IBM SAN Volume Controller Performance Configuration Guidelines for Implementing Oracle Databases with Automatic Storage Management (ASM)

Bob Gonzalez
IBM Systems and Technology Group
Open Systems Lab, San Jose, California
June 2009
# Table of contents

1 Abstract ................................................................................................................................. 1

2 Introduction ............................................................................................................................ 1
   2.1 Outline .............................................................................................................................. 1
   2.2 Assumptions .................................................................................................................... 2
   2.3 Intended Audience .......................................................................................................... 2

3 Storage Area Network Technology Stack .............................................................................. 3
   3.1 IBM System Storage SAN Volume Controller ................................................................. 3
      3.1.1 SVC Description .......................................................................................................... 3
      3.1.2 SVC Terminology and Pictorial Overview ................................................................. 4
   3.2 IBM System Storage DS8000 ............................................................................................. 7
      3.2.1 DS8000 Hardware Terminology ................................................................................ 8
      3.2.2 DS8000 Virtualization Terminology ......................................................................... 9
   3.3 Fibre Channel switch overview ........................................................................................ 9
      3.3.1 Hardware .................................................................................................................. 10
      3.3.2 Software – Fabric Operating System ........................................................................ 10
   3.4 Host and HBA (Host Bus Adapter) hardware and software overview ............................. 10

4 Oracle Automatic Storage Management (ASM) .................................................................. 11
   4.1 ASM Disk Groups ............................................................................................................ 11
   4.2 ASM Disks ...................................................................................................................... 11
   4.3 ASM Striping ................................................................................................................... 12
   4.4 ASM Instances ............................................................................................................... 12

5 I/O Workloads ...................................................................................................................... 14
   5.1 General I/O Workload Types and Associated Metrics ..................................................... 14
      5.1.1 Small Block, random I/O workloads ......................................................................... 14
      5.1.2 Large Block, sequential I/O workloads ..................................................................... 14
   5.2 Oracle I/O Workloads ..................................................................................................... 14
      5.2.1 Oracle random I/O workloads .................................................................................. 14
      5.2.2 Oracle sequential I/O workloads .............................................................................. 15
      5.2.3 Determining the current Oracle I/O profile .............................................................. 15
   5.3 SVC Performance Characteristics .................................................................................. 15

6 SVC Configuration Best Practices ....................................................................................... 16
   6.1 Summary of SAN Volume Controller Configuration Best Practices ............................... 16
      6.1.1 SAN Topology for SVC Deployment ......................................................................... 16
      6.1.2 Zoning of SVC, Storage Controller, and Host ......................................................... 17
      6.1.3 Storage Controller – DS8000 .................................................................................... 18
      6.1.4 MDisk Configuration .................................................................................................. 19
      6.1.5 MDisk Group (MDG) Configuration .......................................................................... 20
      6.1.6 VDisk Configuration .................................................................................................. 22
      6.1.7 Host Configuration ..................................................................................................... 24
   6.2 SAN Volume Controller Logical Configuration ............................................................... 25
      6.2.1 Create the Fibre Channel zoning between SVC and DS8000 .................................... 25
      6.2.2 Create the DS8000 LUNs and make them visible to the SVC as MDisks .................. 26
      6.2.3 Configure the SVC Storage Controller, MDisks, and MDisk groups ...................... 27
      6.2.4 Create the Fibre Channel zoning between SVC and front-end Hosts ....................... 28
1 Abstract

This white paper offers guidelines for optimizing performance when configuring IBM® System Storage™ SAN Volume Controller (SVC) for running an Oracle database with Automatic Storage Management (ASM). It documents the SVC performance best practices which are detailed in various IBM Redbooks®. These configuration best practices are then applied and confirmed via the usage of Oracle's ORION load generation tool.

2 Introduction

This paper discusses some performance configuration guidelines that can be used when deploying IBM® System Storage™ SAN Volume Controller (SVC) for use in an Oracle Automatic Storage Management environment. It attempts to answer some fundamental configuration questions that System Administrators, Storage Administrators, or Oracle DBAs will have when deploying an SVC for Oracle with ASM. The back-end storage used with SVC will be an IBM® System Storage™ DS8000®. Therefore the white paper “IBM DS8000 Performance Configuration Guidelines for Implementing Oracle Databases with Automatic Storage Management” (item 5 in the References) should be considered a companion piece to this white paper and will be referenced throughout.

The load generator used to arrive at the configuration guidelines is ORION - the Oracle I/O Numbers Calibration tool. This tool generates I/O using the same I/O software stack used by the Oracle server software without having to install the server software and create a database. It can simulate various workload types at different load levels to arrive at performance metrics for I/Os per second (IOPS), Latency (Response Time) and Megabytes per second (throughput). It can also simulate the effect of striping performed by ASM.

This paper is not meant as a statement of maximum possible I/O benchmark numbers for SVC. Those numbers are stated in the Storage Performance Council SPC-1 and SPC-2 results posted by IBM on the SPC Web site. Rather, this paper runs controlled experiments on a given set of hardware where only one variable is changed at a time; for example, change the RAID arrays from RAID-5 to RAID-10 while the rest of the SAN configuration stays constant. Thus conclusions can be drawn regarding the effect that changing one configuration item has on performance.

2.1 Outline

A general outline of this paper is as follows:

- Overview of all of the layers of the technology stack that are part of a Storage Area Network which includes the SAN Volume Controller (SVC). The back-end storage used with SVC in this paper is the IBM System Storage DS8000.
- Discussion of the Oracle Automatic Storage Management (ASM) volume manager.
- Discussion of general and Oracle-specific I/O workloads.
- Coverage of the performance characteristics of the SAN Volume Controller.
- Consolidation and summarization of the extensive information that is available regarding best practices for SVC configuration.
Detailed steps for logical configuration of an SVC.

Coverage of the usage of monitoring and modeling tools, IBM TotalStorage® Productivity Center® and IntelliMagic BV’s Disk Magic.

Description of ORION, the Oracle I/O Numbers Calibration tool, which is used to generate I/O load and gather storage performance metrics.

Description of the testing methodology.

Various benchmarking exercises using ORION. These exercises will be detailed further down in the paper. But briefly, the exercises will do the following:

- Compare Oracle I/O performance when using SVC with a back-end DS8000 versus using the same DS8000 without SVC. The SVC test runs will be done using one and two iogrps.
- Determine the effects on SVC performance depending on whether RAID-5 or RAID-10 is being used on the back-end storage controller. Verify whether or not the recommendations which were made in the DS8000 performance configuration guidelines paper (item 5 in the References section) with regard to RAID-5 and RAID-10 still pertain when an SVC is being used.
- Compare the effects of striping at the ASM level only, at the SVC level only, and at both the ASM and SVC level together (striping-on-striping).
- Determine the effect that changing the ASM Allocation Unit (AU) size has on performance.

2.2 Assumptions

This paper will start with the assumption that the best practices that have been documented by the SAN Volume Controller performance architects are relevant for a wide range of applications, including Oracle databases using ASM. The performance testing done for the SVC assumes that one of the most common usages for storage is to run databases such as Oracle and DB2. Therefore benchmark testing always includes the types of general, and indeed Oracle-like, workloads that will be discussed in the paper. However, this will not skew the results of the testing done in this paper. The performance numbers which result from the test matrix will definitely have to speak for themselves.

Another assumption is that the load generation tool which will be used, Oracle’s ORION, is an appropriate tool for determining storage performance for Oracle I/O workloads. A detailed description of ORION will be given in a later section. But briefly, ORION generates I/O using the same I/O software stack used by the Oracle server software without having to install the server software. By using ORION, it is possible to avoid addressing the subject of Oracle database server tuning, a subject which is beyond the scope of this paper. We can then conduct a more controlled experiment where it is assumed that no matter how it would be done by an Oracle database, we are going to be generating Oracle I/O at various load levels and with varying workload mixes.

2.3 Intended Audience

The intended audience of this paper is any Technical Lead, System Administrator, Storage Administrator, or Oracle DBA in a production environment that is part of an effort to deploy the IBM System Storage SAN Volume Controller for running Oracle with ASM. After reading this paper, the technical staff will have at least some starting point, based on documented and tested best practices, with which to make configuration and deployment decisions for the SVC.
3 Storage Area Network Technology Stack

It is important to understand all of the layers of the storage technology stack in a SAN environment before any configuration decisions can be made. This section will include overviews for the concepts related to SVC storage attributes. The back-end storage used in this overview will be the IBM® System Storage™ DS8000.

3.1 IBM System Storage SAN Volume Controller

The IBM System Storage SAN Volume Controller is a block-storage virtualization appliance. SVC brings storage devices together in a virtual pool to make all available storage appear as:

- One logical device to centrally manage and to allocate capacity as needed.
- One solution to help achieve the most effective use of key storage resources on demand.

Therefore SVC allows you to consolidate the heterogeneous SAN storage which may be present in your data center, including many non-IBM storage products, and present it as one pool of storage.

SVC uses an in-band architecture which means that data flowing between a host and a storage controller flows through an SVC node. On the front end, SVC presents an interface to a host which looks like a storage controller, i.e., like a storage device. On the back end, SVC provides an interface to storage controllers that looks like a host to those storage controllers.

Virtualization solutions can be implemented in the storage network, in the server, or in the storage device itself. SVC, the IBM storage virtualization solution, is SAN-based, which helps allow for a more open virtualization implementation. Locating virtualization in the SAN, and therefore in the path of input/output (I/O) activity, helps provide a solid basis for policy-based management. The focus of IBM on open standards means its virtualization solution supports freedom of choice in storage-device vendor selection.

The IBM System Storage SAN Volume Controller is designed to:

- Simplify storage management.
- Reduce IT data storage complexity and costs while enhancing scalability.
- Extend on demand flexibility and resiliency to the IT infrastructure.
- Increase application availability by making changes in the infrastructure without having to shut down hosts.

3.1.1 SVC Description

The following items highlight some of the characteristics of the SVC:

- SVC implements an indirection, or “virtualization”, layer in a Fibre Channel fabric. The virtualization layer mediates between RAID controllers on back-end physical SAN storage (such as a DS8000) and virtual disks in front-end hosts. SVC pools the physical storage from one or more of those back-end storage devices and provides parts of that storage pool to hosts in a way that makes the hosts think they are just using disks in the normal way.
SVC maps the virtual disk seen by the hosts to the physical storage in the pool using a virtualization map. The hosts are unaware of the details of this map.

SVC implements various Copy Services, which are used to duplicate server data by providing instant local copies (FlashCopy®) or copies at a remote site (Metro Mirror and Global Mirror).

SVC can mirror the server data onto multiple local physical disks to improve local availability (VDisk mirroring).

SVC provides the user with the ability to over-allocate the virtualized physical storage to provide sparsely populated virtual disks to the application servers and save physical storage cost (space-efficient VDisks).

SVC is deployed as a cluster of nodes. Nodes are always deployed in pairs called I/O groups. There can be up to four I/O groups (eight nodes) in an SVC cluster. In case of I/O path failure, non-disruptive failover is performed inside the I/O group only, via Subsystem Device Driver Path Control Module (SDDPCM) software on AIX® or the native multipathing software provided by supported operating systems (e.g., Device-Mapper Multipath on Linux®).

Each SVC node is a 1U high rack-mounted appliance based on an IBM System x® server. SVC itself is IBM machine type 2145. A node has at least four Fibre Channel ports and is protected by a UPS (uninterruptible power supply). A node runs a Linux kernel and a specialized Virtualization Storage Software environment that provides proprietary clustering capability. SVC is based on the COMmodity Parts Storage System (Compass) architecture, developed at the IBM Almaden Research Center.

### 3.1.2 SVC Terminology and Pictorial Overview

Table 3.1 below contains some SVC definitions:

<table>
<thead>
<tr>
<th>Node</th>
<th>A single 1U SVC machine (Models: 4F2, 8F2, 8F4, 8G4) which provides virtualization, cache, and copy services for the SAN.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1U</td>
<td>One rack unit, or 1U (less commonly 1RU). A rack unit is the EIA-310 standard measure for the height of a rackmount system and is equal to 1.75 inches (44.45 mm). The standard width is 19 inches. The depth can vary but is typically in the range of 20 to 30 inches.</td>
</tr>
<tr>
<td>I/O group</td>
<td>SVC nodes are always deployed in pairs called I/O groups. The nodes in an I/O group duplicate each other’s write commands; i.e., the write cache is mirrored between these two nodes.</td>
</tr>
<tr>
<td>Cluster</td>
<td>A set of one to four I/O groups managed as a single entity. The SVC cluster is the management entity, the domain name space, and the single point of provisioning.</td>
</tr>
<tr>
<td>Configuration Node</td>
<td>A single node that holds the cluster’s configuration and has been assigned the cluster IP address.</td>
</tr>
<tr>
<td><strong>Managed Disk (MDisk)</strong></td>
<td>A unit of storage from a back-end storage RAID array which has been virtualized by the SVC. Basically, an MDisk is a LUN from the back-end storage which has been made visible to the SVC.</td>
</tr>
<tr>
<td>--------------------------</td>
<td>--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td><strong>Managed Disk Group (MDisk group or MDG)</strong></td>
<td>A pool of MDisks. The MDisks that comprise an MDisk group should have the same back-end disk drive characteristics, i.e., the same drive speed, drive type and RAID type.</td>
</tr>
<tr>
<td><strong>Virtual Disk (Vdisk)</strong></td>
<td>This is the virtual LUN that is presented to the hosts on the front end as usable storage. A VDisk is created from one or more MDisks in an MDisk group.</td>
</tr>
<tr>
<td><strong>Extent</strong></td>
<td>An atomic unit of storage in SVC. An MDisk is divided into extents; a VDisk is formed from sets of extents from one or more MDisks in an MDisk group. Extent sizes can vary from 16 MB to 2 GB in powers of 2.</td>
</tr>
<tr>
<td><strong>Controller --or-- Disk Controller --or-- RAID Controller --or-- Storage Controller</strong></td>
<td>These are all synonyms for the back-end storage device that has been made visible, via zoning and logical configuration, to an SVC cluster.</td>
</tr>
</tbody>
</table>

Table 3.1 - SVC terminology

Figure 3.1 below is a high-level view of the SVC architecture.
SVC Architecture

- Inband virtualization appliance
- Scale-out performance by Adding nodes
- Scale-up performance by Increasing individual node Processing power
- Nodes are paired into I/O Groups for high availability
- Each Virtual Disk is visible through one I/O Group

Figure 3.1 – SAN Volume Controller Architecture – High Level View

Figure 3.2 below is a detailed view of the SVC architecture.
3.2 IBM System Storage DS8000

The IBM System Storage DS8000 series is a high-performance, reliable, and exceptionally scalable enterprise disk storage system. The IBM System Storage DS8000 series is designed to:

- Deliver robust, flexible, and cost-effective disk storage for the mission-critical workloads of medium and large enterprises
- Enable the creation of multiple Storage System Logical Partitions (LPARs) within a single DS8000 Model 9B2, that can be used for completely separate production, test, or other unique storage environments
- Support high availability, storage sharing, and consolidation for a wide variety of operating systems and mixed server environments
- Help increase storage administration productivity with centralized and simplified management

An overview of the IBM System Storage DS8000 hardware, virtualization hierarchy and performance configuration guidelines (with respect to Oracle and ASM) is contained in the white paper “IBM DS8000 Performance Configuration Guidelines for Implementing Oracle Databases with
Automatic Storage Management™ (item 5 in the References). While that paper offers detailed definitions and descriptions of DS8000 features, the two sections below will include a brief glossary.

3.2.1 DS8000 Hardware Terminology

Table 3.1 contains definitions for some of the hardware components of the DS8000.

<table>
<thead>
<tr>
<th>Processor Complexes and Servers</th>
<th>Each DS8000 consists of two N-way symmetric multiprocessor (SMP) System p® POWER5+ processors. They manage all read and write requests to the logical volumes on the disk arrays.</th>
</tr>
</thead>
<tbody>
<tr>
<td>I/O Enclosures</td>
<td>I/O enclosures hold Device Adapters (DAs) and Host Adapters (HAs). The I/O enclosure provides connectivity between these adapters and the Processor Complexes (Servers). Each I/O enclosure has 6 slots. Each slot supports PCI-X adapters running at 64-bit, 133 Mhz. Slots 3 and 6 are used for the DAs. The remaining slots are available to install up to four HAs per I/O enclosure.</td>
</tr>
<tr>
<td>Device Adapters (DAs)</td>
<td>Each DS8000 device adapter (DA) card offers four 2 Gbps FC-AL ports. These ports are used to connect the processor complexes to the disk enclosures. The adapter is responsible for managing, monitoring, and rebuilding the RAID arrays. The DAs are installed in pairs because each storage partition requires its own adapter to connect to each disk enclosure for redundancy.</td>
</tr>
<tr>
<td>Host Adapters (HAs)</td>
<td>A DS8000 Host Adapter is a Fibre Channel card which offers four Fibre Channel ports (port speed of 2 or 4 Gbps depending on the Host Adapter). These ports are connected to a switch and then zoned to either a front-end host or the SVC.</td>
</tr>
<tr>
<td>DDM – Disk Drive Module</td>
<td>Another term for “Disk Drive”.</td>
</tr>
<tr>
<td>Disk Enclosure</td>
<td>Each DS8000 frame contains either 8 or 16 disk enclosures depending on whether it is a base or expansion frame. Half of the disk enclosures are accessed from the front of the frame, and half from the rear. Each DS8000 disk enclosure contains a total of 16 DDMs or dummy carriers. A dummy carrier looks very similar to a DDM in appearance but contains no electronics.</td>
</tr>
</tbody>
</table>

Table 3.2 - DS8000 hardware terminology
### 3.2.2 DS8000 Virtualization Terminology

Table 3.3 contains definitions for the components of the DS8000 virtualization hierarchy.

<table>
<thead>
<tr>
<th>Array sites</th>
<th>An array site is a group of eight DDMs. The DDMs selected for an array site are chosen from two disk enclosures on different loops. The DDMs in the array site are of the same DDM type, which means the same capacity and the same speed (rpm).</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arrays</td>
<td>An array is created from one array site. Forming an array means defining it as a specific RAID type. The supported RAID types are RAID-5, RAID-10, and RAID-6.</td>
</tr>
<tr>
<td>Ranks</td>
<td>In the DS8000 virtualization hierarchy, there is another logical construct: a rank. You have to add an array to a rank. In the DS8000 implementation, a rank is built using just one array. The available space on each rank will be divided into extents. The extents are the building blocks of the logical volumes.</td>
</tr>
<tr>
<td>Extent Pools</td>
<td>An Extent Pool is a logical construct to aggregate the extents from a set of ranks to form a domain for extent allocation to a logical volume. Typically the set of ranks in the Extent Pool should have the same RAID type and the same disk RPM characteristics so that the extents in the Extent Pool have homogeneous characteristics.</td>
</tr>
<tr>
<td>Logical Volumes (LUNs)</td>
<td>A logical volume is composed of a set of extents from one extent pool. On a DS8000, up to 65,280 (65,536 - 256) volumes can be created. The DS8000 can have Fixed Block volumes, CKD volumes, or System i LUNs. Only Fixed Block volumes are used in an Open Systems environment.</td>
</tr>
<tr>
<td>Storage Pool Striping</td>
<td>This allows LUN striping over multi-rank extent pools. It is a feature that was introduced in DS8000 LMC 5.30xx.xx in October 2007. Previous LMC versions allowed multi-rank extent pools, but the ranks were effectively concatenated, not striped.</td>
</tr>
</tbody>
</table>

Table 3.3 - DS8000 virtualization terminology

### 3.3 Fibre Channel switch overview

The Fibre Channel switches used in the lab environment are the IBM System Storage SAN32B-3. This is IBM machine type 2005, model number B5K. The Brocade name is Brocade 5000. The following are brief descriptions of both the hardware and software in the IBM Brocade 2005-B5K.
3.3.1 Hardware


With a flexible architecture, the SAN32B-3 support native E_Port connectivity into Brocade and McDATA products. The SAN32B-3 can operate in FOS native, FOS open, M-EOS native and M-EOS open modes without routing and special partitioning. This enables non-disruptive expansion of any Brocade or McDATA fabric based on these operation modes, and provides the ability to implement non-disruptive software upgrades. A special NI code has to be downloaded to set the SAN32B-3 in any McDATA mode.

3.3.2 Software – Fabric Operating System

The Fabric Operating System (FOS) manages the operation of the switch and delivers the same, and compatible, functionality to all the different models of switches and directors. The switch firmware is designed to make the switches easy to install and use while retaining the flexibility required to accommodate user requirements.

The FOS includes all the basic switch and fabric support software as well as optionally licensed software that is enabled using license keys. It is composed of two major software components: firmware that initializes and manages the switch hardware, and diagnostics. Fabric OS (FOS) Version 5.x is a Linux-based operating system.

3.4 Host and HBA (Host Bus Adapter) hardware and software overview

The discussion in this paper pertains strictly to open systems (Linux/UNIX) hosts. The best practices for configuring the zoning and pathing from the host to storage will be discussed in the section 6.1.7 Host Configuration. But briefly, here are some of the considerations with regard to the tuning, troubleshooting and performance of a host attached to the any SAN storage device, including the SVC:

- Verify the optimal version and configuration of the Host Bus Adapter and disk devices. For example, on AIX the queue depth and max transfer size of either the HBA or hdisks should be set properly.

- Understand the optimal load-balancing and failover configuration of the HBA multipathing software. The multipathing software is Device-Mapper Multipath, or DM, for Linux. For AIX, it is the Subsystem Device Driver Path Control Module, or SDDPCM.

- Be familiar with the host-side I/O performance monitoring tools. The iostat command is the traditional open systems disk performance monitoring tool. In addition, AIX offers the nmon, filemon, and topas commands for disk I/O monitoring.
4 Oracle Automatic Storage Management (ASM)

Oracle Automatic Storage Management (ASM) is a volume manager and a filesystem for Oracle database files that supports single-instance Oracle Database and Oracle Real Application Clusters (Oracle RAC) configurations. Oracle ASM is the Oracle recommended storage-management solution that provides an alternative to conventional volume managers, file systems, and raw devices. ASM is designed to provide the performance of raw I/O with the management capabilities of a filesystem.

4.1 ASM Disk Groups

Oracle ASM uses disk groups to store data files. An Oracle ASM disk group is a collection of disks that Oracle ASM manages as a unit. Within a disk group, Oracle ASM exposes a file-system interface for Oracle database files. The content of files that are stored in a disk group are evenly distributed, or striped, to prevent hot spots and to provide uniform performance across the disks. The performance is comparable to the performance of raw devices.

You can add or remove disks from a disk group while a database continues to access files from the disk group. When you add or remove disks from a disk group, ASM automatically redistributes the file contents and eliminates the need for downtime when redistributing the content.

When you create a disk group, you specify an ASM disk group type based on one of the following three redundancy levels:

- **Normal** for 2-way mirroring
- **High** for 3-way mirroring
- **External** to not use ASM mirroring, such as when you configure hardware RAID for redundancy

The disk group type determines the mirroring levels with which Oracle creates files in a disk group. The redundancy level controls how many disk failures are tolerated without dismounting the disk group or losing data. This paper assumes that **External Redundancy** is being used.

4.2 ASM Disks

ASM disks are the storage devices that are provisioned to ASM disk groups. Examples of ASM disks include:

- A disk or partition from a storage array
- An entire disk or the partitions of a disk
- Logical volumes
- Network-attached files (NFS)

When you add a disk to a disk group, you either assign a disk name or the disk is given an ASM disk name automatically. This name is different from the name used by the operating system. In a cluster, a disk may be assigned different operating system device names on different nodes, but the disk has
the same ASM disk name on all of the nodes. In a cluster, an ASM disk must be accessible from all of the instances that share the disk group.

If the disks are the same size, then ASM spreads the files evenly across all of the disks in the disk group. This allocation pattern maintains every disk at the same capacity level and ensures that all of the disks in a disk group have the same I/O load. Because ASM load balances among all of the disks in a disk group, different ASM disks should not share the same physical drive. The maximum allowable size for an ASM disk is 2 TB.

Every ASM disk is divided into allocation units (AU). An AU is the fundamental unit of allocation within a disk group. A file extent consists of one or more AU. An ASM file consists of one or more file extents.

When you create a disk group in Oracle Database 11g, you can set the ASM AU size to be between 1 MB and 64 MB in powers of two, such as, 1, 2, 4, 8, 16, 32, or 64. Oracle Database 10g AUs are 1 MB, although this can be changed by modifying some Oracle hidden initialization parameters.

### 4.3 ASM Striping

ASM striping has two primary purposes:

- To balance loads across all of the disks in a disk group
- To reduce I/O latency

Coarse-grained striping provides load balancing for disk groups while fine-grained striping reduces latency for certain file types by spreading the load more widely.

To stripe data, ASM separates files into stripes and spreads data evenly across all of the disks in a disk group. The stripes are equal in size to the effective AU. The coarse-grained stripe size is always equal to the AU size. The fine-grained stripe size always equals 128 KB; this provides lower I/O latency for small I/O operations such as redo log writes.

The ASM stripe size is automatically set when a particular type of Oracle file is created since the stripe size is defined in the Oracle ASM file templates. ASM file templates exist for datafiles, online redo logs, archive log files, control files, tempfiles, and parameter files. The Oracle documentation mentioned in the References contains a complete list of file types that are supported.

### 4.4 ASM Instances

An ASM instance is built on the same technology as an Oracle Database instance. An ASM instance has a System Global Area (SGA) and background processes that are similar to those of Oracle Database. However, because ASM performs fewer tasks than a database, an ASM SGA is much smaller than a database SGA. In addition, ASM has a minimal performance effect on a server. ASM instances mount disk groups to make ASM files available to database instances; ASM instances do not mount databases. The ASM instance executes only a small portion of the code in the Oracle kernel, thus it is less likely to encounter failures or contention.

The ASM instance creates an extent map which has a pointer to where each 1 MB extent of the data file is located. When a database (Relational Database Management System or RDBMS) instance
creates or opens a database file that is managed by ASM, the database instance messages the ASM instance and ASM returns an extent map for that file. From that point the RDBMS instance performs all I/O directly to the disks unless the location of that file is being changed. Therefore, during normal operation the ASM instance is not in the I/O path. The three things that might cause the extent map for a database instance to be updated are:

1) Rebalancing the disk layout following a storage configuration change (adding or dropping a disk from a disk group).
2) Opening of a new database file.
3) Extending an existing database file when a tablespace is enlarged.

ASM metadata is the information that ASM uses to control a disk group and the metadata resides within the disk group. The RDBMS instances never directly update ASM metadata. ASM metadata is written only by the ASM instance. ASM metadata includes the following information:

- The disks that belong to a disk group
- The amount of space that is available in a disk group
- The filenames of the files in a disk group
- The location of disk group datafile data extents
- A redo log that records information about atomically changing data blocks

ASM and database instances require shared access to the disks in a disk group. ASM instances manage the metadata of the disk group and provide file layout information to the database instances.
5 I/O Workloads

This section gives a general overview of I/O workload types and the metrics used to gauge performance for these workloads. It then discusses the specifics of Oracle I/O workloads.

5.1 General I/O Workload Types and Associated Metrics

I/O workloads are typically characterized as belonging to one of the following types:

1) Small block, random I/Os with a relatively high transaction rate. This includes OLTP databases, mail servers, Web servers and file servers.

2) Large block, sequential I/Os. This includes Data Warehouse databases, video servers and backup servers.

3) Mixed workloads. This is a combination of the workload types in 1) and 2) above.

5.1.1 Small Block, random I/O workloads

The performance of these OLTP-type workloads is measured in terms of two metrics, I/Os per second (IOPS) and latency (or response time). IOPS are dependent on two characteristics of disk drives, the average seek time and the rotational latency. The response time is the actual service time for a given I/O. Calculations for the IOPS that can be expected from the DS8000 are given in the paper. "IBM DS8000 Performance Configuration Guidelines for Implementing Oracle Databases with Automatic Storage Management" (item 5 in the References)

5.1.2 Large Block, sequential I/O workloads

The performance of sequential I/O workloads is measured in throughput or Megabytes per second (MBPS). Response times are generally not important for sequential I/O workloads as long as throughput objectives are met.

5.2 Oracle I/O Workloads

Oracle I/O workloads can be small-block, random I/Os or large-block sequential I/Os:

5.2.1 Oracle random I/O workloads

The I/O size for Oracle random I/O is the database block size, which is set when the database is created. This is usually set to 8 KB for Oracle databases that are going to be used for OLTP applications or for mixed workloads. However, the database block size can be set up to 32 KB. And starting with Oracle 9i, it became possible to create tablespaces that have block sizes that differ from the base block size set at database creation time. So the block size can be 8 KB for tablespaces with OLTP data and 16 KB or 32 KB for tablespaces containing the data warehouse data.
5.2.2 Oracle sequential I/O workloads

For sequential I/O, Oracle will create an I/O size that is composed of several database blocks. Oracle will use I/O sizes of up to 1 MB for sequential I/O. This is the default size used for the allocation units (AUs) in Oracle Database 10g and 11g ASM. Allocation units are the fundamental unit of allocation within an ASM disk group. In Oracle Database 11g, you can modify the ASM AU size to be between 1 MB and 64 MB in powers of two, i.e., 1, 2, 4, 8, 16, 32, or 64.

5.2.3 Determining the current Oracle I/O profile

Appendix A contains two SQL*Plus scripts which can be used to determine the I/O profile, with regard to IOPS and throughput (MBPS), of a currently running Oracle database. They can be run on either single-instance or RAC databases.

5.3 SVC Performance Characteristics

The Storage Performance Council Benchmark1 (SPC-1) results (OLTP workloads) for SVC 4.3 can be viewed at:


The SPC Benchmark2 (SPC-2) results (Sequential workloads) for SVC 4.2 can be viewed at:

6 SVC Configuration Best Practices

This section consolidates and summarizes the best practices gathered from the SAN Volume Controller Redbooks® listed in the References section. As stated in the introduction, one of the assumptions for this paper is that the best practices documented for the SVC have been arrived at by doing testing which included database-type I/O workloads. Therefore it is critical to review these documented best practices and use them as the starting point in any SVC deployment for Oracle with ASM. They are the foundation upon which any configuration decisions must be made.

6.1 Summary of SAN Volume Controller Configuration Best Practices

This section will discuss the best practices for each level of virtualization in an SVC environment.

6.1.1 SAN Topology for SVC Deployment

This section will give an overview of the recommended SAN topology when an SVC is being deployed.

(1) One of the most basic SVC SAN requirements is to create two (or more) entirely separate SANs (i.e., fabrics) that are not connected to each other over Fibre Channel in any way. The easiest way to do this is to construct two SANs that are mirror images of each other. Technically, the SVC will support using just a single SAN (appropriately zoned) to connect the entire SVC. However, this design is not recommend in any production environment since it offers no redundancy.

In addition, constructing two “logical” SANs, such as with Cisco VSANs or Brocade Traffic Isolation Zones, is not acceptable in a production environment. Those two approaches can create multiple “virtual” fabrics, but they offer no true hardware redundancy.

A single-fabric SAN is also not recommended even for development environments since an outage in that type of environment could have an expensive business impact. But a single-fabric SAN may be acceptable for a platform that is dedicated to testing storage only.

(2) Given that the recommendation in item (1) above is followed, and a dual-fabric topology is being deployed, each fabric would have half of the SVC I/O ports connected to it. Each of these two sets of SVC I/O ports would of course be mutually exclusive. Therefore, the best approach would be to have two out of the four I/O ports on each SVC node going to each of the two fabrics.

(3) The back-end storage subsystem I/O ports should also be split evenly across the two fabrics. Therefore, for the DS8000 the best approach would be to have two out of the four I/O ports on each Host Adapter (HA) going to each of the two fabrics.

(4) Due to the nature of Fibre Channel, it is very important to avoid inter-switch link (ISL) congestion. While Fibre Channel (and the SVC) can, under most circumstances, handle a host or storage array that has become overloaded, the mechanisms in Fibre Channel for dealing with congestion in the fabric itself are not very effective. The problems caused by fabric congestion can range anywhere from dramatically slow response time all the way to storage access loss. These issues are common with all high-bandwidth SAN devices and are inherent to Fibre Channel; they are not unique to the SVC.

(5) All SVC node ports in a cluster must be connected to the same SAN switch as the storage devices with which the SVC cluster is expected to communicate. Thus the SVC I/O ports mentioned in item (2) above and the DS8000 I/O ports mentioned in item (3) above would all go to the same physical switch in a given fabric.
If using a basic core-edge SAN topology, which consists of a switch in the center (usually, a director-class switch) surrounded by other switches, then the core switch should contain all SVC ports, storage ports, and high-bandwidth hosts. The edge switches could contain all of the lower-bandwidth hosts. The core switch will be connected via ISLs to the edge switches.

(6) Storage traffic and inter-node traffic must never transit an ISL, except during migration scenarios.

(7) If a high-bandwidth-utilization server (such as a tape backup server) is using VDisks directly, then it should be on the same SAN switches as the SVC node ports. Putting it on a separate switch can cause unexpected SAN congestion problems. Putting a high-bandwidth server on an edge switch is a waste of ISL capacity.

However, if a tape backup server is backing up filesystem data, then it should ideally be connected through a different fabric and should be isolated from the SVC altogether. Slow draining tape server ports can cause major fabric congestion and can therefore affect SVC performance.

(8) If at all possible, plan for the maximum size configuration that you ever expect your SVC installation to reach. The design of the SAN can change radically for larger numbers of hosts. Modifying the SAN later to accommodate a larger-than-expected number of hosts will either produce a poorly-designed SAN or be very difficult, expensive, and disruptive to your business. This does not mean that you need to purchase all of the SAN hardware initially, just that you need to lay out the SAN while keeping the maximum size in mind.

(9) Always deploy at least one "extra" ISL per switch. Not doing so opens you up to consequences from complete path loss (this is bad) to fabric congestion (this is even worse).

(10) If there are SVC paths from the same SVC node to the same disk array on multiple core switches that are linked together, it is important to zone the SVC so that it will only see the paths on the same switch as the SVC nodes.

(11) For highest performance and availability, a four-fabric SAN can be utilized.

6.1.2 Zoning of SVC, Storage Controller, and Host

This section will describe the best practices for the Fibre Channel zoning of all of the components in the SAN environment – the SVC, the back-end storage controller, and the front-end hosts.

(1) Only use WWPN (worldwide port name) zoning. Do not use WWNN (worldwide node name) or port zoning. It is a common misconception that WWPN zoning provides poorer security than WWNN or port zoning. This is not the case. Modern SAN switches enforce the zoning configuration directly in the switch hardware, and port binding functions can be used to enforce that a given WWPN must be connected to a particular SAN switch port.

There are multiple reasons not to use WWNN zoning. For hosts, it is absolutely a bad idea, because the WWNN is often based on the WWPN of only one of the HBAs. If you have to replace that HBA, the WWNN of the host will change on both fabrics, which will result in access loss. In addition, it also makes troubleshooting more difficult, because you have no consolidated list of which ports are supposed to be in which zone, and therefore, it is difficult to tell if a port is missing.

(2) Use a well-defined naming convention when configuring SVC-related zones. The maintenance of the SVC configuration can become difficult and unwieldy if the zone names do not follow a standard convention and are not self-documenting.

(3) It is highly recommended to use zoning aliases (if your switch hardware supports that). This self-documentation makes maintenance much easier. The following zoning aliases should be created:
(a) One alias that holds all of the SVC node I/O ports on each fabric. So each alias in your dual-fabric setup will contain half of the SVC I/O ports.

(b) One alias that holds all of the back-end storage subsystem I/O ports on each fabric. So each alias in your dual-fabric setup will contain half of the back-end storage subsystem I/O ports.

(c) One alias for each corresponding pair of ports on both nodes of an I/O group; i.e., create an alias that contains port 1 from both nodes in I/O group 0, create an alias that contains port 2 from both nodes in I/O group 0, etc. So each I/O group will have four aliases associated with it, split evenly across the two fabrics. These aliases will be the building blocks of the host-related zones.

(d) Host zoning aliases can be omitted in smaller environments, e.g., in a lab environment.

(4) Create an SVC intra-cluster zone. This zone would contain all of the SVC I/O ports on a given fabric. So, assuming that a dual-fabric topology is being deployed, as was recommended in item (1) of section 6.1.1 SAN Topology for SVC Deployment, there would be one such zone created in each fabric. This zone should consist of the alias that was recommended in item (3)(a) above. The zone cannot have the same name as the alias.

   This zone will overlap with all of the SVC-to-storage zones that are going to be created in item (5) below. This is okay, because this zone is a fail-safe in case any zones in (5) get disabled. A requirement of the SVC is that all nodes be able to see each other over the fabric.

(5) For each back-end storage subsystem, create an SVC-to-storage zone that contains all of that storage subsystem’s I/O ports that are connected to that fabric and the SVC intra-cluster zone. This zone could consist of the aliases that were recommended in items (3)(b) and (3)(a) above. This is the SVC-to-storage zone (for each storage subsystem).

(6) Create a single zone for each host port. It should contain the host port and one port from each SVC node that the host will need to access. Therefore, this zone would contain the WWPN of one host port and one of the aliases from (3)(c) above (which would contain two SVC I/O ports). Assuming two ports per host, each one connected to a different fabric in the dual-fabric setup, this will give four paths to each VDisk. That is the optimal configuration for which both IBM Subsystem Device Driver (SDD) multipathing software and SVC have been tuned.

(7) Avoid zoning different vendor storage subsystems together.

6.1.3 Storage Controller – DS8000

This section will give the guidelines for configuring a DS8000 when it is being used as the storage controller for an SVC. In addition to the following items, the more detailed guidelines given in the paper “IBM DS8000 Performance Configuration Guidelines for Implementing Oracle Databases with Automatic Storage Management”, which is item 5 in the References, should also be followed.

(1) For the DS8000, the concept of preferred path is not used since Fibre Channel cards are outboard of the controllers (called processor complexes, or servers, in the DS8000). Therefore, all Fibre Channel ports are available to access all LUNs regardless of server affinity. While server affinity still exists, the network between the outboard Fibre Channel ports and the processor complexes performs the appropriate “routing” as opposed to the DS4000® and DS6000™ where controller routing is performed by the multipathing driver in the host, such as with IBM Subsystem Device Driver (SDD).
(2) Change the default name of the storage controller as it is first created in SVC. Typically, the controller name will be something like "controller1". So use a standard naming convention which contains the storage controller serial number. For example, a DS8000 with a serial number of 75HP422 could be named "DS8000_75HP422".

(3) There is no array and cache parameter tuning in DS8000. The RAID-5 arrays will be 6+P+S or 7+P and the segment size (stripe width) is always 256 KB for fixed block volumes. Caching for the DS8000 is done on a 64 KB track boundary.

(4) Balance DS8000 ranks which will be used by the SVC across even and odd Extent Pools. This is the same recommendation for the DS8000 as when not using SVC.

(5) Configure a minimum of 8 I/O ports per DS8000. If there are more than 48 ranks being presented to the SVC, then configure 16 ports from the DS8000.

(6) Configure a maximum of two ports from a four-port DS8000 Host Adapter.

(7) Configure adapters across redundant SAN networks from different DS8000 I/O enclosures.

### 6.1.4 MDisk Configuration

This section gives an overview of the recommended MDisk configuration.

(1) Each LUN from the back-end storage that is to be presented to the SVC as an MDisk should take up all of the space on the RAID array; i.e., there should only be one LUN per back-end RAID array. This is more important for midrange storage controllers where multiple LUNs per RAID array have been shown to result in performance degradation. It is less important for the DS8000 (with its large cache size) where testing has shown a negligible performance difference between having one or two LUNs per array. But in general, for both performance reasons and simplicity of administration, one LUN per array is the best approach for all back-end storage.

With the increasing disk sizes available for the DS8000 (450 GB as of this writing), it is possible that one RAID array is larger than the maximum allowable LUN size, which is 2 TB for fixed block volumes. If that is the case, then two equally sized LUNs should be created from such a RAID array and presented as two MDisks to the SVC. These two MDisks should be included in the same MDisk Group, as will be discussed in detail in section 6.1.5 MDisk Group (MDG) Configuration item (4).

The largest allowable MDisk size is 2 TB and, as mentioned in section 4.2 ASM Disks, 2 TB is also the largest allowable ASM disk size.

(2) Rename MDisks from their default assigned SVC name. As always, use a consistent and self-documenting naming convention.

(3) All of the LUNs from a DS8000 that are to be presented as MDisks to an SVC cluster should be visible to only that one SVC cluster.

(4) The performance characteristics of the DS8000 RAID-5 and RAID-10 arrays compare as follows:
(a) For reads from disk, either sequential or random, there is no significant difference between RAID-5 and RAID-10.

(b) For sequential writes, RAID-5 performs better than RAID-10.

(c) For random writes, RAID-10 performs better than RAID-5.

A detailed analysis of these DS8000 RAID characteristics is contained in the white paper “IBM DS8000 Performance Configuration Guidelines for Implementing Oracle Databases with Automatic Storage Management” noted in the References. That paper discusses the performance effect when using a DS8000 without SVC. Section 11.2 Performance Comparison of DS8000 RAID-10 and RAID-5 on SVC further down in this paper contains a detailed analysis of the effect that DS8000 RAID-10 and RAID-5 have on a front-end SVC.

6.1.5 MDisk Group (MDG) Configuration

This section gives an overview of the recommended MDisk Group (MDG) configuration.

(1) The LUNs from the DS8000 which are presented as MDisks to the SVC cluster, and which are going to be added to a given MDisk Group, should follow these guidelines:

(a) They should be from the same back-end storage controller.

(b) They should be the same type, i.e., all fixed block volumes for Open Systems.

(c) They should be the same RAID level (all RAID-5, all RAID-10, or all RAID-6).

(d) They should have the same RAID width, i.e., the same number of physical disks in the array (the DS8000 will always have 8 disks per array).

(e) They should have the same performance, availability, and fault tolerance characteristics.

Combining MDisks in an MDisk Group where the back-end RAID arrays have different performance characteristics will cause the performance of the MDisk Group to be constrained by the speed of the slowest RAID array. And combining MDisks in an MDisk Group where the back-end RAID arrays have different availability characteristics will decrease the overall availability of the MDisk Group. Combining MDisks from different back-end storage controllers, even if the storage controller is from the same vendor and is the same model, is especially inadvisable for that reason.

(2) The capability to stripe across disk arrays is the single most important performance advantage of the SVC; however, the more arrays you stripe across is not necessarily better. It is necessary to weigh performance against availability when deciding how to add MDisks to MDGs. Putting MDisks from many RAID arrays in an MDG may potentially improve performance, but it increases exposure of the MDG to failure due to one of the arrays failing. In addition, SVC performance testing has shown that the performance advantage with striping across a larger number of arrays is not as pronounced as might be expected, i.e., there is not much performance difference between 4 or 8 arrays per MDG.

Given these facts, the following are some guidelines that take both availability and performance into account:
(a) Each storage subsystem must be used with only a single SVC cluster, i.e., all of the LUNs from a DS8000 that are to be presented as MDisks to an SVC cluster should be visible to only one such SVC cluster (this was previously mentioned in Section 6.1.4 MDisk Configuration).

(b) Each storage RAID array must be included in only one MDG.

(c) Each MDG must include MDisks from only one storage subsystem, i.e., only the LUNs from one DS8000 should be in a given MDG.

(d) The number of arrays per MDG for the DS8000 should be 4 to 12.

(3) For workload mixing, here are some guidelines:

(a) It is usually recommended to mix Batch (Sequential) and OLTP workloads, so that maximum resources are available to any workload when needed.

(b) But Batch workloads may simply use up all of the I/O resources available. So it may help to put Batch in its own MDG. However, this does not prevent node or path resources from still being overrun.

(c) It is possible to cap the data rate at which batch volumes are allowed to run by limiting the maximum throughput of a VDisk. This can potentially let online work benefit from periods when the batch load is light while limiting the damage when the batch load is heavy. Obviously, possible negative impacts to the workload running on the throttled down VDisks must be very carefully considered.

(d) If it is not possible to run OLTP and Batch at different times of the day, and there is no cap or throttling rate applied, then segregate OLTP and Batch onto different MDGs that are supported by different back-end resources.

(4) Add MDisks to the MDG such that you start with an even-numbered Extent Pool MDisk, then add an odd-numbered Extent Pool MDisk, then an even-numbered, and so on. This will spread the workload across the back-end DS8000 processor complexes. When you create a striped VDisk (to be discussed in section 6.1.6 VDisk Configuration below), the SVC will stripe in the order that the MDisks were added to the MDG. The starting MDisk is randomly chosen by the SVC.

As was discussed in section 6.1.4 MDisk Configuration item (1), larger DS8300 disk sizes may necessitate creating two equally-sized LUNs on the same back-end RAID array. These two LUNs will of course map to two equally-sized MDisks. And, as was discussed in item (2)(b) above, a given back-end RAID array should only be included in one MDisk Group. Therefore, do not add two MDisks from the same back-end RAID array to an MDG one right after the other. There should be an intervening MDisk which maps to an extent pool with a different mathematical parity, i.e., after adding an MDisk which maps to an even-numbered extent pool, the next MDisk to be added should map to an odd-numbered extent pool. Likewise, after adding an MDisk which maps to an odd-numbered extent pool, the next MDisk to be added should map to an even-numbered extent pool.
(5) Extent size is set during MDG creation. The maximum storage capacity of an SVC cluster is dependent on the extent size. The following shows the relationship between extent size and storage capacity of an SVC cluster:

<table>
<thead>
<tr>
<th>Extent Size</th>
<th>Maximum storage capacity of SVC cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>16 MB</td>
<td>64 TB</td>
</tr>
<tr>
<td>32 MB</td>
<td>128 TB</td>
</tr>
<tr>
<td>64 MB</td>
<td>256 TB</td>
</tr>
<tr>
<td>128 MB</td>
<td>512 TB</td>
</tr>
<tr>
<td>256 MB</td>
<td>1 PB</td>
</tr>
<tr>
<td>512 MB</td>
<td>2 PB</td>
</tr>
<tr>
<td>1 GB</td>
<td>4 PB</td>
</tr>
<tr>
<td>2 GB</td>
<td>8 PB</td>
</tr>
</tbody>
</table>

Extent size does not affect performance. For most clients, extent sizes of 128 MB or 256 MB give a reasonable balance between VDisk granularity and cluster capacity. Having said that, the benchmarks run for this paper showed a slight edge in sequential I/O performance when a 1 GB extent size was used as opposed to a 256 MB extent size. Therefore all benchmark runs used 1 GB extents for the MDG.

(6) Think twice before adding an MDisk to an MDG. This can reduce the reliability characteristics of the MDG. It is better to create a new MDG and add the appropriate MDisks to this new MDG. However, if it has been decided to add an MDisk to an MDG, MDisks are automatically tested for reliable read/write access before being added to that MDG. This is a feature which became available in SVC 4.2.1; prior to SVC 4.2.1 it was necessary to manually test R/W access.

Adding MDisks to existing MDGs can result in reduced performance across the MDG due to the extent imbalance that will occur and the potential to create hot spots within the MDG. After adding MDisks to MDGs, it is recommended that extents be rebalanced across all available MDisks. This can be accomplished by using a script located at http://www.alphaworks.ibm.com/tech/svctools.

### 6.1.6 VDisk Configuration

This section will detail the recommendations for VDisk configuration.

(1) Decide on the VDisk naming convention before starting. It is easier to use the correct names at creation time than change them later, although the `svctask chvdisk` command can be used to accomplish such a name change.

(2) With extremely few exceptions, you must always configure VDisks using striping mode. The one exception to this rule is when the workload is 100% sequential and disk loading across all VDisks is guaranteed to be balanced by the nature of the application. Two examples are video streaming or an environment where there is a high dependency on a large number of flash copies. But electing to use sequential mode over striping mode requires a detailed understanding of the data layout and workload characteristics in order to avoid negatively impacting system performance.
Therefore, in the vast majority of cases, the striping of VDisks is recommended even if the application software offers its own striping (or load balancing), as is the case with Oracle ASM. Striping at the VDisk level takes advantage of full Workload Spreading as was defined in the paper “IBM DS8000 Performance Configuration Guidelines for Implementing Oracle Databases with Automatic Storage Management” (item 5 in the References). It also reduces the administrative work for DBAs and System Administrators since it offers a constant level of load balancing on the back-end without having to spend time on that aspect of performance planning.

(3) Balance VDisks across the I/O groups in the SVC cluster. It is best to get this configuration correct at the time of VDisk creation. Moving a VDisk to another I/O group is a disruptive procedure. The parameter -iogrp is required with the svctask mkvdisk command.

In configurations with large numbers of attached hosts, where it is not possible to zone a host to multiple I/O Groups, it might not be possible to choose to which I/O Group to attach the VDisks. The VDisk has to be created in the I/O Group to which its front-end host belongs.

(4) It is important to distribute the workload evenly on the SVC nodes within an I/O group. By default, the SVC assigns ownership of even numbered VDisks to one node of an I/O group and odd numbered VDisks to the other node. A node that is assigned ownership of a VDisk is known as the preferred node. The SVC implementation of a preferred node is meant to improve cache efficiency and cache usage. So the default automatic distribution of the VDisks across the nodes in an I/O group is a load balancing mechanism.

It is possible for a performance imbalance to occur either because of predominantly even- or odd-numbered VDisks, or because of unbalanced VDisk sizes. To avoid this problem, it is possible to explicitly assign the preferred node during VDisk creation. However, it is not possible to change the preferred node for a VDisk once it has been created. Changing the preferred node for a VDisk currently involves moving the VDisk to another I/O group. So it is disruptive to the host to which the VDisk is mapped. A future release of SVC may have a single command to change the preferred node.

(5) Performance can be better if the access is made on the preferred node to a VDisk. Conversely, in some situations the performance impact of using the non-preferred node can be significant. Under normal operating conditions, it is expected that hosts will, via the multipathing software, access VDisks through the preferred node. But in the event of a path failure, the data can still be accessed by the partner (non-preferred) node in the I/O Group.

(6) The maximum number of VDisks per I/O group is 1024.

(7) The maximum number of VDisks per SVC cluster is 8192 (in an eight node cluster).

(8) Extent size does not affect performance. For most clients, extent sizes of 128 MB or 256 MB give a reasonable balance between VDisk granularity and cluster capacity. Extent size is set during MDisk group creation.

Having said that, the benchmarks run for this paper showed a slight edge in performance for sequential workloads when a 1 GB extent size was used as opposed to a 256 MB extent size.
Therefore, all test runs in this paper used a 1 GB extent size for the MDG. This was mentioned previously in item (5) of the section 6.1.5 MDisk Group (MDG) Configuration.

(9) It is possible to throttle the I/O to a VDisk either by response time (I/Os per second for random, OLTP I/O), or by throughput (MB per second for Sequential I/O).

(10) Ensure that striped VDisks alternate between MDisks that are from even and odd Extent Pools on the DS8000. This can be done in one of two ways:

   (a) Add MDisks to the MDG such that you start with an even-numbered Extent Pool MDisk, then add an odd-numbered Extent Pool Mdisk, then an even-numbered, and so on. When you create a striped VDisk, the SVC will stripe in the order that the MDisks were added to the MDG. The starting MDisk is randomly chosen by the SVC.

   (b) If you did not add the MDisks to the MDG as described in (a), you can specify the order of striping when creating the VDisk.

   (c) The method in (a) may be preferable to (b), because (a) will allow the starting MDisk to be randomly chosen by the SVC and thus it is not necessary to try to keep track of using a different starting MDisk as you might want to do with (b).

6.1.7 Host Configuration

This section gives the guidelines for configuring the front-end hosts when they are using an SVC. Both AIX and Linux are discussed.

(1) Limit the number of host port connections to the SVC to two, on two different physical adapters.

(2) Zone each of these two host ports to two ports on the SVC, one SVC port on each node of an I/O group. This is a total of four paths, which is optimal. This was discussed in item (6) of the section 6.1.2 Zoning of SVC, Storage Controller, and Host.

(3) The two paths from a given host port will both be going to one fabric in the dual-fabric topology. So each host HBA will be connected to a different fabric. This was also discussed in item (6) of the section 6.1.2 Zoning of SVC, Storage Controller, and Host.

(4) Use one host object to represent a cluster of hosts and use multiple worldwide port names (WWPNs) to represent the ports from all the hosts that will share the same set of VDisks.

(5) An I/O Grouping consists of two SVC nodes that share management of VDisks within a cluster. The recommendation is to utilize a single I/O Group (iogrp) for all VDisks allocated to a particular host. This recommendation has many benefits. One major benefit is the minimization of port fanouts within the SAN fabric. Another benefit is to maximize the potential host attachments to the SVC, because maxima are based on I/O Groups. A third benefit is within the host itself, having fewer target ports to manage.

(6) SVC requires the use of multipathing software on the host. For AIX, use the SDDPCM (Subsystem Device Driver Path Control Module) software. For Linux, use the Device-Mapper Multipath software that is now native to kernel level 2.6.

(7) For AIX, be sure that the queue depth setting is correct. This can be set at the individual host adapter level, at the LUN level, or at the physical adapter resource basis. Also properly set the num_cmd_elem, lg_term_dma and max_xfer_size configuration parameters for the fcsX
A discussion of the appropriate settings, and how to implement them with the `chdev` command, is contained in the SVC Best Practices Redbook listed in the References.

(8) For Linux, the queue depth, and other QLogic HBA settings, can be set in the `/etc/modprobe.conf` file. A line would be added that looks something like:

```
options qla2xxx ql2xfailover=0 qlport_down_retry=1 ql2xmaxqdepth=new_queue_depth
ql2xretrycount=5
```

A discussion of the Linux settings is also contained in the SVC Best Practices Redbook listed in the References.

## 6.2 SAN Volume Controller Logical Configuration

Given the SVC configuration best practices discussed in the previous section, we can now detail all of the configuration steps required to make storage available to a host when using SVC.

As previously mentioned, the controller (back-end storage) used for this paper will be the IBM System Storage DS8000 and the Fibre Channel switch is an IBM System Storage SAN32B-3 (Brocade 5000). The examples will show the command-line interface for some of the SVC and DS8000 commands.

The Brocade configuration is often more easily done using the Brocade browser-based GUI since it gives you a visual overview of the SAN topology. The Reference section lists the command line interface manuals (accessing the Brocade manual requires registering on **Brocade Connect**, which is available to Brocade customers).

A very high-level view of the SVC configuration process is:

- Create the Fibre Channel zoning between the SVC and DS8000 I/O ports.
- Create LUNs on the DS8000 and make them visible to the SVC as MDisks.
- Configure the SVC storage controller, the MDisks and the MDisk groups.
- Create the Fibre Channel zoning between SVC and the front-end host(s).
- Create VDisks on the SVC and make them available to the front-end hosts.
- Configure the multipathing software on the front-end hosts to make the VDisks visible and usable on those hosts. The multipathing software is Device-Mapper Multipath on Linux and SDDPCM (Subsystem Device Driver Path Control Module) on AIX.

Now let's drill down a bit into each of these bullet points.

### 6.2.1 Create the Fibre Channel zoning between SVC and DS8000

As was previously mentioned, each SVC node has four Fibre Channel I/O ports. Since there can be up to eight nodes in an SVC cluster, this would allow for up to 32 I/O ports on an SVC.

The DS8000 has 4 Fibre Channel I/O ports on each of its Host Adapters (HAs). It can have up to 8 I/O enclosures with 4 HAs per I/O enclosure. Thus the DS8000 can have a maximum of 32 HAs for a total of 128 Fibre Channel I/O ports.

The steps for zoning between the SVC and DS8000 are as follows:
(1) On each fabric in your dual-fabric topology, create an alias that contains all of the SVC I/O ports on the fabric.

(2) On each fabric in your dual-fabric topology, create a single alias that contains all of the DS8000 I/O ports on the fabric.

(3) On each fabric in your dual-fabric topology, create a single zone which consists of the alias from item (1) above. This is the SVC intra-cluster zone.

(4) On each fabric in your dual-fabric topology, create a single zone which consists of the aliases from items (1) and (2) above. This is the SVC-to-storage zone.

6.2.2 Create the DS8000 LUNs and make them visible to the SVC as MDisks

The steps below will configure the DS8000 virtualization hierarchy as discussed in the white paper "IBM DS8000 Performance Configuration Guidelines for Implementing Oracle Databases with Automatic Storage Management" noted in the References. The resulting LUNs will then be made available to the SVC via DS8000 hostconnects. Some of the commands below use the DS8000 dscli command-line tool and some use the SVC CLI tool:

(1) Create arrays (RAID-10, RAID-5 or RAID-6) on the unused, “blank-slate” arraysites. The following lsarraysite command will determine which arraysites do not have an array assigned to them yet, i.e., which arraysites have a State of “Unassigned”. The mkarray command will then create a RAID-10 array on arraysite S1 (assuming it was in the Unassigned state):

```
scli> lsarraysite
scli> mkarray -raidtype 10 -arsite S1
```

(2) Create extent pools. The command below will create a fixed block volume extent pool on rankgroup 0 (i.e., this extent pool will have an affinity for DS8000 server0) and will give it a description of “Extent Pool 0”. So the assumption is that this is the first extent pool being created. That being the case, it will automatically get assigned an ID of P0:

```
scli> mkextpool -rankgrp 0 -stgtype fb "Extent Pool 0"
```

(3) Create a fixed block volume rank on an array and simultaneously assign it to an extent pool:

```
scli> mkrank -array A0 -stgtype fb -extpool P0
```

(4) Create a volume group. For SVC, the type must be scsimask. Use a self-documenting naming convention; we will use the name SVC11. The volume group IDs get assigned automatically by the DS8000 and start with the first unused ID; so the first volume group ID assigned by the DS8000 would be V0, and then V1, etc.:

```
scli> mkvolgrp -type scsimask SVC11
```

(5) Set the DS8000 I/O port topologies correctly (otherwise the ports will not be visible on the switches). The I/O port topology must be scsi-fcp if connecting to an SVC. The following will set the topology for port I0203 on the DS8000:
(6) Create the fixed block volumes (LUNs) and simultaneously assign them to the appropriate extent pool (P0 created in step (2)) and volume group (V0 created in step (4)). The following command will create a fixed block volume with ID 0000. It will be given a name, SVC11_P0, which denotes the front-end SVC that will use this LUN (SVC11), and the fact that it is from extent pool P0. Be aware of the fact that the first two hexadecimal digits in the volume id (known as the logical subsystem, or LSS, in the DS8000 virtualization hierarchy) must be even-numbered if the extent pool number is even-numbered and they must be odd-numbered if the extent pool number is odd-numbered:

```
dscli> mkfbvol -extpool P0 -cap 100 -volgrp V0 -name SVC11_P0 0000
```

(7) You need to determine the WWPNs for the I/O ports on the SVC. So you will first need to list all of the SVC nodes and then get the detail for a specific node and make note of the four port id entries. The following assumes that one of the SVC nodes returned by the first command has a node ID of 1:

```
SVC11:admin> svcinfo lsnode
SVC11:admin> svcinfo lsnode 1
```

(8) Create the DS8000 host connections for the volume group. The WWPN in the following command is from one of the SVC I/O ports determined in step (7). Use a self-documenting naming convention. The hostconnect name SVC11_N1_P1_F1 denotes the fact that volume group V0 will have a connection to port 1 (P1) on node 1 (N1) of SVC11 and that port is connected via fabric1 (F1). Add all four of the ports for each node so that the volume group has all of the WWPNs for that node, regardless of which fabric the port is connected to:

```
dscli> mkhostconnect -dev IBM.2107-75HP422 -wwname 50050768014028D2 -profile "San Volume Controller" -volgrp V0 SVC11_N1_P1_F1
```

(9) Run a scan on the SVC to make the DS8000 LUNs visible as MDisks. The LUNs should now be visible because the SVC-to-DS8000 zoning has been set up (in Section 6.2.1 Create the Fibre Channel zoning between SVC and DS8000) and because the DS8000 volume groups have had hostconnects created in step (8) above:

```
SVC11:admin> svctask detectmdisk
```

(10) List the MDisks to verify that the SVC can now see the back-end LUNs:

```
SVC11:admin> svcinfo lsmdisk
```

### 6.2.3 Configure the SVC Storage Controller, MDisks, and MDisk groups

The following commands will all be run on the SVC:

(1) List the names of the storage controllers and rename them from the default so that a good naming convention is being used. One suggestion is to use the storage type and its serial number as part of the name:
(2) On the SVC, list the MDisks:

SVC11:admin> svctask chcontroller -name DS8000_75HP422 controller0

(3) Now rename the MDisks from the default mdiskXX name so that a good naming convention is used. One possibility is to use the controller type, the order in which this controller type was added to the SVC, and the back-end array as part of the name. The name DS8K_1_A0 used below denotes that this is a DS8K, that it was the first DS8K added to the SVC (the number 1), and that the MDisk is on back-end RAID array A0. Assume that mdisk0 was one of the MDisks returned by the svcinfo lsmdisk command in step (2):

SVC11:admin> svctask chmdisk -name DS8K_1_A0 mdisk0

(4) Create the MDisk groups and simultaneously assign the MDisks to those MDGs. One possible naming convention for the MDGs is to use the storage controller type, the order in which it was added to the SVC, and the RAID type in the name. The following will create the MDG with four MDisks and an extent size of 1 GB:

SVC11:admin> svctask mkmdiskgrp -name DS8K_1_RAID10 -mdisk
DS8K_1_A0:DS8K_1_A1:DS8K_1_A2:DS8K_1_A3 -ext 1024

6.2.4 Create the Fibre Channel zoning between SVC and front-end Hosts

The steps for zoning between the SVC and the front-end hosts are as follows:

(1) On each fabric in the dual-fabric topology, let's call the two fabrics fabric1 and fabric2, create two aliases. Each alias will contain two members which are the corresponding pair of SVC I/O ports on both nodes of an I/O group, where those I/O ports are connected to that particular fabric. So, assume that ports 1 and 3 from both nodes in I/O group 0 are connected to fabric1, and ports 2 and 4 from both nodes in I/O group 0 are connected to fabric2. Then:

(a) Create an alias on fabric1 that contains port 1 from both nodes in I/O group 0.
(b) Create an alias on fabric1 that contains port 3 from both nodes in I/O group 0.
(c) Create an alias on fabric2 that contains port 2 from both nodes in I/O group 0.
(d) Create an alias on fabric2 that contains port 4 from both nodes in I/O group 0.

Thus each SVC I/O group will have four aliases associated with it, and these four aliases will be split evenly across the two redundant fabrics.

(2) For each host port, create one zone that contains the WWPN for that host port and only one of the two aliases available to it on the fabric to which it is connected. For example, create a zone which contains the WWPN for host port 1 and the alias from either item (1)(a) or (1)(b) (but not both) above. Likewise, create another zone which contains the WWPN for host port 2 and the alias from either item (1)(c) or (1)(d) (but not both). This
will create two paths from each host port to only one of the two fabrics and will create a total of four possible paths, spread across the two fabrics, to any VDisk on the SVC.

6.2.5 Create the SVC VDisks and make them available to front-end Hosts as LUNs

The following commands will all be run on the SVC:

(1) Create the VDisks using the appropriate MDisk groups and using a good naming convention. The following command will create a 100 GB VDisk striped across all four of the MDisks previously added to the MDisk group in step (4) of Section 6.2.3 Configure the Storage Controller, MDisks, and MDisk groups. It will specify that the VDisk should use iogrp 0 and it will assign a name to the VDisk which indicates that it is the first ASM disk in the Oracle ASM diskgroup dedicated to database datafiles. It will format the VDisk asynchronously before the VDisk becomes available for use. The svcinfo lsvdiskprogress command can be used to monitor the progress of the VDisk formatting:

```
SVC11:admin> svctask mkvdisk -mdiskgrp DS8K_1_RAID10 -iogrp 0 -vtype striped -size 100 -unit gb -fmtdisk -name ORA_DG_DATA1
SVC11:admin> svcinfo lsvdiskprogress
```

If the SVC has more than one iogrp, then the next VDisk to be created can either use iogrp 0 again or it can use iogrp 1 to spread the workload. The very significant advantages of spreading the VDisks across iogrps will be shown in the benchmark sections of this paper. For example, the following command will create a second VDisk which uses iogrp 1:

```
SVC11:admin> svctask mkvdisk -mdiskgrp DS8K_1_RAID10 -iogrp 1 -vtype striped -size 100 -unit gb -fmtdisk -name ORA_DG_DATA2
```

(2) List the information for the VDisks that have been created in step (1) above. Take note of the vdisk_UID value for each VDisk. These are the VDisk serial numbers and will be used later when configuring the hosts:

```
SVC11:admin> svcinfo lsvdisk
```

(3) Wait until the svcinfo lsvdiskprogress command shows that formatting is complete. Then verify that the host ports are visible on the SVC. They should be visible because the zoning from the host ports to the SVC was done in 6.2.4 Create the Fibre Channel zoning between SVC and front-end Hosts:

```
SVC11:admin> svcinfo lshbaportcandidate
```

(4) Create the SVC host definitions. The WWPNs used in the following commands were determined from step (3) above to belong to the host ports visible to the SVC. The following commands will create two such connections from two different hosts named NODE1 and NODE2 and will include both host port WWPNs for each host. Host port WWPNs can always later be added to a host definition with the svctask addhostport command:

```
SVC11:admin> svcinfo lsxmlportcandidate
```
SVC11:admin> svctask mkhost -name NODE1_PORT1 -hbawwpn
2100001B32114E8B:2101001B32314E8B

SVC11:admin> svctask mkhost -name NODE2_PORT1 -hbawwpn
210000E08B941606:210100E08BB41606

(5) List the newly created host connection(s):
SVC11:admin> svcinfo lshost

(6) Now make the VDisk visible and available for use to the front-end host using the SVC-
to-host connections created in step (4) above:
SVC11:admin> svctask mkvdiskhostmap -host NODE1_PORT1
ORA_DG_DATA1

In an Oracle RAC environment, you would want to make the VDisk usable to more than
one host node. This can be accomplished with the -force option. NODE2 in the
following command denotes that the connection is for the second host node:
SVC11:admin> svctask mkvdiskhostmap -force -host NODE2_PORT1
ORA_DG_DATA1

(7) List the just-created VDisk hostmaps for VDisk ORA_DG_DATA1:
SVC11:admin> svcinfo lsvdiskhostmap ORA_DG_DATA1

6.2.6 Configure front-end Host multipathing software

The following commands will all be run on the front-end hosts. Both AIX and Linux commands
will be shown. The steps assume that the multipathing software, Device-Mapper Multipath (DM)
on Linux and Subsystem Device Driver Path Control Module (SDDPCM) on AIX, has already
been configured and is already being used to manage some SAN storage. Therefore we just
need to modify the configuration to include the newly created VDisks that are to be used on the
host. Documentation containing detailed instructions for the installation and initial configuration
of the multipathing software can be found in the References section in item (8) “Host
Attachment for SDDPCM on AIX”, item (9) “Using Device-Mapper Multipath Configuration
and Administration 5.2”, and item (10) “Configuring Device Mapper Multipath Tool
(DMMP) for hosts running the Linux operating system”. Device-Mapper Multipath software
is part of the standard Linux installation rpm files.

Linux (using Device-Mapper Multipath):

(1) Edit the /etc/multipath.conf file and add a stanza in the multipaths section that creates
an alias for the VDisk serial numbers that showed up in step (2) of the Section 6.2.5
Create the SVC VDisks and make them available to front-end Hosts as LUNs. The
following stanza would create a LUN alias that is the same name as the VDisk name on
the SVC cluster. Notice that the serial number has to have the number “3” prepended to
the VDisk serial number (vdisk_UID):

multipath {
  wwid  36005076801920146780000000000008

alias ORA DG DATA1
path_grouping_policy group by prio
prio_callout "/sbin/mpath_prio_alua /dev/%n"
features "1 queue_if_no_path"
path_checker tur
path_selector "round-robin 0"
fallback immediate

(2) View the currently configured and recognized SCSI devices:
[root@arcx3550fxhw6 ~]# cat /proc/scsi/scsi

(3) Now determine the HBA Host IDs for the system:
[root@arcx3550fxhw6 ~]# ls /sys/class/fc_host
This will return something like “host1” and “host2”.

(4) Rescan the SCSI buses. For each one of the HBA Host IDs returned in step (3), run the following commands. There must be a space between the minus signs in the echo commands:
[root@arcx3550fxhw6 ~]# echo "- - -" > /sys/class/scsi_host/host1/scan
[root@arcx3550fxhw6 ~]# echo "- - -" > /sys/class/scsi_host/host2/scan

(5) View the newly recognized SCSI devices. The SCSI device corresponding to the ORA DG DATA1 LUN should now be visible:
[root@arcx3550fxhw6 ~]# cat /proc/scsi/scsi

(6) Device-Mapper Multipath should automatically create the new alias. Verify that the alias has indeed been created by running the following command:
[root@arcx3550fxhw6 ~]# multipath –ll | grep -i ORA DG DATA1

AIX (using SDDPCM):

(1) Run cfgmgr to detect any new attached SAN disks:
root@/home/root> cfgmgr -v

(2) Verify that the new VDisks are visible as hdiskXX devices. Assume that hdisk15 shows up as a new hdisk device from either the pcmpath command or the lsdev | grep hdisk command. Run lscfg to verify that the hdisk15 Serial number corresponds to the back-end VDisk serial number (vdisk_UID):
root@/home/root> pcmpath query device
root@/home/root> lsdev | grep hdisk
root@/home/root> lscfg -vl hdisk15 | grep Serial
7 Monitoring and Modeling tools

This section describes two important tools that should be used in any SAN environment which includes the DS8000. TotalStorage Productivity Center (TPC), which as of this writing is at Version 3.3.2, is IBM’s tool for monitoring all layers of the technology stack in a SAN environment. It is not a monitoring tool in the traditional sense, where polling is constantly taking place for the hardware and software in the technology stack. Rather it collects configuration and performance data at regularly scheduled intervals so that the data can be analyzed and reported on.

Disk Magic is a modeling tool which has been licensed to IBM by the owner of the product, IntelliMagic BV. This tool can be used to predict the effect that storage configuration changes will have on performance.

7.1 TotalStorage Productivity Center (TPC) Version 3.3.2

TotalStorage Productivity Center (TPC) is an essential tool in an IBM SAN environment to ensure the health of the data center. TPC offers the following features:

1) It presents a graphical overview of the entire data center topology, from hosts to Fibre Channel switches to storage.

2) It allows a drill-down into each object in the topology. For example, you can select a given DS8000 and expand it to view all of the layers of the virtualization hierarchy.

3) It collects very detailed performance data on LUNs, RAID arrays, switches, etc. For example, for a given LUN over a specified time period you can see the IOPS, the response time, the throughput and the read or write cache hit ratio.

4) It offers a wide range of reports that can be used to analyze the collected performance data.

TPC is simply the only way to monitor and report on the all of the layers of the technology stack in the IBM SAN environment and is a critical component of the setup in a data center.

7.2 Disk Magic

Disk Magic is a storage configuration modeling tool. It can be used to predict the performance impact of changes in the configuration of IBM SAN storage. It also supports storage from other vendors. Disk Magic can be used both before deploying a new DS8000, to predict the outcome of moving data to the DS8000, or also after deployment to predict the consequences of configuration changes. Disk Magic is an internal IBM modeling tool which is available from IBM Sales and Business Partners. It is also available for purchase from the vendor, IntelliMagic BV.

These are some of the storage configuration changes for which Disk Magic can predict a performance impact:

- Move the current I/O load to a different disk subsystem.
- Merge the I/O load of multiple disk subsystems into a single one.
- Insert a SAN Volume Controller into an existing disk configuration.
- Increase the current I/O load.
- Increase the disk subsystem cache size.
- Change to large DDMs.
- Use fewer or more LUNs.

It is advisable to run Disk Magic in any installation that is about to deploy an SVC to get an idea of the best way to configure the storage.
8 ORION - Oracle I/O Numbers Calibration tool

The ORION tool can be used to generate Oracle-specific I/O load and to gather the resulting performance statistics. The overview and description of the input parameters for ORION given below is copied directly from the ORION User’s Guide. That guide should be referenced for the most complete understanding of the tool.

8.1 ORION tool overview

ORION is a tool for predicting the performance of an Oracle database without having to install Oracle or create a database. Unlike other I/O calibration tools, ORION is expressly designed for simulating Oracle database I/O workloads using the same I/O software stack as Oracle. It can also simulate the effect of striping performed by ASM.

The following types of I/O workloads are supported:

1) Small Random I/O: OLTP applications typically generate random reads and writes whose size is equivalent to the database block size, typically 8 KB. Such applications typically care about the throughput in I/Os Per Second (IOPS) and about the average latency (I/O turn-around time) per request. These parameters translate to the transaction rate and transaction turn-around time at the application layer.

ORION can simulate a random I/O workload with a given percentage of reads vs. writes, a given I/O size, and a given number of outstanding I/Os. The I/Os are distributed across all disks.

2) Large Sequential I/O: Data warehousing applications, data loads, backups, and restores generate sequential read and write streams composed of multiple outstanding 1 MB I/Os. Such applications are processing large amounts of data, like a whole table or a whole database and they typically care about the overall data throughput in MegaBytes Per Second (MBPS).

ORION can simulate a given number of sequential read or write streams of a given I/O size with a given number of outstanding I/Os. ORION can optionally simulate ASM striping when testing sequential streams.

3) Large Random I/O: A sequential stream typically accesses the disks concurrently with other database traffic. With striping, a sequential stream is spread across many disks. Consequently, at the disk level, multiple sequential streams are seen as random 1 MB I/Os, which we also call Multi-User Sequential I/O.

4) Mixed Workloads: ORION can simulate 2 simultaneous workloads: Small Random I/O and either Large Sequential I/O or Large Random I/O. This enables you to simulate, for example, an OLTP workload of 8 KB random reads and writes with a backup workload of 4 sequential read streams of 1 MB I/Os.

For each type of workload, ORION can run tests at different levels of I/O load to measure performance metrics like MBPS, IOPS and I/O latency. Load is expressed in terms of the number of outstanding
asynchronous I/Os. Internally, for each such load level, the ORION software keeps issuing I/O requests as fast as they complete to maintain the I/O load at that level. For random workloads (large and small), the load level is the number of outstanding I/Os. For large sequential workloads, the load level is a combination of the number of sequential streams and the number of outstanding I/Os per stream. Testing a given workload at a range of load levels helps the user understand how performance is affected by load.

8.2 ORION input parameters

This section lists mandatory and optional input parameters to the ORION tool.

**MANDATORY INPUT PARAMETERS**

**run**: Test run level. This option provides simple command lines at the simple and normal run levels and allows complex commands to be specified at the advanced level. If not set as -run advanced, then setting any other non-mandatory parameter (besides -cache_size or -verbose) will result in an error.

**simple**: Generates the Small Random I/O and the Large Random I/O workloads for a range of load levels. In this option, small and large I/Os are tested in isolation. The only optional parameters that can be specified at this run level are -cache_size and -verbose. This parameter corresponds to the following invocation of ORION:

```
# ./orion -run advanced -testname mytest \
   -num_disks NUM_DISKS \n   -size_small 8 -size_large 1024 -type rand \n   -simulate concat -write 0 -duration 60 \n   -matrix basic
```

**normal**: Same as -simple, but also generates combinations of the small random I/O and large random I/O workloads for a range of loads. The only optional parameters that can be specified at this run level are -cache_size and -verbose. This parameter corresponds to the following invocation of ORION:

```
# ./orion -run advanced -testname mytest \
   -num_disks NUM_DISKS \n   -size_small 8 -size_large 1024 -type rand \n   -simulate concat -write 0 -duration 60 \n   -matrix detailed
```

**advanced**: Indicates that the test parameters will be specified by the user. Any of the optional parameters can be specified at this run level.

**testname**: Identifier for the test. The input file containing the disk or file names must be named <testname>.lun. The output files will be named with the prefix <testname>_.

**num_disks**: Actual number of physical disks used by the test. Used to generate a range for the load.

**OPTIONAL INPUT PARAMETERS**

**help**: Prints ORION help information. All other options are ignored when help is specified.

**size_small**: Size of the I/Os (in KB) for the Small Random I/O workload. (Default is 8).

**size_large**: Size of the I/Os (in KB) for the Large Random or Sequential I/O workload. (Default is 1024).

**type**: Type of the Large I/O workload. (Default is rand):
**rand:** Large Random I/O workload.

**seq:** Large Sequential I/O workload.

**num_streamIO:** Number of outstanding I/Os per sequential stream. Only valid for -type seq. (Default is 1).

**simulate:** Data layout to simulate for Large Sequential I/O workload.

- **concat:** A virtual volume is simulated by serially chaining the specified LUNs. A sequential test over this virtual volume will go from some point to the end of one LUN, followed by the beginning to end of the next LUN, etc.

- **raid0:** A virtual volume is simulated by striping across the specified LUNs. The stripe depth is 1M by default (to match the Oracle ASM stripe depth) and can be changed by the -stripe parameter.

**write:** Percentage of I/Os that are writes; the rest being reads. This parameter applies to both the Large and Small I/O workloads. For Large Sequential I/Os, each stream is either read-only or write-only; the parameter specifies the percentage of streams that are write-only. The data written to disk is garbage and unrelated to any existing data on the disk. **WARNING: WRITE TESTS WILL OBLITERATE ALL DATA ON THE SPECIFIED LUNS.**

**cache_size:** Size of the storage array's read or write cache (in MB). For Large Sequential I/O workloads, ORION will warm the cache by doing random large I/Os before each data point. It uses the cache size to determine the duration for this cache warming operation. If not specified, warming will occur for a default amount of time. If set to 0, no cache warming will be done. (Default is not specified, which means warming for a default amount of time).

**duration:** Duration to test each data point in seconds. (Default is 60).

**matrix:** Type of mixed workloads to test over a range of loads. (Default is detailed).

- **basic:** No mixed workload. The Small Random and Large Random/Sequential workloads will be tested separately.

- **detailed:** Small Random and Large Random/Sequential workloads will be tested in combination.

- **point:** A single data point with S outstanding Small Random I/Os and L outstanding Large Random I/Os or sequential streams. S is set by the num_small parameter. L is set by the num_large parameter.

- **col:** Large Random/Sequential workloads only.

- **row:** Small Random workloads only.

- **max:** Same as detailed, but only tests the workload at the maximum load, specified by the num_small and num_large parameters.

**num_small:** Maximum number of outstanding I/Os for the Small Random I/O workload. Can only be specified when matrix is col, point, or max.

**num_large:** Maximum number of outstanding I/Os for the Large Random I/O workload or number of concurrent large I/Os per stream. Can only be specified when matrix is row, point, or max.
**verbose**: Prints progress and status information to standard output.

The offsets for I/Os are determined as follows:

For Small Random and Large Random workloads:

- The LUNs are concatenated into a single virtual LUN (VLUN) and random offsets are chosen within the VLUN.

For Large Sequential workloads:

- With striping (-simulate raid0). The LUNs are used to create a single striped VLUN. With no concurrent Small Random workload, the sequential streams start at fixed offsets within the striped VLUN. For n streams, stream i will start at offset \(\text{VLUNsize} \times (i + 1) / (n + 1)\), except when n is 1, in which case the single stream will start at offset 0. With a concurrent Small Random workload, streams start at random offsets within the striped VLUN.

- Without striping (-simulate CONCAT). The LUNs are concatenated into a single VLUN. The streams start at random offsets within the single VLUN.
9  Lab Setup

This section lists the hardware and software configurations that were used in the lab exercises.

9.1.1 SAN Volume Controller

Table 10.1 describes the SVC hardware used in the exercises.

<table>
<thead>
<tr>
<th>Storage name</th>
<th>OSL_SVC_11</th>
</tr>
</thead>
<tbody>
<tr>
<td>Storage IP</td>
<td>9.1.113.35</td>
</tr>
<tr>
<td>Node Model</td>
<td>2145-8G4</td>
</tr>
<tr>
<td>Micro code level</td>
<td>4.3.1.2 (build 9.14.0811282000)</td>
</tr>
<tr>
<td>Logical configuration</td>
<td>4 node cluster (2 I/O groups)</td>
</tr>
</tbody>
</table>

9.1.2 System Storage DS8000

Table 10.2 describes the DS8000 image used in the exercises. This DS8000 storage unit has two separate images, and behaves like two separate DS8000s. However, only one image was used for the testing.

<table>
<thead>
<tr>
<th>IBM System Storage DS8000 Model</th>
<th>DS8300 Model 932</th>
</tr>
</thead>
<tbody>
<tr>
<td>Storage Unit</td>
<td>IBM.2107-75HP420</td>
</tr>
<tr>
<td>Storage Image</td>
<td>Image2</td>
</tr>
<tr>
<td>Storage id</td>
<td>IBM.2107-75HP422</td>
</tr>
<tr>
<td>Code levels</td>
<td></td>
</tr>
<tr>
<td>License Machine Code (LMC)</td>
<td>5.3.1.116</td>
</tr>
<tr>
<td>Storage Manager</td>
<td>6.1.3.20080826.1</td>
</tr>
<tr>
<td>DSCLI</td>
<td>5.3.1.236</td>
</tr>
<tr>
<td>WWNN</td>
<td>5005076308FFCCB7</td>
</tr>
<tr>
<td>DDMs</td>
<td>300 GB, 15K RPM</td>
</tr>
<tr>
<td>Number of arraysites</td>
<td>8</td>
</tr>
<tr>
<td>Number of DDMs</td>
<td>64 (8 per arraysite)</td>
</tr>
<tr>
<td>Non-Volatile Storage</td>
<td>2.0 GB</td>
</tr>
<tr>
<td>Cache Memory</td>
<td>54.3 GB</td>
</tr>
<tr>
<td>Processor Memory</td>
<td>62.7 GB</td>
</tr>
</tbody>
</table>

Table 10.2 - DS8000 Storage Unit

9.1.3 Brocade switches

Table 10.3 describes the four IBM Brocade switches used in the exercises. The switches are configured into two fabrics with two switches per fabric.

<table>
<thead>
<tr>
<th>IBM name</th>
<th>IBM System Storage SAN32B-3</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBM machine type</td>
<td>2005</td>
</tr>
</tbody>
</table>
### IBM machine model
B5K

### Brocade name
Brocade 5000

### Ports per switch
thirty-two 4 Gbps ports

### Kernel
2.6.14

### Fabric OS
v5.3.1

### BootProm
4.6.5

<table>
<thead>
<tr>
<th>IBM machine model</th>
<th>B5K</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brocade name</td>
<td>Brocade 5000</td>
</tr>
<tr>
<td>Ports per switch</td>
<td>thirty-two 4 Gbps ports</td>
</tr>
<tr>
<td>Kernel</td>
<td>2.6.14</td>
</tr>
<tr>
<td>Fabric OS</td>
<td>v5.3.1</td>
</tr>
<tr>
<td>BootProm</td>
<td>4.6.5</td>
</tr>
</tbody>
</table>

**Table 10.3 - Brocade switches**

### 9.1.4 Host nodes

Table 10.4 describes the host nodes that will be used in the exercises.

<table>
<thead>
<tr>
<th>Server type</th>
<th>IBM System x3550</th>
</tr>
</thead>
<tbody>
<tr>
<td>Processor</td>
<td>4 x Dual-Core Intel® Xeon® processor 5160 @ 3.00 GHz</td>
</tr>
<tr>
<td>Memory</td>
<td>8 GB</td>
</tr>
<tr>
<td>Host bus adapter (HBA) model</td>
<td>QLE2462</td>
</tr>
<tr>
<td>Operating system</td>
<td>Red Hat Enterprise Linux® (RHEL) AS4 U6</td>
</tr>
<tr>
<td>Kernel version</td>
<td>2.6.9-67.ELsmp</td>
</tr>
<tr>
<td>Multipath software</td>
<td>device-mapper-multipath-0.4.5-27.RHEL4</td>
</tr>
<tr>
<td>HBA driver</td>
<td>8.01.07-d4</td>
</tr>
<tr>
<td>HBA firmware</td>
<td>4.00.150</td>
</tr>
</tbody>
</table>

**Table 10.4 - host nodes**

### 9.1.5 ORION version

ORION Version 11.1.0.0.0 for linux_x86-64 was used in the testing.
10 Testing Methodology

This section describes the methodology used for testing performance. It will note which metrics will be gathered from both ORION and TPC and how they will be correlated. It will also list the matrix of possible variations with respect to the following:

(1) The parameters used for each ORION test run.

(2) The storage configuration used for each run in (1).

The permutations of (1) and (2) that are actually used for the testing will be noted in the introduction to the exercises.

10.1 ORION test run descriptions

The following describes the individual ORION test runs that will be considered and the workload that the runs are simulating. The -advanced parameter will always be used so that all of the parameters which are used in a run are explicitly stated:

10.1.1 OLTP only

This will use 8 KB block sizes and varying percentage combinations of reads and writes. The combinations tested will be:

- 70% reads and 30% writes
- 50% reads and 50% writes
- 30% reads and 70% writes

The combination of 70% reads and 30% writes is often cited as fairly representative of a typical OLTP workload. But the particular profile for any running database must be determined by running the SQL given in Appendix A. The parameters used with ORION for RAID-10 with 70% reads and 30% writes would be:

```
# ./orion -run advanced -testname oltp_raid10_read70_write30 \
-num_disks 4 -size_small 8 -cache_size 8000 \
-num_large 0 -write 30 -matrix row -duration 60 --verbose
```

The –cache_size parameter is set to 8000 to correspond to the 8 GB cache of the SVC.

10.1.2 Data Warehouse only (large sequential reads and writes)

This will use 1 MB block sizes and varying percentage combinations of reads and writes. The combinations tested will be:

- 100% reads
- 70% reads and 30% writes
- 50% reads and 50% writes
- 30% reads and 70% writes
• 100% writes

The parameters used with ORION for RAID-5 with 100% reads would be:

```
# ./orion -run advanced -testname seq_raid5_read100_write0 \
    -num_disks 4 -type seq -size_large 1024 -simulate raid0 \ 
    -stripe 1024 -write 0 -num_small 0 -matrix col -duration 60 \ 
    -verbose
```

10.1.3 Mixed OLTP and Data Warehouse

This will simulate a mixed OLTP and Data Warehouse load. The block size for the small, random I/O will be 8 KB and the block size for the large, sequential I/O will be 1 MB. The combinations tested will be:

• 70% reads and 30% writes
• 50% reads and 50% writes

The parameters used with ORION for RAID-10 with 70% reads and 30% writes would be:

```
# ./orion -run advanced -testname mixed_raid10_read70_write30 \
    -num_disks 4 -size_small 8 -size_large 1024 -type rand \ 
    -simulate raid0 -stripe 1024 -write 30 -matrix detailed \ 
    -duration 60 -verbose
```

As noted in the ORION User’s Guide, in a mixed environment with both small-block, random I/Os and large-block, sequential I/Os, the large-block sequential I/Os actually get interpreted by the storage as large-block, random I/Os. That is the reason that the `-type rand` parameter will be used in this test as opposed to the `-type seq` parameter that was used in the Data Warehouse only test.

10.2 Storage configuration variations for ORION test runs

Below are the storage configurations which will be considered for the SVC, DS8000 and ASM.

10.2.1 DS8000 Configurations

The paper “IBM DS8000 Performance Configuration Guidelines for Implementing Oracle Databases with Automatic Storage Management” (item 5 in the References) discussed in detail the effects of various DS8000 LUN configurations on overall Oracle I/O performance. In summary, those storage configurations were as follows:

1. Four 100 GB LUNs in one extent pool which consists of one rank. Therefore all LUNs are actually on only one RAID array.
2. Four 100 GB LUNs in four separate extent pools that each consist of one rank. This puts each LUN on a separate RAID array and striping across arrays is done at the ASM level.
3. Four 100 GB LUNs from one multi-rank extent pool that is using storage pool striping. This automatically stripes LUNs from that extent pool across back-end arrays.

The paper showed that configurations (2) and (3) were vastly superior to configuration (1). In addition, storage configuration (2) was superior in performance to storage configuration (3), except for pure OLTP workloads where the performance was similar. The conclusion was then drawn that in an Oracle ASM
environment, with pure OLTP workloads, storage configuration (3) may be preferable for administrative reasons - if space needs to get added to an ASM diskgroup, a LUN could simply be created from the multi-rank extent pool in storage configuration (3) and it would automatically be using multiple RAID arrays. Storage configuration (2) would require that one additional LUN be created from each of the four arrays and then all four LUNs be added as ASM disks so that the I/O load would continue to be balanced across arrays and no array hot spots would be generated.

However, with the use of SVC, VDisk striping across back-end storage RAID arrays can be accomplished by creating MDisk Groups which consist of multiple MDisks where each MDisk is mapped to a different back-end storage RAID array. Thus storage configuration (3), storage pool striping, becomes unnecessary for all workloads and it will therefore not be used for any of the benchmark runs in this paper. Storage configuration (2) will be used for all benchmark runs, although the LUN sizes will be such that the one LUN which is created on each back-end DS8000 RAID array will take up all of the space on that array. Storage configuration (2) with one LUN using all of the space on the extent pool is, by the way, the SVC best practice documented in section 6.1.4 MDisk Configuration, item (1).

In summary, these are the DS8000 configurations that will be used:

1. The DS8000 will first be configured to connect directly to the front-end host(s). Four 100 GB LUNs will be created on four separate single-rank extent pools and be presented to the front-end host(s). The LUNs must alternate between even-numbered and odd-numbered extent pools so that both DS8000 processor complexes (and also alternating Device Adapters) are being used. The ORION test runs for this configuration will show the performance numbers when an SVC is not being used.

2. The DS8000 will then be configured to connect to the SVC and will present to the SVC four LUNs from the same four single-rank extent pools (and RAID arrays) used in step (1). Each LUN will consume all of the space on a given extent pool. These LUNs will be visible as MDisks to the SVC. VDisks will then be created from these MDisks and presented to front-end hosts as LUNs. The ORION test runs for this configuration will show the performance effect of using the in-band SVC with a back-end DS8000.

3. All test runs in (1) and (2) will be done using both RAID-5 and RAID-10.

### 10.2.2 SAN Volume Controller Configurations

The basic setup of the SVC will follow all of the guidelines given in 6.1 Summary of SAN Volume Controller Configuration Best Practices. This includes everything from the use of a dual-fabric SAN topology, to the recommendations for zoning aliases and zones, to the proper configuration of MDisks, MDisk Groups, VDisks and Hosts. There would be little purpose in writing this paper if the SVC configuration guidelines documented herein were not followed to the letter. Therefore, there is strict adherence to all SVC configuration best practices. In fact, all of the commands listed in the section 6.2 SAN Volume Controller Logical Configuration were created via “copy-paste” from the DS8000 and SVC command-line tools when they were run in the lab.

The following are specific details regarding some of the SVC configuration variations:
(1) Benchmarks will be run using an MDisk Group extent size of 1 GB. As was noted in item (5) of section 6.1.5 MDisk Group (MDG) Configuration, extent sizes of 1 GB showed a slight edge in performance for sequential workloads over 256 MB extent sizes.

(2) Each DS8000 LUN presented to the SVC will be from a single-rank extent pool and will consume all of the space on that extent pool. The four extent pools will alternate between even-numbered and odd-numbered and will thus be evenly split across both server complexes in the DS8000 (two extent pools per processor complex).

(3) VDisk sizes will be 100 GB for all test runs, with one exception. The exception is the test where no ASM striping will be used. In that test, one 400 GB VDisk will be presented to ORION. This is discussed in item (1) of the next section, 10.2.3 ASM Configurations.

(4) Benchmarks will be run where VDisks are using only one I/O group and then again where VDisks are using two I/O groups.

### 10.2.3 ASM Configurations

Section 11.3 Comparison of ASM and SVC striping configurations discusses the effect of ASM striping in conjunction with SVC striping. Benchmarks in that section will be run with the following three configurations:

(1) Striping only at the SVC level. This will simulate an ASM configuration where there is only one ASM disk in the diskgroup by running ORION on only one 400 GB VDisk which is striped across all four MDisks in the MDisk Group. As discussed in section 4 Oracle Automatic Storage Management (ASM), ASM automatically stripes across all ASM disks in a diskgroup. Therefore, presenting only one VDisk to ORION is effectively eliminating ASM striping.

(2) Striping only at the ASM level. This will use four 100 GB VDisks where each VDisk is using only one MDisk. Each MDisk of course maps to a separate back-end DS8300 RAID array.

(3) Striping at both the ASM and SVC levels. This is the default configuration in an ASM on SVC environment and is the recommended best practice, as will be discussed.

### 10.3 Metrics collected for each ORION test run

The following are the metrics that will be collected for each ORION run:

- **ORION metrics:**
  - IOPS
  - MBPS
  - Latency

As discussed in the excerpt from the ORION User’s Guide, IOPS and Latency will only be generated if random I/Os are part of the ORION run and MBPS will only be generated if large sequential I/Os are part of the ORION run.

- **TPC metrics** – the following will be collected at the RAID array level. If an array contains more than one LUN used as a target in the ORION test run, the sum of the statistics from all of the LUNs in a given array always equals the statistics gathered for that array as a whole:
  - IOPS:
10.4 Analysis of metrics collected for each ORION test run

There will be no way to correlate the data points from ORION exactly with the 5-minute interval average data collection points of TPC. Therefore, the comparison of the collected metrics will be done as follows:

- **Overall Summary Comparison** – There will be a comparison of the metrics for the entire run period. ORION provides a <testname>_summary.txt file that contains the peak metrics observed over the entire run time. This is measured at a discrete data point. The reports in TPC show average metrics per 5 minute sample period. So the peak average metric collected by TPC over the entire run time will be compared to the peak discrete data point metric for ORION.

  This is not a perfect basis for comparison, since it is comparing a discrete data point (ORION) to a maximum average over a 5 minute time period (TPC), but it is the best comparison available.

- **Comparison over a continuum during the same time period** – One way to compensate for the fact that a comparison cannot be done at exactly each data point for ORION and TPC is to graph the results and do a visual comparison over the entire run time period. However, since the sample data points for ORION do not map exactly to TPC, and since ORION collects data at more data points than TPC for the run period, a graph will only be included for ORION. As mentioned above, the TPC metrics are averaged over 5 minute intervals whereas the ORION metrics are gathered at discrete data points.

- **Comparison by absolute metric number, not by percentage difference** – The most meaningful comparison of the collected metrics at a given data point is to compare by the difference in absolute numbers. In other words, it is more helpful to be able to say for a given data point “RAID-10 offers 1500 more IOPS at this data point than RAID-5” as opposed to saying “RAID-10 offers 10% more IOPS than RAID-5 at this data point”. The same goes for the comparisons of Latency and MBPS. This approach can contribute in a more meaningful manner to being able to do a cost-benefit analysis when making a purchasing decision. You are buying “horsepower” in absolute terms, not by comparison of percentage differences.

The metrics gathered from ORION will be included in this document in graphical format. The TPC statistics will not be displayed. TPC was used to verify that the metrics reported by ORION were accurate. This was, in fact, the case. So while a data center may be collecting I/O performance data at the host level, it is still important that TPC be installed and running when deploying an SVC or a DS8000. Otherwise, it is extremely difficult to determine the root cause of any storage performance problems that may arise.
11 ORION Test Run Results

11.1 Comparison of SVC to non-SVC Oracle I/O performance

This section will compare the I/O performance effects of having an in-band SVC between the hosts and back-end DS8300 as opposed to having the hosts directly using the DS8300.

The testing was done by:

1. Natively connecting (i.e., without an SVC) four 100 GB DS8300 LUNs to the front-end host and running ORION using storage configuration (2) as discussed in Section 10.2.1 DS8000 Configurations. This was done with both RAID-10 and RAID-5 configurations on the back-end DS8300.

2. Those DS8300 LUNs were then removed and the exact same DS8300 arrays were used to create four new LUNs which took all of the space on the array, as per the best practices discussed in 6.1.4 MDisk Configuration. This means, of course, that those arrays were using the same RAID level that was used in step (1). These LUNs were then presented to the SVC as MDisks, as was discussed in Section 10.2.2 SAN Volume Controller Configurations.

3. An MDisk Group was then created which included all four of those MDisks.

4. Four 100 GB VDisks were then created by striping across all four MDisks as discussed in section 6.1.6 VDisk Configuration, item (2).

5. The tests were run on the VDisks in step (4) two times, once when the VDisks had been created using one iogrp and once when the VDisks had been created with two iogrps.

6. The workloads used were:
   - OLTP (Random) I/O with Read/Write ratios of 70%/30%, 50%/50% and 30%/70%.
   - Data Warehouse (Sequential) I/O with Read/Write ratios of 100%/0%, 70%/30%, 50%/50%, 30%/70% and 0%/100%.
   - Mixed OLTP and Data Warehouse I/O with Read/Write ratios of 70%/30% and 50%/50%.

11.1.1 Summary for comparison of SVC to non-SVC Oracle I/O performance

This subsection gives a summary of the ORION test results for OLTP, Data Warehouse (Sequential) and Mixed workloads.

11.1.1.1 OLTP Workload Summary

For the OLTP workloads:

- whether SVC used RAID-10 or RAID-5 on the back-end DS8300
- whether SVC used one or two iogrps
- for all Read/Write ratios that were tested
- at anything but the very lowest load levels

SVC had significantly better performance for IOPS than native DS8300. As the load level and the proportion of random writes increased, the SVC performance improvement became even more
significant. At the very lowest load levels for all workload types, the difference between SVC and DS8300 was fairly insignificant.

- For the OLTP Read/Write workloads of 70%/30% with both RAID-10 and RAID-5, the difference in Latency between SVC with 1 iogrp, SVC with 2 iogrp and the DS8300 was fairly insignificant. For 50%/50% and 30%/70% with both RAID-10 and RAID-5, the Latency was significantly better (i.e., lower latency) for SVC with 1 iogrp and SVC with 2 iogrp compared to DS8300. The overall tendency was for the SVC to increase its advantage in Latency numbers as the load level increased.

- For the OLTP workloads, the use of two iogrp on the SVC improved performance significantly over the use of one iogrp for both RAID-10 and RAID-5 on the back-end DS8300. Two iogrp also tended to “smooth out” some of the curves as the load level increased. At the Read/Write ratios of 50%/50% and 30%/70%, the curves for both one iogrp and two iogrp for both RAID-10 and RAID-5 on the back-end DS8300 tend to “bounce” up and down a bit as the load level increases. However, some of the largest performance advantages of the SVC were attained with those Read/Write ratios, anywhere from 250% - 400% at the highest load levels. The overall look of the upward curve should be irrelevant as long as the SVC performance is so much better than native DS8300 at each load level.

11.1.1.2 Data Warehouse (Sequential) Workload Summary

- For the Data Warehouse workloads:
  - with either RAID-10 or RAID-5 on the back-end DS8300
  - using either 1 or 2 iogrp on the SVC
  - at Read/Write ratios of 100%/0% and 0%/100%
  - when the results are compared over the full range of load levels

there was not a significant difference in throughput (MBPS) between native DS8300 and SVC. However, using RAID-5 on the back-end DS8300 tended to show the smallest difference in results between the DS8300 and SVC with either 1 or 2 iogrp.

- For the Data Warehouse workloads:
  - with either RAID-10 or RAID-5 on the back-end DS8300
  - using either 1 or 2 iogrp on the SVC
  - at Read/Write ratios of 70%/30%, 50%/50% and 30%/70%
  - at anything but the very lowest load levels

SVC showed a very significant MBPS advantage over native DS8300. At some load levels for the SVC with 2 iogrp, the advantage was close to a 100% improvement over DS8300. The SVC with 2 iogrp had significantly better performance than SVC with 1 iogrp at all but the lowest load levels.

11.1.1.3 Mixed OLTP and Data Warehouse (Sequential) Workload Summary

- Read/Write 70%/30%
  - RAID-10 on back-end DS8300
  - IOPS:
SVC with 1 iogrp is faster than the DS8300 at low load levels but significantly slower at high load levels.
SVC with 2 iogrps is significantly faster than DS8300 at all but a few load levels.
SVC with 1 iogrp and SVC with 2 iogrps are close at low load levels but SVC with 2 iogrps is significantly better at all other load levels.

- **Latency**
  
  SVC with 1 iogrp has better numbers than the DS8300 at low load levels but is slower at high load levels.
  SVC with 2 iogrps has better numbers than the DS8300 except for at a couple of load levels.
  SVC with 2 iogrps is significantly better than SVC with 1 iogrp except at a couple of low load levels where they are close.

- **MBPS**
  
  There is no significant difference between the DS8300, SVC with 1 iogrp and SVC with 2 iogrps at any load level.

- **RAID-5 on back-end DS8300**
  
  - **IOPS:**
    
    SVC with 1 iogrp is significantly faster than the DS8300 at low load levels but significantly slower at high load levels.
    SVC with 2 iogrps is significantly faster than DS8300 at all but a very few load levels.
    SVC with 1 iogrp is significantly faster than SVC with 2 iogrps at low load levels but SVC with 2 iogrps is significantly better at higher load levels.
  
  - **Latency**
    
    The DS8300, SVC with 1 iogrp and SVC with 2 iogrps all have similar numbers although the SVC with 2 iogrps has a spike in Latency at the lowest load levels.
  
  - **MBPS**
    
    There is no significant difference between the DS8300 and SVC with 1 iogrp. SVC with 2 iogrps is significantly better than the DS8300 at the middle range of load levels but dips just a bit below the DS8300 at the highest load levels.

- **Read/Write 50%/50%**
  
  - **RAID-10 on back-end DS8300**
    
    - **IOPS:**
      
      The DS8300 is significantly faster at anything but the very lowest load levels. SVC with 1 iogrp and SVC with 2 iogrps are close to each other at all load levels.
    
    - **Latency**
      
      The DS8300 has Latency around 2 ms/op at all load levels while the SVC with both 1 and 2 iogrps has Latency in the 4 – 5 ms/op range at all load levels.
    
    - **MBPS**
There is no significant difference between the DS8300, SVC with 1 iogrp and SVC with 2 iogrps. They all have a range of 250 – 400 MBPS over the full test run and are very close to each other at each data point.

- **RAID-5 on back-end DS8300**
  
  - **IOPS**: SVC with 1 iogrp and the DS8300 have similar numbers at all load levels. SVC with 2 iogrps is significantly faster than the DS8300 at some of the highest load levels but they are close at the lower load levels.
  
  - **Latency**: The DS8300, SVC with 1 iogrp and SVC with 2 iogrps all have similar numbers.
  
  - **MBPS**: Both SVC with 1 iogrp and SVC with 2 iogrps performed better than the DS8300 at all but the lowest load levels. At the highest load levels, they both performed significantly better than the DS8300.

**11.1.2 Detailed Analysis for OLTP Workloads**

This subsection contains the detailed analysis and graphs which compare performance between the DS8300, SVC with 1 iogrp, and SVC with 2 iogrps for OLTP workloads.

**11.1.2.1 OLTP RAID-10 70% Reads and 30% Writes**

<table>
<thead>
<tr>
<th></th>
<th>IOPS</th>
<th>Latency (ms/op)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS8300</td>
<td>378 – 17,235</td>
<td>0.87 – 2.64</td>
</tr>
<tr>
<td>SVC with 1 iogrp</td>
<td>253 – 19,807</td>
<td>0.75 – 3.94</td>
</tr>
<tr>
<td>SVC with 2 iogrps</td>
<td>245 – 28,798</td>
<td>0.75 – 4.07</td>
</tr>
</tbody>
</table>

- SVC with 1 iogrp offers an improvement at higher load levels of around 2,500 to 3,000 IOPS over native DS8300 with the largest improvement being 3,187 IOPS at one data point.
- SVC with 2 iogrps offers an improvement at higher load levels of around 1,500 to 11,000 IOPS over native DS8300 with the largest improvement being 11,563 IOPS at one data point.
- SVC with 2 iogrps offers an improvement at higher load levels of around 1,200 to almost 9,000 IOPS over SVC with 1 iogrp, with the largest improvement being 8,991 IOPS at one data point.
- Except at the very lowest load level, the differences in Latency between the DS8300 and SVC with either 1 or 2 iogrps was insignificant. As can be seen in the graph, at the very lowest load levels the SVC configurations had slightly higher Latency.
11.1.2.2 OLTP RAID-5 70% Reads and 30% Writes

<table>
<thead>
<tr>
<th></th>
<th>IOPS</th>
<th>Latency (ms/op)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS8300</td>
<td>370 – 11,746</td>
<td>0.99 – 2.7</td>
</tr>
<tr>
<td>SVC with 1 iogr</td>
<td>261 – 18,895</td>
<td>0.93 – 3.83</td>
</tr>
</tbody>
</table>
• SVC with 1 iogrp offers an improvement at higher load levels of around 1,100 to 8,000 IOPS over native DS8300 with the largest improvement being 8,213 IOPS at one data point.

• SVC with 2 iogrps offers an improvement at higher load levels of around 1,200 to 15,000 IOPS over native DS8300 with the largest improvement being 15,349 IOPS at one data point.

• SVC with 2 iogrps offers an improvement at higher load levels of around 1,600 to 9,800 IOPS over SVC with 1 iogrp, with the largest improvement being 9,877 IOPS at one data point.

• Except at the very lowest load level, the difference in Latency between the DS8300 and SVC with either 1 or 2 iogrps was usually less than 1 ms/op and thus insignificant. At the lowest load level the SVC had slightly higher Latency, as can be seen in the graph.

Orion OLTP IOPS RAID-5 - 70% Reads and 30% Writes

### SVC Performance Configuration Guidelines for Implementing Oracle Databases with ASM

http://www.ibm.com/support/techdocs

© Copyright IBM Corporation, 2009
11.1.2.3 OLTP RAID-10 50% Reads and 50% Writes

<table>
<thead>
<tr>
<th></th>
<th>IOPS</th>
<th>Latency (ms/op)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS8300</td>
<td>459 – 10,226</td>
<td>0.91 – 2.18</td>
</tr>
<tr>
<td>SVC with 1 iogrp</td>
<td>380 – 20,498</td>
<td>0.74 – 2.63</td>
</tr>
<tr>
<td>SVC with 2 iogrps</td>
<td>416 – 25,398</td>
<td>0.66 – 2.4</td>
</tr>
</tbody>
</table>

- SVC with 1 iogrp offers an improvement at higher load levels of around 1,200 to 10,000 IOPS over native DS8300 with the largest improvement being 10,743 IOPS at one data point.
- SVC with 2 iogrps offers an improvement at higher load levels of around 2,000 to 15,000 IOPS over native DS8300 with the largest improvement being 15,643 IOPS at one data point.
- SVC with 2 iogrps offers an improvement at higher load levels of around 200 to 13,000 IOPS over SVC with 1 iogrp, with the largest improvement being 13,230 IOPS at one data point.
- The differences in Latency between the DS8300 and SVC with either 1 or 2 iogrps was insignificant, although at the highest load level the DS8300 was 1 ms/op slower than both SVC configurations.
Orion OLTP IOPS RAID-10 - 50% Reads and 50% Writes

Orion OLTP Latency RAID-10 - 50% Reads and 50% Writes

11.1.2.4 OLTP RAID-5 50% Reads and 50% Writes

<table>
<thead>
<tr>
<th></th>
<th>IOPS</th>
<th>Latency (ms/op)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS8300</td>
<td>448 – 7,254</td>
<td>1.05 – 3.1</td>
</tr>
<tr>
<td>SVC with 1 iogrp</td>
<td>362 – 18,207</td>
<td>0.91 – 2.76</td>
</tr>
</tbody>
</table>
- SVC with 1 iogrp offers an improvement at higher load levels of around 1,000 to 11,000 IOPS over native DS8300 with the largest improvement being 11,749 IOPS at one data point.
- SVC with 2 iogrps offers an improvement at higher load levels of around 1,500 to 13,000 IOPS over native DS8300 with the largest improvement being 13,764 IOPS at one data point.
- SVC with 2 iogrps offers an improvement at higher load levels of around 1,000 to 9,000 IOPS over SVC with 1 iogrp, with the largest improvement being 9,893 IOPS at one data point. There was one data point at a higher load level where the SVC with 2 iogrps was essentially equivalent to SVC with 1 iogrp, as can be seen in the graph.
- The difference in Latency between the DS8300 and SVC with either 1 or 2 iogrps was usually around 1 ms/op at the higher load levels and reached 2 ms/op at the highest load level.
- The difference in Latency between SVC with 2 iogrps and SVC with 1 iogrp was insignificant, although as can be seen in the graph there was one data point where SVC with 2 iogrps was 1 ms/op faster.
11.1.2.5 OLTP RAID-10 30% Reads and 70% Writes

<table>
<thead>
<tr>
<th></th>
<th>IOPS</th>
<th>Latency (ms/op)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS8300</td>
<td>762 – 7,220</td>
<td>0.84 – 2.9</td>
</tr>
<tr>
<td>SVC with 1 iogrp</td>
<td>746 – 15,605</td>
<td>0.66 – 2.22</td>
</tr>
<tr>
<td>SVC with 2 iogrps</td>
<td>686 – 19,846</td>
<td>0.52 – 1.87</td>
</tr>
</tbody>
</table>

- SVC with 1 iogrp offers an improvement at higher load levels of around 200 to 8,000 IOPS over native DS8300 with the largest improvement being 8,714 IOPS at one data point.
- SVC with 2 iogrp offers an improvement at higher load levels of around 1,500 to 12,000 IOPS over native DS8300 with the largest improvement being 12,970 IOPS at one data point.
- SVC with 2 iogrps is usually faster than SVC with 1 iogrp except for at six data points. Excluding those six data points, SVC with 2 iogrps offers an improvement of around 2,000 to 10,000 IOPS over SVC with 1 iogrp with the largest improvement being 10,301 at one data point. The performance advantage of SVC with 1 iogrp over SVC with 2 iogrps at those six data points was anywhere from 60 to 6,677 IOPS.
- The difference in Latency between the DS8300 and SVC with either 1 or 2 iogrps was mostly insignificant, although at the two highest load levels the DS8300 was almost 2 ms/op slower than the two SVC configurations.
11.1.2.6  OLTP RAID-5 30% Reads and 70% Writes

<table>
<thead>
<tr>
<th></th>
<th>IOPS</th>
<th>Latency (ms/op)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS8300</td>
<td>706 – 5,354</td>
<td>0.97 – 4.28</td>
</tr>
<tr>
<td>SVC with 1 iogrp</td>
<td>797 – 14,378</td>
<td>0.62 – 2.87</td>
</tr>
</tbody>
</table>
• SVC with 1 iogrp offers an improvement at higher load levels of around 1,100 to 9,000 IOPS over native DS8300 with the largest improvement being 9,713 IOPS at one data point.

• SVC with 2 iogrps offers an improvement at higher load levels of around 1,200 to 14,000 IOPS over native DS8300 with the largest improvement being 14,018 IOPS at one data point.

• SVC with 2 iogrps is faster than SVC with 1 iogrp except for at one data point. Excluding that one data point, SVC with 2 iogrps offers an improvement of around 200 to 8,400 IOPS over SVC with 1 iogrp with the largest improvement being 8,443 IOPS at one data point. The performance advantage of SVC with 1 iogrp over SVC with 2 iogrps at the one data point was 8,415 IOPS.

• At the lower load levels the difference in Latency between the DS8300 and SVC with either 1 or 2 iogrps was insignificant. At the higher load levels, the SVC with 1 iogrp and SVC with 2 iogrps had a significant advantage in Latency over the DS8300 where it was sometimes 2.5 to 3 ms/op.

• The difference in Latency between SVC with 2 iogrps and SVC with 1 iogrp was insignificant except at a couple of data points. At those data points, the difference was a bit less than 2 ms/op. At one of the data points, SVC with 1 iogrp had the lower (better) Latency.
11.1.3 Detailed Analysis for Data Warehouse (Sequential) Workloads

This section contains the detailed analysis and graphs which compare performance between the DS8300, SVC with 1 iogrp and SVC with 2 iogrps for Data Warehouse (Sequential) workloads.

### 11.1.3.1 Data Warehouse (Sequential) RAID-5 100% Reads and 0% Writes

<table>
<thead>
<tr>
<th></th>
<th>MBPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS8300</td>
<td>437.65 – 667.47</td>
</tr>
<tr>
<td>SVC with 1 iogrp</td>
<td>386.42 – 599.92</td>
</tr>
<tr>
<td>SVC with 2 iogrps</td>
<td>383.45 – 556.4</td>
</tr>
</tbody>
</table>

- Native DS8300 offered an improvement over SVC with 1 iogrp at five load levels. The improvement in general ranged from 30 – 70 MBPS with the largest improvement being 152.62 MBPS at one data point. The SVC with 1 iogrp showed an improvement over the native DS8300 at three data points with the advantage ranging from 45 to 85 MBPS.

- Native DS8300 offered an improvement over SVC with 2 iogrps at five load levels. The improvement ranged from 60 – 155 MBPS with the largest improvement being 155.59 MBPS at one data point. The SVC with 2 iogrps showed an improvement over the native DS8300 at three load levels with the advantage ranging from 31 to 86 MBPS.
The SVC with 1 iogrp always outperformed the SVC with 2 iogrps with the improvement being in the range of 13 – 75 MBPS. The reason that the SVC with 1 iogrp configuration is outperforming the SVC with 2 iogrps configuration for this workload mix is perhaps related to the fact that this controlled experiment kept the number of MDisks, storage controllers and ports constant. With the addition of two nodes, SVC may be expecting the number of MDisks, storage controllers and ports to increase. Therefore the queue depth per MDisk will have been reduced by adding more nodes.

### Orion Sequential MBPS RAID-5 - 100% Reads and 0% Writes

**Graph:**

- **DS8000**
- **SVC with 1 iogrp**
- **SVC with 2 iogrps**

### 11.1.3.2 Data Warehouse (Sequential) RAID-10 100% Reads and 0% Writes

<table>
<thead>
<tr>
<th></th>
<th>MBPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS8300</td>
<td>413.34 – 573.38</td>
</tr>
<tr>
<td>SVC with 1 iogrp</td>
<td>395.9 – 590.35</td>
</tr>
<tr>
<td>SVC with 2 iogrps</td>
<td>378.53 – 482.16</td>
</tr>
</tbody>
</table>

- SVC with 1 iogrp offered an improvement over native DS8300 at four load levels. The improvement ranged from 24 – 177 MBPS. The DS8300 showed an improvement over the SVC with 1 iogrp at four other load levels with the advantage ranging from 8 to 136 MBPS.
- Native DS8300 outperformed SVC with 2 iogrps except at seven data points. The improvement ranged from 18 to 153 MBPS. The SVC with 2 iogrps was 48 MBPS faster than the DS8300 at the one exceptional data point.
The SVC with 1 iogrp always outperformed the SVC with 2 iogrps with the improvement being in the range of 17 - 128 MBPS. The reason that the SVC with 1 iogrp configuration is outperforming the SVC with 2 iogrps configuration for this workload mix is perhaps related to the fact that this controlled experiment kept the number of MDisks, storage controllers and ports constant. With the addition of two nodes, SVC may be expecting the number of MDisks, storage controllers and ports to increase. Therefore the queue depth per MDisk will have been reduced by adding more nodes.

### Orion Sequential MBPS RAID-10 - 100% Reads and 0% Writes

<table>
<thead>
<tr>
<th>Outstanding Async I/O's</th>
<th>DS8000</th>
<th>SVC 1 iogrp</th>
<th>SVC 2 iogrps</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>300</td>
<td>500</td>
<td>600</td>
</tr>
<tr>
<td>1</td>
<td>400</td>
<td>550</td>
<td>700</td>
</tr>
<tr>
<td>2</td>
<td>500</td>
<td>650</td>
<td>800</td>
</tr>
<tr>
<td>3</td>
<td>600</td>
<td>800</td>
<td>950</td>
</tr>
<tr>
<td>4</td>
<td>700</td>
<td>950</td>
<td>1100</td>
</tr>
</tbody>
</table>

#### 11.1.3.3 Data Warehouse (Sequential) RAID-5 70% Reads and 30% Writes

<table>
<thead>
<tr>
<th></th>
<th>MBPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS8300</td>
<td>416.56 – 611.23</td>
</tr>
<tr>
<td>SVC with 1 iogrp</td>
<td>365.52 – 682.17</td>
</tr>
<tr>
<td>SVC with 2 iogrps</td>
<td>389.82 – 791.6</td>
</tr>
</tbody>
</table>

- SVC with 1 iogrp offered an improvement over native DS8300 except at the two lowest load levels. The improvement ranged from 49.11 - 255.17 MBPS. The two data points where the DS8300 showed an advantage over SVC with 1 iogrp were 27.05 and 179.44 MBPS.
- SVC with 2 iogrps offered an improvement over native DS8300 except at the two lowest load levels. The improvement ranged from 86.12 – 333.53 MBPS. The two data points where the DS8300 showed an advantage over SVC with 2 iogrps were 30.79 and 155.14 MBPS.
- The SVC with 2 iogrps outperformed the SVC with 1 iogrp except at two load levels where the difference was insignificant. The improvement was in the range of 24.3 – 145.05 MBPS.
Orion Sequential MBPS RAID-5 - 70% Reads and 30% Writes

11.1.3.4 Data Warehouse (Sequential) RAID-10 70% Reads and 30% Writes

<table>
<thead>
<tr>
<th></th>
<th>MBPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS8300</td>
<td>405.92 – 620.63</td>
</tr>
<tr>
<td>SVC with 1 iogrp</td>
<td>395.03 – 541.2</td>
</tr>
<tr>
<td>SVC with 2 iogrps</td>
<td>379.5 – 664.98</td>
</tr>
</tbody>
</table>

- SVC with 1 iogrp offered an improvement half of the time over native DS8300 with the improvement ranging from 5.76 to 114.78 MBPS. The other half of the time the DS8300 offered an improvement over SVC with 1 iogrp ranging from 10.89 to 193.69 MBPS.
- SVC with 2 iogrps offered an improvement over native DS8300 except at three load levels. The improvement ranged from 25.53 – 222.69 MBPS. The three data points where DS8300 outperformed SVC with 2 iogrps ranged from 26.42 to 141.69 MBPS.
- The SVC with 2 iogrps outperformed the SVC with 1 iogrp except at the three lowest load levels. The improvement was in the range of 107.91 – 167.29 MBPS. The SVC with 1 iogrp improvement over SVC with 2 iogrps ranged from 15.53 to 60.68 MBPS.
### Orion Sequential MBPS RAID-10 - 70% Reads and 30% Writes

![Graph](image)

### 11.1.3.5 Data Warehouse (Sequential) RAID-5 50% Reads and 50% Writes

<table>
<thead>
<tr>
<th></th>
<th>MBPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS8300</td>
<td>386.8 – 540.6</td>
</tr>
<tr>
<td>SVC with 1 iogrp</td>
<td>384.69 – 622.15</td>
</tr>
<tr>
<td>SVC with 2 iogrps</td>
<td>400.37 – 730.34</td>
</tr>
</tbody>
</table>

- SVC with 1 iogrp offered an improvement over native DS8300 except at the lowest load level. The improvement ranged from 83.13 – 235.35 MBPS. The DS8300 showed an advantage of 155.91 MBPS at the one exceptional load level.
- SVC with 2 iogrps offered an improvement over native DS8300 except at the lowest load level. The improvement ranged from 180.96 – 332.11 MBPS. The DS8300 showed an advantage of 140.23 MBPS at the lowest load level.
- The SVC with 2 iogrps always outperformed the SVC with 1 iogrp with the improvement in the range of 15.68 – 106.33 MBPS.
11.1.3.6 Data Warehouse (Sequential) RAID-10 50% Reads and 50% Writes

<table>
<thead>
<tr>
<th></th>
<th>MBPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS8300</td>
<td>383.63 – 514.97</td>
</tr>
<tr>
<td>SVC with 1 iogrp</td>
<td>393.36 – 542.08</td>
</tr>
<tr>
<td>SVC with 2 iogrps</td>
<td>381.45 – 615.28</td>
</tr>
</tbody>
</table>

- SVC with 1 iogrp offered an improvement over native DS8300 at four load levels. The improvement ranged from 8.74 – 154.91 MBPS. The DS8300 showed a significant advantage at only one data point where it was 93.91 MBPS.

- SVC with 2 iogrps offered an improvement over native DS8300 except at the lowest load level. The improvement ranged from 44.83 – 226.33 MBPS. The DS8300 showed an advantage of 107.18 MBPS at the lowest load level.

- The SVC with 2 iogrps outperformed the SVC with 1 iogrp, except at the lowest load level, with the improvement in the range of 71.42 – 131.02 MBPS. The lowest load level showed an insignificant difference between the two.
Orion Sequential MBPS RAID-10 - 50% Reads and 50% Writes

11.1.3.7 Data Warehouse (Sequential) RAID-5 30% Reads and 70% Writes

<table>
<thead>
<tr>
<th></th>
<th>MBPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS8300</td>
<td>356.49 – 591.64</td>
</tr>
<tr>
<td>SVC with 1 iogr</td>
<td>392.03 – 598.43</td>
</tr>
<tr>
<td>SVC with 2 iogrps</td>
<td>386.21 – 706.73</td>
</tr>
</tbody>
</table>

- SVC with 1 iogr offered an improvement over native DS8300 except at the lowest load level. The improvement ranged from 64.32 – 241.94 MBPS. The DS8300 showed an advantage of 199.61 MBPS at the lowest load level.
- SVC with 2 iogrps offered an improvement over native DS8300 except at the lowest load level. The improvement ranged from 169.07 – 344.14 MBPS. The DS8300 showed an advantage of 205.43 MBPS at the lowest load level.
- The SVC with 2 iogrps outperformed the SVC with 1 iogr, except at the lowest load level, with the improvement in the range of 65.91 – 112.61 MBPS. The lowest load level showed an insignificant difference between the two.
### Orion Sequential MBPS RAID-5 - 30% Reads and 70% Writes

![Chart showing MBPS vs Outstanding Async I/O's for Orion Sequential RAID-5](chart.png)

#### 11.1.3.8 Data Warehouse (Sequential) RAID-10 30% Reads and 70% Writes

<table>
<thead>
<tr>
<th></th>
<th>MBPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS8300</td>
<td>347.41 – 547.08</td>
</tr>
<tr>
<td>SVC with 1 iogrp</td>
<td>378.57 – 514.47</td>
</tr>
<tr>
<td>SVC with 2 iogrps</td>
<td>382.94 – 601.3</td>
</tr>
</tbody>
</table>

- SVC with 1 iogrp offered an improvement over native DS8300 at four load levels. The improvement ranged from 44.66 – 112.23 MBPS. The DS8300 showed an advantage at four load levels that ranged from 24.75 to 165.04 MBPS.
- SVC with 2 iogrps offered an improvement over native DS8300 at five load levels. The improvement ranged from 103.28 – 242.34 MBPS. The DS8300 showed an advantage at three load levels that ranged from 18.17 to 160.67 MBPS.
- The SVC with 2 iogrps always outperformed the SVC with 1 iogrp with the improvement in the range of 4.37 to 130.11 MBPS.
Orion Sequential MBPS RAID-10 - 30% Reads and 70% Writes

11.1.3.9 Data Warehouse (Sequential) RAID-5 0% Reads and 100% Writes

<table>
<thead>
<tr>
<th></th>
<th>MBPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS8300</td>
<td>285.34 – 446.14</td>
</tr>
<tr>
<td>SVC with 1 iogrp</td>
<td>403.94 – 429.6</td>
</tr>
<tr>
<td>SVC with 2 iogrps</td>
<td>373.61 – 414.33</td>
</tr>
</tbody>
</table>

- SVC with 1 iogrp offered an improvement over native DS8300 except at two load levels where the difference was insignificant. The improvement ranged from 10.95 – 129.44 MBPS.
- SVC with 2 iogrps offered an improvement over native DS8300 except at three load levels where the difference was insignificant. The improvement ranged from 3.75 – 121.49 MBPS.
- The SVC with 1 iogrp always outperformed the SVC with 2 iogrps. But the difference was never more than 30 MBPS and therefore was insignificant at each load level.
11.1.3.10 Data Warehouse (Sequential) RAID-10 0% Reads and 100% Writes

<table>
<thead>
<tr>
<th></th>
<th>MBPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS8300</td>
<td>289.38 – 372.72</td>
</tr>
<tr>
<td>SVC with 1 iogrp</td>
<td>253.05 – 327.43</td>
</tr>
<tr>
<td>SVC with 2 iogrps</td>
<td>281.42 – 333.04</td>
</tr>
</tbody>
</table>

- Native DS8300 offered an improvement over SVC with 1 iogrp at six load levels. The improvement ranged from 2.16 to 98.78 MBPS. The SVC with 1 iogrp outperformed the DS8300 at two load levels where the differences were 10.57 and 31.58 MBPS. In general, there was not a great amount of difference between the DS8300 and SVC with 1 iogrp.

- Native DS8300 offered an improvement over SVC with 2 iogrps at six load levels. The improvement ranged from 5.5 to 70.41 MBPS. The SVC with 2 iogrps outperformed the DS8300 at two load levels where the differences were 16.18 and 36.37 MBPS. In general, there was not a great amount of difference between the DS8300 and SVC with 2 iogrps.

- The differences between SVC with 1 iogrp and SVC with 2 iogrps were insignificant at each load level.
11.1.4 Detailed Analysis for Mixed OLTP and Data Warehouse (Sequential) Workloads

This section contains the detailed analysis and graphs which compare performance between the DS8300, SVC with 1 iogrp and SVC with 2 iogrps for Mixed OLTP and Data Warehouse (Sequential) workloads.

11.1.4.1 Mixed OLTP and Data Warehouse (Sequential) RAID-10 70% Reads and 30% Writes at Sequential Data Point 5

<table>
<thead>
<tr>
<th></th>
<th>IOPS</th>
<th>Latency (ms/op)</th>
<th>MBPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS8300</td>
<td>479 - 8,521</td>
<td>1.87 – 2.49</td>
<td>402.74 – 484.21</td>
</tr>
<tr>
<td>SVC with 1 iogrp</td>
<td>415 – 5,945</td>
<td>0.88 – 4.19</td>
<td>400.43 – 453.02</td>
</tr>
<tr>
<td>SVC with 2 iogrps</td>
<td>1,134 - 11,968</td>
<td>0.9 – 3.64</td>
<td>340.85 – 518.41</td>
</tr>
</tbody>
</table>

- SVC with 1 iogrp IOPS are significantly better than DS8300 at low load levels, reaching 3,041 at one data point. At higher load levels, SVC with 1 iogrp IOPS are significantly lower than DS8300 and was 3,715 lower at one point.

- SVC with 2 iogrps IOPS are significantly better than DS8300 at everything but two load levels. The SVC with 2 iogrps advantage over the DS8300 ranged from 655 to 4,184 IOPS. At the two exceptional data points the advantage of the DS8300 over SVC with 2 iogrps was 3,024 and 4,137 IOPS.
- SVC with 1 iogrp Latency is slightly lower than DS8300 at the lowest load levels. At higher load levels, SVC with 1 iogrp tends to have Latency a bit less than 2 ms/op more than DS8300.
- SVC with 2 iogrps Latency is usually 2.5 ms/op lower than DS8300 except at two load levels where it spiked a bit over the DS8300.
- DS8300, SVC with 1 iogrp and SVC with 2 iogrps show very close MBPS at all load levels, although the SVC with 2 iogrps has the slightly highest numbers at almost all load levels.
- SVC with 2 iogrps outperforms SVC with 1 iogrp at almost all load levels for all three metrics, and significantly so for IOPS and Latency.

**Orion Mixed OLTP and Data Warehouse RAID-10 - 70% Reads and 30% Writes IOPS at Sequential Data Point 5**

**Mixed OLTP and Data Warehouse RAID-10 - 70% Reads and 30% Writes Latency at Sequential Data Point 5**
### 11.1.4.2 Mixed OLTP and Data Warehouse (Sequential) RAID-5 70% Reads and 30% Writes at Sequential Data Point 5

<table>
<thead>
<tr>
<th></th>
<th>IOPS</th>
<th>Latency (ms/op)</th>
<th>MBPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS8300</td>
<td>560 – 8,331</td>
<td>1.71 – 2.49</td>
<td>405.72 – 485.11</td>
</tr>
<tr>
<td>SVC with 1 iogrp</td>
<td>449 – 8,662</td>
<td>0.91 – 2.82</td>
<td>395.78 – 528.31</td>
</tr>
<tr>
<td>SVC with 2 iogrps</td>
<td>138 – 10,403</td>
<td>1.85 – 7.2</td>
<td>274.74 – 565.83</td>
</tr>
</tbody>
</table>

- SVC with 1 iogrp IOPS are always significantly better than DS8300 except at the highest load levels where the difference is insignificant. The performance advantage of SVC with 1 iogrp ranged from 224 to 3,596 IOPS.
- SVC with 2 iogrps IOPS are very close to DS8300 at the lower load levels but are significantly better at the highest load levels, where the advantage ranged from 32 to 2,317 IOPS.
- SVC with 1 iogrp Latency is slightly better than DS8300 at the lowest load levels. At higher load levels the difference is insignificant.
- SVC with 2 iogrps Latency is very close to the DS8300 except at the lowest load levels where it spikes a bit over the DS8300. The lowest load level has a 7.2 ms/op Latency for the SVC with 2 iogrps.
- DS8300, SVC with 1 iogrp and SVC with 2 iogrps show very close MBPS at all load levels although the SVC with 2 iogrps shows a dip at the highest load levels as can be seen from the graph.

- SVC with 2 iogrps performs significantly slower for IOPS than SVC with 1 iogrp at lower load levels but performs significantly better at the highest load levels. Each could show more than a 3,000 IOPS advantage over the other depending on the load level. SVC with 2 iogrps MBPS outperform SVC with 1 iogrp at most load levels but dips below SVC with 1 iogrp at the highest load levels.

Orion Mixed OLTP and Data Warehouse RAID-5 - 70% Reads and 30% Writes IOPS at Sequential Data Point 5

Orion Mixed OLTP and Data Warehouse RAID-5 - 70% Reads and 30% Writes Latency at Sequential Data Point 5
11.1.4.3 Mixed OLTP and Data Warehouse (Sequential) RAID-10 50% Reads and 50% Writes at Sequential Data Point 5

<table>
<thead>
<tr>
<th></th>
<th>IOPS</th>
<th>Latency (ms/op)</th>
<th>MBPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS8300</td>
<td>449 – 6,970</td>
<td>1.87 – 2.49</td>
<td>402.74 – 484.21</td>
</tr>
<tr>
<td>SVC with 1 iogrp</td>
<td>280 – 4,287</td>
<td>0.88 – 4.19</td>
<td>400.43 – 453.02</td>
</tr>
<tr>
<td>SVC with 2 iogrps</td>
<td>230 – 3,959</td>
<td>0.9 – 3.64</td>
<td>340.85 – 518.41</td>
</tr>
</tbody>
</table>

- DS8300 IOPS are significantly better than both SVC with 1 iogrp and SVC with 2 iogrps at all load levels. The advantage reaches 2,600 to 3,000 IOPS at the highest load levels.
- SVC with 1 iogrp slightly outperforms SVC with 2 iogrps at all load levels with the range being 50 to 453 IOPS.
- DS8300 had better Latency numbers than both SVC with 1 iogrp and SVC with 2 iogrps at all load levels. The advantage for the DS8300 was mostly in the range of 1.5 to 2 ms/op.
- There is no significant difference in Latency between SVC with 1 iogrp and SVC with 2 iogrps.
- There is no significant difference in MBPS at any load level between the DS8300, SVC with 1 iogrp and SVC with 2 iogrps.
Orion Mixed OLTP and Data Warehouse RAID-10 - 50% Reads and 50% Writes IOPS at Sequential Data Point 5

Orion Mixed OLTP and Data Warehouse RAID-10 - 50% Reads and 50% Writes Latency at Sequential Data Point 5
11.1.4.4 Mixed OLTP and Data Warehouse (Sequential) RAID-5 50% Reads and 50% Writes at Sequential Data Point 5

<table>
<thead>
<tr>
<th></th>
<th>IOPS</th>
<th>Latency (ms/op)</th>
<th>MBPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS8300</td>
<td>434 – 4,921</td>
<td>2.28 – 4.06</td>
<td>202.09 – 426.06</td>
</tr>
<tr>
<td>SVC with 1 iogrp</td>
<td>543 – 4,221</td>
<td>1.33 – 4.91</td>
<td>259.61 – 470.13</td>
</tr>
<tr>
<td>SVC with 2 iogrps</td>
<td>730 – 6,900</td>
<td>1.21 – 3.95</td>
<td>298.97 – 457.78</td>
</tr>
</tbody>
</table>

- SVC with 1 iogrp IOPS are better than DS8300 at the four lowest load levels with the improvement ranging from 109 to 1,041 IOPS. At all other load levels, the DS8300 had better performance in the range of 86 to 852 IOPS.
- SVC with 2 iogrps IOPS were better than the DS8300 at the lowest and the highest load levels with the improvement ranging from 105 to 1,979 IOPS. In the mid-range load levels the DS8300 had better performance in the range of 33 to 630 IOPS.
- SVC with 2 iogrps IOPS were better than SVC with 1 iogrp IOPS except for at three load levels. The SVC with 2 iogrps advantage was in the range of 60 to 2,831 IOPS. SVC with 1 iogrp had a 127, 637 and 1,151 IOPS advantage at the three exceptional data points.
- There was no significant difference for Latency between SVC with 1 iogrp and the DS8300. The SVC with 2 iogrps and DS8300 were very close at all load levels except for two, where one time the DS8300 had a ~1 ms/op disadvantage and one time it had a ~1 ms/op advantage.
- SVC with 1 iogrp MBPS outperformed DS8300 at all but the lowest load level. The SVC with 1 iogrp advantage was usually around 50 MBPS.
- SVC with 2 iogrps MBPS outperformed the DS8300 at all but the three lowest load levels. The SVC with 2 iogrp advantage was in the range of 15.31 to 123.55 MBPS.
- There was no overall significant difference in MBPS between SVC with 1 iogrp and SVC with 2 iogrps.
This section will discuss the effects that RAID-10 and RAID-5 LUNs on the back-end DS8300 have on the front-end SVC. The testing will be done when SVC is using both 1 and 2 iogrps. Basically, the tests will determine whether or not the performance effects are similar to the results that were documented for a native DS8000 environment in the paper “IBM DS8000 Performance Configuration Guidelines for Implementing Oracle Databases with Automatic Storage Management” (item 5 in the References section). In particular, the questions that will be answered are:

- Do the benefits of RAID-10 still accrue in pure OLTP (random) I/O SVC environments?
- Do the benefits of RAID-5 still accrue in pure Data Warehouse (Sequential) SVC environments?
- What are the overall effects in Mixed OLTP and Data Warehouse (Sequential) SVC environments of RAID-10 or RAID-5?
- What is the effect of adding an iogrp to a given configuration, i.e., for a given workload mix, what is the performance difference between RAID-10 with 1 iogrp and RAID-10 with 2 iogrps and what is the difference between RAID-5 with 1 iogrp and RAID-5 with 2 iogrps? This effect was already discussed in detail in 11.1 Comparison of SVC to non-SVC Oracle I/O Performance, but it will be covered again here.
- Assuming that in an SVC environment DS8300 RAID-10 is shown to still have a significant advantage in OLTP environments, and that DS8300 RAID-5 is shown to still have a significant advantage in Data Warehouse (Sequential) environments, does introducing a second iogrp somewhat change the overall comparison between the RAID levels? In other words, assuming that RAID-10 with 1 iogrp still shows a significant advantage over RAID-5 with 1 iogrp in an OLTP environment, does adding an iogrp to the RAID-5 configuration “overtake” the RAID-10 with 1
iogrp configuration? Similarly, if RAID-5 with 1 iogrp is shown to still have an advantage over RAID-10 with 1 iogrp in a heavy sequential write environment, does adding an iogrp to the RAID-10 configuration improve its performance over RAID-5 with 1 iogrp?

11.2.1 OLTP Workload Summary

- **IOPS for Read/Write ratio of 70%/30%:**
  - RAID-10 with 1 iogrp was significantly better than RAID-5 with 1 iogrp with the advantage ranging from 1,500 to 5,000 IOPS.
  - RAID-10 with 2 iogrps was significantly better than RAID-5 with 2 iogrps with the advantage ranging from 1,000 to 3,500 IOPS.
  - RAID-10 with 2 iogrps was significantly better than RAID-10 with 1 iogrp with the advantage ranging from 1,200 to 9,000 IOPS.
  - RAID-5 with 2 iogrps was significantly better than RAID-5 with 1 iogrp with the advantage ranging from 1,600 to 9,800 IOPS.
  - RAID-5 with 2 iogrps was significantly better than RAID-10 with 1 iogrp with the advantage ranging from 1,000 to 6,200 IOPS.

- **Latency for Read/Write ratio of 70%/30%:**
  - There was no significant difference in Latency between any of the configurations.

- **IOPS for Read/Write ratio of 50%/50%:**
  - RAID-10 with 1 iogrp was significantly better than RAID-5 with 1 iogrp at all but two load levels with the advantage ranging from 1,000 to 3,200 IOPS. The two exceptional load levels had a 3,300 and 3,900 IOPS advantage for RAID-5 with 1 iogrp.
  - RAID-10 with 2 iogrps was significantly better than RAID-5 with 2 iogrps at all but one load level with the advantage ranging from 800 to 6,800 IOPS. The one exceptional load level had a 4,600 IOPS advantage for RAID-5 with 2 iogrps.
  - RAID-10 with 2 iogrps was significantly better than RAID-10 with 1 iogrp with the advantage ranging from 200 to 13,000 IOPS.
  - RAID-5 with 2 iogrps was significantly better than RAID-5 with 1 iogrp with the advantage ranging from 100 to 9,800 IOPS.
  - RAID-5 with 2 iogrps was significantly better than RAID-10 with 1 iogrp, at all but four data points, with the advantage ranging from 260 to 7,500 IOPS. Three of the exceptions had an advantage for RAID-10 with 1 iogrp between 200 to 400 IOPS and one exception was 3,100 IOPS.

- **Latency for Read/Write ratio of 50%/50%:**
  - There was no overall significant difference in Latency between any of the configurations.

- **IOPS for Read/Write ratio of 30%/70%:**
  - RAID-10 with 1 iogrp was significantly better than RAID-5 with 1 iogrp 75% of the time with the advantage ranging from 200 to 9,000 IOPS. The exceptions had an advantage for RAID-5 with 1 iogrp of 3,100 to 4,800 IOPS.
  - RAID-10 with 2 iogrps was significantly better than RAID-5 with 2 iogrps with the advantage ranging from 1,500 to 7,000 IOPS except for two load levels where RAID-5 with 2 iogrps had an approximately 1,000 and 4,000 IOPS advantage.
RAID-10 with 2 iogrps was significantly better than RAID-10 with 1 iogrp 75% of the time with the advantage ranging from 1,000 to 10,300 IOPS. The exceptions had an advantage for RAID-10 with 1 iogrp of approximately 800 to 6,600 IOPS.

- RAID-5 with 2 iogrps was significantly better than RAID-5 with 1 iogrp at all but one load level with the advantage ranging from 1,700 to 8,443 IOPS. The one exception had an advantage for RAID-5 with 1 iogrp of 8,415 IOPS.
- RAID-5 with 2 iogrps was significantly better than RAID-10 with 1 iogrp at all but three load levels with the advantage ranging from 200 to 9,100 IOPS. The three exceptions had an advantage for RAID-10 with 1 iogrp of 5,700 to 9,500 IOPS.

- Latency for Read/Write ratio of 30%/70%:
  - There was no overall significant difference between any of the configurations although the RAID-5 configurations had a 2 ms/op disadvantage at just one load level.

11.2.2 Data Warehouse (Sequential) Workload Summary

- MBPS for Read/Write ratio of 100%/0%:
  - The overall order of performance from fastest to slowest was RAID-5 with 1 iogrp, RAID-10 with 1 iogrp, RAID-5 with 2 iogrps and RAID-10 with 2 iogrps.
  - However the difference between RAID-5 with 1 iogrp, RAID-10 with 1 iogrp and RAID-5 with 2 iogrps was fairly negligible over the full load range.
  - RAID-10 with 2 iogrps was the one configuration significantly below the others with a throughput 80 to 150 MBPS slower than RAID-10 with 1 iogrp.

- MBPS for Read/Write ratio of 70%/30%:
  - RAID-5 with 2 iogrps performed the best. Next in performance were RAID-5 with 1 iogrp and RAID-10 with 2 iogrps which had very close numbers over the full load range. RAID-10 with 1 iogrp had the lowest overall throughput.
  - RAID-5 with 2 iogrps was by far the best with a 25 - 145 MBPS advantage over both RAID-5 with 1 iogrp and RAID-10 with 2 iogrps.
  - RAID-10 with 2 iogrps had significant 125 to 170 MBPS advantage over RAID-10 with 1 iogrp.

- MBPS for Read/Write ratio of 50%/50%:
  - RAID-5 with 2 iogrps performed the best. Next in performance were RAID-5 with 1 iogrp and RAID-10 with 2 iogrps which had very close numbers over the full load range. RAID-10 with 1 iogrp had the lowest overall throughput.
  - RAID-5 with 2 iogrps was by far the best with a 15 - 106 MBPS advantage over both RAID-5 with 1 iogrp and RAID-10 with 2 iogrps.
  - RAID-10 with 2 iogrps had significant 70 to 130 MBPS advantage over RAID-10 with 1 iogrp.

- MBPS for Read/Write ratio of 30%/70%:
  - RAID-5 with 2 iogrps performed the best. Next in performance were RAID-5 with 1 iogrp and RAID-10 with 2 iogrps which had very close numbers over the full load range. RAID-10 with 1 iogrp had the lowest overall throughput.
  - RAID-5 with 2 iogrps was by far the best with a 65 - 112 MBPS advantage over both RAID-5 with 1 iogrp and RAID-10 with 2 iogrps.
- RAID-10 with 2 iogrps had significant 86 to 128 MBPS advantage over RAID-10 with 1 iogr.

**MBPS for Read/Write ratio of 0%/100%:**
- RAID-5 with 2 iogrps and RAID-5 with 1 iogr performed best and were very close in throughput over the full load level range.
- RAID-10 with 2 iogrps and RAID-10 with 1 iogr were very close in throughput over the full load level range.
- The RAID-5 configurations had a throughput advantage of 50 to 120 MBPS over the RAID-10 configurations.

### 11.2.3 Mixed OLTP and Data Warehouse (Sequential) Workload Summary

**Read/Write 70%/30%**

- **IOPS:**
  
  RAID-10 with 2 iogrps had by far the best overall performance, with just two exceptions where there is a downward spike.

  RAID-5 with 1 iogr and RAID-5 with 2 iogrps are next in performance. RAID-5 with 1 iogr greatly outperformed RAID-5 with 2 iogrps at lower load levels and RAID-5 with 2 iogrps greatly outperformed RAID-5 with 1 iogr at higher load levels.

  RAID-10 with 2 iogrps, RAID-10 with 1 iogr and RAID-5 with 1 iogr are all very close at the lowest load levels, but RAID-10 with 1 iogr drops dramatically lower than all of the other configurations at higher load levels.

- **Latency**
  
  RAID-10 with 2 iogrps had by far the best overall Latency numbers, with just two exceptions where the Latency spikes up.

  RAID-5 with 1 and RAID-5 with 2 iogrps are close at the higher load levels but RAID-5 with 1 iogr has a significant advantage at the lower load levels.

  RAID-10 with 2 iogrps, RAID-10 with 1 iogr and RAID-5 with 1 iogr are all very close at the lowest load levels. But RAID-10 with 1 iogr spikes at higher load levels where it is about 2.5 ms/op slower than RAID-10 with 2 iogrps.

- **MBPS**
  
  RAID-5 with 2 iogrps is the best overall performer except at four of the higher load levels where it spikes below all of the other configurations by about 100 MBPS. Otherwise it has approximately a 50 - 130 MBPS advantage over the other configurations.

  RAID-5 with 1 iogr, RAID-10 with 1 iogr and RAID-10 with 2 iogrps are all close in throughput over the full load range.

**Read/Write 50%/50%**

- **IOPS:**
  
  RAID-5 with 1 iogr and RAID-5 with 2 iogrps are the best overall and are close to each other except at the three highest load levels where RAID-5 with 2 iogrps has 1,000 to 2,800 IOPS advantage.

  RAID-10 with 1 iogr and RAID-10 with 2 iogrps are almost exactly the same over the full load range.
The RAID-5 configurations have approximately a 1,000 IOPS advantage over the RAID-10 configurations at almost the full load level range.

- **Latency**
  
  RAID-5 with 1 iogrp and RAID-5 with 2 iogrps Latency numbers are the best overall and are close to each other except at one load level where RAID-5 with 2 iogrps has a 2.74 ms/op disadvantage and one load level where it has a 2 ms/op advantage.
  
  RAID-10 with 1 iogrp and RAID-10 with 2 iogrps are very close over the full load range.
  
  The RAID-5 configurations in general have a 1 to 3 ms/op advantage over the RAID-10 configurations.

- **MBPS**
  
  Overall, all four configurations are close in throughput.
  
  RAID-10 with 1 iogrp and RAID-10 with 2 iogrps are virtually indistinguishable over the full load range.
  
  RAID-5 with 1 iogrp and RAID-5 with 2 iogrps are close at all load levels.
  
  The RAID-10 configurations are a bit slower than the RAID-5 configurations at load levels below the midpoint load level but they are faster at load levels above the midpoint load level.

### 11.2.4 Graphs for OLTP - IOPS and Latency at Read/Write 70%/30%, 50%/50% and 30%/70%

![Orion OLTP IOPS - 70% Reads and 30% Writes](chart)

- **RAID5 1 iogrp**
- **RAID5 2 iogrps**
- **RAID10 1 iogrp**
- **RAID10 2 iogrps**
11.2.5 Graphs for Data Warehouse (Sequential) - MBPS at Read/Write 100%/0%, 70%/30%, 50%/50%, 30%/70% and 0%/100%
Orion Data Warehouse (Sequential) MBPS - 70%
Reads and 30% Writes

Outstanding Async I/O's

Orion Data Warehouse (Sequential) MBPS - 50%
Reads and 50% Writes

Outstanding Async I/O's
Orion Data Warehouse (Sequential) MBPS - 30% Reads and 70% Writes

Orion Data Warehouse (Sequential) MBPS - 0% Reads and 100% Writes

11.2.6 Graphs for Mixed OLTP and Data Warehouse (Sequential) - IOPS, Latency and MBPS Read/Write 70%/30% and 50%/50%
Orion Mixed OLTP and Data Warehouse
(Sequential) IOPS - 70% Reads and 30% Writes at Sequential Data Point 5

Orion Mixed OLTP and Data Warehouse
(Sequential) Latency - 70% Reads and 30% Writes at Sequential Data Point 5
Orion Mixed OLTP and Data Warehouse
(Sequential) MBPS - 70% Reads and 30% Writes at Sequential Data Point 5

Orion Mixed OLTP and Data Warehouse
(Sequential) IOPS - 50% Reads and 50% Writes at Sequential Data Point 5
11.3 Comparison of ASM and SVC striping configurations

This section will discuss various subjects pertaining to both ASM and SVC VDisk striping. In addition, benchmarks will be run which compare the effects of striping at the ASM level only, at the SVC level only, and at both the ASM and SVC levels together (striping-on-striping). The two primary references which will be used are the Oracle Press book “Oracle Automatic Storage Management – Under-the-Hood and Practical Deployment Guide” (item 15 in the References) and the IBM Redbook SAN Volume Controller: Best Practices and Performance Guidelines (item 2 in the References).
The following subjects will be covered:

- What are the purposes of Oracle ASM coarse and fine-grained striping?
- What is an Oracle ASM Allocation Unit (AU)? What is the purpose of changing the AU size?
- What is the purpose of SVC VDisk striping?
- What is the effect of using Oracle ASM striping “on top of” SVC VDisk striping; i.e., is the overall effect beneficial, negative, or neutral? In other words, what are the comparisons between
  (1) striping only at the SVC level
  (2) striping only at the ASM level
  (3) striping at both the ASM and SVC levels in tandem (striping-on-striping)

To simulate the various possible combinations of striping at only one or both of the ASM and SVC levels, the following will be done:

(a) For striping only at the SVC level and not at the ASM level, one 400 GB VDisk will be created which is striped across all four of the MDisks in the MDisk Group. This VDisk will be presented as one ASM disk to ORION. As discussed in section 4 Oracle Automatic Storage Management (ASM), ASM automatically stripes across all ASM disks in a diskgroup. Therefore, presenting only one VDisk to ORION is effectively eliminating ASM striping.

This configuration is not the stated best practice for ASM. It is safe to assume that, in a typical production environment, an ASM diskgroup will need to have space added at some point. As noted in Chapter 5 of the Oracle ASM Under-the-Hood book, in the section titled Resizing a Physical Disk or LUN and the ASM Diskgoup, it is best practice to add disks of similar size to increase the size of a diskgroup as opposed to resizing existing disks. Resizing an existing LUN (a VDisk in our case) which is an ASM disk, while it may be appropriate in certain circumstances, may require, depending on the OS, a system reboot for the new LUN size to be recognized at the OS level (it should be noted that if it is decided to shrink LUNs associated with ASM disks, then the appropriate “ALTER DISKGROUP … RESIZE …” command must be issued prior to shrinking those LUNs at the storage level). Also, in Oracle 10.2.0.3 and earlier, ASM instances need to be restarted for ASM to recognize the new LUN size. Oracle 11g ASM can detect new LUN sizes without an instance restart.

In addition, as noted in Chapter 3 of the Oracle ASM Under-the-Hood book, in the section titled ASM and Storage Arrays, a single LUN configuration makes it difficult to grow or shrink the diskgroup incrementally (that is, if LUN resizing is not being considered as an option). If it is ever decided to grow the diskgroup by adding another LUN, the additional LUN would have to be the same size as the existing LUN (which may have grown quite large by that time), and therefore the diskgroup would be doubling in size, which is probably not an optimal approach to storage provisioning.

Another problem with a single-LUN configuration is that the OS may not allow enough I/O to be simultaneously queued to one LUN. The OS may assume that it is only one “spindle” and thus will limit the number of simultaneous I/O operations.

One more problem with a single-LUN/single-VDisk configuration is that it does not allow for load balancing multiple VDisks across multiple iogrps. As has been previously shown, except
for rare exceptions, spreading VDisks across multiple iogrps improves performance very significantly. This is especially true for OLTP random I/O environments. Thus load balancing across multiple iogrps should always be an available scalability option in an SVC environment.

Finally, as was discussed in section 6.1.6 VDisk Configuration item (4), a VDisk automatically gets assigned a preferred node in a given iogrp. Even-numbered VDisks get assigned to one node in an iogrp and odd-numbered VDisks get assigned to the other node in the same iogrp. Thus a single-LUN/single-VDisk configuration means that all of the load will go to only one node in the iogrp to which that VDisk has been assigned.

For all of these reasons, the best practice for Oracle ASM is to have multiple LUNs in one diskgroup.

(b) For striping only at the ASM level and not at the SVC level, four 100 GB VDisks will be created, each of which is mapped to only one MDisk in the MDisk Group. The `svctask mkvdisk` command has a `–mdisk` option which allows a VDisk to be mapped to one or more specific MDisks in an MDisk Group. These four VDisks will be presented as four individual ASM disks to ORION. As noted in section 6.1.6 VDisk Configuration item (2), this is not an SVC best practice. Striping at the VDisk level takes advantage of full Workload Spreading on the back-end DS8300. It also reduces the administrative work for DBAs and System Administrators since it offers a constant level of load balancing on the back-end without having to spend time on that aspect of performance planning.

Therefore the best practice for SVC is to use both application-level striping (ASM striping in our case) and SVC striping in conjunction with each other.

(c) For the striping-on-striping configuration, i.e., the ASM striping on top of SVC striping, the configuration is exactly the same as in all of the previous benchmarking sections of this paper. Therefore, those test results have already been collected. As has been previously described, this configuration entailed creating four 100 GB VDisks, each of which was striped across all four MDisks in the MDisk Group. These were presented as four ASM disks to ORION.

The possibility that application-level striping (ASM striping in our case) on top of SVC striping may have a negative overall effect when compared to striping at only one level is discussed in detail in Chapter 10 of the SVC Best Practices Redbook in the section 10.4 When the application does its own balancing of I/Os. That section of the Redbook mentions the possibility that striping on striping will create a divergent “beat” effect similar to the harmonics of music where one striping method “undoes” the benefits of another striping method. It goes on to say that the effect is easy to avoid by ensuring a wide difference in stripe granularities. The benchmarks which will be discussed further down in this section will uncover whether or not there is any such antagonistic effect with ASM striping on top of SVC striping for any possible Oracle I/O workload.

- Given the many technological advantages of both Oracle ASM and SVC individually (aside from the significant benefits of striping, since there are many additional benefits provided by both technologies), and the significant performance benefits that were demonstrated in section 11.1 Comparison of SVC to non-SVC Oracle I/O Performance when they are used together as opposed to a native DS8300 configuration, does the investigation of striping-on-striping really
matter? In other words, if a decision has been made to use both of these technologies together, in accordance with the stated best practices for each of them individually, should not the question simply be “how do we make them work together in an optimal manner”? That is in fact the question that should be answered by running the benchmarks since special cases may be uncovered where striping should be done only at the ASM level, or only at the SVC level, but not at both levels simultaneously.

### 11.3.1 ASM coarse and fine-grained striping

As was stated in the section 4.3 ASM Striping, coarse striping provides load balancing across all of the LUNs presented as ASM disks to an ASM diskgroup. This is meant to improve throughput for sequential workloads. To stripe data, ASM separates files into stripes and spreads data evenly across all of the disks in a disk group. The stripes are always equal in size to the effective Allocation Unit (AU). An AU is the fundamental unit of allocation within a disk group. An ASM file consists of one or more file extents and a file extent consists of one or more AUs. The default size of an Allocation Unit is 1 MB.

Fine-grained striping reduces latency for certain file types by spreading the load more widely. The fine-grained stripe size always equals 128 KB. This provides lower I/O latency for small I/O operations such as redo log writes.

When you create a disk group in Oracle Database 11g, you can set the ASM AU size to be between 1 MB and 64 MB in powers of two, such as, 1, 2, 4, 8, 16, 32, or 64. Oracle Database 10g AUs are 1 MB, although this can be changed by modifying some Oracle hidden initialization parameters.

Item 14 in the References section, "Oracle Database Storage Administrator’s Guide, 11g Release 1 (11.1)”, which is part of the Oracle documentation, states in Chapter 1, in the section Allocation Units, that “Larger AU sizes typically provide performance advantages for data warehouse applications that use large sequential reads”. Also, as noted in Chapter 5 of the Oracle ASM Under-the-Hood book, in the section titled Setting Larger AU Sizes for VLDBs, it may be beneficial for very large databases (10 TB or larger) to change the default AU size for the following purposes:

- Reduce the memory required in the Oracle System Global Area (SGA) for managing the extent maps in the RDBMS instance
- Increase the file size limits
- Reduce the database open time, because very large databases usually have many big datafiles

Section 11.3.4 Performance comparisons of different ASM Allocation Unit sizes further down will show some benchmark runs where the AU size takes on all possible values from 1 MB to 64 MB in powers of two.

### 11.3.2 SVC VDisk striping

SVC VDisk striping is simply meant to spread the I/O workload across more than one MDisk. If the MDisks are mapped to separate back-end DS8300 RAID arrays, as they should be if the recommendations made in this paper with regard to SVC configurations are being followed, then the VDisks are taking advantage of full Workload Spreading as was defined in the paper “IBM DS8000 Performance Configuration Guidelines for Implementing Oracle Databases with Automatic Storage Management” (item 5 in the
References. The very significant performance advantages of full Workload Spreading were detailed in that paper.

11.3.3 Performance comparisons for ASM and SVC striping configurations

This section will compare the test results from the three striping configurations mentioned above:

(1) striping only at the SVC level (denoted as SVC-only)
(2) striping only at the ASM level (denoted as ASM-with-1-iogrps or ASM-with-2-iogrps)
(3) striping at both the ASM and SVC levels simultaneously (denoted as ASM/SVC-with-1-iogrps or ASM/SVC-with-2-iogrps)

11.3.3.1 OLTP Workload Analysis

This section gives an analysis of OLTP workloads.

- **RAID-10**

  - **Read/Write 70%/30% IOPS**
    - ASM/SVC-with-2-iogrps is best over the full load level range, and significantly better than all other configurations at higher load levels.
    - ASM/SVC-with-1-iogrps, ASM-with-1-iogrps, ASM-with-2-iogrps and SVC-only are next and are close at all load levels except that SVC-only is best at the top three load levels.

  - **Read/Write 70%/30% Latency**
    - There is no significant difference in Latency between any of the configurations.

  - **Read/Write 50%/50% IOPS**
    - ASM/SVC-with-2-iogrps is significantly the best over the full load level range with one exception where it equals the other configurations.
    - ASM/SVC-with-1-iogrps is next in performance over lower load levels.
    - SVC-only is next over higher load levels.
    - ASM-with-1-iogrps and ASM-with-2-iogrps trail all others and are close to each other at all load levels.

  - **Read/Write 50%/50% Latency**
    - There is no significant difference in latency between any of the configurations; the differences were always less than 1 ms/op.

  - **Read/Write 30%/70% IOPS**
    - ASM/SVC-with-2-iogrps is overall significantly the best although it dips at some load levels.
    - ASM-with-1-iogrps and SVC-only are next, but SVC-only lags ASM-with-1-iogrps at higher load levels.
    - ASM-with-2-iogrps and ASM/SVC-with-1-iogrps are next in performance.

  - **Read/Write 30%/70% Latency**
    - There is no significant difference in Latency between any of the configurations, except for two very minor exceptions. Latency for all configurations at all load levels was 2 ms/op or less.

- **RAID-5**
- **Read/Write 70%/30% IOPS**
  - ASM/SVC-with-2-iogrps is best over the full load level range.
  - ASM/SVC-with-1-iogrp, ASM-with-1-iogrp, ASM-with-2-iogrps and SVC-only are close over low load levels. ASM/SVC-with-1-iogrp is better at the highest load levels and ASM-with-2-iogrps dips below the others at high load levels.

- **Read/Write 70%/30% Latency**
  - There is no significant difference in Latency between any of the configurations.

- **Read/Write 50%/50% IOPS**
  - ASM/SVC-with-2-iogrps is significantly the best over the full load level range.
  - ASM/SVC-with-1-iogrp is next and is significantly better than the remaining configurations at most load levels.
  - ASM-with-1-iogrp and SVC-only are next and are close over the full load level range.
  - ASM-with-2-iogrps is close to ASM-with-1-iogrp and SVC-only at the lowest load levels but lags significantly behind at higher load levels.

- **Read/Write 50%/50% Latency**
  - In general, there is no significant difference in Latency between any of the configurations. At a couple of load levels, ASM-with-2-iogrps spikes a bit but no Latency is ever over 3 ms/op.

- **Read/Write 30%/70% IOPS**
  - ASM/SVC-with-2-iogrps is overall the best although it dips at some load levels.
  - ASM/SVC-with-1-iogrp is next in performance.
  - ASM-with-1-iogrp, ASM-with-2-iogrps and SVC-only are next in performance.

- **Read/Write 30%/70% Latency**
  - In general, there is no significant difference in Latency between any of the configurations. At a couple of higher load levels, there is a difference between the fastest and slowest configurations of 2 ms/op, but that is rare. No Latency is ever over 4 ms/op and the vast majority are between 1 - 2 ms/op.

11.3.3.2 **OLTP Summary**

The results are very clear:
- For both RAID-10 and RAID-5 on the back-end DS8300
- at all Read/Write workload ratios
- whether one or two iogrps are being used
- when taking IOPS performance results into account
- when taking Latency performance results into account
- when taking ASM administration best practices into account
- when taking SVC administration best practices into account

both ASM and SVC striping should always be used together for all OLTP workloads, i.e., striping-on-striping is the recommended configuration.

11.3.3.3 **Data Warehouse (Sequential) Workload Analysis**

This section gives an analysis of Data Warehouse workloads.

- **RAID-5**
  - **Read/Write 100%/0% MBPS**
11.3.3.4 Data Warehouse (Sequential) Summary

RAID-10

- Read/Write 100%/0% MBPS
  - ASM-with-2-iogrps is the best overall.
  - ASM/SVC-with-1-iogrp is next in performance.
  - ASM/SVC-with-2-iogrps is next in performance, although it's close to ASM-with-1-iogrp at most load levels.
  - SVC-only is way behind all the others, although it's close to ASM/SVC-with-1-iogrp at some load levels.

- Read/Write 0%/100% MBPS
  - ASM-with-1-iogrp and ASM-with-2-iogrps are the best overall and are significantly better than all other configurations at all load points.
  - ASM/SVC-with-1-iogrp, ASM/SVC-with-2-iogrps and SVC-only are next and are pretty close to each other at all load levels, although SVC-only bounces around a bit.

- Read/Write 70%/30%, 50%/50% and 30%/70% MBPS
  - ASM/SVC-with-2-iogrps and ASM-SVC-with-2-iogrps are the best overall.
  - ASM/SVC-with-1-iogrp and ASM-with-1-iogrp are next in performance.
  - SVC-only is way behind all the other configurations, 40% less than the top performer.

- Read/Write 0%/100% MBPS
  - ASM/SVC-with-1-iogrp, ASM/SVC-with-2-iogrps, ASM-with-1-iogrp and ASM-with-2-iogrps are best overall and are close at all load levels.
  - SVC-only is way behind all the other configurations, 40% less than the top performer.

RAID-5

- If two iogrps are available, then ASM/SVC-with-2-iogrps should be used. If only one iogrp is available, then ASM/SVC-with-1-iogrp should be used.

- For the Read/Write ratios of 100%/0% and 0%/100%, the ASM-only striping can match the performance of the ASM/SVC striping configurations. But they are so close overall that
it makes no sense to deviate from the best practices previously discussed, and therefore both ASM and SVC striping should be used simultaneously, i.e. the striping-on-stripping configuration should be used.

- For the Read/Write ratios of 70%/30%, 50%/50% and 30%/70%, the ASM/SVC-with-2-iogrps and ASM-with-2-iogrps are very close to each other. The ASM/SVC-with-1-iogrp and ASM-with-1-iogrp are very close to each other and lag significantly behind the 2-iogrp configurations. Therefore the difference in performance with these workloads is dependent on how many iogrps are available. Again, whether one iogrp is used or two iogrps are used, there is no performance reason to deviate from best practices and therefore both ASM and SVC striping should be used simultaneously.

- The SVC-only configuration, which uses VDisk striping but no ASM striping, should never be used. This configuration lags very significantly behind all of the other configurations.

RAID-10:

- Overall, the ASM-only striping configurations tend to perform best and often exceed the ASM/SVC-stripping configurations by 100 – 150 MBPS. If two iogrps are available, then both of them should be used, resulting in an ASM-with-2-iogrp configuration.

- The SVC-only configuration, which uses VDisk striping but no ASM striping, should never be used. Most of the time, this configuration lags very significantly behind all of the other configurations.

11.3.3.5 Mixed OLTP and Data Warehouse (Sequential) Workload Analysis

This section gives an analysis of mixed OLTP and Data Warehouse workloads.

- **RAID-10 Read/Write 70%/30% at Sequential Data Point 5**
  - IOPS
    - ASM/SVC-with-2-iogrps, ASM-with-1-iogrp and SVC-only and are significantly the best from lowest to highest load levels.
    - ASM/SVC-with-2 iogrps dips sharply a couple of times at higher load levels.
    - ASM-with-1-iogrp is close to above three configurations but dips sharply at highest load levels.
    - ASM-with-2 iogrps is next.
    - ASM/SVC-with-1-iogrp lags all others at everything but the lowest load levels.
  - Latency
    - ASM/SVC-with 2 iogrps, ASM-with-1-iogrp, ASM-with-2-iogrps and SVC-only are close at most load levels.
    - ASM-with-1-iogrp, ASM-with-2-iogrps and ASM/SVC-with-2-iogrps spike at some higher load levels.
    - ASM/SVC-with-1-iogrp lags others by about 2 ms/op at the middle to highest load levels.
  - MBPS
    - ASM/SVC-with-1-iogrp, ASM/SVC-with-2-iogrps and ASM-with-1-iogrp perform the best and are fairly close to each other at most load levels.
    - ASM-with-2-iogrps is close to the above three configurations at the lowest load levels but drops significantly below them at middle to high load levels.
    - SVC-only lags all other configurations quite significantly.

- **RAID-10 Read/Write 50%/50% at Sequential Data Point 5**
  - IOPS
- SVC-only, ASM-with-1-iogrp and ASM-with-2-ioggrps are significantly the best at all 
  load levels.
- ASM/SVC-with-1-iogrp and ASM/SVC-with-2-ioggrps are very close to one another 
  at all load levels and lag all the other configurations significantly at all load levels.

  - Latency
    - SVC-only, ASM-with-1-iogrp and ASM-with-2-ioggrps are significantly the best at all 
      load levels and hover around 2 ms/op.
    - ASM/SVC-with-1-iogrp and ASM/SVC-with-2-ioggrps are very close to one another 
      at all load levels and lag all the other configurations significantly at all load levels, but 
      they never exceed 5 ms/op latency.

  - MBPS
    - ASM-with-1-iogrp and ASM-with-2-ioggrps are significantly better than all other 
      configurations at almost all load levels except at the very highest levels.
    - ASM/SVC-with-1-iogrp, ASM/SVC-with-2-ioggrps and SVC-only are all close to 
      each other over the full load level range and are about 100 MBPS behind the 
      ASM-striping-only configurations, except at the highest load levels where they all 
      converge.

- RAID-5 Read/Write 70%/30% at Sequential Data Point 5
  - IOPS
    - ASM-with-1-iogrp and SVC-only are the best over full load level range.
    - ASM/SVC-with-1-iogrp is next, but it drops at higher load levels.
    - ASM/SVC-with-2-ioggrps and ASM-with-2-ioggrps are close to each other at lower to 
      middle load levels and are significantly less than all the other configurations, 
      except that ASM/SVC-with-2-ioggrps spikes and is the best at the highest load 
      levels.
  
  - Latency
    - ASM/SVC-with-2-ioggrps and ASM-with-2-ioggrps are significantly worse than all of 
      the other configurations at the lowest load levels, but all of the configurations 
      converge at the middle to highest load levels.

  - MBPS
    - All of the configurations, except for SVC-only, are close at most load levels, 
      although ASM/SVC-with-2-ioggrps drops significantly below the others at the 
      highest load levels
    - SVC-only is significantly lower than all of the other configurations at almost all 
      load levels.

- RAID-5 Read/Write 50%/50% at Sequential Data Point 5
  - IOPS
    - All of the configurations are very close to each other over the full load level range; 
      although at the highest load levels ASM/SVC-with-2-ioggrps and ASM-with-2-
      ioggrps have significantly better numbers than the rest.
  
  - Latency
    - There is no significant difference in any of the configurations over the full load 
      level range. The Latencies are mostly between 2 - 3 ms/op and get close to, but 
      never reach, 5 ms/op at the highest load levels.
- **MBPS**
  - All of the configurations, except SVC-only, are close at almost all of the load levels.
  - SVC-only lags behind all other configurations by close to 100 MBPS at some of the middle load levels.

### 11.3.3.6 Mixed OLTP and Data Warehouse (Sequential) Workload Summary

**RAID-10 Read/Write 70%/30%:**

The ASM/SVC-with-2-iogrps and ASM-with-1-iogrp are the best overall choice since they are in the top performing group for all of IOPS, Latency and MBPS. SVC-only does well for IOPS and Latency but, as was also shown in the Data Warehouse workloads, it performs so poorly with respect to MBPS that it should not be considered.

**RAID-10 Read/Write 50%/50%:**

The ASM-with-1-iogrp and ASM-with-2-iogrp configurations perform the best overall with respect to IOPS, Latency and MBPS. Therefore ASM-only striping should be considered for this RAID-10 mixed workload.

**RAID-5 Read/Write 70%/30% and 50%/50%:**

When all of the following factors are considered:

- that the benchmark results for a specific configuration should take into consideration the combination of IOPS, Latency and MBPS (i.e., not just one metric in isolation)
- that following best practices (i.e., striping-on-stripping configurations) deserves a little bit of extra consideration as long as the corresponding benchmark numbers are not tremendously lower for all three of the metrics (IOPS, Latency and MBPS) at most load levels

then the best configuration choices for these two RAID-5 workload mixes are the striping-on-stripping configurations, i.e., ASM/SVC-with-1-iogrp and ASM/SVC/with-2-iogrps.

### 11.3.3.7 Performance comparison graphs for ASM and SVC striping configurations

This section will include some graphs that are representative of the results from the OLTP, Data Warehouse (Sequential), and Mixed OLTP and Data Warehouse (Sequential) striping comparison benchmark runs.

**Striping Comparison RAID-10 OLTP IOPS Read/Write 70%/30%**
Striping Comparison RAID-10 OLTP IOPS - 70% Reads and 30% Writes

Striping Comparison RAID-10 OLTP Latency Read/Write 70%/30%

Striping Comparison RAID-5 OLTP IOPS Read/Write 70%/30%
Striping Comparison RAID-5 OLTP IOPS - 70% Reads and 30% Writes

Striping Comparison RAID-5 OLTP Latency Read/Write 70%/30%

Striping Comparison RAID-10 OLTP IOPS Read/Write 50%/50%
Striping Comparison RAID-10 OLTP IOPS - 50% Reads and 50% Writes

Striping Comparison RAID-10 OLTP Latency Read/Write 50%/50%

Striping Comparison RAID-5 OLTP IOPS Read/Write 50%/50%
Striping Comparison RAID-5 OLTP IOPS - 50%
Reads and 50% Writes

Striping Comparison RAID-5 OLTP Latency Read/Write 50%/50%

Striping Comparison RAID-5 Data Warehouse (Sequential) MBPS Read/Write 100%/0%
Striping Comparison RAID-5 Data Warehouse
MBPS - 100% Reads and 0% Writes

Striping Comparison RAID-10 Data Warehouse (Sequential) MBPS Read/Write 100%/0%

Striping Comparison RAID-5 Data Warehouse (Sequential) MBPS Read/Write 50%/50%
Striping Comparison RAID-5 Data Warehouse
MBPS - 50% Reads and 50% Writes

Striping Comparison RAID-10 Data Warehouse (Sequential) MBPS Read/Write 50%/50%

Striping Comparison RAID-5 Data Warehouse (Sequential) MBPS Read/Write 0%/100%
Striping Comparison RAID-5 Data Warehouse
MBPS - 0% Reads and 100% Writes

Striping Comparison RAID-10 Data Warehouse (Sequential) MBPS Read/Write 0%/100%

Striping Comparison RAID-10 Mixed OLTP and Data Warehouse (Sequential) IOPS Read/Write 70%/30%
Striping Comparison RAID-10 Mixed OLTP and Data Warehouse IOPS - 70% Reads and 30% Writes

- SVC striping only: 1 VDisk/1 iogrp/1 node
- ASM striping only: 1 iogrp
- ASM striping only: 2 iogrps
- ASM/SVC striping: 1 iogrp
- ASM/SVC striping: 2 iogrps

Striping Comparison RAID-10 Mixed OLTP and Data Warehouse (Sequential) Latency Read/Write 70%/30%

- SVC striping only: 1 VDisk/1 iogrp/1 node
- ASM striping only: 1 iogrp
- ASM striping only: 2 iogrps
- ASM/SVC striping: 1 iogrp
- ASM/SVC striping: 2 iogrps

Striping Comparison RAID-10 Mixed OLTP and Data Warehouse (Sequential) MBPS Read/Write 70%/30%
Striping Comparison RAID-10 Mixed OLTP and Data Warehouse MBPS - 70% Reads and 30% Writes

Striping Comparison RAID-10 Mixed OLTP and Data Warehouse (Sequential) IOPS Read/Write 50%/50%

Striping Comparisons RAID-10 Mixed OLTP and Data Warehouse IOPS - 50% Reads and 50% Writes

Striping Comparison RAID-10 Mixed OLTP and Data Warehouse (Sequential) Latency Read/Write 50%/50%
Striping Comparisons RAID-10 Mixed OLTP and Data Warehouse Latency - 50% Reads and 50% Writes

Striping Comparison RAID-10 Mixed OLTP and Data Warehouse (Sequential) MBPS Read/Write 50%/50%

Striping Comparisons RAID-10 Mixed OLTP and Data Warehouse MBPS - 50% Reads and 50% Writes

Striping Comparison RAID-5 Mixed OLTP and Data Warehouse (Sequential) IOPS Read/Write 70%/30%
Striping Comparison RAID-5 Mixed OLTP and Data Warehouse IOPS - 70% Reads and 30% Writes

Striping Comparison RAID-5 Mixed OLTP and Data Warehouse (Sequential) Latency Read/Write 70%/30%

Striping Comparison RAID-5 Mixed OLTP and Data Warehouse (Sequential) MBPS Read/Write 70%/30%
Striping Comparison RAID-5 Mixed OLTP and Data Warehouse MBPS - 70% Reads and 30% Writes

Striping Comparison RAID-5 Mixed OLTP and Data Warehouse (Sequential) IOPS Read/Write 50%/50%

Striping Comparison RAID-5 Mixed OLTP and Data Warehouse (Sequential) Latency Read/Write 50%/50%
11.3.4 Performance comparisons of different ASM Allocation Unit sizes

This section will determine the performance effect on both the native DS8300 and SVC of varying the ASM Allocation Unit size. As mentioned previously, in the section 11.3.1 ASM coarse and fine-grained
striping, resizing the AU from the default 1 MB size is meant to improve performance for larger sequential I/O. It can also do the following for very large databases (i.e., 10 TB or larger):

- Reduce the memory required in the Oracle System Global Area (SGA) for managing the extent maps in the RDBMS instance
- Increase the file size limits
- Reduce the database open time, because very large databases usually have many big datafiles

Oracle Database 11g minimizes the latency caused by the opening of files by reading extent maps on demand. And to reduce the amount of memory required by the extent maps, Oracle Database 11g automatically employs Variable Sized Extents.

To take advantage of larger ASM AU sizes, it is important to also set the initialization parameter `db_file_multiblock_read_count` to an appropriate value so that it corresponds to the AU size. This will cause the Oracle server to actually issue the larger I/Os. It is true that the OS will break up very large I/O requests into several smaller I/O requests, but the requests will be sent to the storage in parallel so that the disks can serve them all in one seek. `db_file_multiblock_read_count` specifies the maximum number of database blocks read in one I/O operation during a sequential scan. Therefore, assuming that an Oracle database is using a block size of 8 KB, setting `db_file_multiblock_read_count` to 128 will allow Oracle to issue I/Os up to 1 MB in size (8 KB x 128 = 1 MB). The maximum allowable value for `db_file_multiblock_read_count` is dependent on the operating system. If an attempt is made to set it to a value that is larger than what is allowed on a particular OS, then Oracle automatically sets it to the largest valid value. For example, with Oracle Server version 11.1.0.6.0 on AIX 6.1 and on 64-bit Red Hat Enterprise Linux versions 4 and 5, when the database block size is set to 8192, the `db_file_multiblock_read_count` can be set to a maximum of 4096. This corresponds to a maximum I/O size of 32 MB. For Oracle Server version 10.2.0.4.0 on 32-bit Red Hat Enterprise Linux 4, when the database block size is 8192, then `db_file_multiblock_read_count` can be set to a maximum of 128, which corresponds to 1 MB.

With ORION, the way to simulate a particular AU stripe size is with the `–stripe` parameter. And the parameter `–size_large` can be used to create a specific sequential I/O size since it specifies the size of the I/Os in KB. Therefore, the `–size_large` parameter can be set to a value that would simulate a particular `db_file_multiblock_read_count` value. For example, the ORION command line that will be used for a 100% Read/0% Write ratio with both an ASM AU and sequential I/O size of 8 MB is:

```
orion -run advanced -testname seq_raid5_read100_write0_8M -num_disks 4 -type seq -size_large 8192 -simulate raid0 -stripe 8192 -write 0 -num_small 0 -matrix col -duration 60 -verbose
```

Since the benchmarks were run on 64-bit Red Hat Enterprise Linux 4, ORION would eventually return an error when the `–size_large` parameter was set to a value larger than 32768 (32 MB). Therefore the runs for AU sizes of 64 MB used a `–size_large` value of 32768 instead of 65536.

The following benchmarks were run on both native DS8300 and the SVC with RAID-5 at Read/Write ratios of 100%/0%, 70%/30%, 50%/50%, 30%/70% and 0%/100%. The AU sizes used were 1, 2, 4, 8, 16, 32 and 64 MB. The storage configurations that were used were the configurations discussed in Sections 10.2.1 DS8000 Configurations and 10.2.2 SAN Volume Controller Configurations:
• the native DS8300 runs used four 100 GB fixed block volumes where each LUN was on a different single-rank extent pool; the LUNs alternated between even- and odd-numbered extent pools

• the SVC runs used four 100 GB VDisks which were striped across all four MDisks in an MDisk Group where each MDisk mapped to a separate back-end DS8300 single-rank extent pool

11.3.4.1 Summary of performance comparisons for varying ASM AU sizes

• Varying the ASM AU size had more impact on native DS8300 performance than on SVC performance.

• For the DS8300, the 16 MB, 32 MB and 64 MB AU sizes were consistently the best performers and were significantly better than the other AU sizes. The differences tended to be more pronounced with higher Write ratios. For example, with a 30% Read/70% Write workload, the 64 MB AU performance at some load levels was approximately 700 MBPS while the 1 MB AU performance was around 350 MBPS.

• For the SVC, the 2 MB, 4 MB and 8 MB AU sizes were consistently the best performers for the 100%/0%, 70%/30%, 50%/50% and 30%/70% Read/Write ratios. The performance improvement was often in the range of 100 – 200 MBPS. The 0% Read/100% Write workload showed a negligible performance difference between all of the different AU sizes.

11.3.4.2 Variable ASM AU Size Comparison Graphs

DS8300 Data Warehouse (Sequential) RAID-5 MBPS Read/Write 100%/0% - ASM AU Sizes 1 MB – 64 MB
SVC Data Warehouse (Sequential) RAID-5 MBPS Read/Write 100%/0% - ASM AU Sizes 1 MB – 64 MB

SVC Data Warehouse (Sequential) RAID-5 MBPS - 100% Reads and 0% Writes - ASM AU Sizes 1MB - 64MB

DS8300 Data Warehouse (Sequential) RAID-5 MBPS Read/Write 70%/30% - ASM AU Sizes 1 MB – 64 MB

DS8300 Data Warehouse (Sequential) RAID-5 MBPS - 70% Reads and 30% Writes - ASM AU Sizes 1MB - 64MB

SVC Data Warehouse (Sequential) RAID-5 MBPS Read/Write 70%/30% - ASM AU Sizes 1 MB – 64 MB
SVC Data Warehouse (Sequential) RAID-5 MBPS - 70% Reads and 30% Writes - ASM AU Sizes 1MB - 64MB

Outstanding Async I/O’s

DS8300 Data Warehouse (Sequential) RAID-5 MBPS Read/Write 50%/50% - ASM AU Sizes 1 MB – 64 MB

SVC Performance Configuration Guidelines for Implementing Oracle Databases with ASM

http://www.ibm.com/support/techdocs

© Copyright IBM Corporation, 2009

Version June 8, 2009
SVC Data Warehouse (Sequential) RAID-5 MBPS - 50% Reads and 50% Writes - ASM AU Sizes 1MB - 64MB

DS8300 Data Warehouse (Sequential) RAID-5 MBPS Read/Write 30%/70% - ASM AU Sizes 1 MB – 64 MB

SVC Data Warehouse (Sequential) RAID-5 MBPS Read/Write 30%/70% - ASM AU Sizes 1 MB – 64 MB
SVC Data Warehouse (Sequential) RAID-5 MBPS - 30% Reads and 70% Writes - ASM AU Sizes 1MB - 64MB

DS8300 Data Warehouse (Sequential) RAID-5 MBPS Read/Write 0%/100% - ASM AU Sizes 1 MB – 64 MB

SVC Data Warehouse (Sequential) RAID-5 MBPS Read/Write 0%/100% - ASM AU Sizes 1 MB – 64 MB

SVC Performance Configuration Guidelines for Implementing Oracle Databases with ASM
http://www.ibm.com/support/techdocs
© Copyright IBM Corporation, 2009
SVC Data Warehouse (Sequential) RAID-5 MBPS -
0% Reads and 100% Writes - ASM AU Sizes 1MB -
64MB

Outstanding Async I/O's

MBPS

ASM_1M
ASM_2M
ASM_4M
ASM_8M
ASM_16M
ASM_32M
ASM_64M
12 Summary

This paper has discussed performance configuration guidelines that should be followed when configuring IBM SAN Volume Controller for running Oracle with Automatic Storage Management. The paper has shown the following:

- the performance benefits that result from using SVC with a back-end DS8300 in an Oracle ASM environment as opposed to using the DS8300 directly
- the performance effect that back-end DS8300 RAID-10 and RAID-5 have on the SVC when running ASM
- the effects that ASM and SVC striping, whether used individually or in conjunction with each other, have on Oracle ASM performance
- the effect that various ASM Allocation Unit sizes have on sequential I/O for both native DS8300 and SVC-with-backend-DS8300 configurations
13 Appendix A: Oracle I/O profile scripts

The following SQL*Plus scripts can be run on an Oracle database to determine the I/O profile.

This first script is from the Oracle OpenWorld 2007 presentation cited in the References and which originated from Luca Canali at CERN. It can be run on Oracle Database 10g and above. If it is run on a RAC database, it will automatically sum the results for all member instances. It has been modified for formatting and to include the read/write ratios:

```
set lines 120
set pages 9999
col Begin_Time heading 'Begin Time' justify center
col End_Time heading 'End Time' justify center
col Physical_Read_IOPS heading 'Read IOPS' justify center
col Physical_Read_IOPS format 99,999
col Physical_Write_IOPS heading 'Write IOPS' justify center
col Physical_Write_IOPS format 99,999
col IOPS_Read_Percentage heading 'IOPS Read|Percentage' justify center
col Physical_Read_Total_Bps heading 'Read Bytes|per Second' justify center
col Physical_Read_Total_Bps format 9,999,999,999
col Physical_Write_Total_Bps heading 'Write Bytes|per Second' justify center
col Physical_Write_Total_Bps format 9,999,999,999
col Bps_Read_Percentage heading 'Bps Read|Percentage' justify center

alter session set nls_date_format='MM/DD/YYYY HH24:MI';
spool io_profile.lst
select min(begin_time) Begin_Time,
       max(end_time) End_Time,
       sum(case metric_name
           when 'Physical Read Total IO Requests Per Sec'
           then round(average) end) Physical_Read_IOPS,
       sum(case metric_name
           when 'Physical Write Total IO Requests Per Sec'
           then round(average) end) Physical_Write_IOPS,
       round((sum(case metric_name
                      when 'Physical Read Total IO Requests Per Sec'
                      then round(average) end) * 100) /
              (sum(case metric_name
                      when 'Physical Read Total IO Requests Per Sec'
                      then round(average) end) +
               sum(case metric_name
                      when 'Physical Write Total IO Requests Per Sec'
                      then round(average) end))) IOPS_Read_Percentage,
       sum(case metric_name
           when 'Physical Read Total Bytes Per Sec'
           then round(average) end) Physical_Read_Total_Bps,
       sum(case metric_name
           when 'Physical Write Total Bytes Per Sec'
           then round(average) end) Physical_Write_Total_Bps,
       -- Additional scripts for analyzing other metrics...
```
The following SQL*Plus script is from James Koopmann at [http://www.jameskoopmann.com/?p=9](http://www.jameskoopmann.com/?p=9). It has been modified to work properly on a RAC database and also for some content and formatting:

```sql
round((sum(case metric_name
    when 'Physical Read Total Bytes Per Sec'
    then round(average) end) * 100) /
  (sum(case metric_name
    when 'Physical Read Total Bytes Per Sec'
    then round(average) end) +
  sum(case metric_name
    when 'Physical Write Total Bytes Per Sec'
    then round(average) end))) Bps_Read_Percentage
from dba_hist_sysmetric_summary
  group by snap_id
  order by snap_id;
spool off

--- This script is modified for formatting and some content
--- from the following:
--# Script : wrh_sysstat_ioworkload_ALL.sql
--# Author : thecheapdba
--# Tested : Oracle 10.2
--# Version : 2007/08/07
--# Purpose : Report on IOPS & MBPS over a period of time as seen by DB.
--# ----------------------------------------------------------------------
--# NOTES : cut and paste into a worksheet for graphing
--# ----------------------------------------------------------------------
```

set echo off
set feedback off
set linesize 120
set pagesize 55
set verify off
set termout off
column rpt new_value rpt
  select name || '_wrh_sysstat_ioworkload.lst' rpt
    from v$database;
set termout on
prompt
prompt ^^^^^^^^^^^^^
prompt Report Name : &&rpt
prompt ^^^^^^^^^^^^^
spool &&rpt
column end_time format a20
column end_time justify center head "Snap End Time"
column beg_id format 9,999
column beg_id justify center head "Snap ID"
column instance_name format A9
column instance_name justify center head "Instance"
column sri justify center head "Small|Read|IOPS" noprint
column swi justify center head "Small|Write|IOPS" noprint
column tsi justify center head "Total|Small IOPS"
column srp justify center head "Read I/O%|Small IOPS"
column swp justify center head "Small|Write I/O%" noprint
column lri justify center head "Large|Read IOPS"
column lwi justify center head "Large|Write IOPS" noprint
column tli justify center head "Total|Large IOPS"
column lrp justify center head "Read I/O%|Large IOPS"
column lwp justify center head "Large|Write I/O%" noprint
column tr justify center head "Total|Read MBPS"
column tw justify center head "Total|Written|MBPS"
column tm justify center head "Total MBPS"

SELECT to_char(end_time, 'MM/DD/YYYY HH24:MI:SS') end_time,
       beg_id,
       instance_name,
       ROUND(sr/inttime, 3) sri,
       ROUND(sw/inttime, 3) swi,
       ROUND((sr+sw)/inttime, 3) tsi,
       ROUND(sr/DECODE((sr+sw), 0, 1, (sr+sw))*100, 3) srp,
       ROUND(sw/DECODE((sr+sw), 0, 1, (sr+sw))*100, 3) swp,
       ROUND(lr/inttime, 3) lri,
       ROUND(lw/inttime, 3) lwi,
       ROUND((lr+lw)/inttime, 3) tli,
       ROUND(lr/DECODE((lr+lw), 0, 1, (lr+lw))*100, 3) lrp,
       ROUND(lw/DECODE((lr+lw), 0, 1, (lr+lw))*100, 3) lwp,
       ROUND((tbr/inttime)/1048576, 3) tr,
       ROUND((tbw/inttime)/1048576, 3) tw,
       ROUND(((tbr+tbw)/inttime)/1048576, 3) tm
FROM (SELECT beg.snap_id beg_id,
          beg.instance_number,
          beg.instance_name,
          end.snap_id end_id,
          beg.begin_interval_time,
          beg.end_interval_time,
          end.begin_interval_time begin_time,
          end.begin_interval_time end_time,
          (extract(day  from (end.end_interval_time - end.begin_interval_time)) * 86400) +
          (extract(hour  from (end.end_interval_time - end.begin_interval_time)) * 3600) +
          (extract(minute from (end.end_interval_time - end.begin_interval_time)) * 60) +
          (extract(second from (end.end_interval_time - end.begin_interval_time)) * 01) inttime,
          decode(end.startup_time, end.begin_interval_time, end.sr, (end.sr - beg.sr))  sr,
          decode(end.startup_time, end.begin_interval_time, end.sw, (end.sw - beg.sw))  sw,
          decode(end.startup_time, end.begin_interval_time, end.lr, (end.lr - beg.lr))  lr,
          decode(end.startup_time, end.begin_interval_time, end.lw, (end.lw - beg.lw))  lw,
          decode(end.startup_time, end.begin_interval_time, end.tbr, (end.tbr - beg.tbr)) tbr,
          decode(end.startup_time, end.begin_interval_time, end.tbw, (end.tbw - beg.tbw)) tbw
FROM (SELECT dba_hist_snapshot.snap_id,
          gv$instance.instance_number,
          gv$instance.instance_name,
          dba_hist_snapshot.startup_time,
          begin_interval_time,
          end_interval_time,
          begin_interval_time begin_time,
(SELECT
  decode(stat_name, 'physical read total IO requests', value, 0) -
  decode(stat_name, 'physical read total multi block requests', value, 0) sr,
  decode(stat_name, 'physical write total IO requests', value, 0) -
  decode(stat_name, 'physical write total multi block requests', value, 0) sw,
  decode(stat_name, 'physical read total bytes', value, 0) tbr,
  decode(stat_name, 'physical write total bytes', value, 0) tbw
FROM wrh$_sysstat, wrh$_stat_name, dba_hist_snapshot, gv$instance
WHERE wrh$_sysstat.stat_id = wrh$_stat_name.stat_id
  AND wrh$_sysstat.snap_id = dba_hist_snapshot.snap_id
  AND wrh$_sysstat.instance_number = dba_hist_snapshot.instance_number
  AND gv$instance.instance_number = dba_hist_snapshot.instance_number
GROUP BY dba_hist_snapshot.snap_id, gv$instance.instance_number,
  gv$instance.instance_name, dba_hist_snapshot.startup_time,
  begin_interval_time, end_interval_time)
  beg,
(SELECT
  decode(stat_name, 'physical read total IO requests', value, 0) -
  decode(stat_name, 'physical read total multi block requests', value, 0) sr,
  decode(stat_name, 'physical write total IO requests', value, 0) -
  decode(stat_name, 'physical write total multi block requests', value, 0) sw,
  decode(stat_name, 'physical read total bytes', value, 0) tbr,
  decode(stat_name, 'physical write total bytes', value, 0) tbw
FROM wrh$_sysstat, wrh$_stat_name, dba_hist_snapshot, gv$instance
WHERE wrh$_sysstat.stat_id = wrh$_stat_name.stat_id
  AND wrh$_sysstat.snap_id = dba_hist_snapshot.snap_id
  AND wrh$_sysstat.instance_number = dba_hist_snapshot.instance_number
  AND gv$instance.instance_number = dba_hist_snapshot.instance_number
GROUP BY dba_hist_snapshot.snap_id, gv$instance.instance_number,
  gv$instance.instance_name, dba_hist_snapshot.startup_time,
  begin_interval_time, end_interval_time)
  end
WHERE beg.snap_id + 1 = end.snap_id
  AND beg.instance_number = end.instance_number
)
order by 2, 3;
14 Appendix B: References

1. Implementing the IBM System Storage SAN Volume Controller V4.3
   http://www.redbooks.ibm.com/abstracts/sq246423.html?Open


3. IBM System Storage SAN Volume Controller resource web page
   http://www-03.ibm.com/systems/storage/software/virtualization/svc/

4. IBM System Storage SAN Volume Controller
   Command-Line Interface User’s Guide
   Version 4.3.1

5. IBM DS8000 Performance Configuration Guidelines for Implementing Oracle Databases with
   Automatic Storage Management (ASM)
   http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP101375

6. IBM System storage DS8000 Series: Architecture and Implementation

7. Implementing an IBM/Brocade SAN

8. Host Attachment for SDDPCM on AIX

9. Using Device-Mapper Multipath
   Configuration and Administration 5.2

10. Configuring Device Mapper Multipath Tool (DMMP) for hosts running the Linux operating system
    http://publib.boulder.ibm.com/infocenter/svicic/v3r1m0/index.jsp?topic=/com.ibm.storage.svc.conso
ele.doc/svc_dmmpsetuptoolsinux431_3yd13v.html

11. Monitoring Your Storage Using TotalStorage Productivity Center

12. Performance Metrics in TotalStorage Productivity Center Performance Reports

13. Oracle ORION download site (includes ORION User's Guide)

14. Oracle Database Storage Administrator’s Guide
    11g Release 1 (11.1)
Part Number B31107-04
http://download.oracle.com/docs/cd/B28359_01/server.111/b31107/toc.htm

15. Oracle Automatic Storage Management
Under-the-Hood & Practical Deployment Guide
Implementing Robust Storage Solutions with Oracle ASM
Nitin Vengurlekar, Murali Vallath, Rich Long
Oracle Press
ISBN 978-0-07-149607-0

16. Automatic Storage Management Technical Overview
An Oracle White Paper
November 2003

17. Oracle OpenWorld 2007 Presentation -
RAC PACK
Back-of-the-Envelope Database Storage Design
Nitin Vengurlekar

18. The Optimal Oracle Configuration: Using the Oracle ORION Workload Tool to Accurately
Configure Storage
Technical White Paper – LSI Corporation
James F. Koopmann
http://wwwlsi.com/DistributionSystem/AssetDocument/documentation/storage/business_solutions/25034-
00_RevA_Oracle_Optimal_Config.pdf

19. Oracle – Complete history of IOPS & MBPS from workload repository history (snapshots)
James Koopmann
http://www.jameskoopmann.com/?p=9
15 Trademarks and special notices

© Copyright IBM Corporation 2009. All rights Reserved.

References in this document to IBM products or services do not imply that IBM intends to make them available in every country.

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both. If these and other IBM trademarked terms are marked on their first occurrence in this information with a trademark symbol (® or ™), these symbols indicate U.S. registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of IBM trademarks is available on the Web at “Copyright and trademark information” at www.ibm.com/legal/copytrade.shtml.

The following are trademarks of the International Business Machines Corporation in the United States and/or other countries: System Storage, DS8000®, IBM®, IBM (logo)®, System p®, System x®. For a complete list of IBM Trademarks, see www.ibm.com/legal/copytrade.shtml.

Intel, Intel Inside (logos), MMX, and Pentium are trademarks of Intel Corporation in the United States, other countries, or both.

Linux is a trademark of Linus Torvalds in the United States, other countries, or both.

Other company, product, or service names may be trademarks or service marks of others.

Any performance data contained in this document was determined in a controlled environment. Actual results may vary significantly and are dependent on many factors including system hardware configuration and software design and configuration. Some measurements quoted in this document may have been made on development-level systems. There is no guarantee these measurements will be the same on generally-available systems. Users of this document should verify the applicable data for their specific environment.

Information is provided "AS IS" without warranty of any kind.

All customer examples described are presented as illustrations of how those customers have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics will vary depending on individual customer configurations and conditions.

Information concerning non-IBM products was obtained from a supplier of these products, published announcement material, or other publicly available sources and does not constitute an endorsement of such products by IBM. Sources for non-IBM list prices and performance numbers are taken from publicly available information, including vendor announcements and vendor worldwide homepages. IBM has not tested these products and cannot confirm the accuracy of performance, capability, or any other claims related to non-IBM products. Questions on the capability of non-IBM products should be addressed to the supplier of those products.

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.