UCRL-CR-137914

SAN/CXFS Test Report to LLNL

T.M. Ruwart and A. Elder

January 1, 2000

U.S. Department of Energy

Lawrence
Livermore
National
Laboratory

DISCLAIMER

This document was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor the University of California nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or the University of California, and shall not be used for advertising or product endorsement purposes.

Work performed under the auspices of the U. S. Department of Energy by the University of California Lawrence Livermore National Laboratory under Contract W-7405-Eng-48.

SAN/CXFS Test Report to LLNL
By

Thomas M. Ruwart and Alex Elder
University of Minnesota
Laboratory for Computational Science and Engineering

January 2000

Objectives

The primary objectives of this project were to evaluate the performance of the SGI CXFS File System in a Storage Area Network (SAN) and compare/contrast it to the performance of a locally attached XFS file system on the same computer and storage subsystems. The University of Minnesota participants were asked to verify that the performance of the SAN/CXFS configuration did not fall below 85% of the performance of the XFS local configuration.

Test Configuration

There were two basic hardware test configurations constructed from the following equipment:

- Two Onyx 2 computer systems each with two Qlogic-based Fibre Channel/XIO Host Bus Adapter (HBA)
- One 8-Port Brocade Silkworm 2400 Fibre Channel Switch
- Four Ciprico RF7000 RAID Disk Arrays populated Seagate Barracuda 50GB disk drives

The Operating System on each of the ONYX 2 computer systems was IRIX 6.5.6.

The first hardware configuration consisted of directly connecting the Ciprico arrays to the Qlogic controllers without the Brocade switch. The purpose for this configuration was to establish baseline performance data on the Qlogic controllers / Ciprico disk “raw” subsystem. This baseline performance data would then be used to demonstrate any performance differences arising from the addition of the Brocade Fibre Channel Switch. Furthermore, the performance of the Qlogic controllers could be compared to that of the older, Adaptec-based XIO dual-channel Fibre Channel adapters previously used on these systems. It should be noted that only “raw” device tests were performed on this configuration. No file system testing was performed on this configuration.

The second hardware configuration introduced the Brocade Fibre Channel Switch (see figure 1). Two FC ports from each of the ONYX2 computer systems were attached to four ports of the switch and the four Ciprico arrays were attached to the remaining four.

Raw disk subsystem tests were performed on the SAN configuration in order to demonstrate the performance differences between the direct-connect and the switched configurations.

After this testing was completed, the Ciprico arrays were formatted with an XFS file system and performance numbers were gathered to establish a File System Performance Baseline. Finally, the disks were formatted with CXFS and further tests were run to demonstrate the performance of the CXFS file system. A summary of the results of these tests follows.

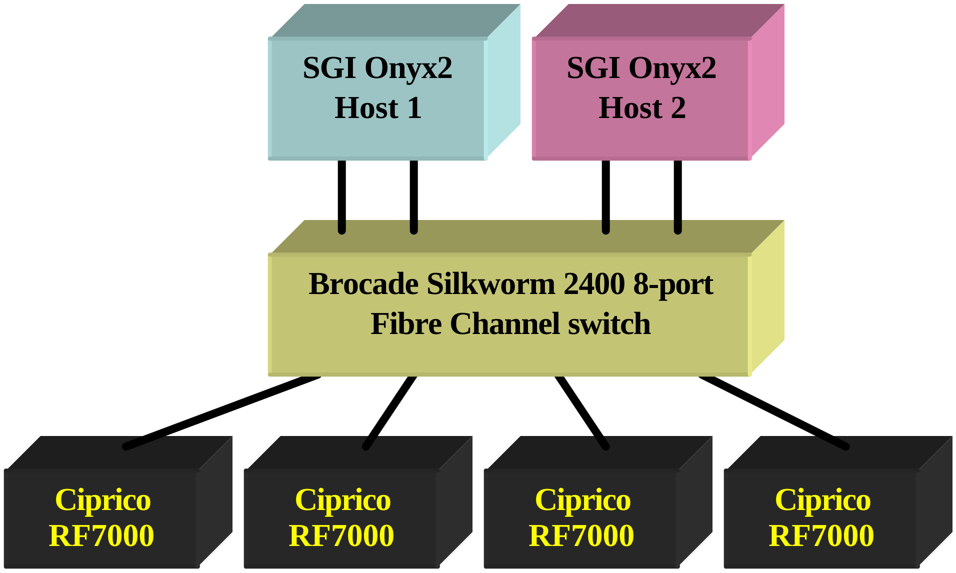
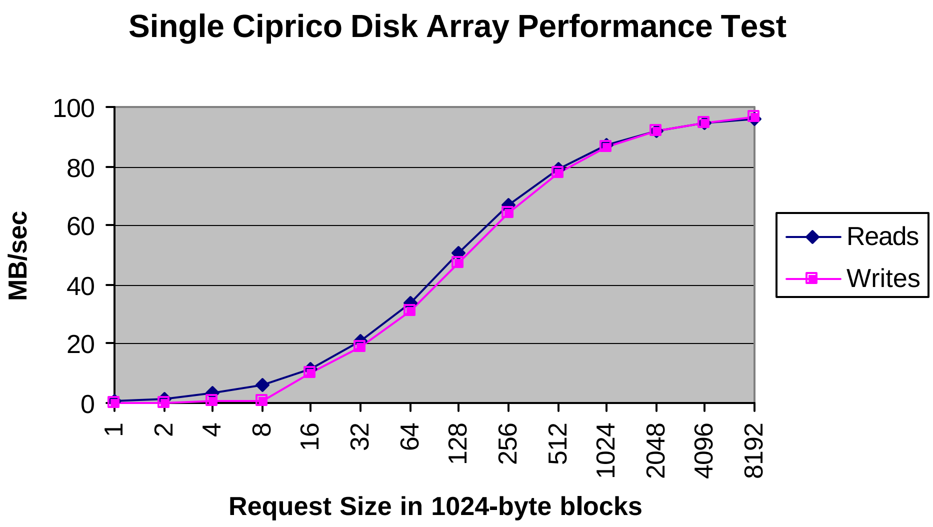


Figure 1. The Storage Area Network (SAN) configuration.

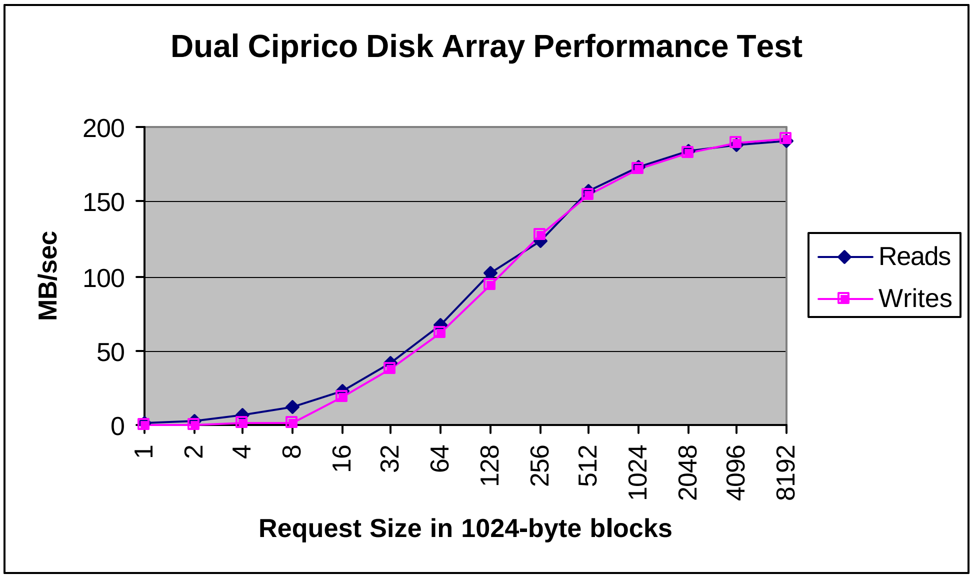
The benchmark program used to generate the results in this paper has been specifically developed over the past several years at the University of Minnesota and contains features necessary to this testing. This program is called *xdd*. Xdd is used to measure many of the disk device performance characteristics as well as helping to identify many of the performance anomalies that appear in more complex configurations.

Test Results

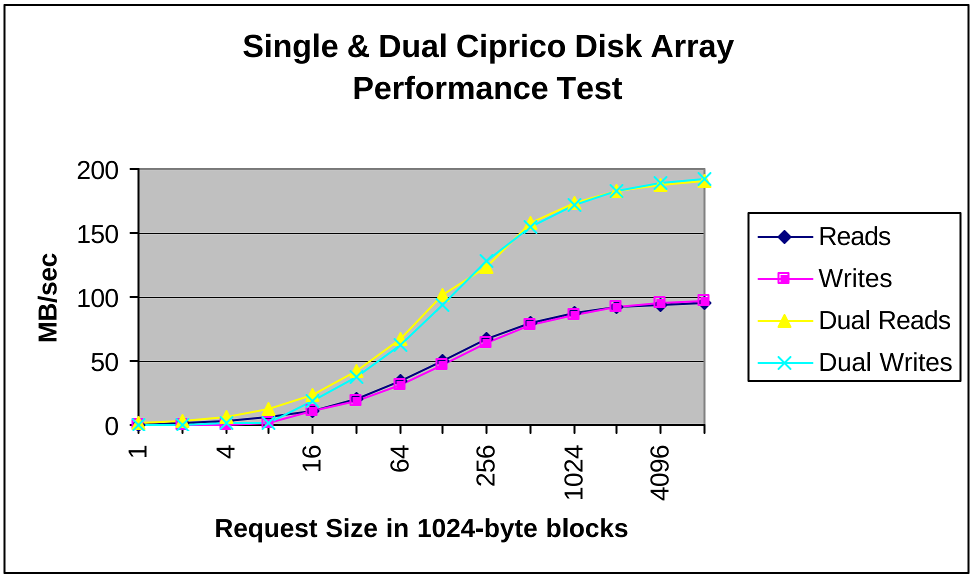
The first test involved a single Ciprico disk array. The performance of read and write operation demonstrate that there is no appreciable difference between read and write operations on this device. It should be noted that the read and write caches were enabled on the disk drives that populated the disk arrays. The peak performance for sequential read or write operations peaked at 96 MB/sec using 8MB data transfers.

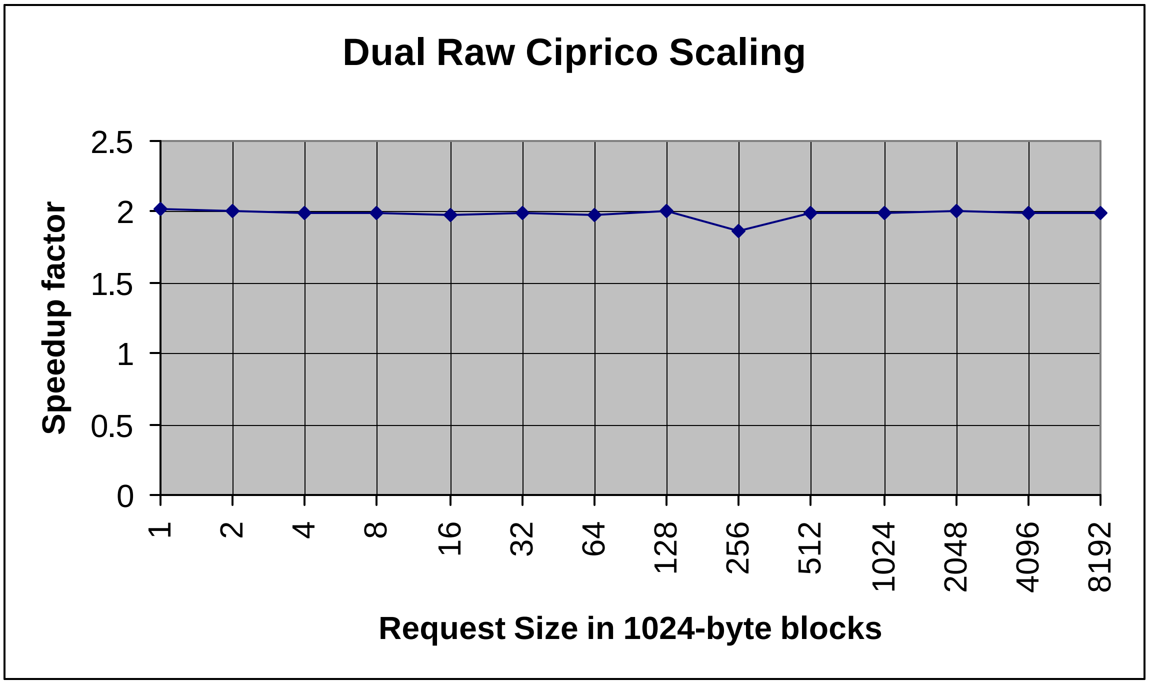


The next test involved read and write performance tests to two Ciprico disk arrays simultaneously from a single machine utilizing two Qlogic host bus adapter, each connected to one of the Ciprico arrays. These tests were run to insure that there were no scaling issues with respect to performance. The results of these tests show that read and write operations peak at about 191 MB/sec using 8MB data transfers.

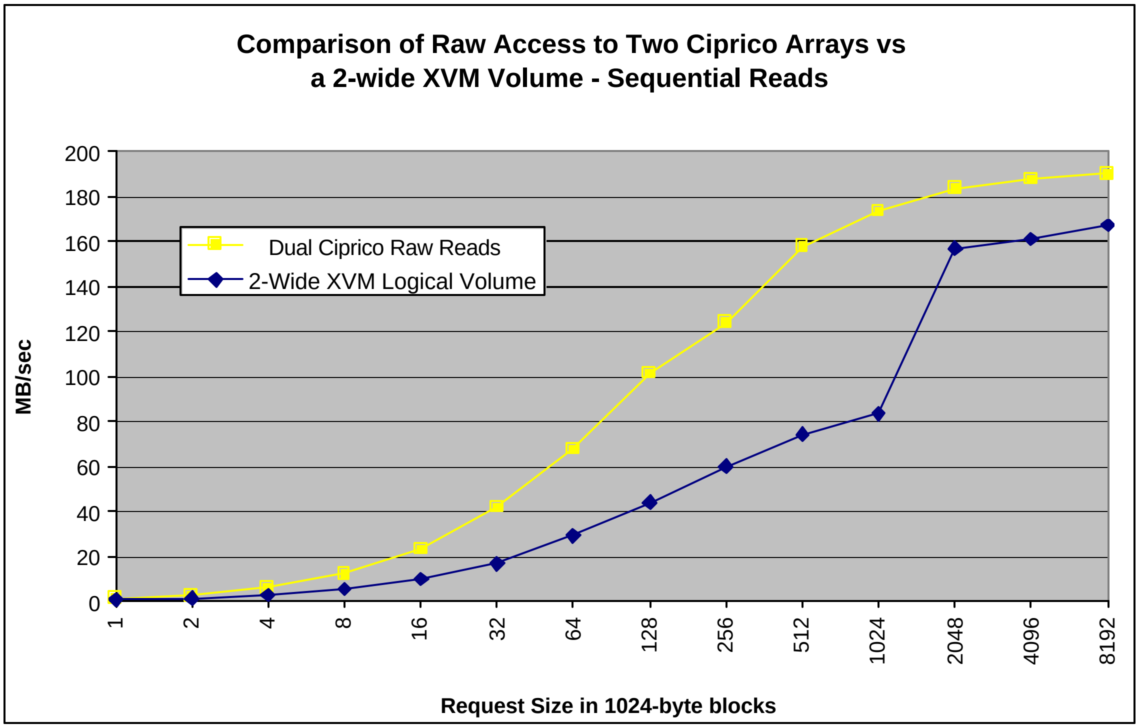


The scaling from one to two Ciprico arrays is also shown in the following two graphs. These graphs demonstrate that there is an effect 2x scaling when accessing two raw Ciprico arrays from a single machine.

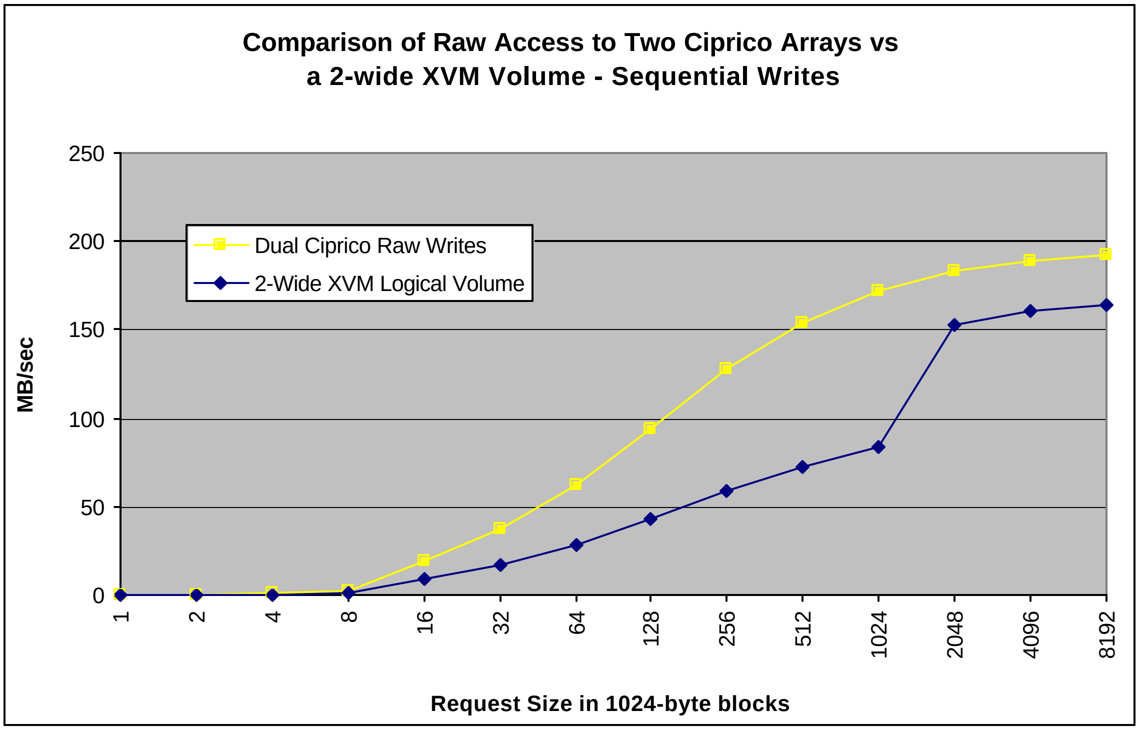




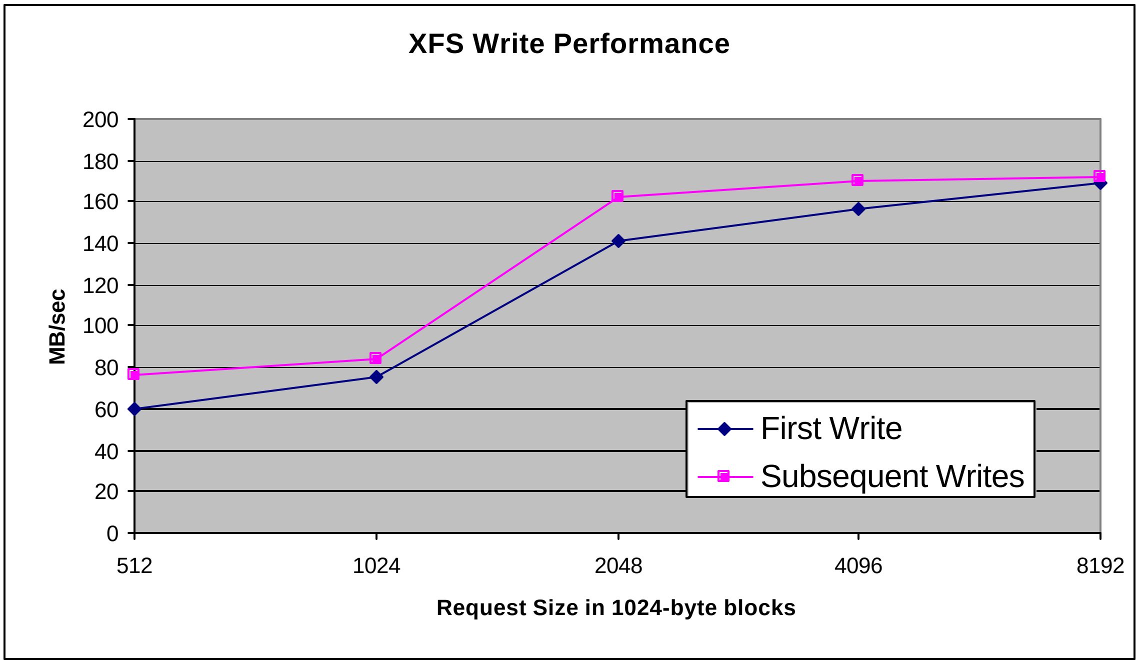
The next test involved creating an XVM logical volume across the same two Ciprico disk arrays used for the preceeding single and dual Ciprico disk arrays tests. This logical volume used a stripe-unit of 2048-blocks or 1MB (1 block = 512 bytes). The following graphs show the relative performance of the XVM logical volume compared to that of the two raw disk arrays. The "jump" in the graph (below) from request sizes 1024 to 2048 is a result of the overall request being able to effectively access both disk arrays in the logical volume simultaneously. The peak performance of the XVM volume for read operations is 166 MB/sec. This represents a performance drop of about 13%.

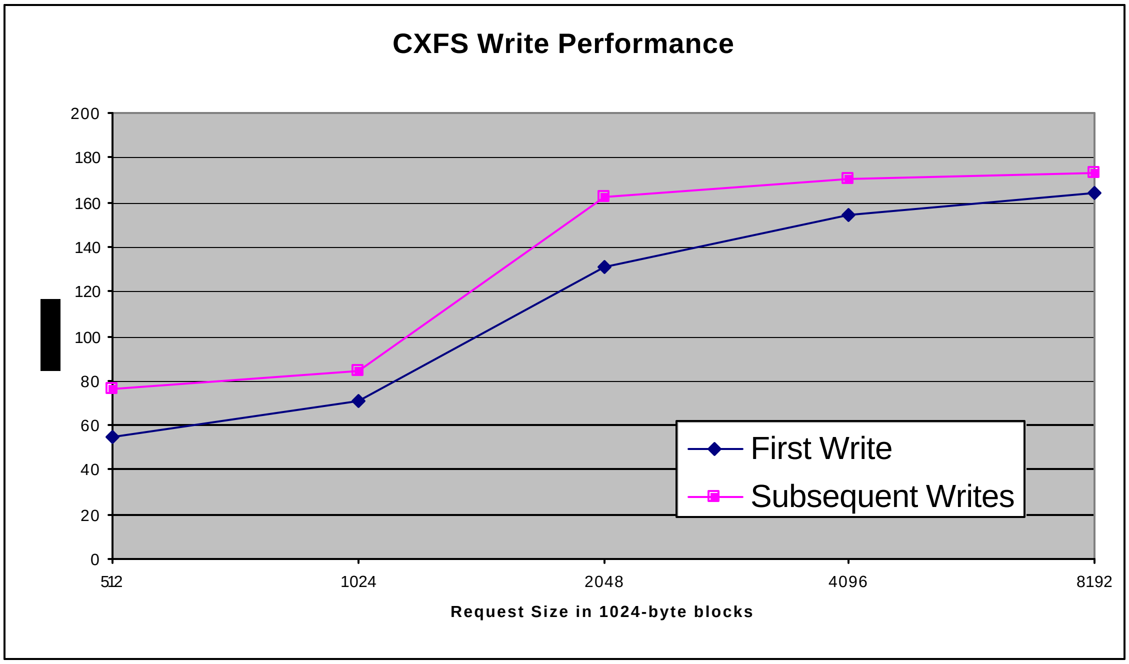


It should be noted that the XVM performance drop is an expected effect. There is a certain amount of overhead involved in managing a logical volume that manifests itself as a performance drop over the theoretical (empirical) peak performance of the underlying hardware. The observed performance drop is better than the expected range of a 15-25% drop. The write operations shown below show a drop of about 15%.

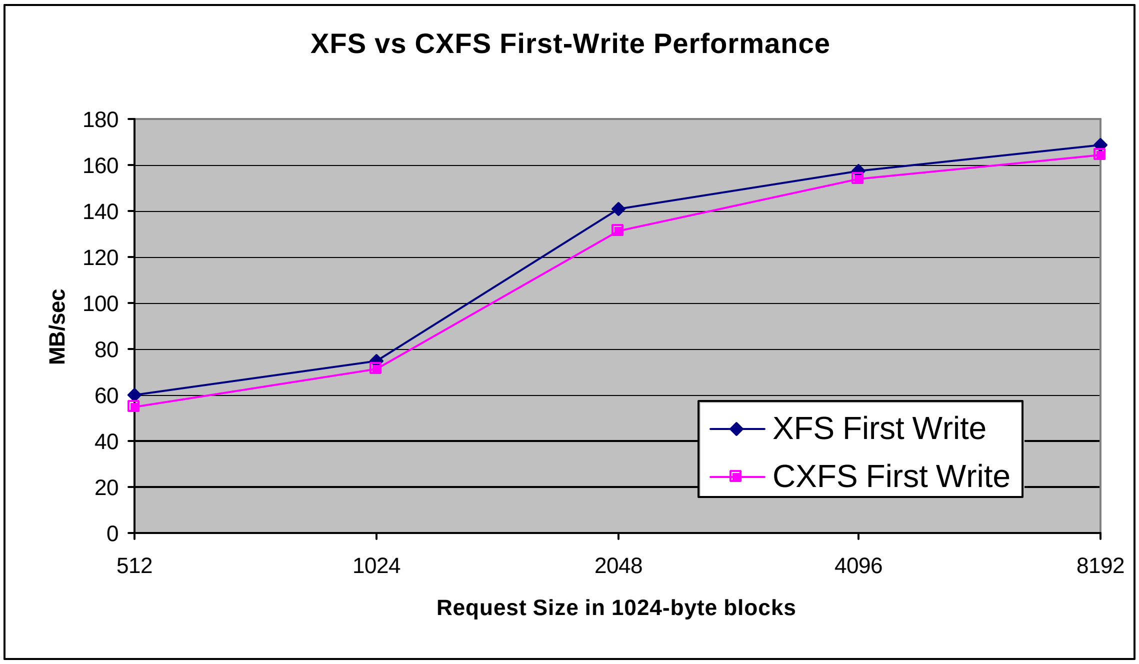


The next test phase involves sequential write and read performance from an XFS and a CXFS file system on the 2-wide XVM logical volume used in the previous series of tests. The first graph in this series shows the difference between the first and subsequent (or second) write operations to a file. The reason for this difference is that when a file is written for the first time there is a certain amount of overhead involved in allocating the space on the disk for the blocks to be written. This overhead results in a lower write speed when compared to writing the same file on subsequent tests. For very large request sizes however, this effect is minimal. The following two graphs demonstrate this effect for XFS and CXFS respectively.

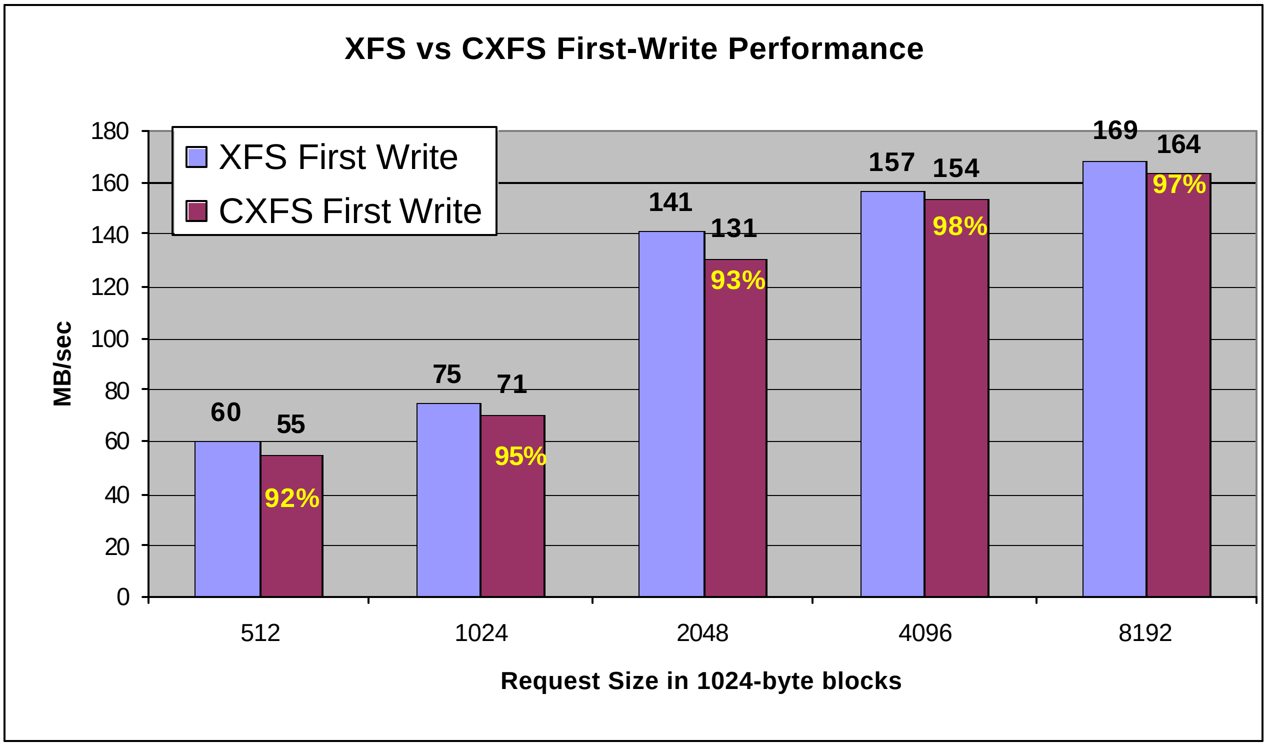


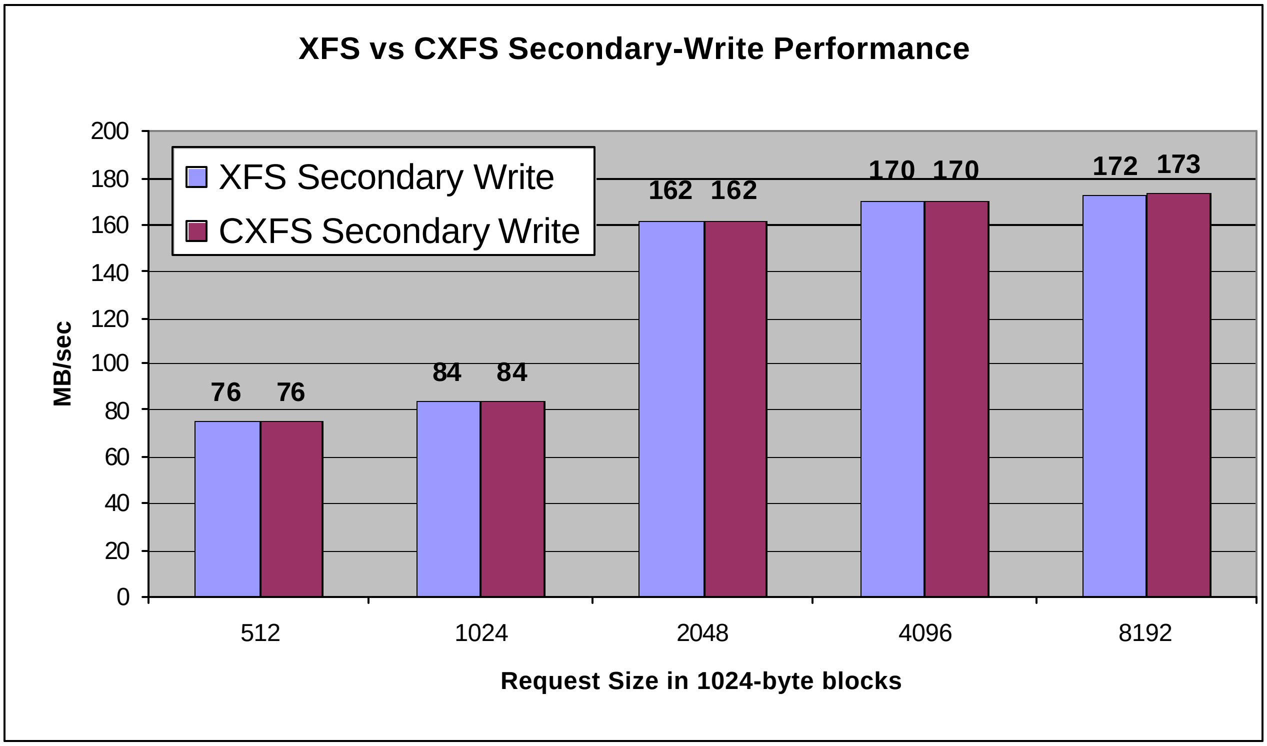


The following graphs compare the “first write” operation performance between XFS and CXFS. The comparison shows that a native XFS file system is slightly faster than the CXFS file system. This is an expected result but the point of interest is how large is this performance gap.

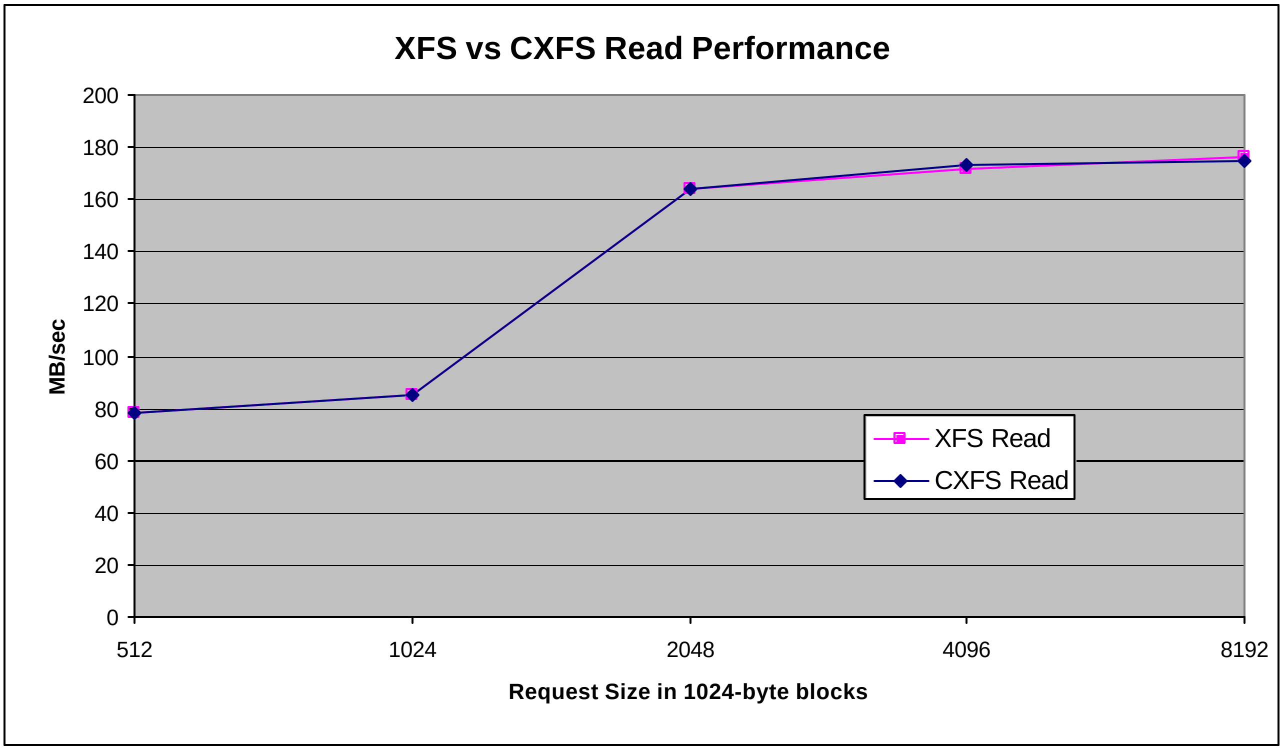


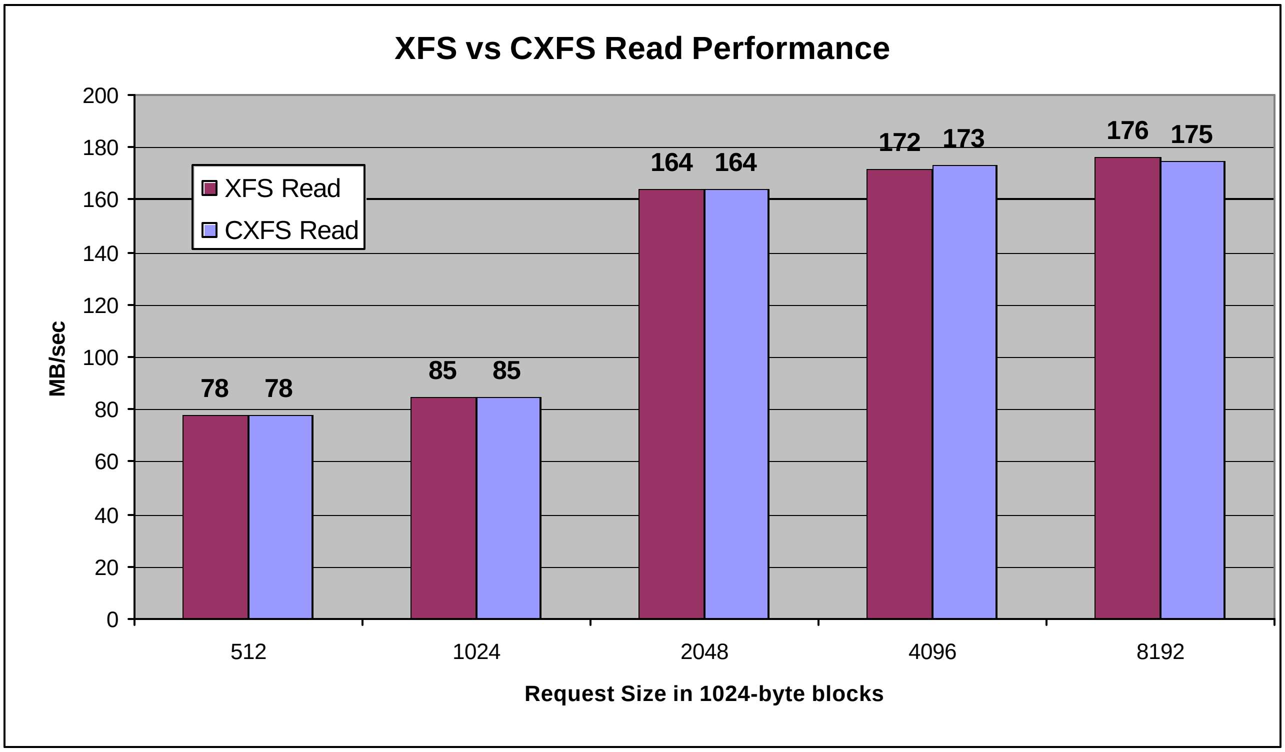
As is shown quite clearly in the following bar-graph, the performance difference between XFS and CXFS for First Write operations is well above the target of 85%. Furthermore, the Secondary (or subsequent) Write operations to the same file are not significantly different between XFS and CXFS.


XFS vs CXFS First-Write Performance
XFS First Write
CXFS First Write
MB/sec
180
160
140
120
100
80
60
40
20
0
60
55
92%
75
71
95%
141
131
93%
157
154
98%
169
164
97%
512
1024
2048
4096
8192
Request Size in 1024-byte blocks

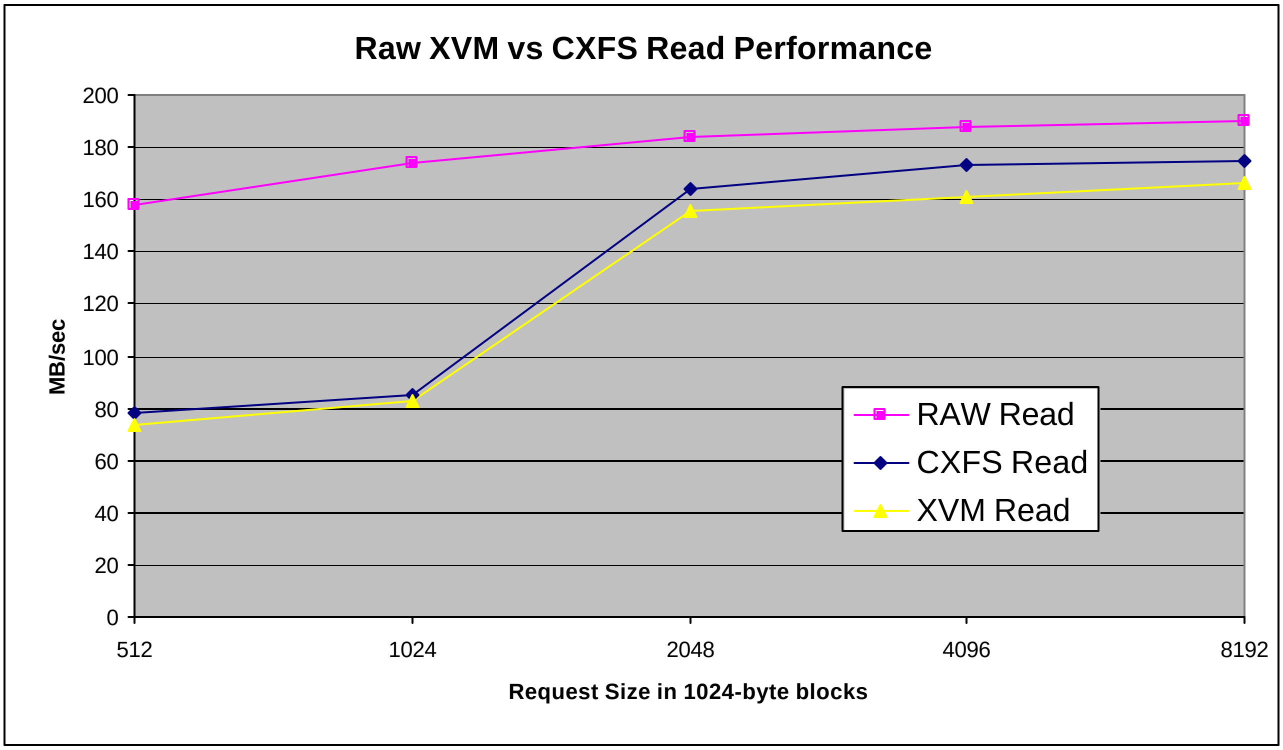

XFS vs CXFS Secondary-Write Performance
XFS Secondary Write
CXFS Secondary Write
MB/sec
200
180
160
140
120
100
80
60
40
20
0
76
76
84
84
162
162
170
170
172
173
512
1024
2048
4096
8192
Request Size in 1024-byte blocks

Finally, the following two graphs show that there is no significant performance difference between XFS and CXFS for sequential read operations of a large file. Again, this reiterates the fact that the CXFS performance falls well within the requirement that the CXFS performance be no less than 85% of the performance of the XFS file system.





The last graph shows the relative performance of the CXFS file system compared to that of the raw XVM Logical Volume and the raw disk arrays themselves. An interesting effect here is that the performance of the file system is actually slightly better than that of the XVM logical volume. It is believed that this is due to alignment effects of how the data was distributed across the disks. However, it does show that there is no performance drop between CXFS and the underlying XVM logical volume.



Conclusions

The testing performed by the Minnesota/SGI/LLNL team demonstrated that there was no measurable difference in performance with the introduction of a Fibre Channel Switch in the data path between the Ciprico disk array and the Qlogic host bus adapter. It was then shown that there was no significant difference in the performance of the CXFS file system versus a native XFS file system. The testing that we were able to perform demonstrated that the write and read performance of the CXFS file system was able to meet the criteria set forth by LLNL that it not be less than 85% of the performance of a native XFS file system on the same hardware.

Acknowledgements
This work was supported in part by the Department of Energy through the ASCI Data Visualization Corridor Program under contract #W-7405-ENG-48 and the ASCI/VIEWS and U. of Minn., Standard Research Subcontract # B347914.