



SBOD vs. JBOD

Xyratex Performance Evaluation



Executive Summary

The purpose of this report is to measure, compare, and analyze the performance of SBOD vs. JBOD under various workloads. The results demonstrate that SBOD clearly has tremendous performance advantages over JBOD, allowing up to a 4x increase of throughput and 2x increase of IOP performance. Combined with the Reliability, Availability, and Serviceability aspects that are the focus of other reports, SBOD technology should be seriously considered to replace and/or augment any existing Fibre Channel JBOD implementations.

Overview

The basic fibre channel loop architecture limits the number of devices to 128. As the number of devices on the fibre channel loop increases, the challenge increases to provide scalable performance to all devices as they compete for communication resources. To reduce this resource starvation while increasing array sizes to their maximum, Xyratex has developed a product referred to as SBOD (Switched Bunch of Disks). This product is designed to be a plug and play replacement for existing Fibre Channel JBOD (Just a Bunch of Disks) technology, with much better RAS and performance capabilities.

The SBOD device includes an internal cross-bar switch to provide device communication while simulating the presentation of all FC-AL devices connected to it. This reduces the number of hops required for host-to-device communication, thus creating an almost host-to-device direct communication path. Another unique feature of the SBOD is the ability to ‘trunk’ two FC links within on simulated FC-AL loop such that inter-enclosure effective loop bandwidth is doubled. Through the elimination of the “physical loop” architecture, each and every node in the system is connected via a dedicated link, allowing full error management and diagnostics to be applied without disruption to the remainder of the devices in the “logical” loop. This enables vastly improved reliability and serviceability.

In summary, the data shown by the actual results is consistent with those results predicted by previous theoretical models. The performance using SBOD devices with “trunking” technology is highly consistent and shows no significant degradation in performance until loop saturation limits are achieved. The non-blocking nature of the SBODs allows concurrent initiator access to the “logical” loop demonstrating total throughput to peak at 1,550 MB/s, while the maximum performance in an equivalent JBOD configuration peaks at 384MB/s.

Results using a simulated SPC1-type workload, which typically include 8KB block transfers and a highly random access workload, shows how the loop performance effects start to influence the results in a standard JBOD configuration with a larger numbers of drives. This behavior is eliminated by Xyratex’s SBOD technology.

Attempts to explore the upper limits for the IOPs performance of this technology, using 512Byte data transfers have been hampered by CPU and other limits in the host systems. Peaks of over 200K IOPs have been achieved on SBOD configurations equivalent to twice that seen with JBODs. Xyratex believes with further host-side tuning that these limits can be overcome and significantly increased.

Configuration Information

Hardware/Software

The performance evaluation configuration involved two DELL 1750 Dual XEON (2.4GHZ/533) systems running Windows 2003 Enterprise Server with 2GB of RAM and 2 QL2342 2Gbps Fibre Channel 64-bit HBAs sharing a single PCI-X bus.

Qlogic HBA specifications and settings are as follows:

- BIOS 1.34
- Firmware version: 3.02.14
- Microsoft Windows 32-bit driver version 8.2.3.11 (W32VI)
- Advanced adapter setting, execution throttle was set to 256

(All other adapter settings were left to default)

This configuration involved test runs from 1 HDD [1 enclosure] to 112 HDDs [7 enclosures] using Seagate 146GB 10K RPM HDDs, running Seagate firmware 0006. SBOD firmware 0A.F8 was used for the SBOD trunking performance testing.

Load Generation

The loads that were generated and measured represent what can be considered the 4 corners of performance testing: Sequential Reads, Sequential Writes, Random Reads, and Random Writes. In addition, a SPC-1 Simulation was run to represent a mixed bag of all of the above loads that is used to benchmark overall performance. The results are organized in the following categories:

- Large Block(64k) Sequential – Reads
- Large Block(64k) Sequential – Writes
- Small Block(512 Bytes) Sequential – Reads
- Small Block(512 Bytes) Sequential – Writes
- Small Block(512 Bytes) Random – Reads
- Small Block(512 Bytes) Random – Writes
- SPC-1 Simulation

Each test was run 9 times while varying the queue depth from 1 to 256 at incremental increase by a power of 2.

IOMeter v2003.12.16 [originally developed by the Intel Corporation and announced at the Intel Developers Forum (IDF) on February 17, 1998, and is distributed under the terms of the Intel Open Source License.] was used to perform to create the simulated I/O loads and all measurements of I/O created for the JBOD and SBOD configurations. The ICF files used to configure IOMETER for each load are included in the Appendix A.

Test Configurations

Due to the complex nature of the large targeted configurations and the importance of predictable access patterns to optimize performance, a method of diagramming paths and access patterns was developed. This

method was used consistently in the diagrams below for planning and describing the results for both JBOD and SBOD testing. This method is described below.

The hosts are the topmost rectangles and described simply as Host A and Host B'. The prime character suffix for Host B' is used as a shorthand method to indicate paths and disks associated with links connected to Host B'. This is especially useful when printing the report in black and white. In all tests, 2 hosts were configured. Each link describes a 2Gbps FC connection from an HBA port on the host to an FC port on the enclosure. Links are also color/style coordinated for ease of tracking and printing.

The matrix of 16 slots refers to the front view of the RS-1600 16-bay 3U enclosure, which is used in all tests. The number in each slot indicates the link from which the drive in that slot is accessed. For example, the red 1's listed in the top row of the enclosure in Figure 1.0 signify disks that are accessed from Host A via Link 1. The green 2's listed in the 2nd row of the enclosure signify disks that are accessed from Host B' via Link 2'.

Varied colors and line styles are used to visually designate loop connections and provide clear configuration information even when printed in black and white. Connections that are physically connected together to form complete loops have a consistent line type (i.e. dotted, dashed, or solid line).

The "X7" to the right of the enclosure in Figure 1.0 is shorthand for indicating that there are 7 enclosures of identical configuration in this diagram. A set of enclosures oriented vertically are referred to as a 'stack'.

In the SBOD diagrams, the term 'string' will be used as a label for inter-enclosure links. This indicates an inter-enclosure link that is predominantly used by the external link of the same color and designator. This is described in further detail below.

JBOD Configuration

This testing focused on using the Xyratex RS-1600-FC Fibre Channel JBOD systems. Please refer to Figure 1.1 below for this configuration noting that the number of enclosures used in the test is increased from 1 to 7 (112 devices addressable).

Initial JBOD performance testing for up to 4 HDDs was completed from a single host. When 8 HDDs were involved, the second host participated in the performance test. This configuration and method was chosen to eliminate drive contention between multiple hosts and to optimally balance the load. Drive access was organized as shown in the above diagram, Figure 1.1. (i.e. HDD slots marked with the number 1 are accessed by host link 1, etc.)

Links 1 / 4' connect into the first FC-AL Loop (Loop A) and links 2' / 3 connect into the second FC-AL Loop (Loop B). Connections to the 'top' and the 'bottom' of the stack allow for access to all

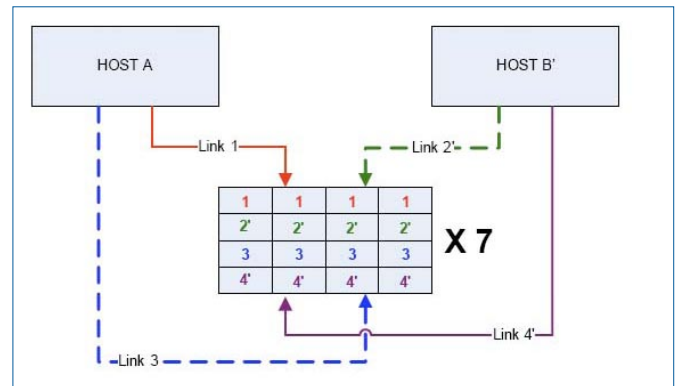


Figure 1: Example Diagram

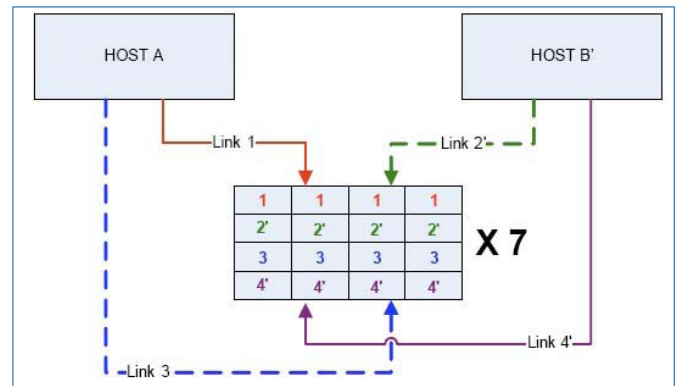


Figure 1.1: JBOD Configuration with 4 Initiators



enclosures in the stack even if both power supplies fail or are turned off in a single enclosure between, a dual point of failure that some OEMs are concerned about eliminating.

SBOD Configuration

This testing focused on using the Xyratex RS-1600-FC-SBD Fibre Channel systems. Note these are the same chassis and drives as used in the JBOD testing, but with the SBOD cards installed. All described results refer to configurations with trunking enabled firmware installed to demonstrate the increase bandwidth and functionality of this technology. Please refer to the diagrams below for these configurations.

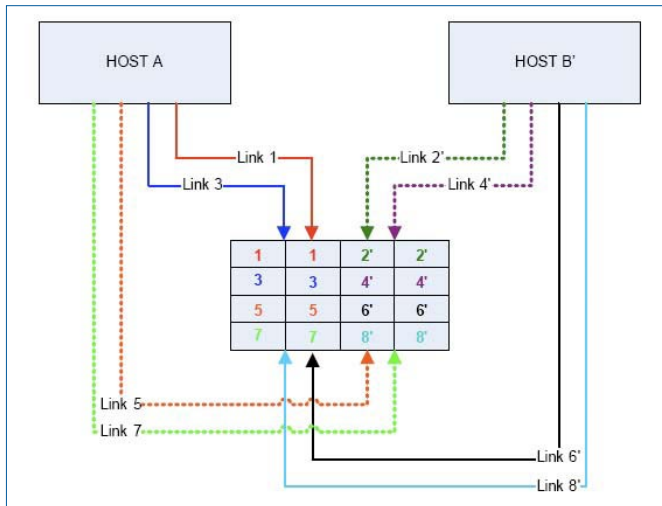


Figure 2.1: SBOD Single Enclosure Configuration with 8 Initiators

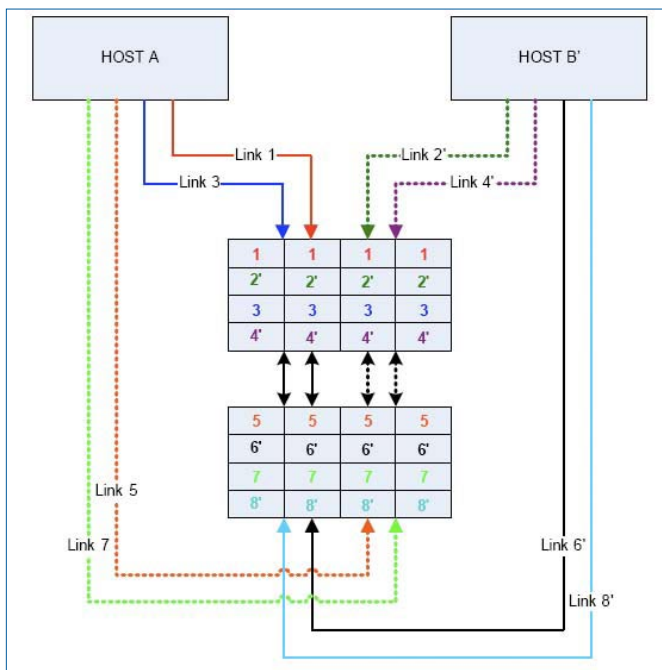


Figure 2.2: SBOD Dual Enclosure Configuration with 8 Initiators

Initial SBOD performance testing for up to 2 HDDs was completed from a single host. When 4 HDDs were involved the second host participated in the performance test. This configuration and method was chosen to eliminate drive contention between multiple hosts. Drive access was organized as shown in the above diagram, Figure 2.1. (i.e. HDD slots marked with the number 1 are accessed by host link 1, etc.)

Links 1, 3, 6' and 8' are the four host initiators connected into the first FC-AL Loop (Loop A), while links 2', 4', 5 and 7 are the four host initiators connected into the second FC-AL Loop (Loop B).

As the number of HDDs increases to 32 (which doubles the number of enclosures to two), the configuration described in Figure 2.2 below is selected to provide optimal throughput by limiting interchassis communication.

As the number of HDDs increases to 48 and beyond (which increases the number of enclosures to three up to a maximum of seven), the configuration described in Figure 2.3 is selected which will provide the most beneficial throughput provided by utilizing SBOD controller trunking.

The configuration described in Figure 2.3, which applies to configurations of 3 or more enclosures, takes maximum advantage of the trunking features of SBOD controllers. Traffic can be initiated from the top and the bottom of the array increasing the overall I/O capabilities of this configuration. However, strategic mapping of the primary path to disks can offer dramatic performance gains. For example, in Figure 2.3 above, it is especially important to note that the connections to the bottom of the stack only access the disks in the bottom-most enclosure of the stack. Any disk can

be accessed from any link, but this mapping minimizes string connections and allows the maximum throughput to approach the maximum theoretical entitlement performance milestone of 1600MB/sec over 8 2Gbps Fibre Channel links.

Summary Results

Large Block (64k) Sequential I/O Results

These results clearly demonstrate the advantage of the increased number of ports and interlinks available with SBOD technology. Also, there is very little degradation of performance as the number of HDDs is increased to the maximum allowed.

Small Block (512 byte) Sequential I/O Results

These results clearly demonstrate the decreased number of hops on the loop with SBOD technology, resulting in much higher IOP numbers with a large number of disks. The decrease in performance from its maximum peak is associated with host-based overheads and high CPU utilization.

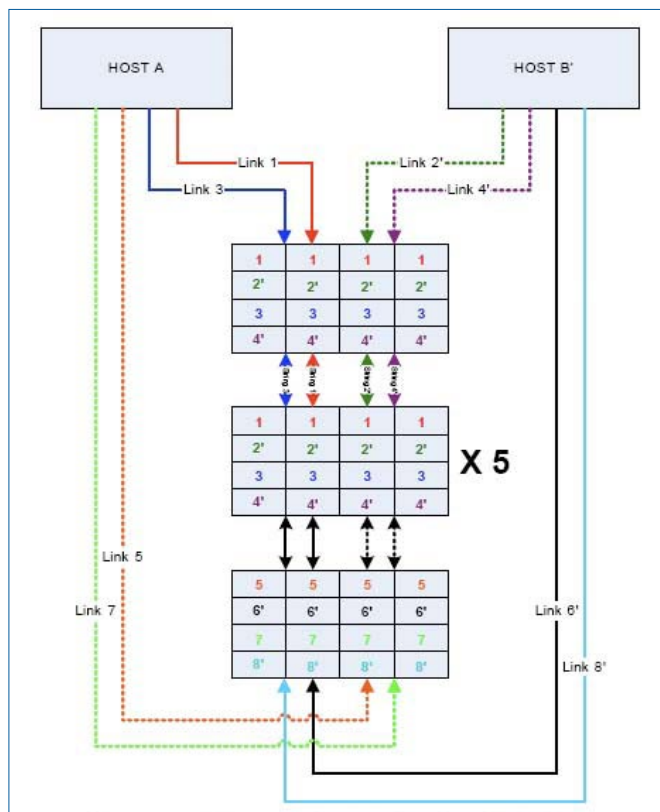


Figure 2.3: Multiple SBOD Enclosure Configuration with 8 Initiators

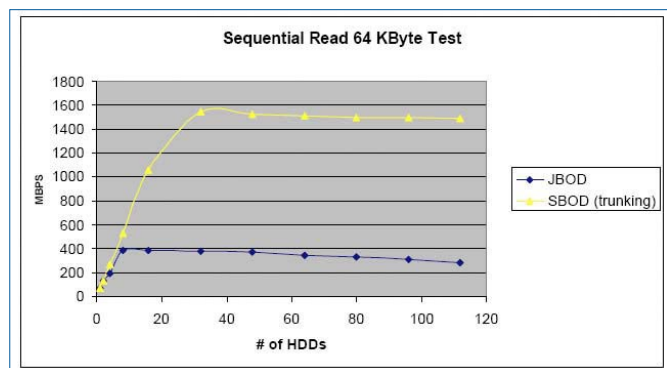


Figure 3.1: SBOD v. JBOD Sequential Read 64KByte Results

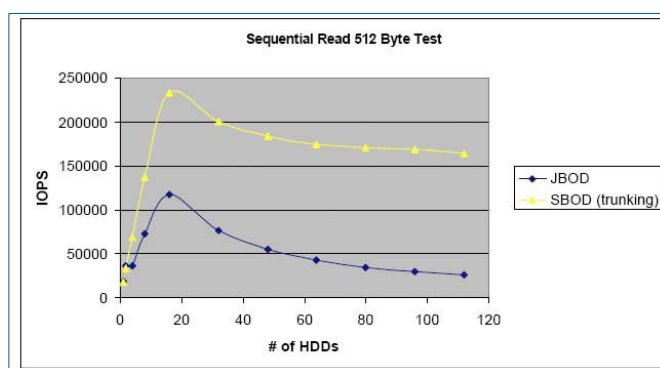


Figure 3.3: SBOD v. JBOD Sequential Read 512Byte Results

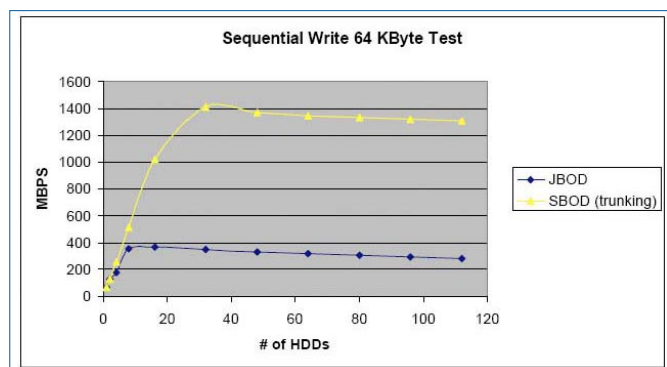


Figure 3.2: SBOD v. JBOD Sequential Write 64KByte Results

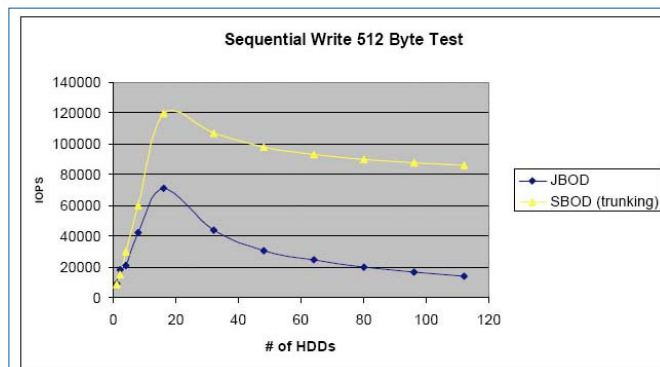


Figure 3.4: SBOD v. JBOD Sequential Write 512Byte Results

Small Block (512 byte) Random I/O Queue Depth Analysis Results

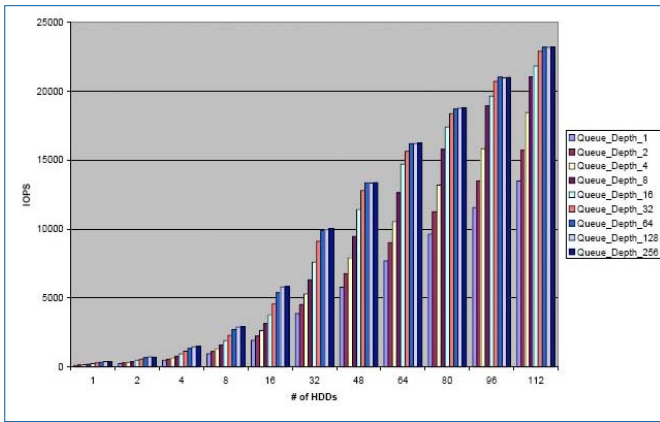


Figure 3.5: SBOD Random Read 512Byte Queue Depth Analysis

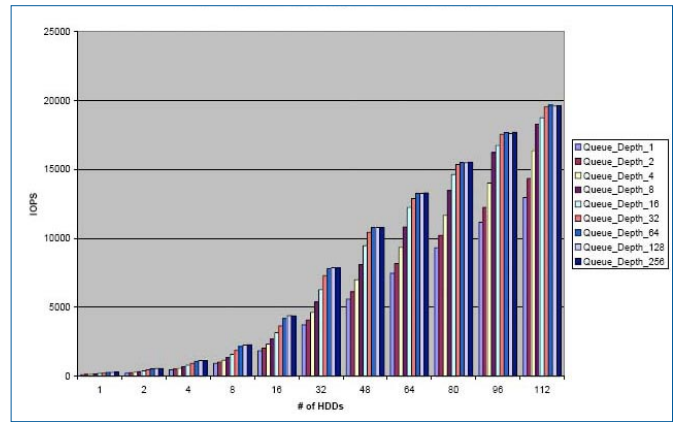


Figure 3.6: SBOD Random Write 512Byte Queue Depth Analysis

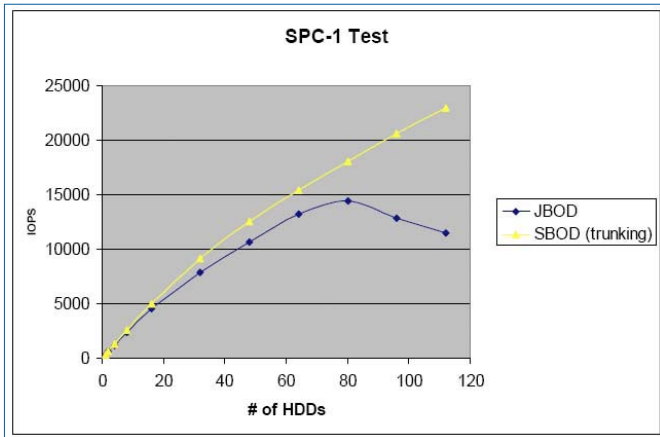


Figure 3.7: SBOD vs. JBOD Simulation Results Comparing IOPS and the Number of HDDs

SPC1 Simulation I/O Results

These results clearly demonstrate the increased scalability of an SBOD populated loop. The roll off in JBOD performance is due to loop arbitration saturation, the point at which this inflection point occurs will reduce to fewer disks when higher performance (15K RPM) disks are used or as the load changes to produce higher disk performance. In other words, as the system is populated with higher performance components or as the system is tuned for maximum performance, less drives are needed to realize the dramatic performance and reliability benefits SBOD technology.

Additional Summary Results – Detailed Queue Depth Analysis

The following results are included to help the reader assess the effects of queue depth and an increased process thread count on the measured performance. These results are a more detailed version of the summary results posted earlier in this document.

Large Block (64k) Sequential I/O Queue Depth Analysis Results

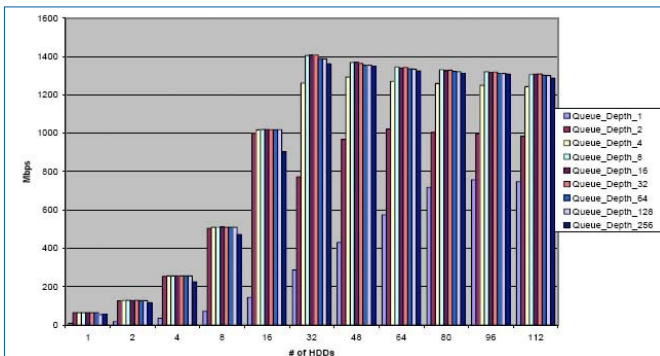


Figure 4.1: SBOD Sequential Read 64K Queue Depth Analysis

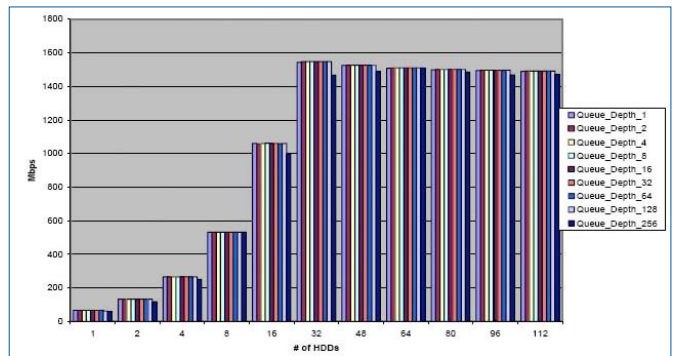


Figure 4.2: SBOD Sequential Write 64K Queue Depth Analysis

Small Block (512 byte) Sequential I/O Queue Depth Analysis Results

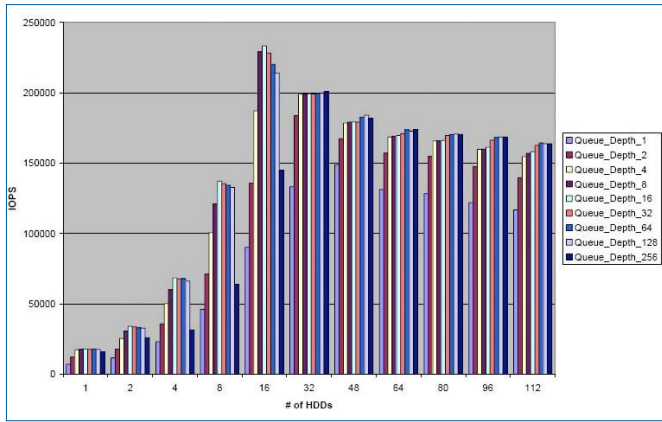


Figure 4.3: SBOD Sequential Read 512bytes Queue Depth Analysis

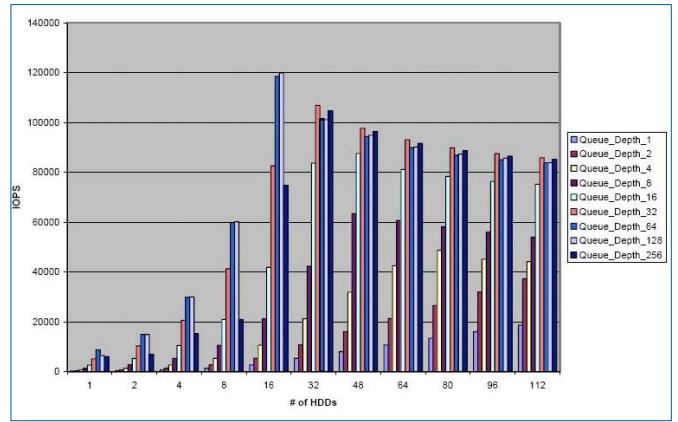


Figure 4.4: SBOD Sequential Write 512bytes Queue Depth Analysis

Small Block (512 byte) Random I/O Queue Depth Analysis Results

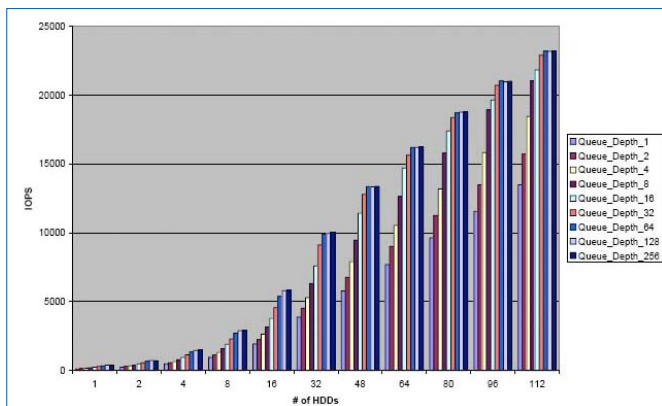


Figure 4.5: SBOD Random Read 512bytes Queue Depth Analysis

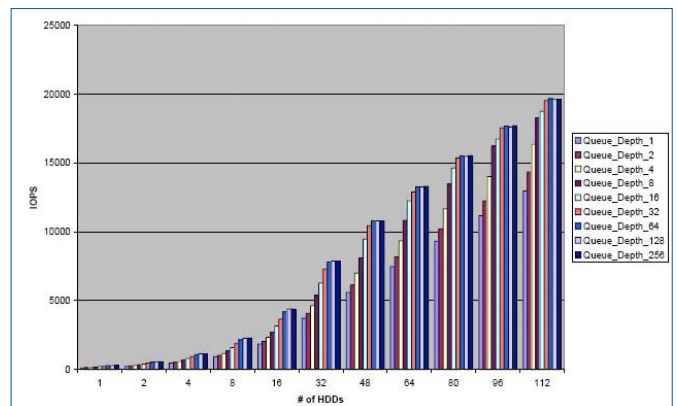


Figure 4.6: SBOD Random Write 512bytes Queue Depth Analysis

SPC-1 Simulation I/O Queue Depth Analysis Results

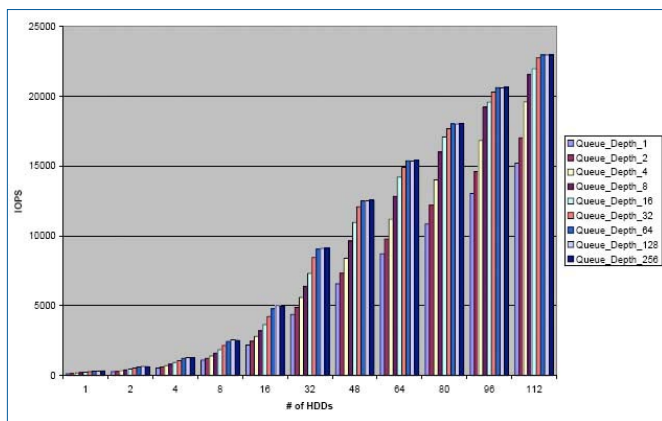


Figure 4.7: SPC-1 Simulation I/O Queue Depth Analysis Results

