Comparison Test: Storage Vendor Drive Rebuild Times and Application Performance Implications

Introduction

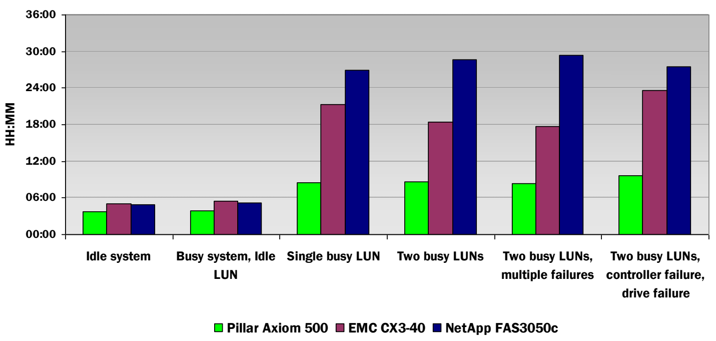
Today's datacenters are migrating towards virtualized servers and consolidated storage. As this happens, more and more importance is being placed on availability and performance of each asset. In the case of storage, disk drives are increasing in capacity at increasing rates; the failure of a drive impacts the performance and availability of far more data for longer periods of time. The failure of a drive should not be a major event within a modern storage array.

In evaluating the relative merits of competing storage vendors, one important consideration is system availability in the event of a drive failure. The time during which a drive is being rebuilt onto a hot spare is critical in that data loss becomes a possibility depending on the RAID configuration. Additionally, as system resources are dedicated to rebuilding a drive, performance degradation for other supported hosts becomes a real concern. Pillar Data Systems commissioned Demartek to compare the drive rebuild times to hot spares of the Pillar Data Systems Axiom 500, EMC® CX3-40 and Network Appliance FAS3050c. Due to the interest in drive rebuild times of storage systems with large-capacity drives, the storage systems tested were populated with four drive shelves of 500-GB SATA disk drives. Each system had at least 52 500-GB disk drives.

Evaluation Summary

Demartek's general findings were that the Pillar Axiom 500 performed the fastest drive rebuilds of the three systems tested, using 500-GB SATA disk drives. These test results were a bit unusual in that normally one product does not perform the best in every category.

LUN Rebuild Time (Lower is Better)



For drive rebuilds on idle LUNs:

- The Pillar Axiom system was up to 30% faster than the EMC CX3-40.
- The Pillar Axiom system was up to 26% faster than the NetApp FAS3050c.

For drive rebuilds on busy LUNs:

- The Pillar Axiom system was up to 60% faster than the EMC CX3-40.
- The Pillar Axiom system was up to 71% faster than the NetApp FAS3050c.

The results of our testing show that the Pillar Axiom design is very efficient and results in rebuild times that are relatively quick compared to other RAID-4 and RAID-5 systems. We believe that RAID-6 systems would be more complex and costly without necessarily improving drive rebuild times.

Technical Focus of This Report

The tests run for this report focused on measuring the drive rebuild times of the three storage systems under various I/O load conditions. The rebuild tests were run by failing one or more drives (all drives were "hot-swap") and noting the time each system took to rebuild (fail over to a hot spare) the data that had been written to that drive. At the completion of the rebuild, the removed drives were re-inserted into the drive slot and the time to "copyback" the data was noted. The times for rebuild and copyback were taken from the respective event logs for each system. It should be noted that the Network Appliance system does not perform a copyback, but simply marks the newly inserted drive as a new hot spare drive.

Test Description

Configuration

Each storage system contained four disk shelves populated with 500-GB SATA disk drives. Each storage system was configured into RAID-5 disk groups (RAID-4 for Network Appliance) of six disk drives, leaving a few drives for hot spare drives.

All configurations of the EMC CX3-40 system require the first five disk drives of the first shelf to be fibre-channel disk drives. The particular configuration used for these tests had no other disk drives in the first shelf, but had a fifth shelf with enough drives to make a complete RAID-5 disk group plus a spare drive (7 total drives). All the other shelves were completely populated with 500-GB SATA disk drives (15 drives per shelf), for a total of 52 500-GB disk drives. The fibre-channel disk drives were not used for any of the I/O workloads in this testing.

The Network Appliance FAS3050c had four drive shelves completely populated with 500-GB SATA disk drives, for a total of 56 disk drives (14 drives per shelf). However, three drives on the first shelf and three drives on the third shelf were dedicated to the controllers, so 50 total drives were available for testing. Each of the three drives allocated for the controllers were a 1 data drive RAID-DP™ disk group (1 data drive plus 2 parity drives).

The Pillar Data Systems Axiom 500 had four drive shelves completely populated with 500-GB SATA disk drives, for a total of 52 disk drives (13 drives per shelf).

Each system had a different usable capacity for the 500 GB disk drives. These are noted in the table below.

Each storage system was configured with as many 600-GB LUNs as possible, with a few LUNs of "leftover" sizes, depending on the particular system. The number of 600-GB LUNs for each system is noted in the table below.

The RAID type, number of disks in the disk group, the LUN size and in some cases specific disks for the disk groups were explicitly specified. All other storage system defaults were accepted during the configuration phase.

Configuration Data	EMC CX3-40	NetApp FAS3050c	Pillar Axiom 500
Number of 500 GB disk drives	52	56	52
Usable capacity of 500 GB drive	458.5 GB	413 GB	464.6 GB
Number of 600 GB LUNs	28	16	30

Initial Test Setup

To fill each LUN to approximately 80% full, each of the 600-GB LUNs on each storage system was written with 64K random writes, aligned at 4K boundaries. This was accomplished using the same IOMeter script, run for approximately three days for each system.

Find Fastest Performing LUNs

In order to get the best performance from each storage system, a LUN-to-LUN comparison was done, looking for the best performing pair of 600-GB LUNs. The LUNs were divided into even and odd pairs (LUN 0 and 1, 2 and 3, etc.). Two servers performed a 100% Random, 80/20 (80% read, 20% write) workload on each LUN pair, one LUN per server, measuring the performance. The best performing pair of LUNs on each storage system were used for all subsequent tests.

Rebuild Tests

All of the rebuild tests followed this basic formula:

1. Fail a drive (pull it from its drive shelf).
2. Record the time to rebuild to a hot spare.
3. Replace the drive (re-insert it into the drive shelf).
4. Record the time to perform a copyback from the hot spare.

A series of tests were run using the basic formula above but with different I/O workloads. The workloads and conditions are listed below. Except for case number 1, the workloads were run for at least 10 minutes before the failure condition and continued until the failure condition was resolved and after the storage system state returned to normal. The purpose of these tests is to discover the effects of various I/O loads on the storage system during a rebuild.

1. No load on any LUN (idle storage system with no host workload)
2. Single LUN busy with Random I/O, fail one idle LUN
3. Single LUN busy with Random I/O, fail one busy LUN
4. Two LUNs busy with Random I/O, fail one busy LUN
5. Two LUNs busy with Random I/O, fail one busy LUN on each shelf
6. Two LUNs busy with Random I/O, fail a controller, then fail one busy LUN

Each of the above tests yields two sets of measurements, the rebuild time and the copyback time. In addition, I/O measurements of IOPS (I/O per second) and MBPS (megabytes per second) were also taken.

For the tests that fail one drive on each shelf, the drive pulls occurred within one minute of each other and the drive re-inserts occurred within one minute of each other. The time measurements began with the first drive pull or reinsert and ended when the last drive recovery was complete. In all cases, the storage systems performed the drive rebuilds in parallel.

The IOMeter workloads used for these tests are defined as follows:

- Random I/O
    - Block size: 4K
    - Read/write mix: 80% read, 20% write
    - Random/sequential mix: 100% random
    - Workers: 2 per server (a total of 4)
    - Queue depth: 256
    - Ramp-up: 5 minutes

Test Results

Rebuild and Copyback Results

The results of the rebuild tests for the storage systems are listed in the table below.

Rebuild Time	EMC CX3-40	NetApp FAS3050c	Pillar Axiom 500
Idle System			
No load, idle system	5 hrs 1 min	4 hrs 53 mins	**3 hrs 48 mins**
Random I/O Workloads – One LUN			
Single LUN busy, fail idle LUN	5 hrs 29 mins	5 hrs 11 mins	**3 hrs 49 mins**
Single LUN busy, fail busy LUN	21 hrs 19 mins	26 hrs 58 mins	**8 hrs 28 mins**
Random I/O Workloads – Two LUNs			
Two LUNs busy, fail busy LUN	18 hrs 25 mins	28 hrs 40 mins	**8 hrs 36 mins**
Two LUNs busy, fail busy LUN on each shelf	17 hrs 45 mins	29 hrs 26 mins	**8 hrs 24 mins**
Two LUNs busy, fail busy controller then busy LUN	23 hrs 33 mins	27 hrs 32 mins	**9 hrs 42 mins**

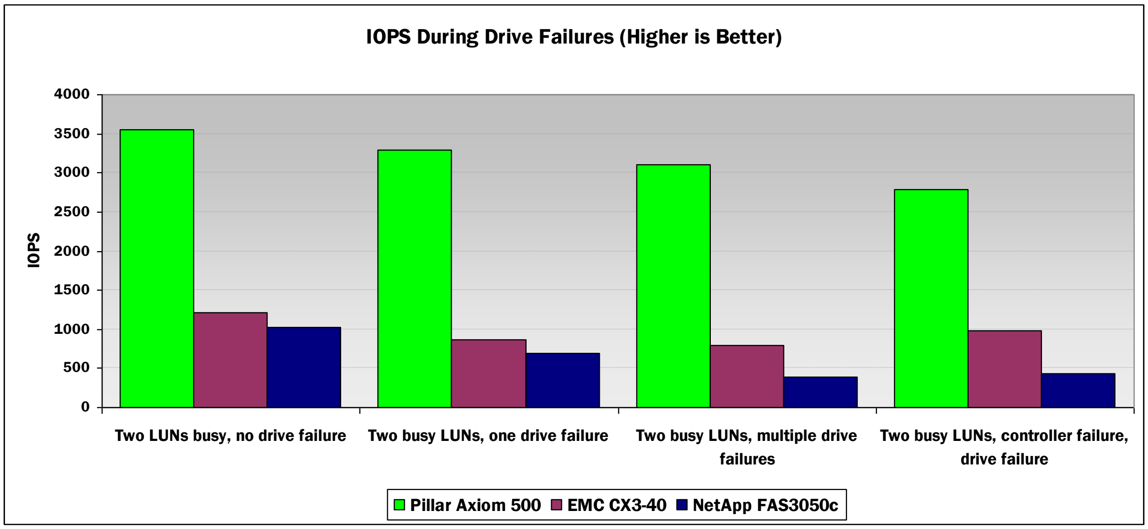
Copyback Time	EMC CX3-40	NetApp FAS3050c	Pillar Axiom 500
Idle System			
No load, idle system	3 hrs 26 mins	N/A	**2 hrs 54 mins**
Random I/O Workloads – One LUN			
Single LUN busy, fail idle LUN	3 hrs 25 mins	N/A	**2 hrs 52 mins**
Single LUN busy, fail busy LUN	17 hrs 10 mins	N/A	**9 hrs 3 mins**
Random I/O Workloads – Two LUNs			
Two LUNs busy, fail busy LUN	16 hrs 35 mins	N/A	**9 hrs 7 mins**
Two LUNs busy, fail busy LUN on each shelf	16 hrs 49 mins	N/A	**7 hrs 46 mins**
Two LUNs busy, fail busy controller then busy LUN	18 hrs 15 mins	N/A	**8 hrs 47 mins**

The Network Appliance system does not perform a copyback after the drive is replaced. It simply marks the new drive as a new hot spare. It should be noted that for every test, the combined rebuild and copyback times for the Pillar Axiom 500 took significantly less time than the NetApp FAS3050c rebuild time alone. The most important aspect of this procedure is keeping the data protected. Data is at some level of risk during the rebuild on any system, but is protected once the rebuild completes. The copyback times are far less important from a data protection standpoint. As the test results show, performance of all systems is impacted during rebuild and copyback functions.

IOPS Results

Although this set of tests was not an I/O performance test, the following I/O's per second (IOPS) were noted during the rebuild tests. The baseline performance for two busy LUNs before any drive failure testing began is also included in these tables.

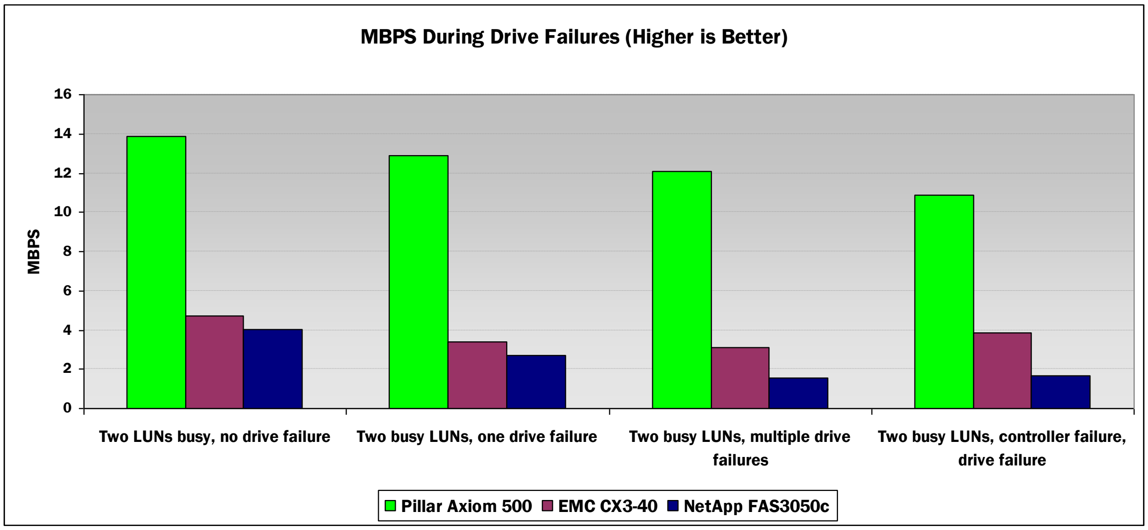
IOPS	EMC CX3-40	NetApp FAS3050c	Pillar Axiom 500
Idle System			
No load, idle system	N/A	N/A	N/A
Random I/O Workloads – One LUN			
Single LUN busy, fail idle LUN	576.5	532.2	**1642.8**
Single LUN busy, fail busy LUN	331.4	264.2	**1386.6**
Random I/O Workloads – Two LUNs			
Two LUNs busy, no drive failure	1206.9	1025.6	**3552.0**
Two LUNs busy, fail busy LUN	871.7	699.3	**3293.8**
Two LUNs busy, fail busy LUN on each shelf	789.4	393.5	**3100.9**
Two LUNs busy, fail busy controller then busy LUN	987.4	427.3	**2782.5**



MBPS Results

Although this set of tests was not an I/O performance test, the following megabytes per second (MBPS) were noted during the rebuild tests. The baseline performance for two busy LUNs before any drive failure testing began is also included in these tables.

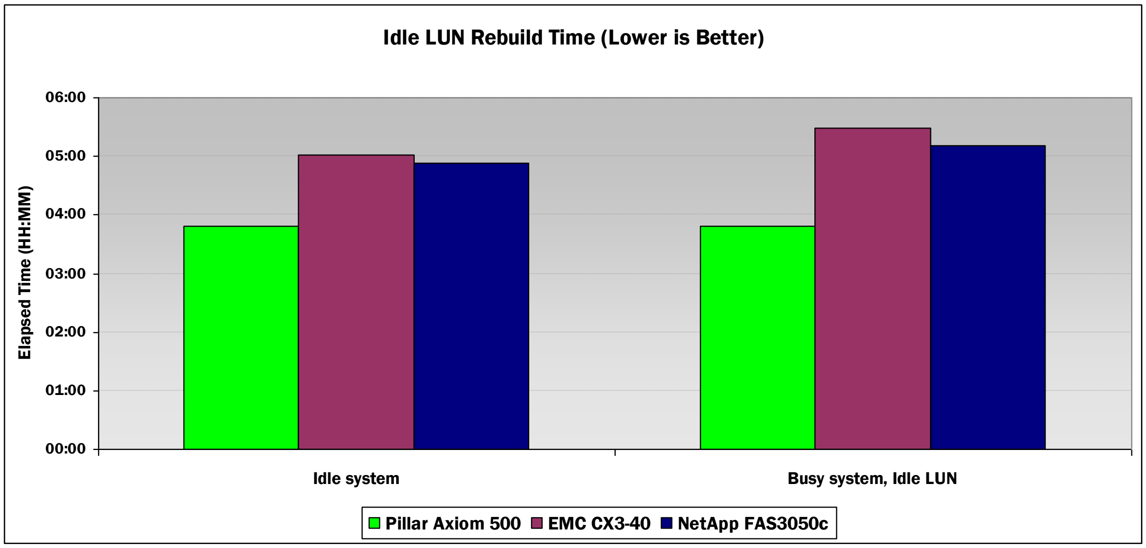
MBPS	EMC CX3-40	NetApp FAS3050c	Pillar Axiom 500
Idle System			
No load, idle system	N/A	N/A	N/A
Random I/O Workloads – One LUN			
Single LUN busy, fail idle LUN	2.25	2.08	**6.41**
Single LUN busy, fail busy LUN	1.29	1.03	**5.42**
Random I/O Workloads – Two LUNs			
Two LUNs busy, no drive failure	4.71	4.01	**13.87**
Two LUNs busy, fail busy LUN	3.4	2.73	**12.87**
Two LUNs busy, fail busy LUN on each shelf	3.08	1.54	**12.11**
Two LUNs busy, fail busy controller then busy LUN	3.85	1.67	**10.87**

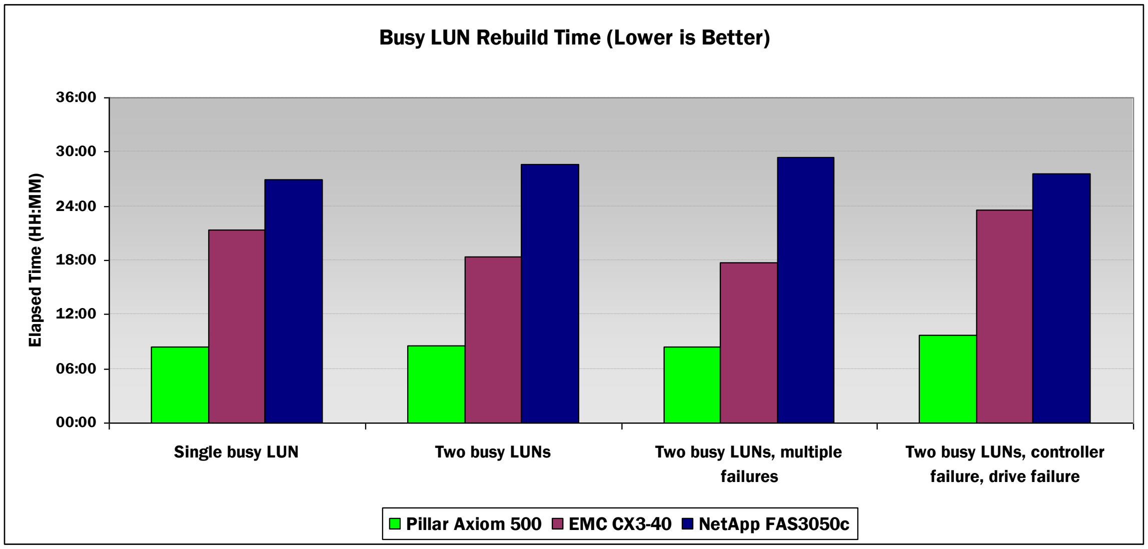


Summary and Conclusion

Drive rebuild time is an important factor to include when considering a disk storage subsystem. Rebuild times for idle and busy LUNs need to be considered.

The graphs below summarize the drive rebuild performance differences between the storage systems during the tests.

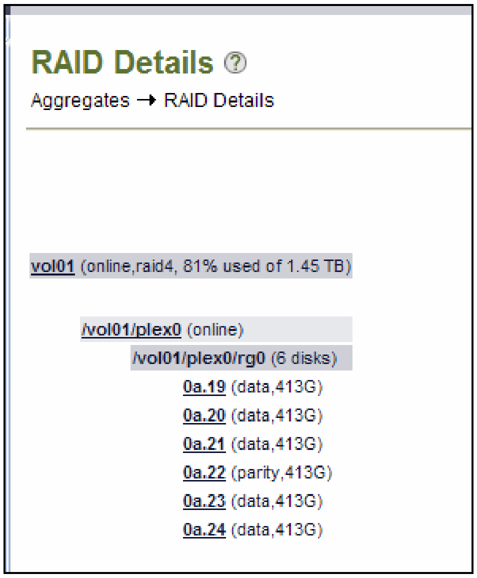




Appendix A – Observations

Capacity

As previously noted in the configuration data, the EMC CX3-40 and Pillar Axiom 500 systems provide similar usable capacity per 500 GB disk but the NetApp FAS3050c provides only 88% of the usable capacity that the Pillar system provides, approximately 51.6 GB less per 500 GB disk. The NetApp FAS3050c had 56 disks in this configuration. Six were allocated to the operating system (by default) leaving 50 disks for user data. Each FAS3050c RAID-4 disk group had 5 data drives which one might expect to result in approximately 2TB of raw capacity. The screen shot taken from the NetApp FAS3050C console below shows only 1.45TB of useable capacity per disk group, or volume. This contrasts to the 2.27TB available per disk group on the Pillar system.



RAID Details

Aggregates → RAID Details

vol01 (online,raid4, 81% used of 1.45 TB)

/vol01/plex0 (online)

/vol01/plex0/rg0 (6 disks)

0a.19 (data,413G)

0a.20 (data,413G)

0a.21 (data,413G)

0a.22 (parity,413G)

0a.23 (data,413G)

0a.24 (data,413G)

RAID Rebuild Process

In these systems a failed drive event is handled by immediately reconstructing the data of the failed drive onto a hot spare. The duration of this rebuild operation is important because a quick rebuild lowers the probability that data will be lost through a second drive failure before the rebuild completes.

The Pillar architecture moves this rebuild functionality to individual drive enclosures or “Bricks”. This ensures that the rebuild can proceed without impact to the storage system controller (head end) performance. The results presented clearly show the effect of these design decisions. Pillar’s rebuild operations are notably quicker than the competing systems that perform the rebuild function in the shared controller unit. The impact on the performance of ongoing host data transfers is reduced as well. While not explicitly tested here, a reasonable supposition can be formed that the Pillar architecture will be able to scale to larger numbers of bricks without serious RAID controller induced performance degradations. Such a test is suggested for follow on work.

Storage System Design and Implementation Impacts on the Testing

Pillar Axiom systems are pre-configured from the factory with two six-drive RAID-5 disk groups, plus a hot spare, per brick, so the tests were conducted using similar six-drive RAID disk groups for the other systems tested. If more disks in each RAID-4 or RAID-5 disk group are used, we would expect rebuild times to be longer.

By design, LUNs on the Pillar system are automatically spread across multiple bricks, and are not limited to a single RAID disk group. When LUNs are created on the other systems, the default process forces each LUN to be created within a single RAID disk group. It would be reasonable to expect better capacity utilization with the Pillar design. Having the data spread across multiple RAID groups can account for some of the performance differences seen in the no failure cases. Other performance differences may have resulted from the tested systems differing methods of allocation of the data of a LUN resulting in greater or lesser locality.

The Pillar design implements two RAID controllers per brick plus a pair of head-end controllers per system, the other systems tested use two head-end controllers with integrated RAID functions per system. Therefore the Pillar system design allows for higher performance for many functions, including the rebuilds that were performed for this report. (Note: Larger Pillar systems may be configured with additional head-end controllers – the tested configuration did not include this option.)

Appendix B – Storage System Pricing Commentary

One remaining question is the price of the systems tested. Some vendors consider their list pricing to be proprietary information and do not publish list pricing. Publicly available sources for pricing reveal that the new equipment prices for the EMC CX3-40 and the Pillar Axiom 500 in the configuration tested were within approximately 3% of each other, while the new equipment price for the Network Appliance FAS3050c in the tested configuration was considerably higher than the EMC system and the Pillar system. However, prices are negotiable and subject to change and specific models and components may be withdrawn or changed over time.

Appendix C – Evaluation Environment

This evaluation was conducted by Demartek at the Demartek facilities in Arvada, Colorado, using four Dell PowerEdge 2900 servers running Microsoft Windows Server 2003 R2 Enterprise x64 Edition. Each server was configured as follows:

- Dual-processor, quad-core Intel Xeon E5345 (2.33 GHz, 1333 MHz FSB, eight total cores)
- 4 GB RAM
- Internal disk array and controller with 15K RPM SAS disk drives
- Emulex LPe11002, dual-port, PCI-express (x4), 4-Gb. Fibre-channel HBA

The servers and storage were connected to a Brocade Silkworm 200e 16-port, 4-Gb. Fibre-channel switch during these tests. All storage data traffic flowed via fibre-channel connections through this switch.

For the Pillar Axiom testing, two servers and the Pillar Axiom storage system were connected to the fibre-channel switch. Each server was connected to two switch ports and the Pillar Axiom was connected to four switch ports, for a total of eight active switch ports. The two servers were dedicated to I/O testing for the Pillar Data system. No other applications were running on these servers during this testing.

For the EMC and Network Appliance testing, all four servers and both storage systems were connected to the fibre-channel switch. Each server was connected to two switch ports, the EMC system was connected to four switch ports and the Network Appliance system was connected to four switch ports, for a total of sixteen active switch ports. Two servers were dedicated to I/O testing for the EMC system and the other two servers were dedicated to I/O testing for the Network Appliance system. The switch zoning was set so that one pair of servers was visible to the EMC system and the other pair of servers was visible to the Network Appliance system. No other applications were running on these servers during this testing.

All the servers, storage and the fibre-channel switch were connected to a Dell PowerConnect 2748 48-port Gigabit Ethernet switch for management purposes.

All the servers were discrete physical servers running the operating system natively. No virtual server software was used during these tests.

The I/O loads were generated by using IOMeter, an industry standard, open source I/O load generator, available from Source Forge at http://sourceforge.net/projects/iometer/. Version 2006.07.27 was used for all the tests.