ORION BENCHMARK RESULTS

RESILIENT LOW-COST STORAGE FOR ORACLE DBS

Luca Canali, CERN, Dec 2005.

OVERVIEW

Database services for Physics at CERN are currently based on Oracle 10g RAC on Linux. Database storage is managed by Oracle's ASM, which provides the function of volume manager and cluster filesystem for Oracle database storage. ASM takes care of striping and mirroring data across the available storage, therefore it allows the used of low-cost SATA disks. Disk arrays with fiber channel controllers and a fiber channel network are used to manage the storage.

The following describes tests performed to measure key-performance of such a storage configuration. Care has been taken to use hardware and software configurations that are consistent with the currently deployed Oracle databases. In particular the testing tool used, Oracle ORION, is used to generate I/O workloads similar to what is produced by Oracle 10g on ASM.

TEST CONFIGURATION

The key configuration elements for the tests presented in the following are:

- SATA disks spinning at 7200 rpm are used
- 16 SATA disks are mounted in a disk array with a fiber channel controller (Infortrend)
- No hardware RAID is configured on the array controller, LUNs are mapped directly to physical disks
- The array controller as 1GB battery-backed cache RAM in write behind mode
- 2Gbps SAN network is used to connect the disk array to the server via a Fiber Channel switch
- Qlogic QLA2312 HBA is used
- The test server is a dual 3GHz XEON installed with Linux RHEL 3.
- Storage is mapped to Linux as a set of raw devices. LUNs-to-physical-disks and partitions-to-raw-devices mappings are chosen so that each raw device addresses the external half of each SATA disk in the storage array (16 SATA HD).
- Orion 10.2 is used as the test tool

This hardware and OS configuration is consistent with the current production databases deployed at CERN for the physics database services, where Oracle 10g RAC + Oracle ASM are used. Building Oracle storage as detailed in the list above fits in the framework of Oracle resilient low-cost storage initiative [Ref 2]. Note also that the use of the external partition of each disk to increase I/O performance is currently used in the production databases.

IO PERFORMANCE - READ-ONLY WORKLOAD

In this paragraph the measured throughput, IOPS and latency for read-only workloads are reported. Those performance metrics are displayed as graphs: on the vertical axis the metric value is reported, while the horizontal axis reports the workload value. The workload for Orion tests is represented by the number of outstanding asynchronous I/Os: Orion issues asynchronous I/Os requests at an increasing rate and then measures performance metric values for each workload

value [Ref 1]. Three metrics are reported and for each three different tests are reported with increasing number of disks, as show in the caption. The runtime options used are "-run simple" (Ref. 1).

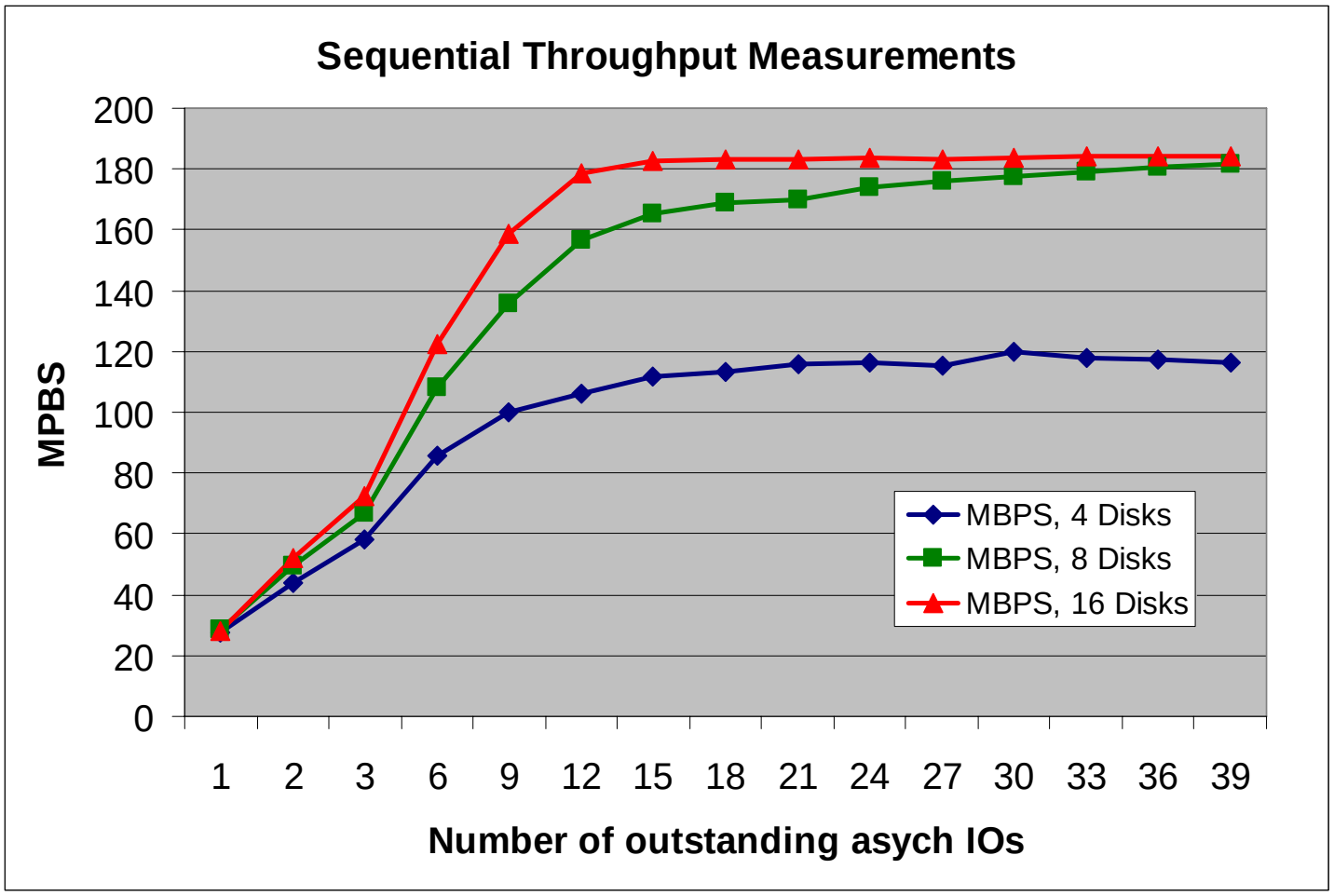
The findings reported here are in good agreement with the paper by J. Loaiza and S.Lee [Ref. 2]:

- The expected maximum throughput for multi-user sequential I/O (using 2Gpbs fiber channel connections and ATA disks) according to Ref. 2 is 180MBPS for arrays with 6 disks or more. The measurements reported below show that the sequential I/O throughput saturates at 180MBPS for 8 SATA disks or more (the 6-disk measurement was not done).
- Ref. 2 indicates an expected rate of 80 IOPS per ATA disks at 7200 rpm (as opposed to the expected 150 IOPS for Fiber channel disks). Measurements reported below show about 100 IOPS per SATA disk in our configuration.

Sequential Throughput Measurements

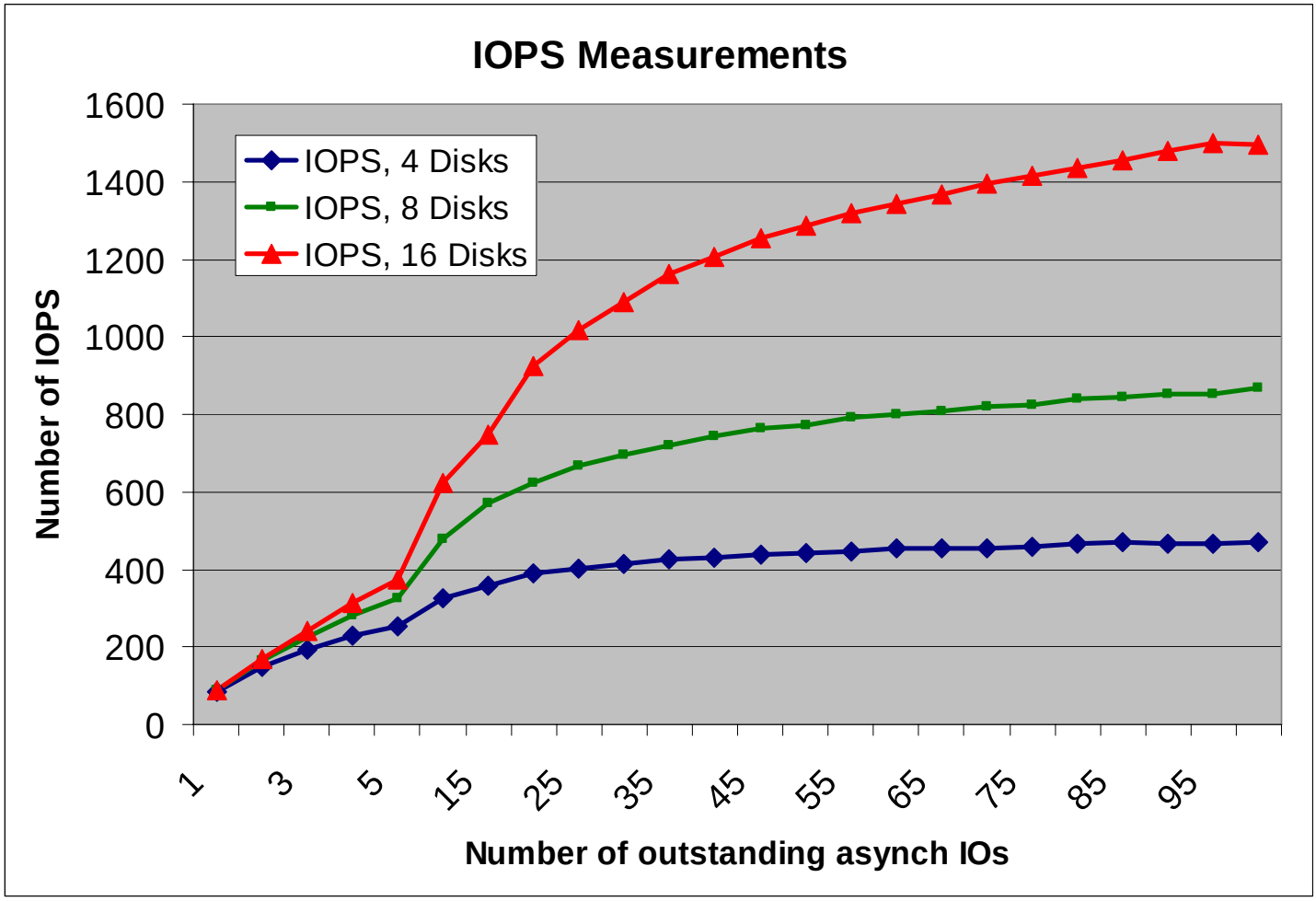
Throughput measurements for read-only sequential workload are reported here. Sequential IO is implemented by ORION as a series of 1MB read requests. The throughput is limited by the Fiber Channel bandwidth (2Gbps). Each disk in the storage arrays has a throughput of about 30MBPS.

Note: further tests performed using QLogic multipathing (LUN load balancing over the 2 HBA ports of the storage array) showed that the saturation threshold for sequential I/O is at 200MB/sec instead of 180 M/s as show here.



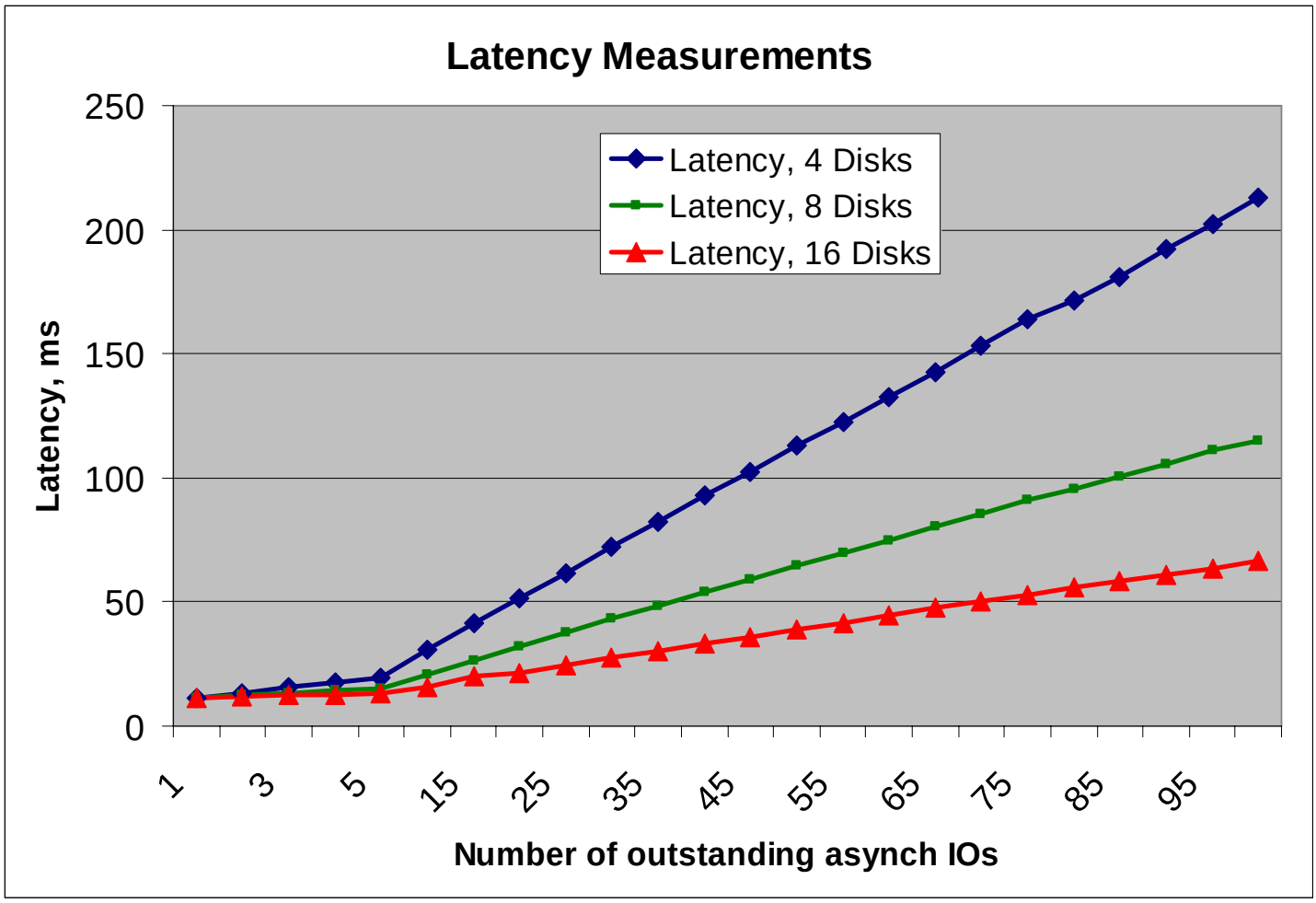
IOPS Measurements

I/O operations per seconds for read-only workload are reported here. Total IOPS are limited by the SATA disks maximum number of operations per seconds: about 100 IOPS per disk.



Latency Measurements

I/O latency measurements for read-only workload are reported here. Latency is lower when more spindles are used, as expected.

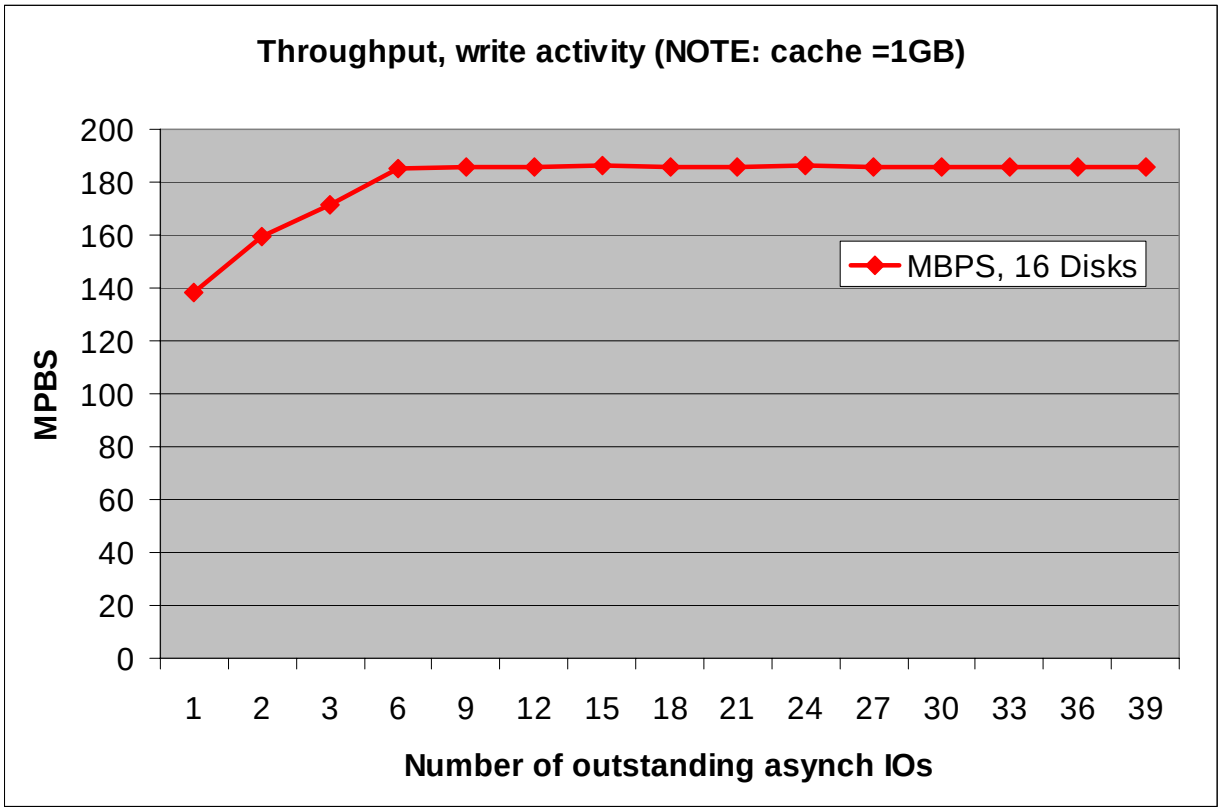


IO PERFORMANCE - WRITE-ONLY WORKLOAD

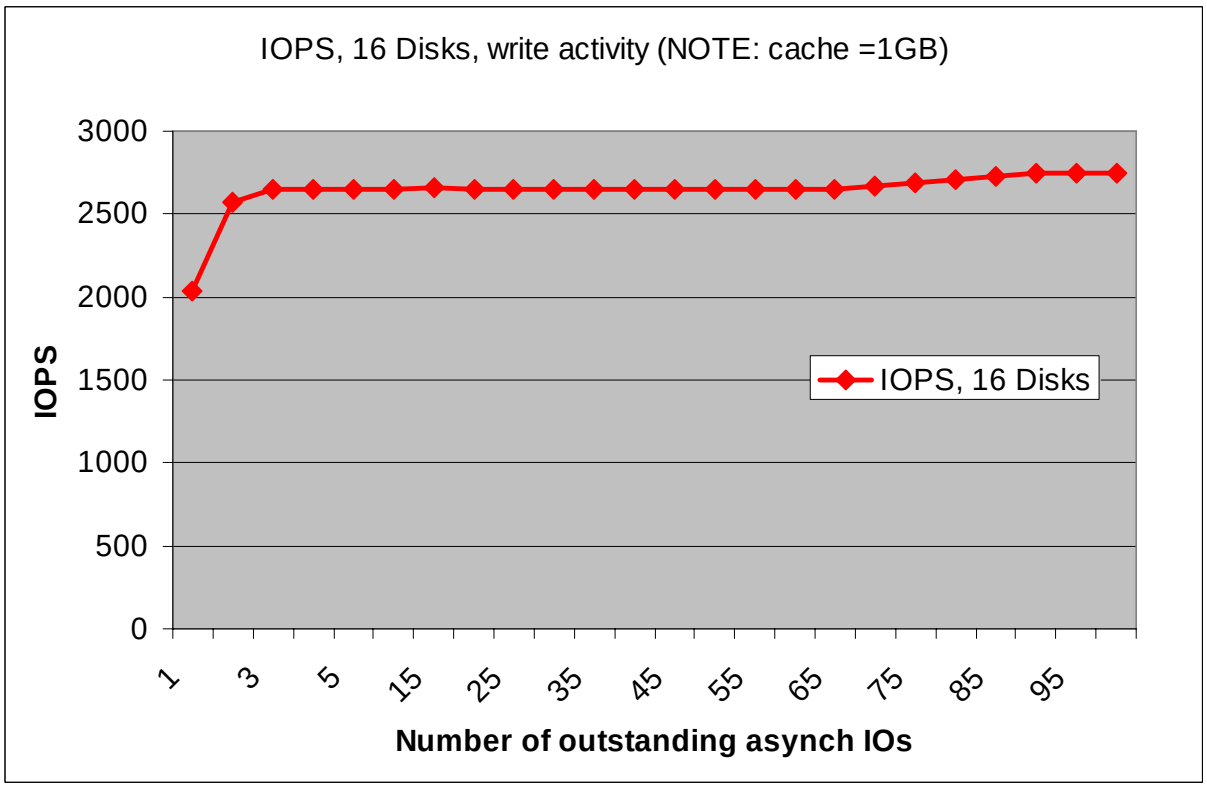
IO metrics measured for write-only workload are presented here. The results are highly influenced by the diskarray controller cache (1GB) and the way tests are performed by ORION. The net effect is that performance metrics for write-only activity measured with Orion are far better that the corresponding values measured for the read-only workload (see above).

We expect that in 'real world' database environments the actual metric values at peak performance may be much lower that what measured here. The findings presented here are consistent with the discussion of J Loaiza and S. Lee [Ref 2].

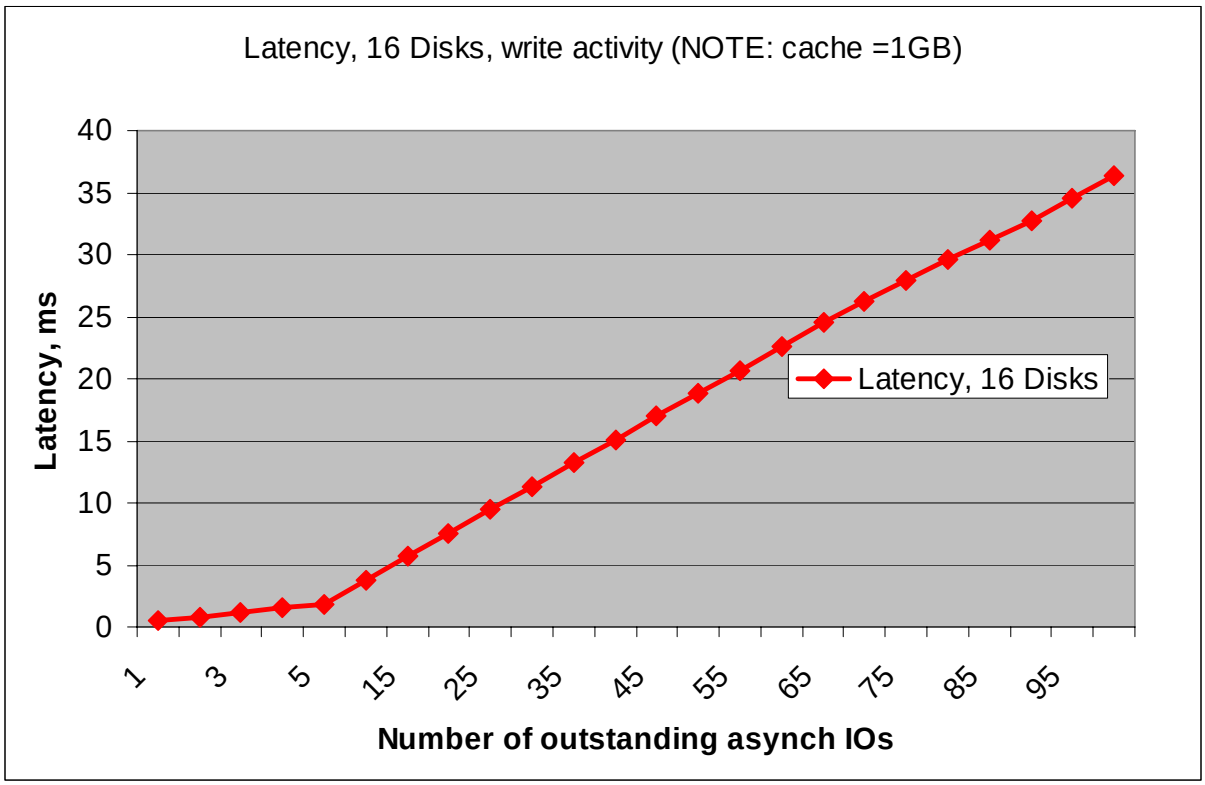
Sequential Throughput Measurements



IOPS Measurements



Latency Measurements



Summary

Performance measurements are reported and discussed for a storage solution built on low-cost SATA disks over a SAN fiber channel infrastructure. Oracle ORION was used to gather the performance data. Storage and server configurations used for these tests are nearly equivalent to the production configuration deployed for the physics database services at CERN. Moreover, ORION I/O workload and tests are designed to simulate Oracle database I/O when ASM is used as volume manager.The most notable difference is that mirroring is not implemented in the tests reported here.

Read-only and write-only workloads have been tested. The results are in good agreement with the expected figures reported by J. Loaiza and S. Lee. For example, read-only throughput was show to reach a saturation level of 180 MBPS (200 MBPS with multipathing) with 8 SATA disks, while IOPS was show to scale up to about 100 IOPS per installed SATA disk. Write-only workload was shown to take good advantage of the controller cache in write behind mode.

References

1. Oracle Orion documentation, http://otn.oracle.com

2. J Loaiza and S Lee, OOW 2005 session 1262, http://www.oracle.com/technology/deploy/availability/pdf/1262_Loaiza_WP.pdf

3. J Loaiza, OTN 1999, Optimal storage configuration made easy, http://www.oracle.com/technology/deploy/availability/pdf/oow2000_same.pdf

4. L. Canali, scalable Oracle 10g architecture, https://twiki.cern.ch/twiki/pub/PSSGroup/HAandPerf/Architecture_description_Feb05.pdf