Dell EqualLogic Best Practices Series

Sizing and Best Practices for Deploying Oracle 11g Transaction Processing Databases on Dell EqualLogic Storage

A Dell Technical Whitepaper

Chidambara Shashikiran

Storage Infrastructure and Solutions
Dell Product Group
September 2010

Table of Contents

1 Introduction

A key challenge for Oracle® database and SAN administrators is to effectively design and manage system storage, especially to accommodate performance, capacity and future growth requirements.

The goal of this paper is to present the results of Oracle I/O performance tests conducted by Dell Labs, and provide sizing guidelines and best practices based on those results for running Oracle pure OLTP workloads on Dell™ PowerEdge™ Servers and Dell™ EqualLogic™ virtualized iSCSI SANs.

The best practices and deployment models presented in this paper focus on supporting OLTP (Online Transaction Processing) workloads running on:

- **Oracle Database 11g Release 2 RAC database**
- **Red Hat® Enterprise Linux® 5.5**
- **Dell EqualLogic PS6010XV and Dell EqualLogic PS6010S iSCSI SANs**

1.1 Audience

This paper is intended for Linux system administrators, database administrators and storage architects interested in deploying Oracle OLTP database solutions on Red Hat Enterprise Linux using Dell EqualLogic PS Series SANs. It is assumed that the readers of this document have familiarity with Red Hat Enterprise Linux system administration and Oracle 11g R2 database installation and administration tasks.

1.2 The Rest of this Paper

The rest of this paper contains the following sections:

2 ORION I/O Profiling

The Oracle I/O Calibration Tool (ORION[1]) uses the Oracle I/O software stack to generate simulated I/O workloads without having to create an Oracle database, load the application and simulate users. ORION is commonly used for benchmarking the I/O performance characteristics of Oracle database storage systems.

A series of ORION tests were executed on PS6010XV and PS6010S arrays to measure the I/O performance profiles for EqualLogic PS Series PS6010XV and PS6010S storage arrays. These tests were focused on measuring the storage I/O performance and system limits for pure OLTP workloads. Most production OLTP systems also contain significant query and reporting workloads. Those workload components were not considered in this test. We tested two different Dell EqualLogic storage array models:

- **EqualLogic PS6010XV**[2]:
  - 14 x 300GB 15K RPM SAS disk drives as RAID 10, with two hot spare disks
  - 10GbE dual-port controller running firmware version 5.0.0
- **EqualLogic PS6010S**[3]:
  - 14 x 100GB Solid State Disk (SSD) drives as RAID 10, with two hot spare disks
  - 10GbE dual-port controller running firmware version 5.0.0

For each of the array models tested a single default pool containing 10 50GB volumes was used, as shown in Figure 1.

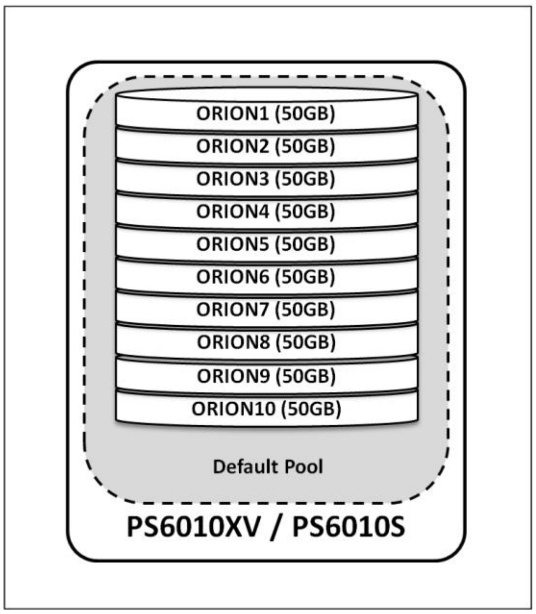

Figure 1 ORION Test Volumes

[1] http://www.oracle.com/technetwork/topics/index-089595.html
[2] EqualLogic PS 6010XV Product details: http://www.equallogic.com/products/default.aspx?id=8973
[3] EqualLogic PS 6010S Product details: http://www.equallogic.com/products/default.aspx?id=8957

2.1 ORION OLTP Test Configuration

Our goal in this test was to measure EqualLogic storage subsystem I/O performance when running Oracle pure OLTP workloads. The ORION tests were performed using the system configuration shown in Figure 2. Component level details for this system configuration are provided in Appendix A .

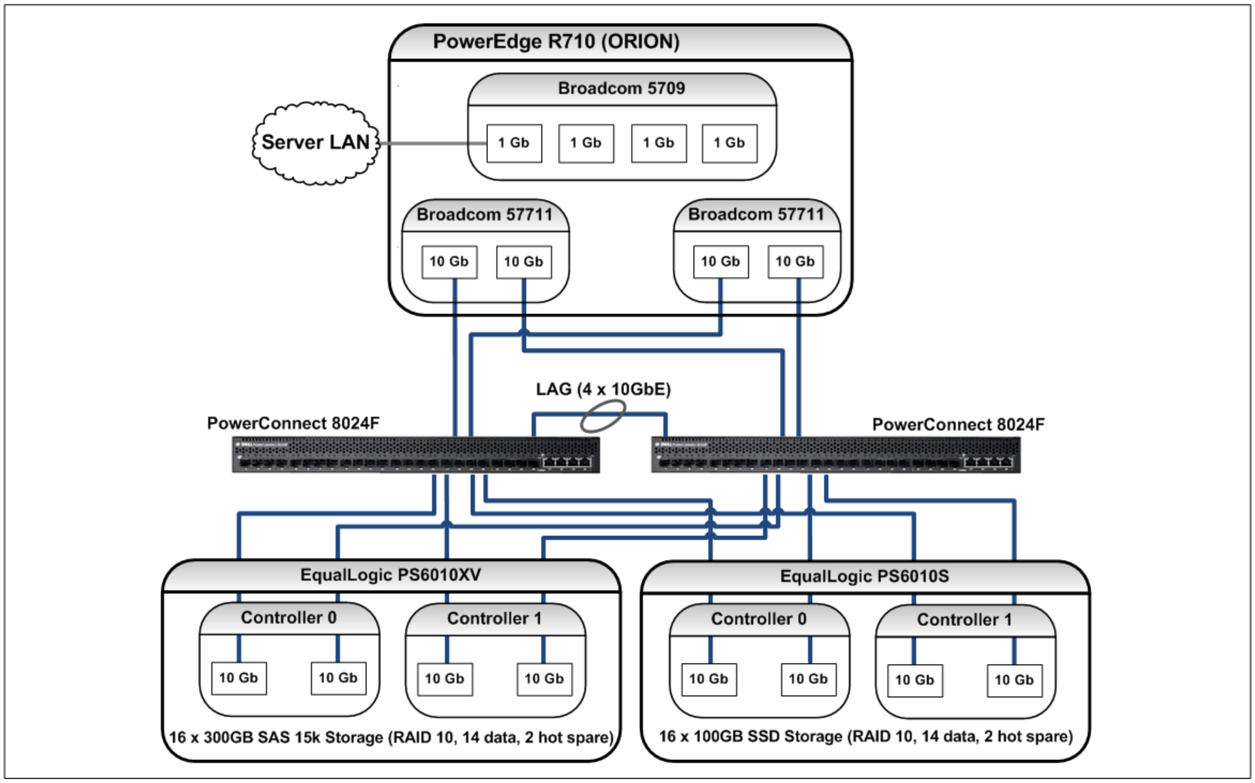

Figure 2 ORION Test System Configuration

In our initial test, we simulated OLTP database transactions using random 8K small I/O at increasing loads to determine the maximum IOPS the storage array could sustain. Even though a mix of 70% read and 30% write random 8K I/O represents majority of Oracle OLTP workloads, there may be OLTP workloads that use higher block sizes. So we also executed ORION random I/O tests using 16K, 32K and 64K block sizes to measure I/O performance under those I/O workloads.

The following ORION test parameters were used:

-oltp	Simulates random 8K I/O at increasing workloads
-write	Set to 30 to create a 70% read and 30% write I/O ratio-
-advanced	Used in conjunction with "-size_small" to simulate other block sizes (16K, 32K and 64K)

Table 1 shows the ORION test sequence.

Test	Database Block Size (KB)	I/O Mix
1	**8**	100% Read[4]
2	**8**	70% Read / 30% Write[5]
3	**16**	70% Read / 30% Write
4	**32**	70% Read / 30% Write
5	**64**	70% Read / 30% Write

Table 1 ORION OLTP Test Sequence

2.2 ORION OLTP Test Results: EqualLogic PS6010XV

The ORION OLTP tests results for the PS6010XV are shown in Table 1. In each test we increased the workload (the queue depth[6]) while measuring IOPS and corresponding disk latency. The workload was increased until a maximum IOPS performance level was reached. In Figure 3 we show a chart of the test data collected from two 8K block size test runs: 100% read, and 70/30 read/write (tests 1 and 2 in Table 2).

Test	Database Block Size (KB)	I/O Mix (%read/%write)	Maximum IOPS at latency <= 20ms
1	**8**	100/0	**4857**
2	**8**	70/30	**4047**
3	**16**	70/30	**3810**
4	**32**	70/30	**3569**
5	**64**	70/30	**3010**

Table 2 PS6010XV ORION OLTP Test Results

[4] 100% read is not a typical OLTP I/O pattern. This test was conducted first to measure maximum small random (read only) IOPS performance for the system.
[5] 70/30 is the generally accepted ratio for read/write mix in typical OLTP workloads
[6] Number of outstanding I/Os

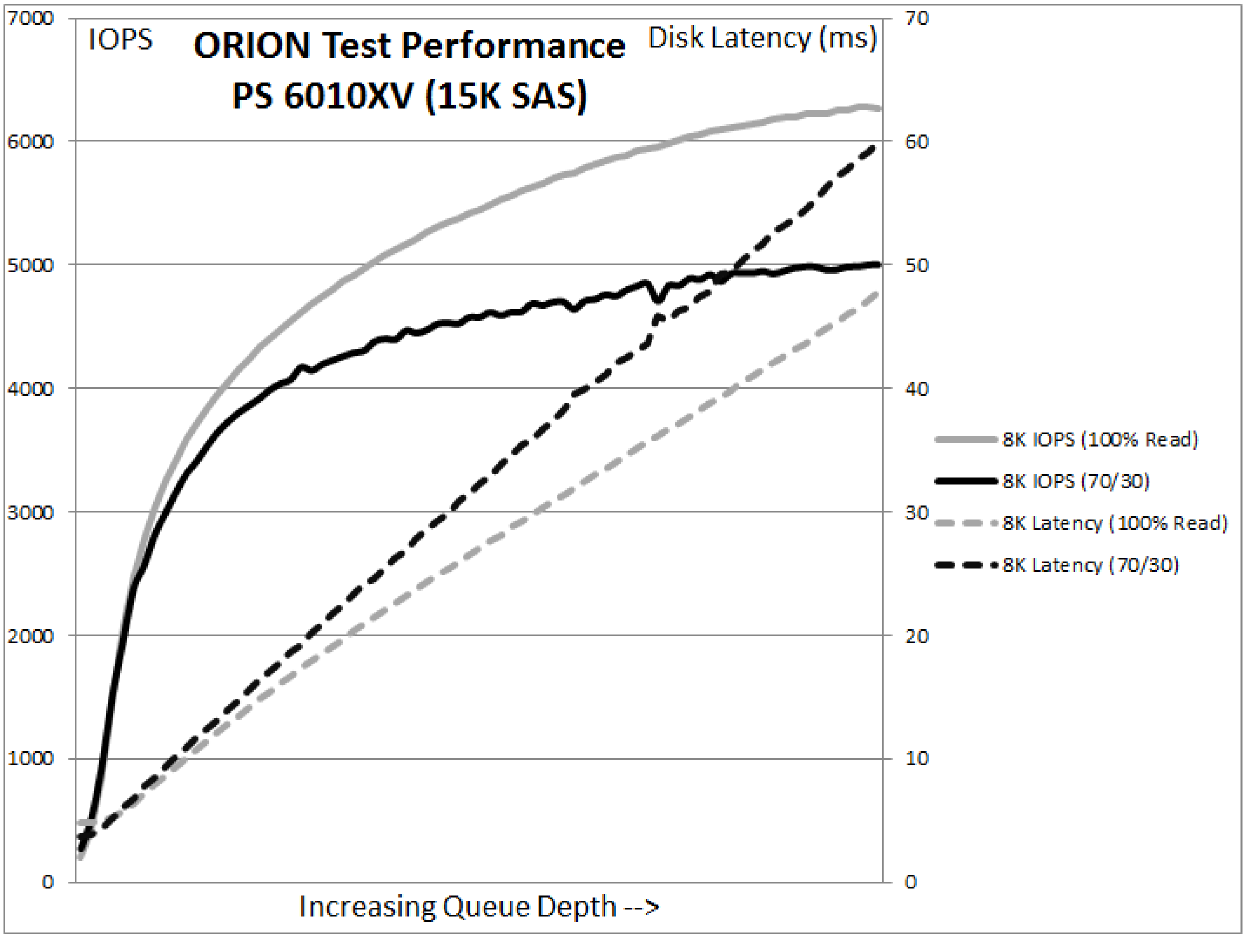

Figure 3 PS6010XV ORION OLTP 8K Test Results

Summary of PS6010XV ORION test results:

- As expected, average IOPS increased with the queue depth.
- The PS6010XV array in a RAID 10 configuration produced approximately 4000 IOPS for OLTP type transactions while staying within the generally accepted disk latency limit of 20ms (for both read and write IOPS).
- For block sizes as large as 64K a single 6010XV array is able to sustain approximately 3000 IOPS for a typical OLTP workload.

2.3 ORION OLTP Test Results: EqualLogic PS6010S

We repeated the same ORION tests using a PS6010S array to measure the performance benefits offered by SSDs as compared to traditional SAS hard disk drives used in the PS6010XV.

The test results are shown in Table 3. As in the previous test, we increased the workload (the queue depth) while measuring IOPS and corresponding disk latency. The workload was increased until a maximum IOPS performance level was reached. In Figure 4 we show a chart of the test data collected from the two 8K block size test runs at 100% read and 70/30 read/write (test #1 and #2 in Table 3)

Test	Database Block Size (KB)	I/O Mix (%read/%write)	Maximum IOPS at latency <= 20ms
1	**8**	100/0	**34393**
2	**8**	70/30	**16116**
3	**16**	70/30	**15549**
4	**32**	70/30	**14648**
5	**64**	70/30	**9923**

Table 3 PS6010S ORION OLTP Test Results

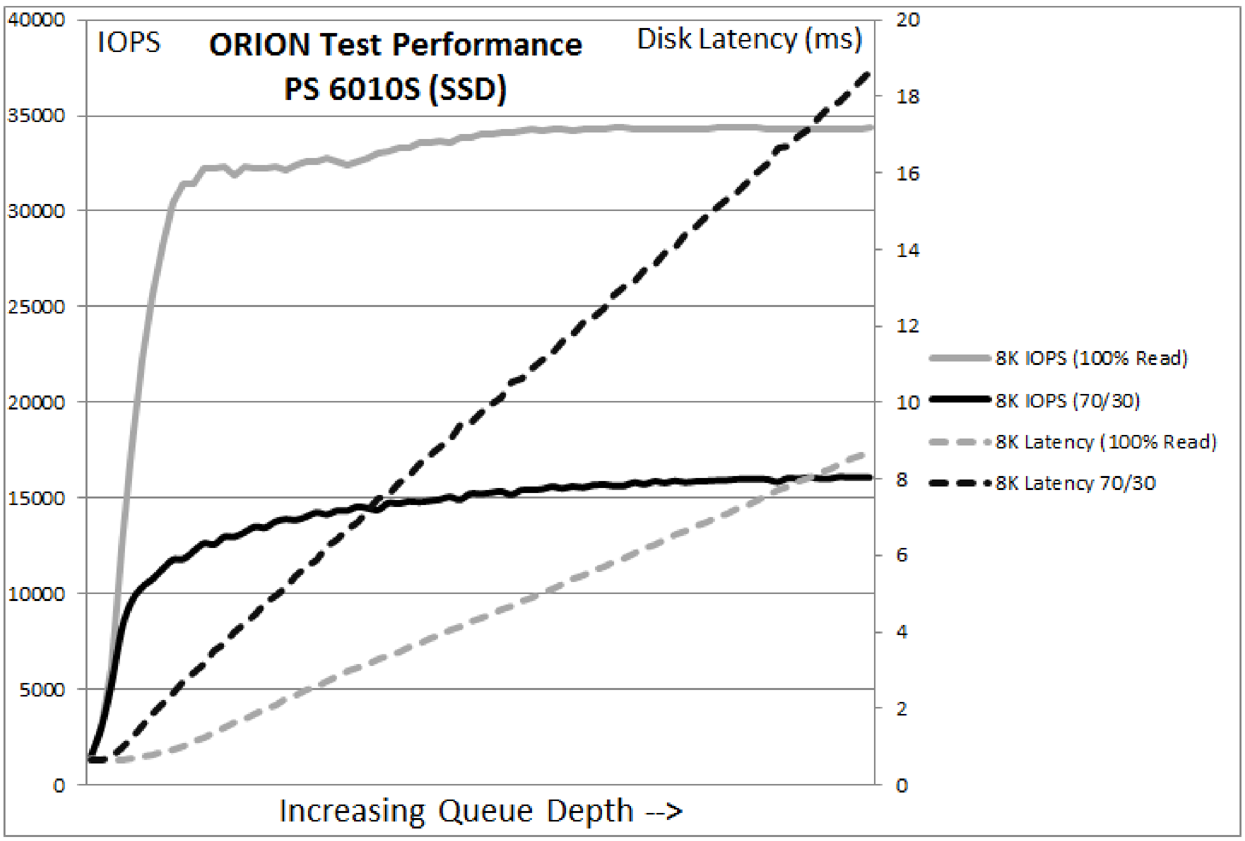

Figure 4 PS6010XV ORION OLTP 8K Test Results

Summary of PS6010S ORION test results:

Similar to the PS6010XV, average IOPS increased with the queue depth. The EqualLogic PS6010S with Solid State Disk (SSD) drives provided high IOPs performance with very low corresponding latencies:

- For 100% read I/O with an 8K block size the PS6010S sustained a maximum of approximately 34000 IOPS.
- For 70/30 read/write I/O with an 8K block size the PS6010S sustained a maximum of approximately 16000 IOPS.
- During both 8K ORION test runs the PS6010S never exceeded the generally accepted disk latency limit of 20ms (for both read and write IOPS), even when the maximum queue depth level for the test was reached.
- At the maximum IOPS level measured for the PS6010XV before exceeding the 20ms disk latency threshold (4047), the disk latency for the PS6010S at the same IOPS level was < 2ms.

Comparing performance: PS6010XV and PS6010S

As shown in Figure 5, the EqualLogic PS6010S configured with SSD drives was able to provide much higher random IOPs performance at much lower corresponding latencies than the PS6010XV configured with 15K SAS drives.

- We measured a 7x IOPS performance increase for the PS6010S over the PS6010XV for 100% read 8K block I/O.
- The PS6010S can sustain 4x IOPS performance increase over the PS6010XV at the 20ms disk latency threshold (16000 vs. 4000) for 70% read 30% write workloads.
- For block sizes as large as 64K the single 6010S array still provided a 3.3x performance increase over the PS6010XV.

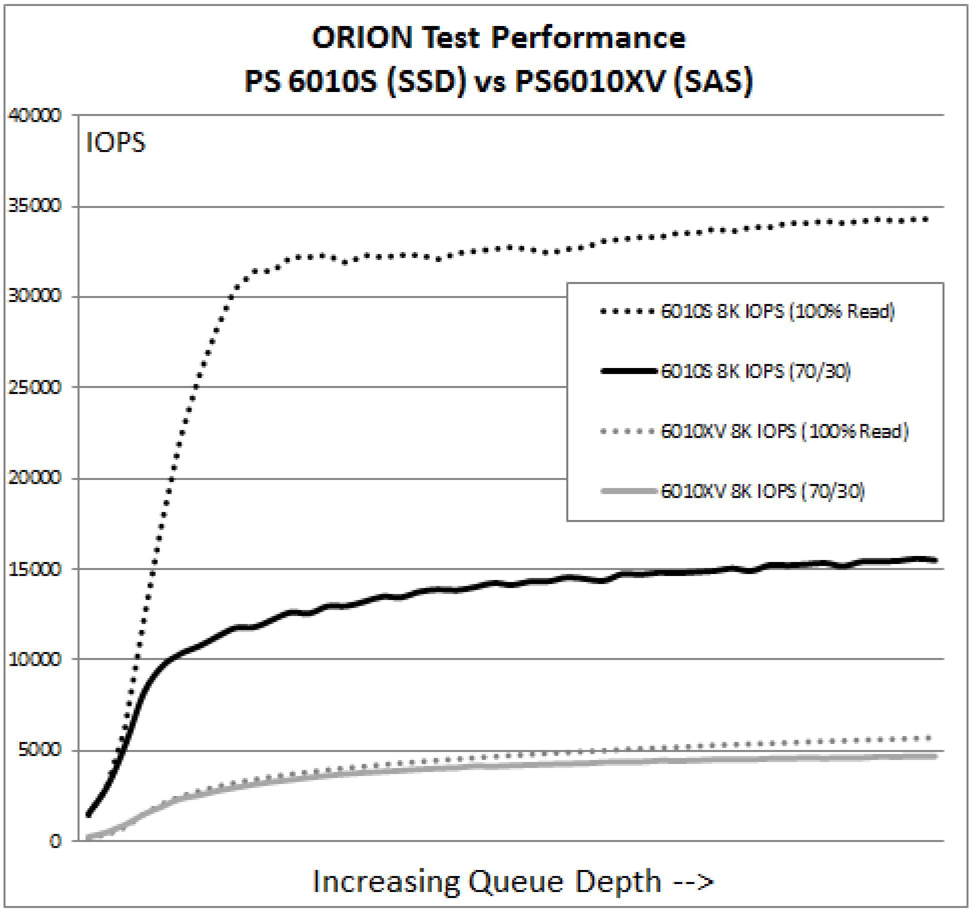

Figure 5 8K ORION IOPS Performance: PS6010S (SSD) vs. PS6010XV (SAS)

3 TPC-C Testing Using Quest Benchmark Factory

Using the ORION test results described in the previous section we calibrated the baseline Oracle OLTP I/O performance levels for PS6010XV and PS6010S arrays (using RAID10). However, ORION does not use an Oracle database. It only simulates the behavior of the Oracle I/O software stack.

In our second testing phase we used Quest Benchmark Factory[7] for Databases to simulate a real TPC-C style workload on our test systems using an actual Oracle 11g R2 RAC database. Quest Benchmark Factory for Databases is a load generator that can simulate OLTP workloads running on a database at various load settings.

Test goals:

A. Simulate a real TPC-C style database transactional workload while gathering Transactions-Per-Second (TPS) and end-to-end transaction response time data under increasing user load.
B. Measure and compare the system performance when increasing the number of EqualLogic volumes used for Oracle database ASM disks.
C. Measure and compare the system performance when using two different dynamic memory management methods:
    - Using Oracle ASMM[8] with manual adjustment of the SGA_TARGET value to automatically tune SGA components.
    - Using Oracle AMM[9] with manual adjustment of the MEMORY_TARGET value to automatically tune SGA and PGA components.

3.1 TPC-C Test System Configuration

The TPC-C tests using Quest Benchmark Factory were performed using the system configuration shown in Figure 6. Component level details for this system configuration are provided in Appendix A .

> **Note:** The Oracle RAC installation was completed using the following Dell publication as a guide:
>
> ➢ *Dell PowerEdge Systems Oracle Database on Enterprise Linux x86_64 Database Setup and Installation Guide Version 1.5*: http://support.dell.com/support/edocs/software/appora11/lin_x86_64/1_4L/multlang/DBIG/DBIGR2.pdf

We used a two-node Oracle 11g R2 RAC database. Some of the key configuration details for the database are described below:

- For this phase of our testing only the EqualLogic PS6010XV array was tested.
- We installed Oracle 11g R2 clusterware on the two Dell PowerEdge R710 "database" servers shown in Figure 6.

---

7 http://www.quest.com/benchmark-factory/
8 Automatic Shared Memory Management feature introduced with Oracle 9i. See: http://download.oracle.com/docs/cd/B28359_01/server.111/b28310/memory004.htm
9 Automatic Memory Management feature introduced with Oracle 11g. See: http://download.oracle.com/docs/cd/B28359_01/server.111/b28310/memory003.htm

- The database elements, redo logs, archive and flash data were managed using Oracle ASM[10] (Automatic Storage Management).
- The Oracle clusterware files (OCR and voting disk) were stored using the ASM cluster file system (ACFS)

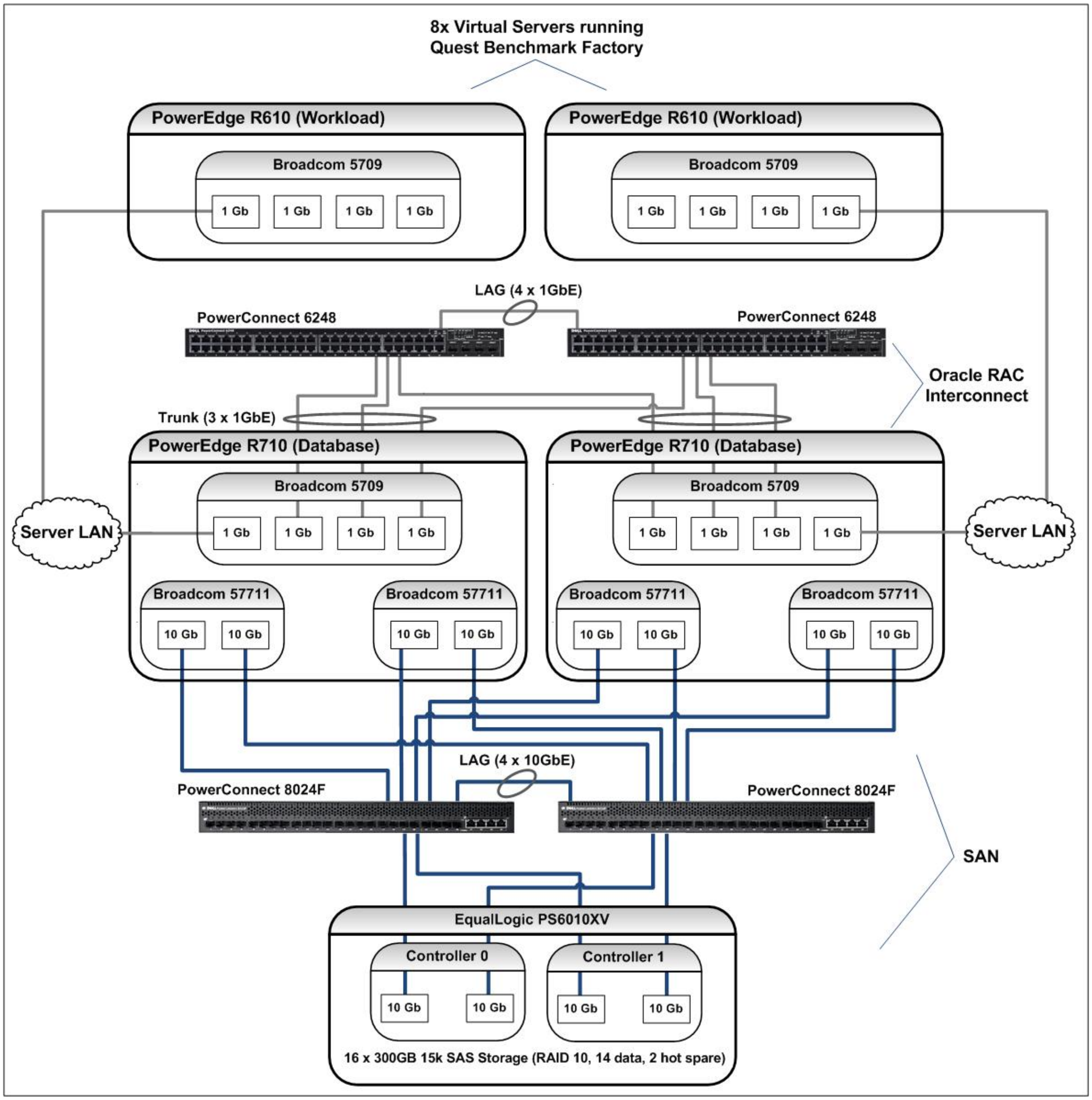

Figure 6 TPC-C Test System Configuration

---

[10] http://www.oracle-base.com/articles/10g/AutomaticStorageManagement10g.php

> **Note:** Oracle recommends using ASM (Automatic Storage Management) for simplified administration and optimal I/O balancing between disk groups. Oracle ASM cluster file system (ACFS) is an extension to ASM introduced with the Oracle 11g R2 release to support general purpose files other than database files. Similar to Oracle ASM, ACFS is the preferred file manager for storing non-database files.

3.1.1 Database Layouts

To compare the system performance when increasing the number of EqualLogic volumes used for Oracle database ASM disk partitions we created two different storage layouts. Both of the storage layouts used the following three ASM disk groups:

ORADB Database files; temporary table space; online redo logs; system related table spaces such as SYSTEM and UNDO

ORAFLASH Archive logs and backup data

ORACRS ASM Cluster file system (ACFS) for storing clusterware related information such as the OCR and voting disks

The ORAFLASH and ORACRS disk group configurations were identical for both storage layouts tested. Table 4 shows the ASM Disk configuration used in each ASM Group for both test layouts.

<table>
<tr><th>ASM Disk Group:</th><th>ORADB</th><th>ORAFLASH</th><th>ORACRS</th></tr>
<tr><td>ASM Disks
Test Layout #1</td><td>ORADB1 (100GB)
ORADB2 (100GB)</td><td rowspan="2">ORAFLASH1 (200GB)
ORAFLASH2 (200GB)</td><td rowspan="2">ORACRS1 (10GB)</td></tr>
<tr><td>ASM Disks
Test Layout #2</td><td>ORADB1 (100GB)
ORADB2 (100GB)
ORADB3 (100GB)
ORADB4 (100GB)</td></tr>
</table>

Table 4 ASM Disk Group Configuration

For each storage layout, a single EqualLogic volume was create and dedicated to each ASM disk. The storage layouts for both test configurations are shown in Figure 7 and Figure 8.

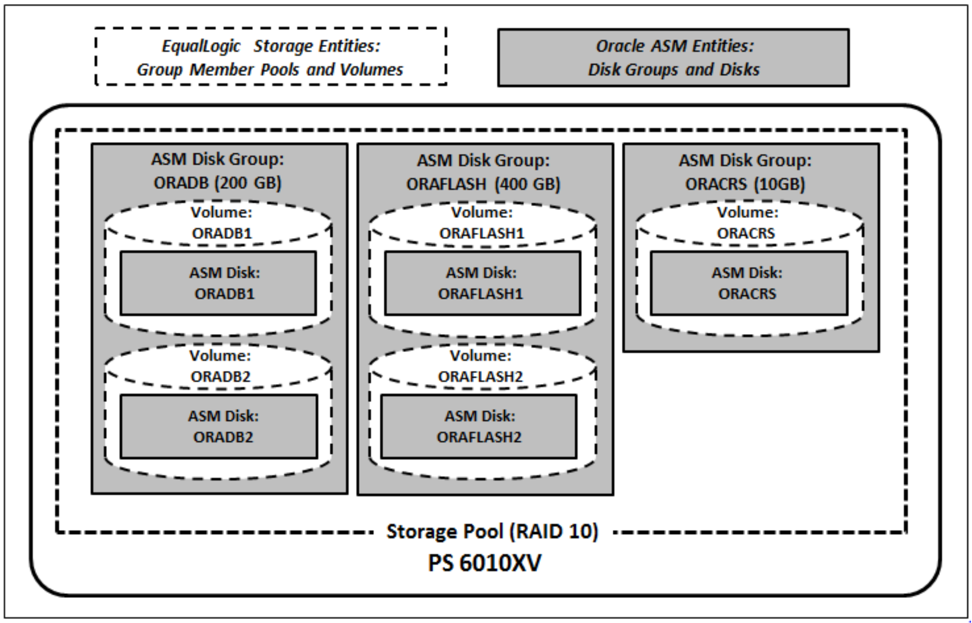

Figure 7 Oracle Database Layout: ORADB Disk Group Using 2 ASM Disks/EqualLogic Volumes

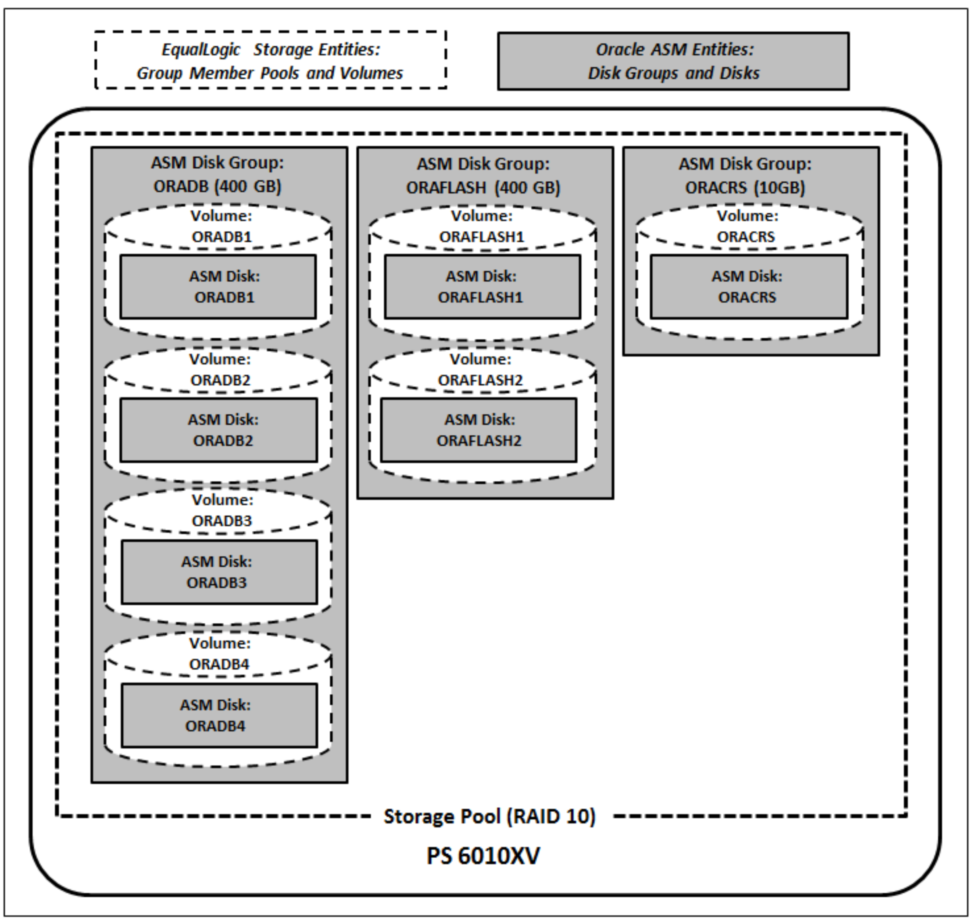

Figure 8 Oracle Database Layout: ORADB Disk Group Using 4 ASM Disks/EqualLogic Volumes

3.2 TPC-C Test Results

For each test, the workload was increased from 1000 to 8000 users until the maximum performance level was reached. Maximum performance corresponds to the highest sustained transactions per second (TPS) achieved before the application response time exceeds 2 seconds. The application response time is measured from the time the SQL statement was issued until the result was received. This is an aggregate measurement that includes response times attributable to all components in the system, including disk I/O response time.

3.2.1 TPS and Response Time

To meet our testing goals we conducted multiple test runs. The best system performance was measured when using the following variations in the test configuration:

- Using Oracle AMM (instead of ASMM) for dynamic memory management
- AMM MEMORY_TARGET setting = 40GB
- Four database ASM disk/EqualLogic Volumes in use (as shown in Figure 8)

Figure 9 shows the TPS and response time profile for the system configured using the above settings. The test run starts at the initial 1000 user load. The user load incrementally increases until the application response time began to exceed the two second threshold, which was reached at a user load of approximately 7000. The maximum sustained TPS achieved in this configuration before exceeding the 2 second response time threshold was 348.

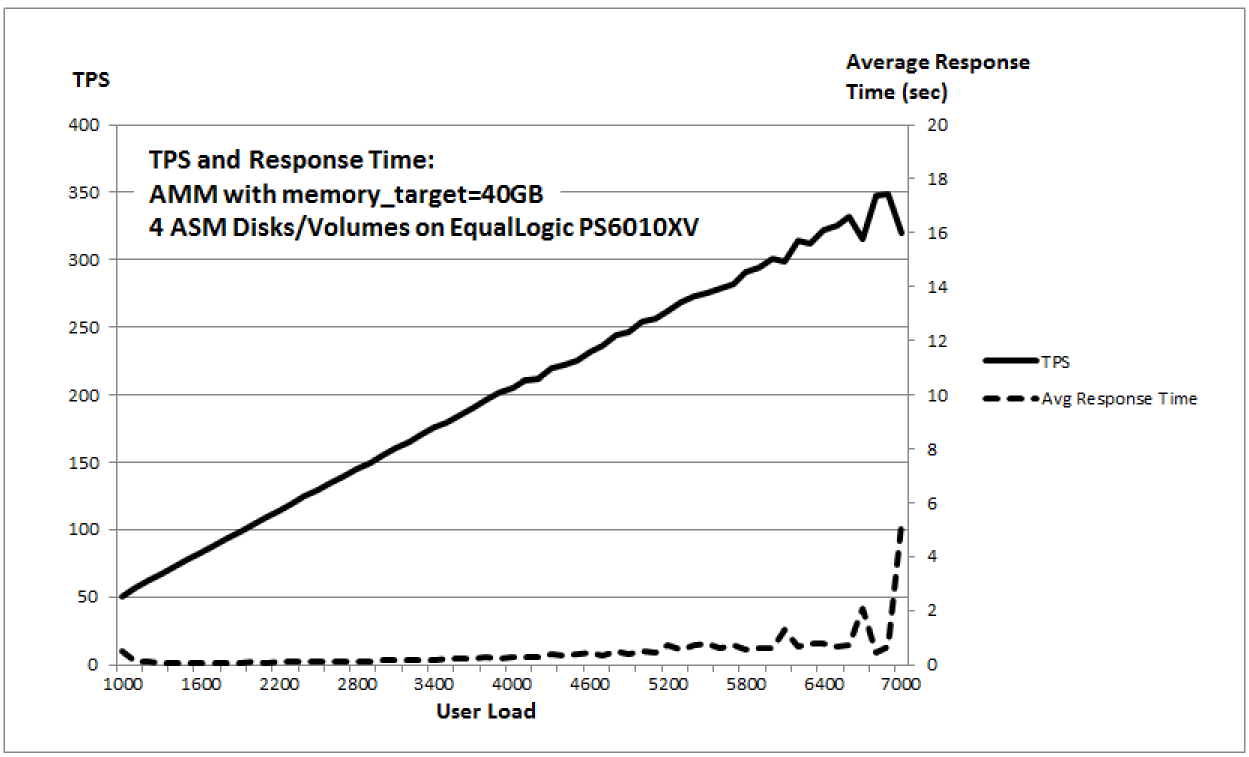

Figure 9 TPC-C Simulation Test Results: AMM MEMORY_TARGET=40GB, Four ASM Disk Volumes

3.2.2 Comparison: Two EqualLogic Volumes vs. Four EqualLogic Volumes

In this test we measured system performance when running the TPC-C user loading test using two database ASM disks contained within two separate EqualLogic volumes. We used the database layout shown in Figure 7. We compared the results of that test against the same test run using four database ASM disks contained within four separate EqualLogic volumes (the layout shown in Figure 8). For both tests we used the same system settings, using Oracle AMM with MEMORY_TARGET=40GB). Table 5 shows the maximum TPS and corresponding average response times for each configuration.

Test	Storage Layout	Maximum Sustained TPS	Application Response Time (seconds)	User Load at Maximum TPS
1	Two ASM Disks, Two Volumes (ref. Table 4, Figure 7)	**294.50**	**0.18**	5700
2	Four ASM Disks, Four Volumes (ref. Table 4, Figure 8)	**348.61**	**0.67**	6900

Table 5 Test results: Two vs. Four EqualLogic Volumes

The results in Table 5 show that increasing the number of EqualLogic volumes used by the Oracle database ASM disk group increased the database performance by 18%. However, additional tests conducted by Dell Labs in similar configurations show minimal increases in system performance once you use more than five EqualLogic volumes for ASM disks.

3.2.3 Dynamic Memory Management Comparison: Oracle ASMM vs. Oracle AMM

The most important and complex aspect of system tuning for maximizing Oracle database performance is memory management. Oracle has introduced many new features to simplify this task. Starting with Oracle 10g, Automatic Shared Memory Management (ASMM) was introduced to help automate configuration of SGA components using a single parameter called SGA_TARGET. With the Oracle 11g release, Automatic Memory Management (AMM) was introduced to help automate and simplify this configuration tuning task by allowing you to manage the SGA and PGA components using a single database initialization parameter called "MEMORY_TARGET".

For this test we measured system performance when using both types of dynamic memory management described above. We ran three tests using ASMM while varying the SGA_TARGET setting between 32GB, 40GB and 48GB. Then we ran the same three tests using AMM and varying the MEMORY_TARGET parameter using the same settings. We used the 4 ASM disk/volume storage lay out shown in Figure 8 for each of the tests. A summary of the test results is shown in Table 6 and the chart in Figure 10 below.

Test	Tuning Parameter	Maximum Sustained TPS	Response Time (seconds)	User Load at Maximum TPS
1	**Using ASMM:** SGA_TARGET = 32GB, HugePages[11] set	**287.60**	**0.068**	5500
2	**Using ASMM:** SGA_TARGET = 40GB, HugePages set	**299.99**	**0.235**	5900
3	**Using ASMM:** SGA_TARGET = 48GB, HugePages set	**275.32**	**0.166**	5400
4	**Using AMM:** MEMORY_TARGET = 32GB	**260.46**	**0.758**	5300
5	**Using AMM:** MEMORY_TARGET = 40GB	**348.61**	**0.671**	6900
6	**Using AMM:** MEMORY_TARGET = 48GB	**284.9**	**0.18**	5500

Table 6 Dynamic Memory Management Test Results: ASMM vs. AMM

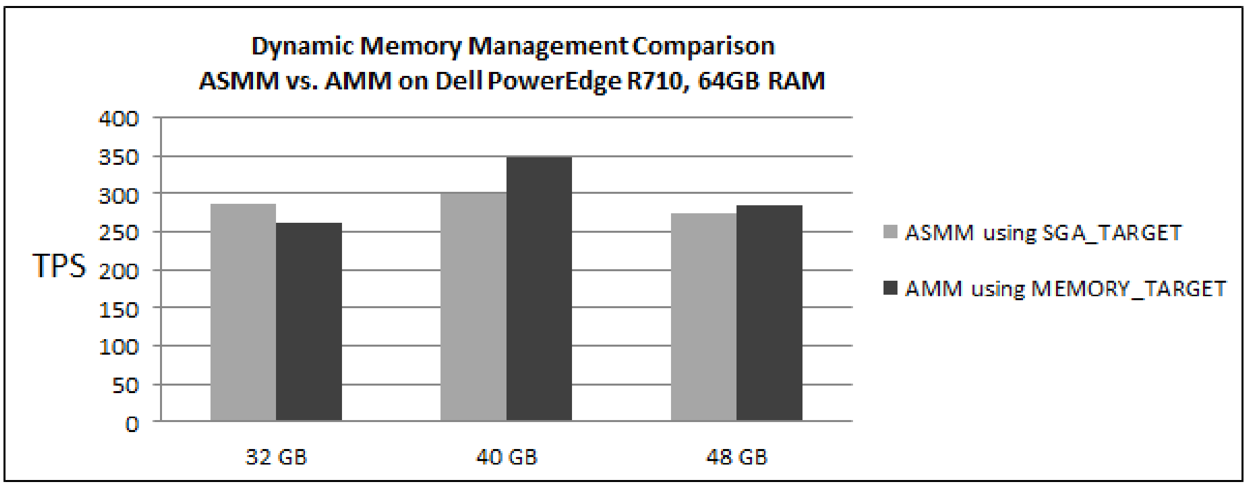

Figure 10 Dynamic Memory Management Comparison: ASMM vs. AMM

These results shown in Table 6 and Figure 10 indicate the following:

- For the Dell PowerEdge R710 server configured with 64GB of RAM, the best TPC-C performance was achieved when SGA_TARGET = 40GB in the first test sequence (using ASMM), and when MEMORY_TARGET = 40GB in the second sequence (using AMM). Using the 40GB setting for both ASMM and AMM clearly indicates a performance sweet spot for this system configuration.

[11] HugePages can optimize system performance when using ASMM in conjunction with large SGA settings.

- At the 40GB sweet spot, when using AMM we measured 16% higher performance (maximum sustained TPS) than when using ASMM for memory management. This correlated to an increase in storage I/O performance of approximately 500 IOPS.

4 Sizing and Deployment Guidelines

This chapter we present general guidelines for sizing and deployment of EqualLogic SANs in support of Oracle OLTP database workloads.

4.1 SSD vs. SAS

Real time transactions are the primary component in OLTP database workloads. A key dependency for real time transaction performance is the maximum IOPS the storage system can sustain while not exceeding operational limits for application response time.

OLTP systems typically generate small random I/O using an 8K block size. As shown in our test results presented in Section 2.2, EqualLogic PS6010XV storage arrays configured with 15K RPM SAS disks are well suited for supporting Oracle database applications, particularly OLTP workloads that generate high volumes of small random I/O.

If you have very high I/O performance requirements then you should consider using the Dell EqualLogic PS6010S (SSD) storage array. The PS6010S array can provide much higher IOPS rates at lower I/O latencies. Figure 11 illustrates the 4x increase in IOPS performance[12] provided by the PS6010S SSD array as compared to the PS6010XV using 15K SAS disks (derived from detailed test results presented in Section 2.3).

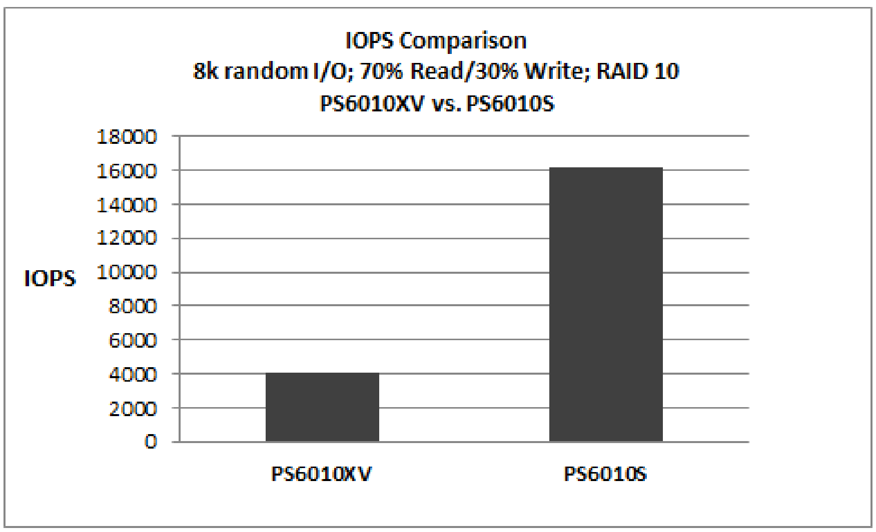

Figure 11 IOPS Comparison: PS6010XV vs. PS6010S

Typically SSD arrays are deployed in a tiered OLTP SAN to host the performance critical database I/O elements, with the remaining database elements stored on SAS based arrays.

---

[12] The data in Figure 11 shows maximum IOPS measured (~4100) before exceeding 20ms disk latency for the PS6010XV. The disk latency at maximum IOPS level for the PS6010S (~16000) was 16.7ms.

4.2 Guidelines for I/O Performance

> **Note:** The sizing guidelines in this section are based on the TPC-C test results presented in Section 3.2. **These sizing guidelines apply only to use of EqualLogic PS Series PS6010XV arrays configured with 15K RPM SAS disk drives hosting Oracle pure OLTP I/O workloads.**

The ORION test results presented in Section 2.2 show that a single PS6010XV (15K SAS) array can support up to 4000 IOPS for 8K small random I/O at a 70/30 read/write ratio (before exceeding the 20ms disk latency threshold). Using the more realistic TPC-C (Quest Benchmark Factory) test, the results presented in Section 3.2 show that the same array was able to support 348 TPS at a 6900 user load, corresponding to a peak IOPS level of 3600.

Based on these results we can conclude the following:

- A single PS6010XV array can support up to 3600 IOPS for typical Oracle OLTP I/O workloads[13].
- For workloads that generate TPC-C like transactions, we can derive the following IOPS Per User loading factor:

 IOPS Per User = 0.52 (3600/6900)

Using this IOPS Per User load factor, you can estimate your IOPS requirement based on the number of users you need to support. For actual IOPS estimates you should also include an overhead factor of 10% to account for additional I/O operations such as backup. You should also include an additional 20% overhead factor in the calculation to support any future database operational requirements. We recommend using the following formula for estimating Oracle OLTP IOPS load on the EqualLogic PS6010XV 15K SAS array:

IOPS = (# of users) * 0.52 * 1.1 * 1.2

Using the above formula, if your expected OLTP user load is 5000 then you can calculate estimated IOPS as follows:

IOPS = 5000 x 0.52 x 1.1 x 1.2 **= 3443**

Note that 5000 users yields an IOPS estimate below the peak IOPS level measured for the array (3600). Using the peak IOPS level measured for this array the recommended maximum number of users is calculated as follows:

3600 / (0.52 * 1.1 * 1.2) = **5245**

Based on this calculation, if your user load is greater than 5245 you will need to add a second array to your SAN to meet IOPS requirements.

---

[13] Includes I/O performed on data files, temp files, redo logs and archive logs as all of these database elements were stored on the same array.

4.3 Sizing Guidelines for Storage Capacity

When sizing storage capacity for an Oracle OLTP system, you need to consider the following Oracle database components:

- Database files
- Online REDO log files
- Archive log files
- Backup/Flash data

Creating an Oracle OLTP database using Quest Benchmark factory with a scaling factor of 1000 warehouses requires approximately 100GB of disk space. A database of this size typically supports up to 10000 users. General capacity sizing guidelines for Oracle OLTP databases are shown in Table 7.

Database Element	Space Requirement	Remarks
Database Files	**DB_size**	Size of the data files used for database storage in a typical OLTP environment.
Temporary Table Space	**DB_size * 0.2**	Allocate 20% of the database size for temporary table space. If your database workload involves any Sequential I/O than you may need a larger allocation.
SYSTEM/UNDO Table Space	**DB_size * 0.05**	You should allocate an additional 5% for the Oracle system table space, UNDO table space and other related files such as parameter files, control files, etc.
REDO Logs	**DB_size * 0.1**	The space requirement for the REDO logs is a function of the I/O transaction profile of the application[14]. A 10GB (10%) space allocation for REDO logs in our test configuration was sufficient.
Archive Logs	**DB_size * 1.0**	Rule of thumb is to allocate the same amount of space as used for the database files. Further tuning depends on your Recovery Time Objective (RTO) and Recovery Point Objective (RPO) for the database.
Backup	**DB_size * 2.0**	We recommended allocation of at least twice the amount of space used for the database files.

Table 7 OLTP Database Capacity Sizing Guidelines

[14] Note: If the REDO log file size setting is too small, then this may result in a performance impact due to frequent log switching. Similarly, if the REDO logs are too large, then in case of events such as a database instance crash some transactions may be lost.

5 Best Practice Recommendations

Each of the component layers in the system stack shown in Figure 12 requires careful design and configuration to ensure optimal system performance.

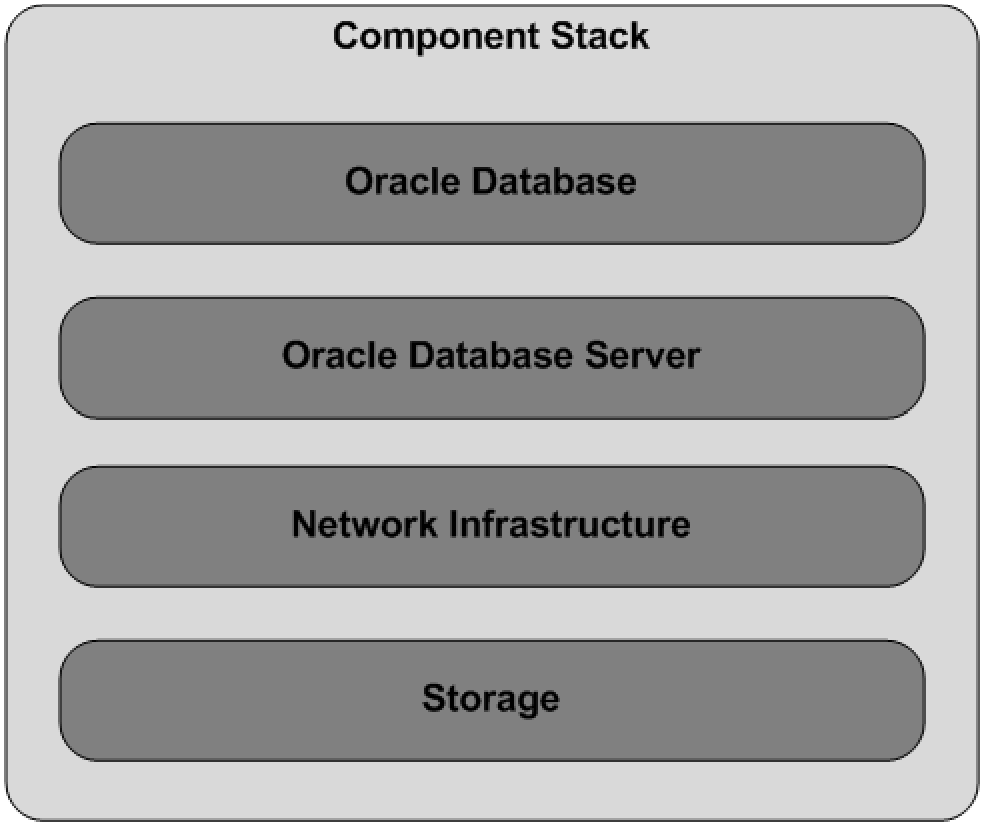

Figure 12 Component Layers in the System Stack

Within each layer of the stack you must consider each of the following design goals:

Availability Ensuring high availability is a critical design guideline for each component in the solution. Redundant sub-components and connection paths must be maintained to avoid single points of failure.

Scalability Scalability is also critical design goal. Your solution architecture should be able to scale to meet performance and capacity needs.

Performance Performance of each component in the stack must be balanced such that all performance requirements for system are met (examples: disk latency, user response times).

5.1 Storage

The following sections provide design considerations and best practice recommendations for EqualLogic storage arrays.

5.1.1 Use High Performance Drives

We present test results illustrating performance differences between SAS and SSD based arrays in Section 2. Further considerations and recommendations for choosing EqualLogic disk drives types are presented in this section.

Over the past few years drive capacities have increased at a faster rate than drive I/O performance. This means that if you focus primarily on I/O performance when selecting drives you should still have plenty of capacity available with whatever drive type you choose.

For our OLTP test configuration we focused on using the two highest performing drive types currently available with EqualLogic arrays: 15K RPM SAS drives and Solid State Disk (SSD) drives. The 15K RPM SAS drives offer excellent performance and reduced seek times as a result of higher rotational speeds[15]. These drives are ideal for supporting Oracle OLTP I/O workloads. SSD drives offer significantly faster random read/write response times compared to traditional drives. SSD drives provide the highest performing option, capable of supporting very high IOPS with exceptionally low latency. However, SSDs are more cost effective when deployed in a tiered storage architecture so that the I/O load can be balanced based on performance requirements.

We recommend you use the latest model 100GB SSD drives available for the EqualLogic PS6010S. As shown in Figure 13, we measured a significant improvement (75% increase) in performance for the 100GB SSDs as compared to the older 50GB SSDs.

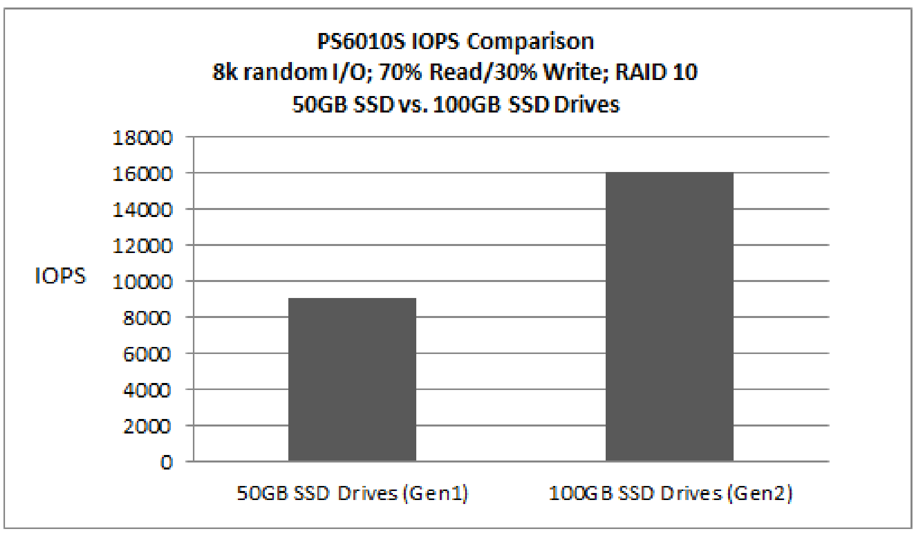

Figure 13 IOPS Comparison: 50GB vs. 100GB SSD

[15] Note: for drives with the same rotational speed, changing the capacity does not significantly change performance.

5.1.2 Use an Optimal Number of EqualLogic Volumes

Our test results indicate that increasing the number of EqualLogic volumes used by an Oracle database ASM disk group can increase database I/O performance. We measured an 18% increase in transactions per second using TPC-C test workload when using four volume containers vs. using two volume containers for the database ASM disks. But when increasing beyond five volume containers the performance difference is minimal. See Section 3.2.2 for details.

5.1.3 Use RAID 10

If performance is your main criteria, then RAID 10 is the most preferred RAID configuration and is most suitable for OLTP I/O workloads.

> **Note:** Oracle recommended RAID configurations are described in the article:
>
> - **I/O Tuning with Different RAID Configurations** *:*
>
>   https://support.oracle.com/CSP/main/article?cmd=show&type=NOT&id=30286.1

5.2 Network Infrastructure

We recommend the following network infrastructure design best practices:

- You should design redundant SAN component architectures. This includes the NICs on the servers and switches for the storage network (including server blade chassis switches and external switches).
- You should also make sure that the server NIC ports and storage NIC ports are connected in a way such that any single component failure in the SAN will not disable access to any storage array volumes.
- Flow control should be enabled on both the server NICs and the switch ports connecting to server and storage ports.
- We recommended you enable jumbo frames on the server ports and the switch ports, for both the SAN and the Oracle RAC interconnect.
- You should disable spanning tree on switch ports connecting to end devices like server ports and storage ports. Portfast should be enabled for these ports.
- You should use NIC bonding to combine at least two 1GbE NICs for RAC interconnectivity. In our test configuration we bonded three NICs on each server for the connections to the RAC interconnect switches.

Note: General recommendations for EqualLogic PS Series array network configuration and performance is provided in the following document.

- *PS Series Array Network Performance Guidelines:*

  http://www.equallogic.com/resourcecenter/assetview.aspx?id=5229

5.3 Oracle Database Server

In this section we describe considerations and best practice recommendations for the server NIC configuration and for the host operating system (Linux).

5.3.1 NIC Configuration

It is important that you properly configure your servers to provide the optimal number of NICs for your Exchange I/O workload. You should consider the following best practices:

- Configure your servers to provide dedicated NICs for client/server LAN I/O, the Oracle RAC interconnect, and separate dedicated NICs/HBAs for iSCSI SAN I/O.
- To provide for redundancy, you should provision at least two NICs per server for each path.

5.3.2 Host Operating System: MPIO and the multipath.conf File

In Linux, multipath I/O relies on a kernel level component call the device-mapper[16]. The device-mapper provides a framework for mapping block devices to one another. For MPIO, it is used for mapping multiple I/O paths into a single virtual I/O path. We recommend using the device-mapper for implementing MPIO for shared EqualLogic iSCSI storage devices on Linux database servers.

Changes were required in **/etc/multipath.conf**[17] to properly configure blacklist settings, persistent device mappings and tuning parameters specific to EqualLogic storage devices.

Note: The changes to /etc/multipath.conf described in this section were made in accordance with guidelines provided by this Dell EqualLogic whitepaper:

- ***Red Hat Linux v5.x Software iSCSI Initiator Configuration, MPIO and Tuning Guide***

  http://www.equallogic.com/resourcecenter/assetview.aspx?id=8727

[16] For details see: http://docs.redhat.com/docs/en-US/Red_Hat_Enterprise_Linux/4/html/DM_Multipath/MPIO_Overview.html

[17] /etc/multipath.conf is the configuration file for the multipath device mapper

Configure the device blacklist

In the multipath configuration file all devices are blacklisted by default. The system internal disks in our configuration were labeled "/sda1" and "/sda2".We modified the ""blacklist" section in "/etc/multipath.conf" as shown in Example 1 below to allow MPIO for all devices except for the system internal disks.

Example 1 Blacklist section modifications in multipath.conf

```
blacklist {
#        devnode "*"
        devnode "^sd[a]$"
}
```

Configure Persistent Device Mapping

Devices using the Red Hat software initiators do not have a persistent naming scheme, and do not guarantee that a device will always have the same device node. Using a persistent naming scheme you can provide reference points for devices that will not change across system reboots.

In our test system we created persistent device mapping entries for all volumes within the "multipaths" section of /etc/multipath.conf. Example 2 shows the multipath sub-sections for the disk devices used by the Oracle ASM disk group "ORADB".

The parameter "rr_min_io" specifies the number of I/O requests to route to a path before switching to the next path in the current path group. As the block size of I/O is small for typical OLTP databases, a smaller value such as 10 to 20 is recommended. A larger value may be better when there are more sequential I/O transactions.

Example 2 Blacklist section modifications in multipath.conf

```
multipath {
wwid         36090a0684076ab62b1c234758d01b0be
alias        oradb1
rr_min_io    20
}
multipath {
wwid         36090a0684076eb6fb1c264758d0180b2
alias        oradb2
rr_min_io    20
}
multipath {
wwid         36090a0684076fb6a8dc324768d0180fa
alias        oradb3
rr_min_io    20
}
multipath {
wwid         36090a06840769b6e8dc354768d0170e1
alias        oradb4
rr_min_io    20
}
```

Configure Recommended Values for EqualLogic devices

We used the "device" section of multipath.conf to set some attribute values that are specific to our EqualLogic PS Series device for OLTP I/O workloads. Modifications are shown in Example 3.

Example 3 Blacklist section modifications in multipath.conf

```
device {
       vendor                     "EQLOGIC"
       product                    "PS6010"
       path_grouping_policy       multibus
       getuid_callout             "/sbin/scsi_id -g -u -s /block/%n"
       features                   "1 queue_if_no_path"
       path_checker               readsector0
       failback                   immediate
       path_selector              "round-robin 0"
       rr_min_io                  20
       rr_weight                  priorities
       }
```

5.3.3 Host Operating System: iSCSI Daemon Configuration

The default configuration file for the iSCSI daemon is **/etc/iscsi/iscsid.conf**.

For an OLTP type workload the parameters "node.session.cmds_max" and "node.session.queue_depth" should be left at their default settings in iscsid.conf, as follows:

node.session.cmds_max = 128
node.session.queue_depth = 32

You may need to increase the settings for these parameters if the value of rr_min_io is set at 200 or greater. Since we set the rr_min_io value to 20 for OLTP I/O in our test system configuration, there was no need to modify these settings.

5.4 Oracle Database

5.4.1 Oracle database initialization parameters

Some of the database initialization parameters which are recommended for optimal performance are listed in Table 8.

Parameter	Syntax	Remarks
Dispatchers	**dispatchers=(protocol=TCP)(LISTENER=SIEORADB)(dispatchers =4)**	This parameter helps in managing multiple incoming session requests to shared server processes. Setting dispatchers=4 helped to achieve better performance.
Disk Async I/O	**DISK_ASYNCH_IO=true**	This parameter should be enabled to ensure that database I/O is asynchronous.
Sessions	**sessions=9024**	This parameter specifies the total number of user and system sessions. This number should be set slightly higher than the processes parameter.
Processes	**processes=6000**	This parameter specifies the number of operating system related user and system processes which connect to the database instance concurrently.
Shared servers	**Shared_servers=50**	To enable a larger number of concurrent transactions, we set Shared_servers= 50. With this feature, a small number of processes can handle a large number of requests.

Table 8 OLTP Database Initialization Parameters

5.4.2 Oracle Memory Management

Memory management is the most critical and complex part of tuning Oracle databases for performance. If you are using Oracle 11g or later Oracle recommends[18] using AMM instead of ASMM for dynamic memory management. In our testing we confirmed using AMM provided better performance, once we were able to determine the correct initial setting for MEMORY_TARGET. See Section 3.2.3.

5.4.3 Oracle ASM Disk groups

Automatic Storage Management (ASM) is a feature introduced in Oracle 10g which simplifies the management of underlying storage and also helps in improving the performance by load balancing. You should use ASM disk groups for managing the entire database. ASM in Oracle 11g R2 also provides the ability to manage Oracle clusterware files. We used Oracle ASM for the entire database, including clusterware.

Some best practice recommendations for ASM configuration are :

- ASM External redundancy is preferred while configuring ASM disk groups. There is no need to use other types of redundancy.

[18] See: http://download.oracle.com/docs/cd/B28359_01/server.111/b28301/instance004.htm

- Preferably use same sized ASM disks within ASM disk groups. This helps Oracle ASM to optimally perform load balancing.

Note: See the following document for more information:

> ➢ ***Performance Implications for Running Oracle ASM with Dell EqualLogic PS Series Storage***
>
> http://www.equallogic.com/resourcecenter/assetview.aspx?id=8743

5.4.4 Oracle ASM and Multipathing

By default, ASM will generate an error if multiple paths for the same disks are discovered. You must configure ASM to use the kernel device mapper. This will ensure that only multipath disks are discovered by ASMLib. This is accomplished by setting the value of the ORACLEASM_SCANORDER parameter to "dm" in the /etc/sysconfig/oracleasm file, as shown in Example 4. For more information see Section 5.3.2.

Example 4 Setting ORACLEASM_SCANORDER

```
# cat /etc/sysconfig/oracleasm | grep ORACLEASM_SCANORDER
# ORACLEASM_SCANORDER: Matching patterns to order disk scanning
#ORACLEASM_SCANORDER=""
ORACLEASM_SCANORDER="dm"
```

5.4.5 Oracle AMM - /dev/shm setting

The Oracle Automatic Memory Management (AMM) feature requires the /dev/shm file system size to be at least equal to the value of the "MEMORY_TARGET" parameter. By default, the size of this file system will be set to 50% of the total physical memory on the system, as shown in Example 5.

Example 5 Default /dev/shm Size

```
# df -k /dev/shm
Filesystem           1K-blocks      Used Available Use% Mounted on
tmpfs                 32990684    228636  32762048   1% /dev/shm
```

It may be necessary to change the size /dev/shm file system to support custom MEMORY_TARGET settings. You can use the **mount** command as shown in Example 6 to accomplish this.

Example 6 Changing Size of /dev/shm

```
# mount -o remount,size=50G /dev/shm
# df -k
Filesystem           1K-blocks      Used Available Use% Mounted on
/dev/sda1             69436796  22231580  43621120  34% /
tmpfs                 52428800    228636  52200164   1% /dev/shm
```

Appendix A Test System Components

Table 9 and Table 10 provide details for the major hardware and software system components used in the test system configuration.

Test Configuration - Hardware Components	
ORION Simulation and Oracle Database Servers	• 2 x Dell PowerEdge R710 Servers: o System BIOS version 2.0.13 o 2 x Quad Core Intel® Xeon® X5570 Processors o 64 GB RAM, 8M Cache, 2.93 GHz o 2 x 146GB 10K SAS internal disk drives o Broadcom 5709 1GbE quad-port NIC (LAN on motherboard) o 2 x Broadcom NetXtreme II 57711 10GbE NIC, Dual-Port
Virtualization Servers	• 2 x Dell PowerEdge R610 Servers: o System BIOS version 2.0.11 o Quad Core Intel® Xeon® X5570 Processor o 64 GB RAM, 8M Cache, 2.93 GHz o 2 x 146GB 10K SAS internal disk drives o Broadcom 5709 1GbE quad-port NIC (LAN on motherboard)
Network	• 2 x Dell PowerConnect 6248 1Gb Ethernet Switch, running firmware version 3.2.0.7 • 2 x Dell PowerConnect 8024F 10Gb Ethernet Switch, running firmware version 3.1.1.9
Storage	• 1 x Dell EqualLogic PS6010XV: o 16 x 300GB 15K SAS disks with dual 2 port 10GbE controllers o 14 SAS disks configured as RAID 10, two hot spares o Controller firmware version 5.0.0.0 (R122845) • 1 x Dell EqualLogic PS6010S: o 16 x 100GB SSD disks with dual 2 port 10GbE controllers o 14 SSDs configured as RAID 10, with two hot spares o Controller firmware version 5.0.0.0 (R122845)

Table 9 Test Configuration Hardware Components

Test Configuration - Software Components	
Database Servers	• Red Hat Enterprise Linux 5.5 o iSCSI software initiator: 6.2.0.871-0.16.el5 o MPIO Enabled: v0.4.7 • Oracle Database 11g R2 Enterprise Edition o Two-Node Oracle RAC database. o ASM for Clusterware and Database.
Virtualization Server:	• VMware vSphere ESX version 4.1 o Bare metal hypervisor directly installed on R610 servers o Four guest VMs on each ESX server o Managed by vCenter.
Virtual Machines:	• 8 x Windows Server 2008 R2 Enterprise Edition • Workload generators (running within VMs): o Quest Benchmark Factory 6.1.1 o 3 agents from each VM o Oracle 11g R2 client installed
Monitoring Tools:	• EqualLogic SAN Headquarters version 2.0 • Oracle OS Watcher utility (installed on database servers) • Oracle 11g R2 Automatic Workload Repository (AWR)

Table 10 Test Configuration - Software Components

Related Publications

The following Dell publications are referenced in this document or are recommended sources for additional information.

- ***Dell EqualLogic PS Series Network Performance Guidelines***

  http://www.equallogic.com/resourcecenter/assetview.aspx?id=5229

- ***Dell PowerEdge Systems Oracle Database on Enterprise Linux x86_64 Database Setup and Installation Guide Version 1.5***

  http://support.dell.com/support/edocs/software/appora11/lin_x86_64/1_4L/multlang/DBIG/DBIGR2.pdf

- ***Performance Implications for Running Oracle ASM with Dell EqualLogic PS Series Storage***

  http://www.equallogic.com/resourcecenter/assetview.aspx?id=8743

- ***Oracle 11g Database Concepts***

  http://download.oracle.com/docs/cd/B28359_01/server.111/b28318/toc.htm

- ***Oracle Database Storage Administrators Guide***

  http://download.oracle.com/docs/cd/B28359_01/server.111/b31107.pdf

- ***Boosting SAN Performance with Dell EqualLogic Solid-state drive arrays***

  http://www.dell.com/Downloads/Global/Power/ps2q09-20090236-Locsin.pdf

- ***Red Hat Linux Software iSCSI initiator, MPIO Configuration and Tuning Guide***

  http://www.equallogic.com/resourcecenter/assetview.aspx?id=5229

- ***Dell EqualLogic PS Series Network Performance Guidelines***

  http://www.equallogic.com/resourcecenter/assetview.aspx?id=8727