Improving File System Performance by Striping

Terance L. Lam
Computer Sciences Corporation
Numerical Aerodynamic Simulation Division
NASA Ames Research Center

March 4, 1992

Abstract

This document discusses the performance and advantages of striped file systems on the SGI 4D workstations. Performance of several striped file system configurations are compared and guidelines for optimal striping are recommended.

1.0 Introduction

The new source machine (wk200) has been installed and is operational. This machine consists of 4 CPUs, 128 Megabytes of main memory, a 1.1 Gigabyte IPI drive and four 1.2 Gigabyte SCSI drives. These SCSI drives have been configured to create a 4-way striped file system using the logical volume technique to take advantage of the two SCSI controllers available on the WKSII. A logical volume is an entity which behaves like a traditional disk partition, but its storage may span several physical devices. Different striping architectures have also been configured on wk200 for evaluation. The purpose of the evaluation is to:

- verify the functionality of the striped file system
- find the optimum disk striping configuration for the new source machine so that the future source build procedures can be performed efficiently.

2.0 Functional Test

Two-way, three-way and four-way striped file systems have been configured on wk200 and mounted onto a remote workstation for network testing. These striped file systems have passed NFS functional verification and have shown a substantial performance enhancement over a non-striped file system in the NFSstones test.

A compilation test has also been performed on these striped file systems. Compiling the SGI operating system on a 4-way striped file system has achieved a 30% improvement over a compilation on a non-striped file system. Since this build is a CPU intensive process, it is expected that a higher performance enhancement should result from an I/O intensive process.

3.0 Raw Data Transfer Rate of IPI and SCSI File Systems

Table 1 is a summary of the data transfer statistics of the high performance IPI drive and the SCSI drive suggested by SGI. The raw data transfer rate of the IPI drive is 6.0 MB per second and its sustained transfer rate (through the file system) is 3.6 MB per second. The raw data transfer rate of the SCSI drive is from 1.25 to 2.5 MB per second and the sustained transfer rate is from 0.5 to 1.5 MB per second. According to these specifications, both the raw and sustained transfer rates of the IPI drive are more than twice as fast as the SCSI drive.

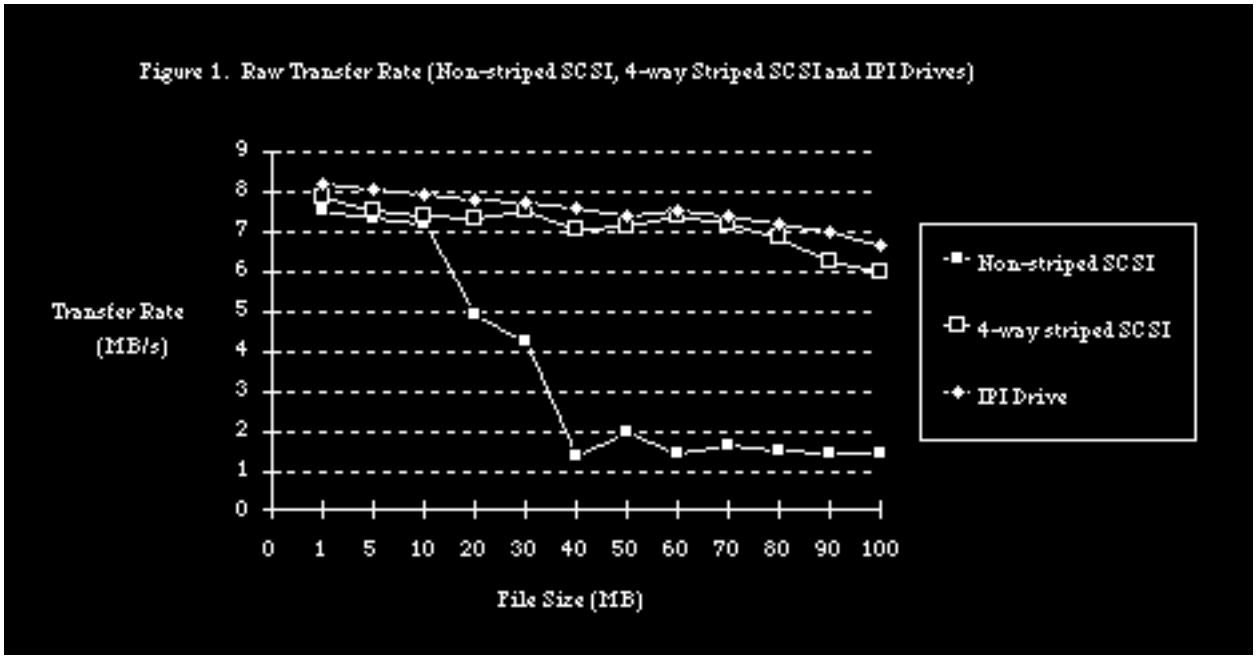
Drive	Raw Data Transfer Rate (MB/s)	Sustained Transfer Rate (MB/s) (Through the File System)
IPI	6.0	3.6
SCSI	1.25 to 2.5	0.5 to 1.5

Table 1.Data Transfer Rate of the IPI drive and the SCSI Drive

A test has been performed to compare the raw data transfer rates of an IPI file system, a non-striped SCSI file system and a 4-way striped SCSI file system using the SIO benchmark. This benchmark program performs read and write on the file systems and returns the raw data transfer statistics. Figure 1 is a plot of the raw data transfer rate of the file systems from Table 2.

File Size (MB)	Non-Striped SCSI	4-way Striped SCSI	IPI Drive
1	7.50	7.88	8.19
5	7.30	7.50	8.05
10	7.20	7.40	7.94
20	4.93	7.30	7.80
30	4.25	7.50	7.70
40	1.35	7.06	7.60
50	2.00	7.14	7.40
60	1.46	7.40	7.54
70	1.62	7.20	7.35
80	1.48	6.87	7.20
90	1.45	6.22	6.98
100	1.43	6.00	6.62

Table 2. Raw Transfer Statistics (Non-striped SCSI, 4-way Striped SCSI and IPI File Systems)



The results suggest that the IPI file system has achieved the highest perfor-

mance; its transfer rate ranges from 6.62 to 8.19 MB per second. The performance of the 4-way striped SCSI file system is very close to that of the IPI drive. The transfer rate ranges from 6.0 MB to 7.88 MB per second. The non-striped SCSI file system achieved the lowest performance rate, from 1.43 to 7.5 MB per second. These results verify SGI's specifications.

The non-striped SCSI file system performs competitively with the other file systems in small file transfers. As data file exceeds 10 MB, its performance degrades dramatically. This dramatic degradation in performance is a result of the large internal memory (128 MB) available on the test system. When more internal memory is available, more memory can be allocated for the cache purposes. In small file (less than 10 MB) transfers, a high percentage of the data can be cached in the buffer space. This buffered data will be written to the disk at a later time, instead of written to the disk immediately. This minimizes the amount of physical disk I/O operations. Since less physical I/O is actually involved, the non-striped SCSI file system shows an exceptionally high performance in small file transfers. But this figure is not the real transfer rate.

As the size of the data file increases (larger than 10 MB), the buffer space become saturated. The cached data has to be transferred to the disk. Paging and swapping occur; physical I/O operations are involved. The data transfer rate of the non-striped SCSI file system degrades to approximately 1.5 MB per second and sustains at this level. Therefore, the realistic data transfer rate of the non-striped SCSI file system should be approximately 1.5 MB per second. This figure matches SGI's specifications.

The same phenomenon applies to the IPI and the 4-way striped SCSI file systems. Since the performance of these two file systems is higher than that of the non-striped SCSI file system, these two file systems are impacted at a lower degree. These two file systems degrade but perform better than SGI's specifications (6.0 MB).

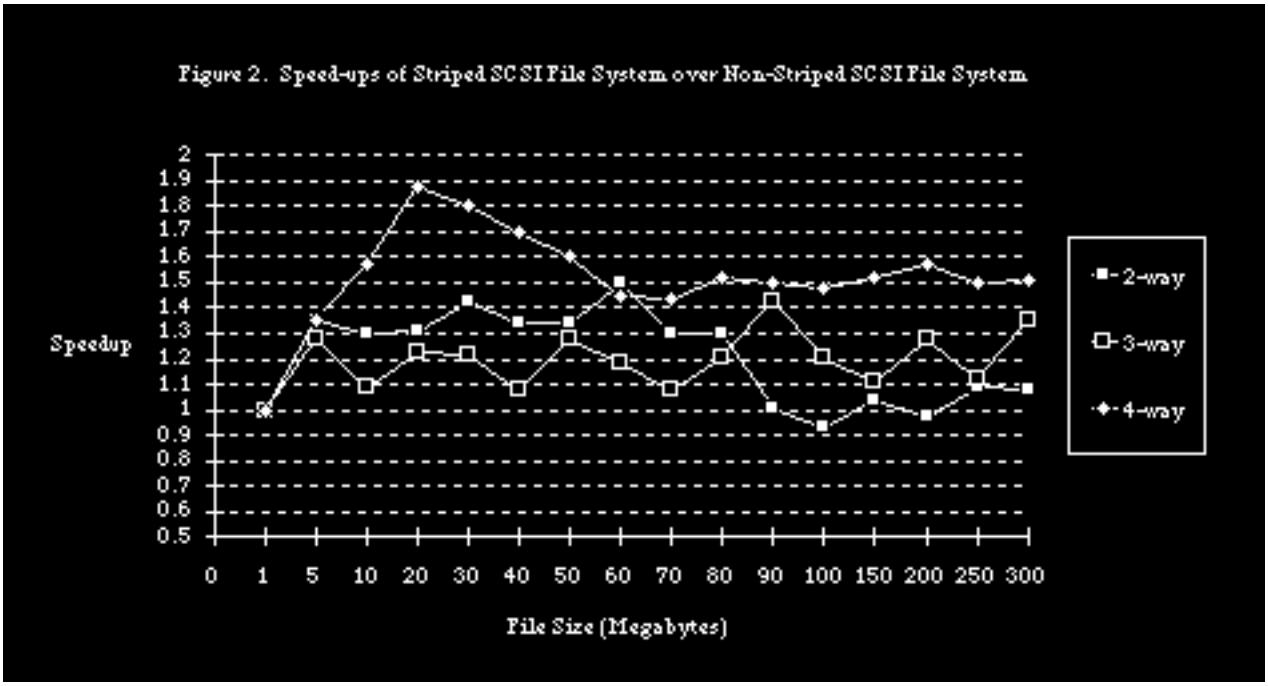
3.1 Performance of the Striped File Systems

In order to understand the realistic performance of the striped SCSI file systems, another performance test has been performed on a number of striped

SCSI file systems. This test involves reading and writing files through file systems of different striping architectures. The speedups (transfer rate of a striped file system/transfer rate of a non-striped file system) of the striped file systems over a non-striped file system are shown in Figure 2 and Table 3.

File Size (MB)	2-way Stripe	3-way Stripe	4-way Stripe
1	1.00	1.00	1.00
5	1.35	1.28	1.35
10	1.30	1.09	1.57
20	1.31	1.23	1.88
30	1.43	1.22	1.80
40	1.34	1.08	1.70
50	1.34	1.28	1.60
60	1.50	1.18	1.45
70	1.30	1.08	1.44
80	1.30	1.21	1.52
90	1.01	1.43	1.50
100	0.94	1.21	1.48
150	1.04	1.11	1.52
200	0.97	1.28	1.57
250	1.09	1.12	1.50
300	1.08	1.35	1.51

Table 3.Speedups of Striped File Systems Over Non-striped File System.



A two-way striped file system shows a maximum speedup of 1.5, a 50% improvement. When the file size reaches 90 Megabytes, the speedup degrades to 1.01. This indicates that a 2-way striped file system is not efficient in large file transfer and performs the same as a non-striped file system.

A three-way striped file system shows that improvement fluctuates between 8% and 45%. The fluctuation results from the fact that there are three SCSI drives connected to two controllers. One of the two SCSI controllers has to service two drives in a sequential order. The second disk has to wait until the first drive is serviced. This is no different than a sequential access of two drives on one controller. The average enhancement is 20%.

A four-way striped file system has achieved the best and most consistent performance in this test. Its improvement ranges from 35% to 88% and sustains a level of 50% in large file transfers. As suggested by SGI, a 50% performance improvement is the optimal enhancement of a striped file system built on SCSI drives. The 88% enhancement in small file duplication is a result of larger internal buffer space available as explained before. Therefore, a 50% enhance-

ment is a more realistic figure.

4.0 Advantages of A Striped File System

The advantages of a striped file system can be summarized as follows:

- Single file system administration is easier on multiple drives compared to multiple file systems on multiple drives.
- The performance of a striped file system is on the average 50% faster than a regular non-striped file system.
- Logical volumes offer a way to extend the size of a file system beyond the physical storage capacity of a disk drive. In this case, the file system on wk200 has been extended to 3.95 Gigabytes (formatted) using four 1.2 Gigabyte (unformatted) SCSI drives. The SGI source tree (2 Gigabytes) which previously could not be stored on any single file system, can now be stored in this logical file system which makes source tree maintenance easy.
- Table 4 is a summary of the cost/performance data between the IPI drive and the 4-way striped SCSI file system. The costs in Table 4 include a 28.5% discount from the ISC contract. Since the price of an IPI drive is more than twice that of a SCSI drive, it becomes more cost effective to purchase multiple SCSI disks for striping than to purchase an IPI drive. It translates to more than 50% savings on disk storage cost (based on a 4-way striped file system) but still achieves the same performance as an IPI drive.

Drive	Transfer Rate (MB/s)	Size (GB) Unformatted	Cost per Drive	Total Cost	Cost per MB
IPI	3.6	1.1	$11,789	$11,789	$10.72
4-way Striped SCSI	3.0	4.8 (4x1.2)	$ 5,005	$20,005	$ 4.17

Table 4.Cost Comparison (Per Megabyte) between IPI drive and SCSI drive.

5.0 Limitations of A Striped File System

Some limitations of the logical volume are:

- The root partition must not be a logical volume because the utilities required for logical volume initialization must reside on the root parti-

tion.

- If a disk drive in a logical volume fails, the complete logical volume built on multiple drives fails. The data in this logical volume is lost.
- It relies on a secure procedure to backup and restore the logical volume on the source machine. The SGI 3.3.2 "dump" backup utility does not support logical volumes; therefore the striped file system on wk200 cannot be backed up using this utility. The Cypress (4.0) Beta "dump" and "restore" utilities have been installed on wk200 and verified functional on logical volumes. The striped file system on wk200 can now be backed up and restored using these 4.0 utilities.
- When striping drives from different manufacturers, the performance of the striped file system is limited by the performance of the slower drive. Therefore, the performance of a striped file system will never be better than twice the performance of the slowest drive.

6.0 Optimal Striping Configuration

To obtain the best performance of disk striping, the following guidelines should be used:

- Disk drives on a striped group should be on a separate controller bus, or separate controllers.
- The number of disk drives in a striping group should be integer multiples of the number of controller buses or controllers such that each controller serves the same number of disk drives.
- Striping of multiple disks per controller should be specified in a way that interleaves access to the two buses or controllers. An example of a properly configured logical volume is:
  devs=/dev/dsk/dks0d0s7,/dev/dsk/dks1d0s7,/dev/dsk/dks0d1s7,/dev/dsk/dks1d1s7
  An example of an improperly configured logical volume is:
  devs=/dev/dsk/dks0d0s7,/dev/dsk/dks0d1s7,/dev/dsk/dks1d0s7,/dev/dsk/dks1d1s7
- Disk partitions which are to be striped together must be the same size; otherwise, all unused disk space is wasted.

7.0 Conclusion

The performance studies have been completed. A 4-way striped file system is

configured on wk200 to take advantage of the two SCSI controllers available. This striped file system has achieved an average of 50% performance enhancement over a non-striped file system, and has performed at levels close to a file system built on an IPI drive. This striping scheme equates to a 50% saving for high performance disk storage.

Disk I/O is one of the bottlenecks hindering the performance of a computer. When disk striping can improve the file system performance by 50%, it is suggested that this technique be considered for the SPS (each is configured with multiple SCSI disks) and future workstations with multiple drives.