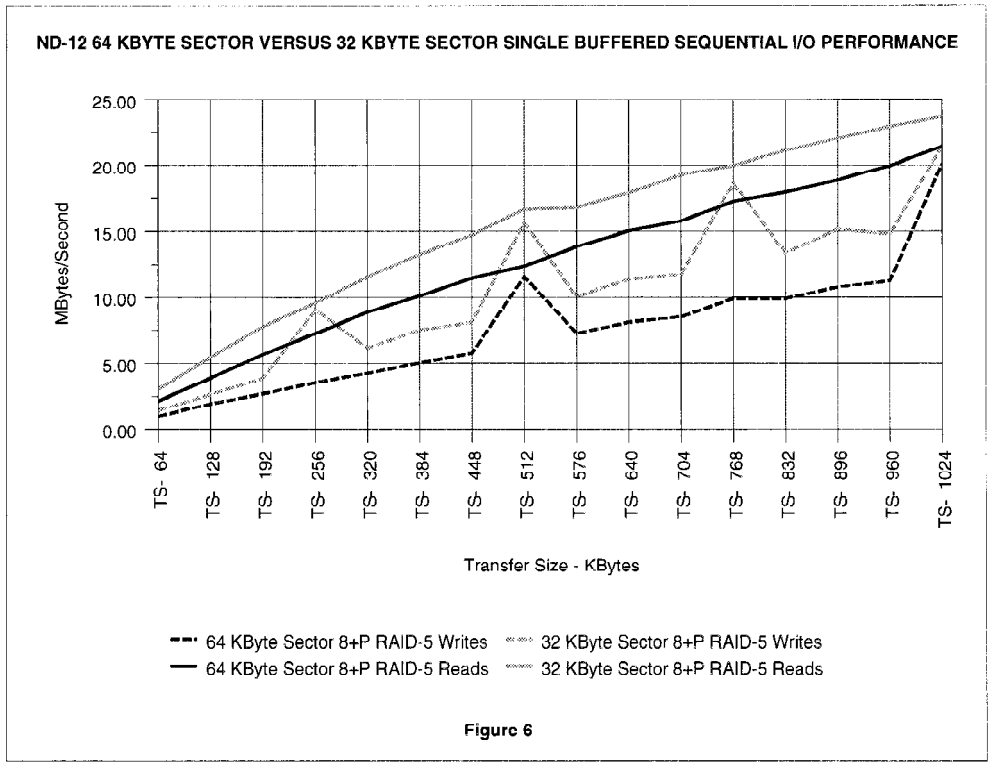

Figure 6

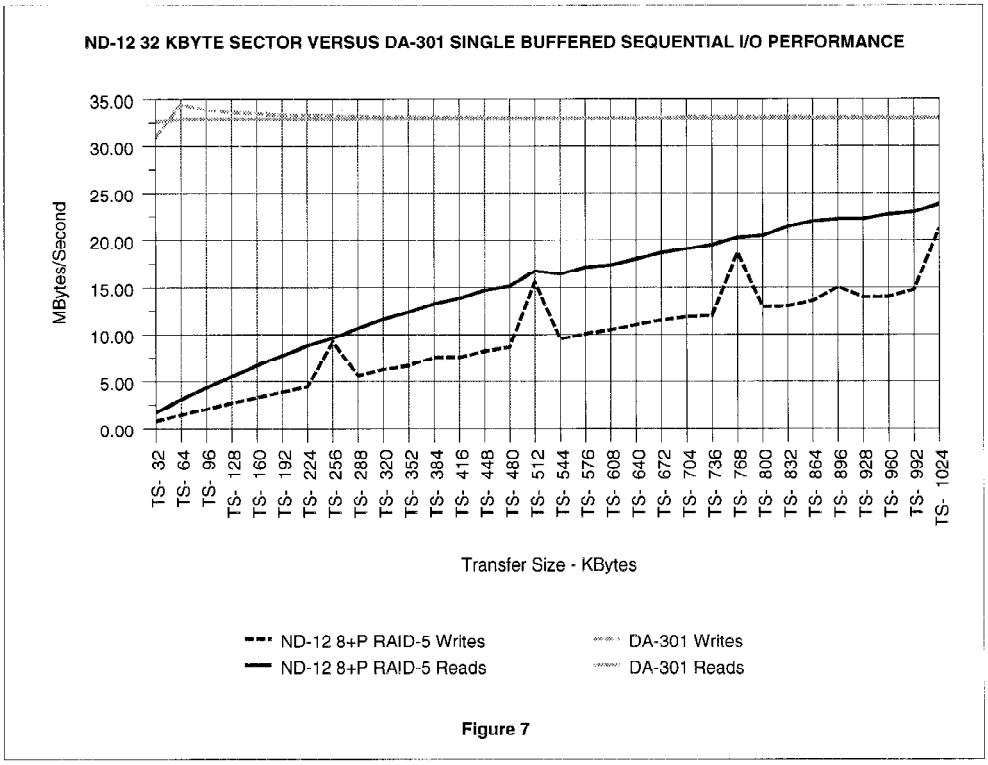

Figure 7

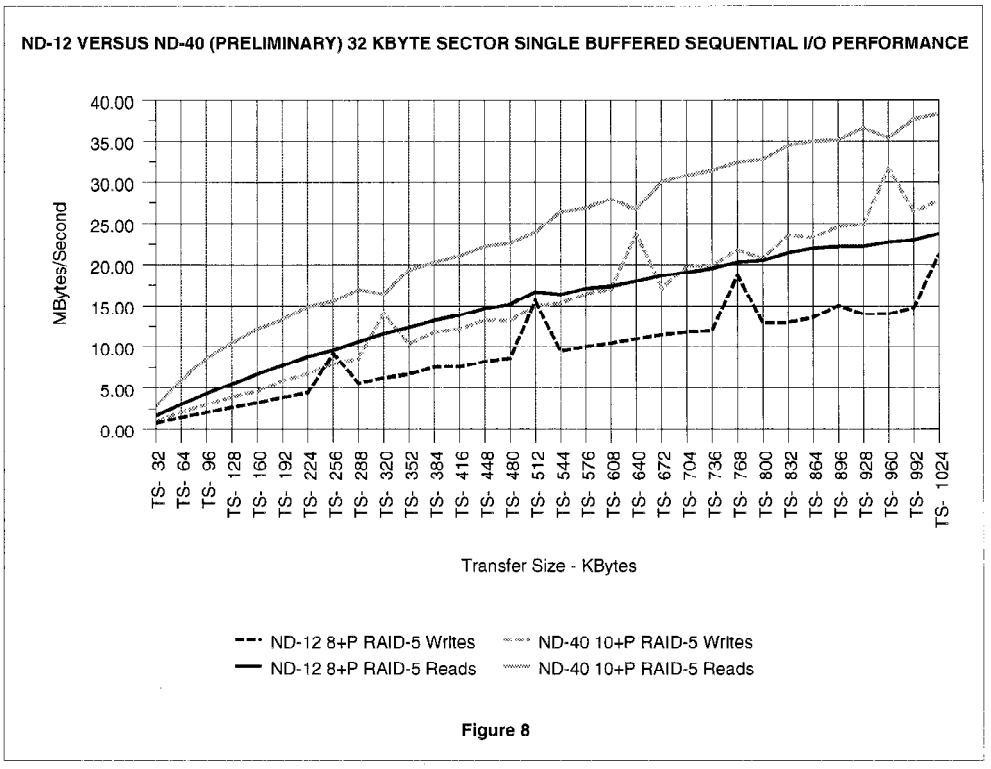

Figure 8