Oracle RAC on VxFlex OS

Optimal Design and Implementation Details

Abstract

This paper describes optimal Oracle configurations and parameter settings when deployed on VxFlex OS. Deployment options, storage resiliency, performance, and data protection for production and development environments are discussed in detail.

December 2018

Revisions

Date	Description
June 2018	Initial Draft
August 2018	Illustrations
December 2018	Added Oracle RDBMS 18c patch update

Acknowledgements

This paper was produced by the following members of the Dell EMC Technical Marketing and Oracle Specialist teams.

Author: Neil Gerren, Consultant Application Architect & Technical Marketing Engineer
Additional Contributors:

Graham Thornton Principal Systems Engineer
Bart Sjerps Principal Systems Engineer
Matt Kaberlein Advisory Systems Engineer

Table of contents

Executive summary

While it is relatively easy to install Oracle RAC storage infrastructure on block-based storage, whether installed on bare metal or virtualized, the permutations of Oracle Database, Grid, ASM, and Storage System settings can seem somewhat daunting.

This document seeks to reduce those permutations, providing the most ideal settings, configurations, and deployment options. It also explains a few back-end features which contribute to storage resiliency. To further expand your understanding of VxFlex OS, you can refer to the User Guide which is included in the product documentation, which is available on support.dell.com.

Information related to VxFlex OS, it's basic deployment configurations and other storage and design considerations which are unique to this modern datacenter storage product are also included. Once you're read this paper, you'll discover there are very few differences in deploying Oracle on VxFlex OS using VxRack Flex and Ready Nodes compared to the other enterprise storage solutions you've been exposed to.

1 Document Scope

This document serves to fill in the blanks related to VxFlex OS deployment configurations when deploying VxFlex OS on VxFlex Ready Nodes. It is not intended to be a step-by-step guide for installing Oracle RAC infrastructure and assumes expert level expertise with Oracle, enterprise storage, and related server administration.

1.1 Deployment Configurations

There is a paper that is dedicated entirely to VxFlex OS Architecture. It's entitled "Dell EMC Flex family and VxFlex OS: The Software Behind the Systems" and can be found here: https://www.dellemc.com/resources/en-us/asset/white-papers/products/converged-infrastructure/DellEMC_VxFlex_OS_Software_Defined_Server_SAN.pdf.

The VxFlex OS storage network is completely TCP/IP based, eliminating the need for any specialized fiber channel networks and switches. Access to data is accomplished through SCSI block-based devices, just like a SAN, but the transport layer is TCP/IP-based. VxFlex OS does not use iSCSI or Fiber Channel storage protocols.

The most important consideration when deploying VxFlex OS is the network. When deploying on VxRack FLEX, the network is already built for you. For other VxFlex OS family solutions, such as VxFlex Ready Nodes, refer to the VxFlex OS Best Practices and Design Considerations white paper. It can be found here: http://www.emc.com/collateral/white-papers/h17332-dellemc-vxflex-os-networking-best-practices.pdf

1.2 The VxFlex OS Logical software components

There are three major logical components within a VxFlex OS cluster.

1.2.1 The Metadata Manager (MDM)

Multiple server nodes are clustered using the VxFlex OS Metadata Server software agent, or MDM. The fundamental metadata server design is that of primary server, one or more standby metadata management servers (to which metadata is concurrently replicated), and one or more tie-breaker nodes. The metadata database is accessible to the other two major VxFlex OS software components which provide and consume storage.

1.2.2 The 3-Node MDM Cluster

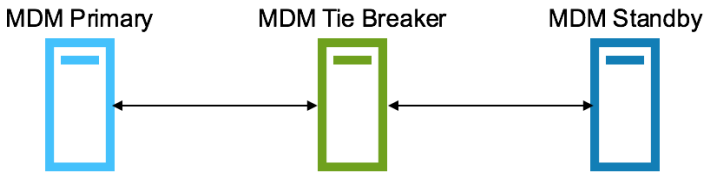

Figure 1 The 3-Node MDM Cluster

VxFlex OS requires a minimum of three nodes to avoid the split-brain problem, preventing simultaneous active Primary MDMs. The Tie Breaker communicates with both the Primary and Standby MDM servers. If the Primary MDM fails to respond, the Standby is promoted by the Tie Breaker.

1.2.3 The 5-Node MDM Cluster

A consideration driving availability is the number of servers comprising the VxFlex OS MDM cluster. While a three node cluster is sufficient for an operational storage system, a five node cluster provides availability meeting or exceeding six-nines. Therefore, development and ancillary systems are better suited for three node clusters, while production systems are best utilizing five nodes.

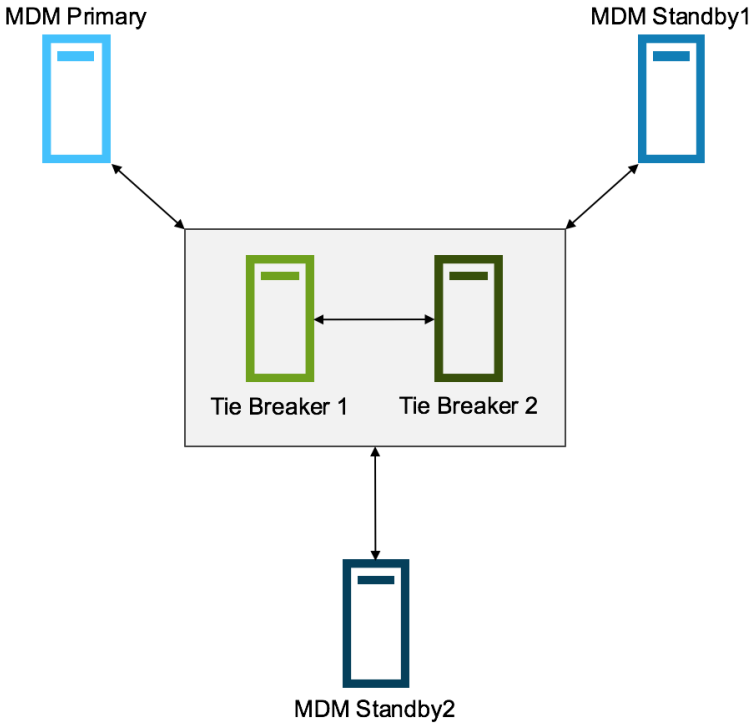

Figure 2 The 5-Node MDM Cluster

Five node clusters include a primary active MDM node, two standby nodes, and two tie-breaker nodes; one active, and one standby. In this configuration, the Tie Breaker nodes not only monitor the Primary and Secondary MDM nodes, but also the other Tie Breaker.

1.2.4 Storage Data Server (SDS)

The second logical component is the data server agent, or Storage Data Server known as the SDS. This service provides access to media devices (SSD, NVMe, HDD) residing on individual servers, combining them into a pool of available storage. For the purposes of availability, there must be at least two data server nodes, each with sufficient capacity to store all the primary and replica copies of data in the system. Best practices actually lead us to using three nodes, since the MDM cluster requires three. As the system grows with additional drives or servers, it can scale up to 1024 SDS systems. This approach enables very fine granular storage growth as your data footprint grows.

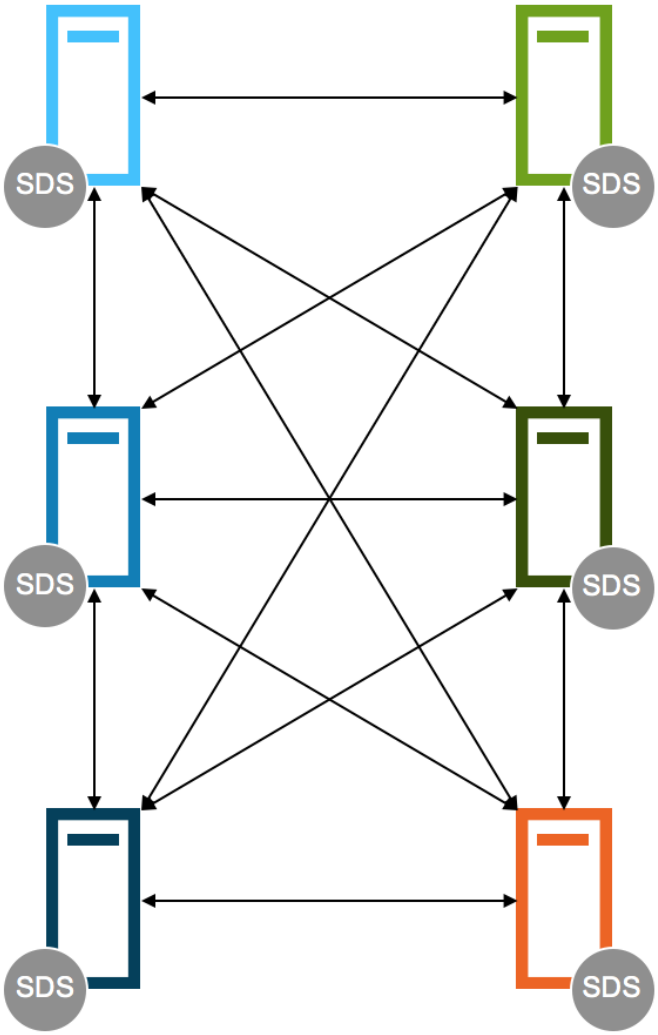

Figure 3 The SDS Cluster

One key responsibility of the SDS is rebuilding content of failed media. Referring to Figure 3 above, you can see that as you add SDS servers, the number of data paths between SDS nodes grows.

Any given device contains both primary and replica data. In the event of a device or node failure, replica data stored on other devices become primary data providers, and that primary data is then replicated across all the remaining devices on all remaining nodes. Therefore, the more SDS nodes we have, the less time device rebuild operations take. These operations typically complete within a few minutes, or even less than a minute depending on the node count. The SDS Service can run on MDM cluster nodes or other nodes not participating in an MDM Cluster.

1.2.5 The Storage Data Client (SDC)

The last major software component is the client agent, or Storage Data Client (SDC) which is installed on any system having access to the storage network that needs to consume storage. The SDC has access to the storage metadata. The MDM is not in the IO path. Instead, the metadata map relevant to each SDC host is maintained within the SDC memory structures.

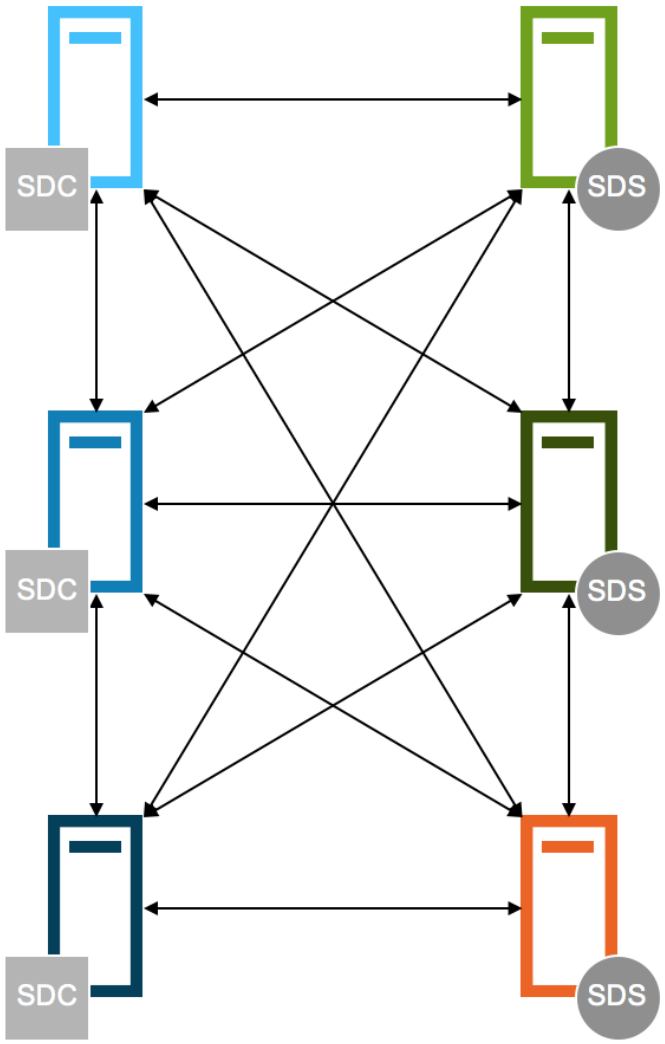

Figure 4 SDC Communications

We can see from Figure 4 that data access on the SDC is facilitated through the SDS, and again, the more SDS nodes we have, the better the system will perform.
Traditional multipath devices are not used, so the multipath daemon can be disabled. Instead, device availability is provided via network bonding, and/or the network layer of VxFlex OS itself, providing network load balancing and high availability across multiple network interfaces.
The SDC can co-reside with the MDM and the SDS services, or run completely independently. We'll see more detail about various configurations in the next section.

1.3 VxFlex OS Deployment Options

There are two major options when deploying a VxFlex storage cluster: Hyper-Converged and 2-Layer.

1.3.1 Hyper-Converged Infrastructure

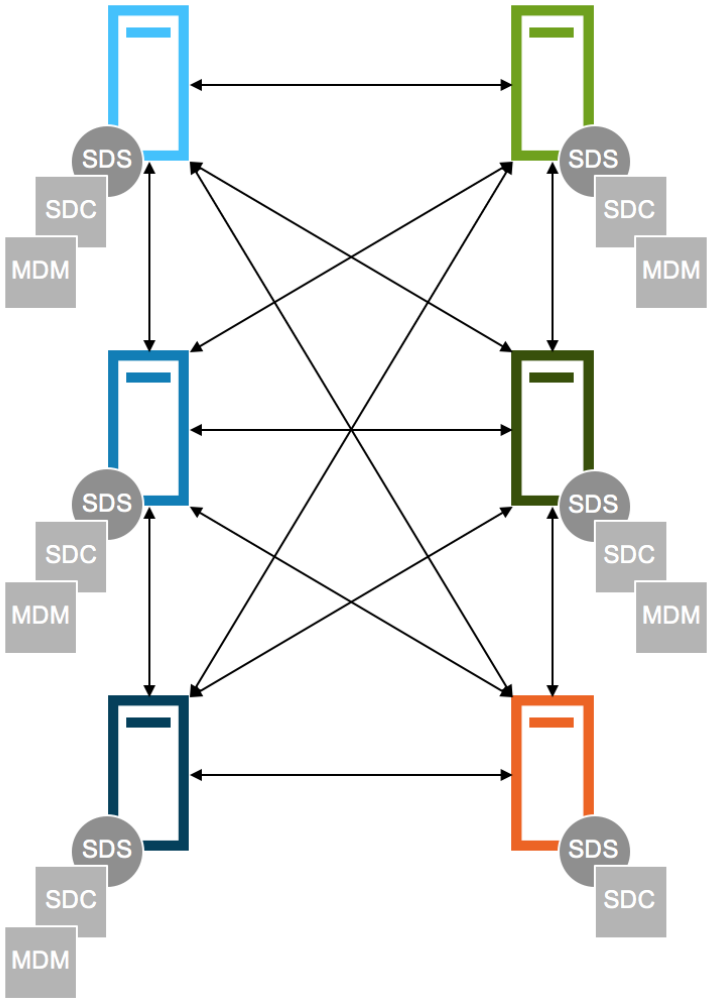

Figure 5 The HCI or CI Deployment

HCI involves a hypervisor providing storage virtual machines and client VMs. Our definition of HCI dictates that the entire stack is virtualized. All server nodes provide and consume storage and participate in the hypervisor cluster, and storage is provided via purpose-built Storage Virtual Machines. When we have more than five nodes in this sort of cluster, only five of them participate in the MDM cluster. The remaining nodes participate in consuming and providing storage via the SDC and SDS clusters respectively. CI, or Converged Infrastructure, which eliminates the hypervisor can also run in this configuration. In this case, Linux or Windows is installed in lieu of a hypervisor.

1.3.2 2-Layer Infrastructure

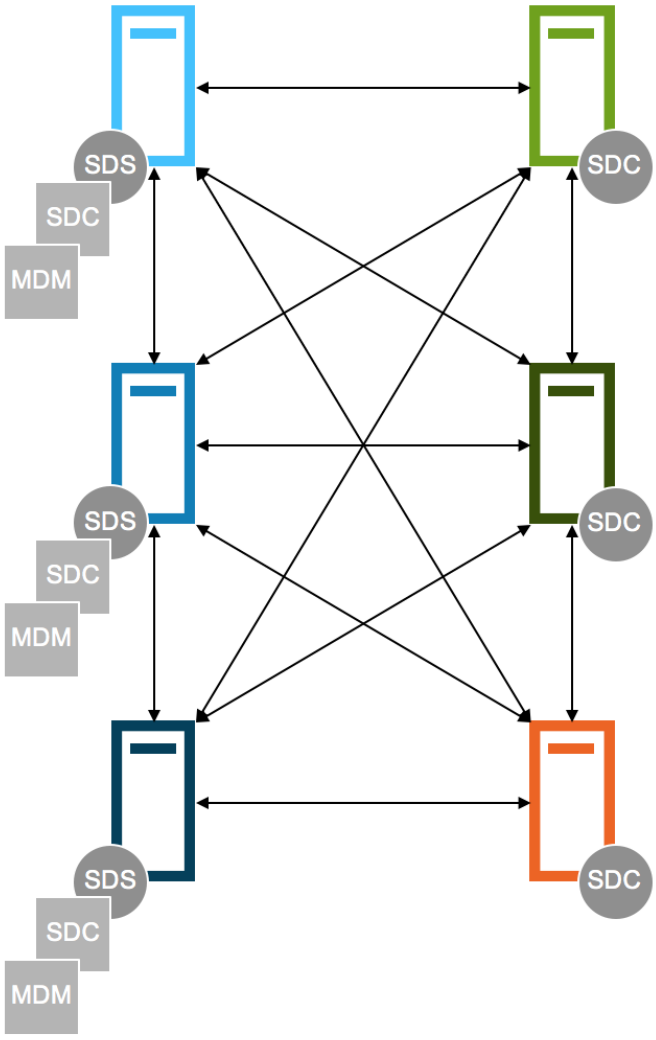

Figure 6 The Two-Layer Deployment

In Figure 6, we have a 3-Node Linux Bare Metal cluster providing storage to three nodes not participating in storage. Consider again, in the case of a clustered application like Oracle RAC, that as we add storage and client nodes, we also increase the number of data paths, thus improving performance of all IO operations.

Why would we elect to use the two-layer deployment in lieu of Hyper-Converged? The most popular use case is in situations where there is a distinct division between storage management and system administration, and related budgets. In this scenario, compute and storage are managed separately.

Why do some customers use a Converged Infrastructure deployment? They start by deploying a pure two-layer storage-only cluster, but eventually become more familiar with the product and its CPU consumption levels. With CPU utilization of VxFlex OS typically running under 10%, they recognize that there is excess CPU capacity available, so they install the SDC on the system storage cluster and deploying applications there to better leverage their hardware investment.

1.4 Performance and Availability Relating to the VxFlex OS Software Components

For solutions demanding the upmost performance and availability, (>2GB/s or >150,000 IOPs per storage server or HCI node) multiple network interfaces are required. Network switches should offer at least 10GbE throughput per port. Network port HA and load balancing can be provided by VxFlex OS natively, or via LACP bonding. For best availability, LACP bonding is not recommended for the network interconnects for the metadata server nodes. Native VxFlex OS bonding should be utilized for that purpose instead, due to the fact that higher LACP fail-over times will unnecessarily force meta-data server fail-over.

When performance is paramount, it is wise to consider dedicating at least a pair of network adapters to storage back-end processing, and another pair to storage front-end and Oracle Grid interconnects. Also, for obvious reasons, SSD or NVMe media are required to drive optimal IO performance.

Testing has confirmed that provisioning four to eight storage volumes per ASM disk group improves client performance, allowing more simultaneous client IO threads with less latency, particularly on Linux. On AIX, IO continues improving well beyond eight volumes per disk group. This holds true for traditional SAN environments as well, so this best practice is not unique to VxFlex OS.

What does VxFlex do better than traditional block-based storage arrays? It SCALES more easily, and effectively, providing IO throughput and operations per second that is extremely difficult to match at the same price point. It's easy to grow the cluster as your workload grows by adding partially or fully populated servers. It's much easier to maintain the storage software, server firmware, and server hardware. And, as the storage cluster grows, performance grows nearly linearly and resiliency increases as well. The CPU efficiency of the VxFlex OS storage cluster is outstanding, with CPU utilization generated by the storage cluster averaging in the single digits.

2 VxFlex OS Optimizations

Whenever installing and configuring VxFlex OS, consult the VxFlex OS Performance Fine-Tuning Technical Notes included in the documentation located on support.dell.com. The document contains pre and post-install recommendations. The recommendations include networking and OS configurations and settings for servers and clients for all approved deployment configurations including 2-layer and virtualized installation.

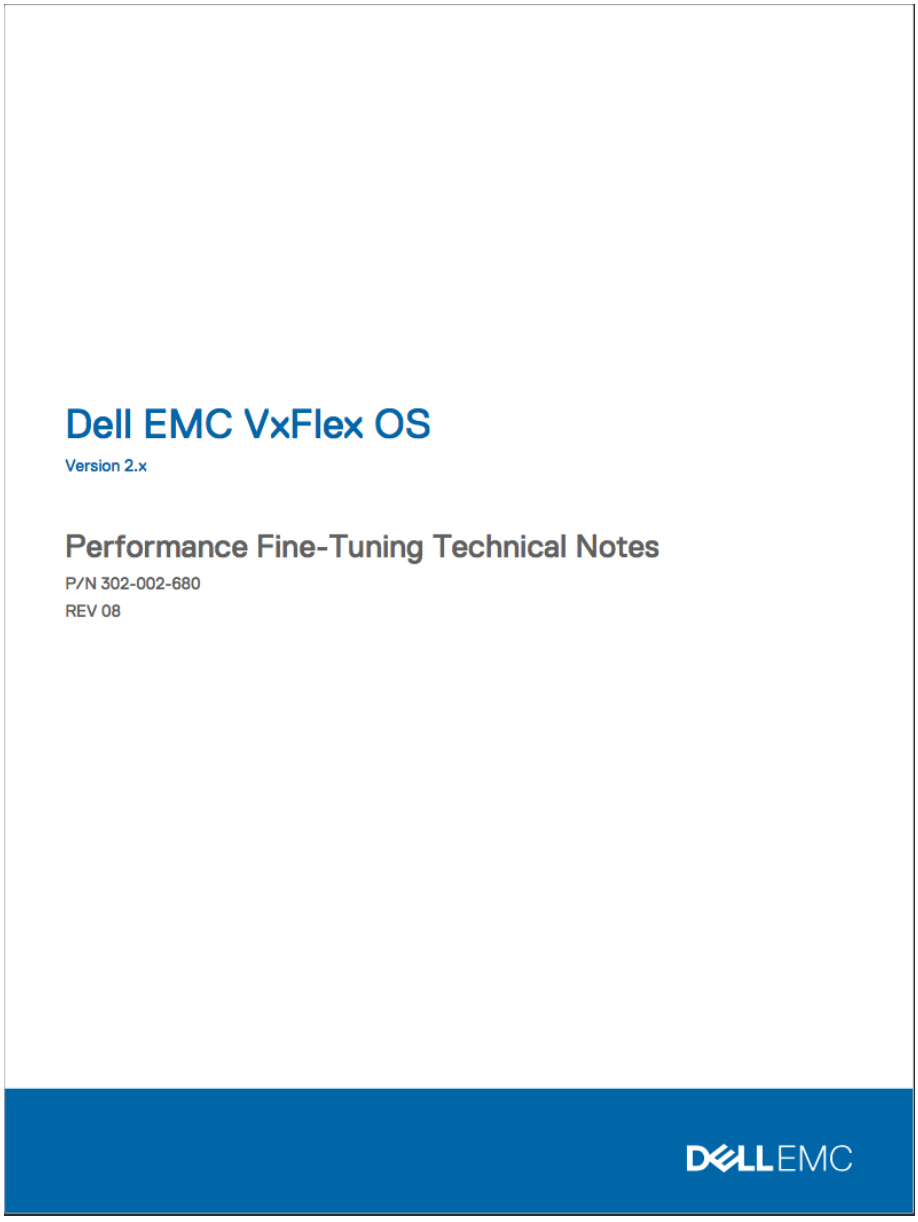

Figure 7 Fine-Tuning Guide

3 ASM and related Oracle Database Considerations

3.1.1 ASM Disk Redundancy

Select external disk redundancy. Just as with any self-protecting block storage, use external redundancy. VxFlex OS optimally replicates writes to secondary locations, eliminating the need for ASM redundancy.

3.1.2 ASM Disk Count

Use 4-8 logical VxFlex OS volume devices for Windows and Linux. AIX supports more IO threads and scales linearly beyond 8 devices per disk group. This is no different than any other SAN-like solution.

3.1.3 ASM Allocation Unit (AU) Size

ASM disks are divided into Allocation Units. Data within a file residing on ASM is broken up into extents. As the AU size increases, the extent count decreases, reducing the number of physical IOs required when performing IO operations. For Data warehouses, 8MB or 16MB AU size is suggested. For OLTP, 1-4MB is sufficient.

3.1.4 Stripesize setting

This parameter in release 10g and greater can have an impact on performance of temp and redo IO. Set the database template of the striped colum for the online log and tempfile to “fine” for better throughput. (Reference the V$ASM_TEMPLATE view, and the “alter diskgroup x alter template” database management commands.)

3.1.5 Use of ASMLIB and UDEV

Either approach is valid to ensure that grid user access to VxFlex OS devices survives a reboot. It is suggested to utilize ASMLIB when selecting Oracle Linux or UEK. For all other OS choices, use UDEV. Details for using UDEV will appear later in this document.

3.1.6 Redo Logs

Make certain you have sufficient redo log groups to avoid any check-pointing errors. It is recommended to maintain at least five log groups per database thread.

3.1.7 Async IO

Set disk_asynch_io=true in cases where Log File Sync activities show up in the top events in AWR.

3.1.8 Filesystemio_options

As with any other block storage solution, be sure to set filesystemio_options=setall.

3.1.9 Hugepages

For optimal memory management, enable hugepages in linux, and disable Automatic Memory Management in the pfile/spfile.

3.1.10 Other Database Parameter Settings

As with any other block storage solution, the following settings are recommended:

archive_lag_target=900
db_block_checksum=FALSE
db_block_checking=FALSE
db_block_size=8192 (or 16384 for OLAP or DW)
db_file_multiblock_read_count – do not set!
db_writer_processes=(CPU count/4)
fast_start_mttr_target=120
_db_block_prefetch_limit=0
_db_block_prefetch_quota=0
_db_file_noncontig_mblock_read_count=0
recyclebin=off

4 VxFlex OS Devices

4.1 Disk Access Persistence

ASMLIB, UDEV, and the ASM Filter driver are all supported, although, we do recommend only using the ASM Filter Driver with 12cR2 or greater. Use the solution that best meets your business and technical needs. ASMLIB is generally aligned with Oracle Linux and/or the Unbreakable Kernel, while Udev is customarily used with other Unix distributions. The new Filter Driver seeks to prevent unintentional out-of-band writes to devices allocated to your database. It also supports TRIM, and will provide integration with other storage APIs as they develop.

Bart Sjerps has developed something he calls asmdisks which simplifies the use of UDEV, generating the UDEV rules for you. It can be found at:

http://outrun.nl/wiki/ASMDisks

Using a simple command of the form: `asm createdisk [DISKNAME] [device]` creates the UDEV rule in the correct location for you, eliminating the opportunity for human error.

4.2 Multipathing

There is no need to be concerned with legacy multipathing. The VxFlex OS Storage Data Client (SDC) provides fault-tolerant access to logical volumes via the IP networking layer as long as there are more than one network interface available.

4.3 Partitioning Your Devices

VxFlexOS block devices appear in the form: /dev/scini?

To properly align your devices, partition them as follows:

[root@node1 /root]# *parted /dev/scinib mklabel gpt*

[root@node1 /root]# *parted /dev/scinib mkpart primary 2048s*

100%

4.4 UDEV Rules with Linux 7

The use of UDEV or ASMLIB is not unique to VxFlex OS. If you've deployed ASM on other block-based storage offerings, you're already familiar with them. You probably already know that they're important for maintaining access of the grid or database user to disk devices, and

preserving that access after the database server is rebooted. If you choose to utilize UDEV, the following describes the steps required to configure them with VxFlex OS.

Use the lsblk command to identify the desired VxFlexOS devices and their corresponding capacity.

As a prerequisite, it is assumed you've already created the grid (or database) user and added it to the asmadmin group.

VxFlex OS volume devices appear in the form: /dev/scini?. You'll see that as you proceed through the steps below:

1. Choose the scini device you wish to use for ASM, and use drv_cfg command to determine it's GUID identifier:

   [root@node1 rules.d]# */opt/emc/scaleio/sdc/bin/drv_cfg --query_block_device_id --block_device /dev/scinib1*

   5506082f2c48a3ef-c62408e300000008
2. Add a UDEV rule as follows, using the displayed GUID identifier, placing the below contents in the /etc/udev/rules.d/9-asm-devices.rules file, making certain the correct device alias, owner and group are entered:

   KERNEL=="scini*", SUBSYSTEM=="block", PROGRAM="/opt/emc/scaleio/sdc/bin/drv_cfg --query_block_device_id --block_device /dev/%k", RESULT=="5506082f2c48a3ef-c62408e300000008",SYMLINK+="oracleasm/disks/OCR1", OWNER="grid", GROUP="asmadmin", MODE="0660"

3. Repeat steps 1 and 2 for all desired devices.

4. Initiate a UDEV rescan:

   [root@node1 rules.d]# /sbin/udevadm control --reload-rules;/sbin/udevadm trigger

5. Verify the rules were correctly applied and provided the expected results:

   [root@node1 rules.d]# ls -la /dev/oracleasm/disks/OCR1

   lrwxrwxrwx. 1 root root 7 May 21 12:55 /dev/oracleasm/disks/OCR1 -> scinib1

   [root@node1 rules.d]# ls -l /dev/scinib1

   brw-rw----. 1 grid asmadmin 248, 17 May 21 12:55 /dev/scinib1

Following these simple instructions will have you well on your way to completing an optimal Oracle installation.

4.5 Volume Performance on VMware

When Oracle is deployed on VMware, and optimal performance is required, the use of RDM is highly recommended when provisioning VxFlex OS client volumes.

4.6 VxFlex OS Device Compatibility with Oracle

There has been a relatively chronic issue with VxFlex OS volume device validation during Oracle installation. The Oracle installer complains that VxFlex OS devices are of unknown type and un-sharable during a RAC installation. R12cR1 generates related warnings, while R12cR2 generates errors. To proceed after installation pre-checks, one is forced to ignore these errors and warnings:

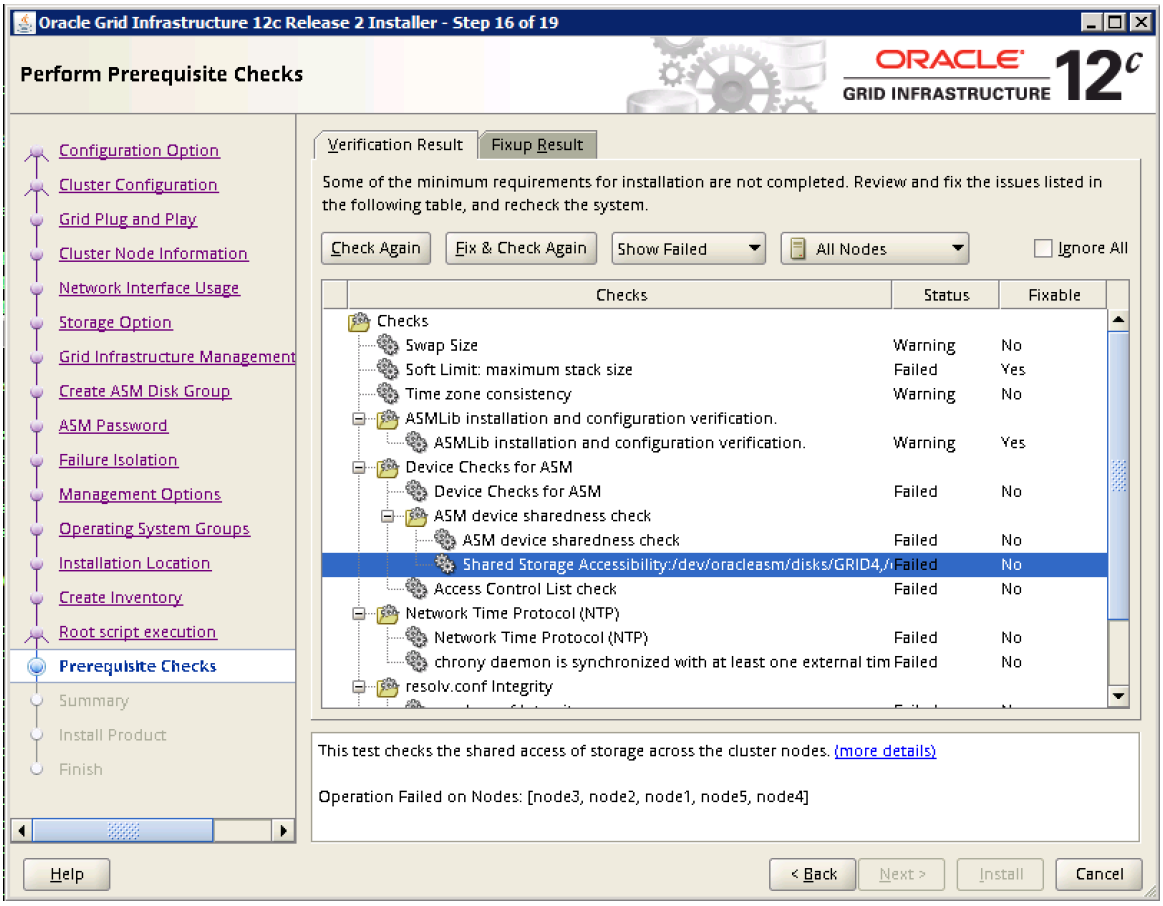

Figure 8 Pre-Install Checks

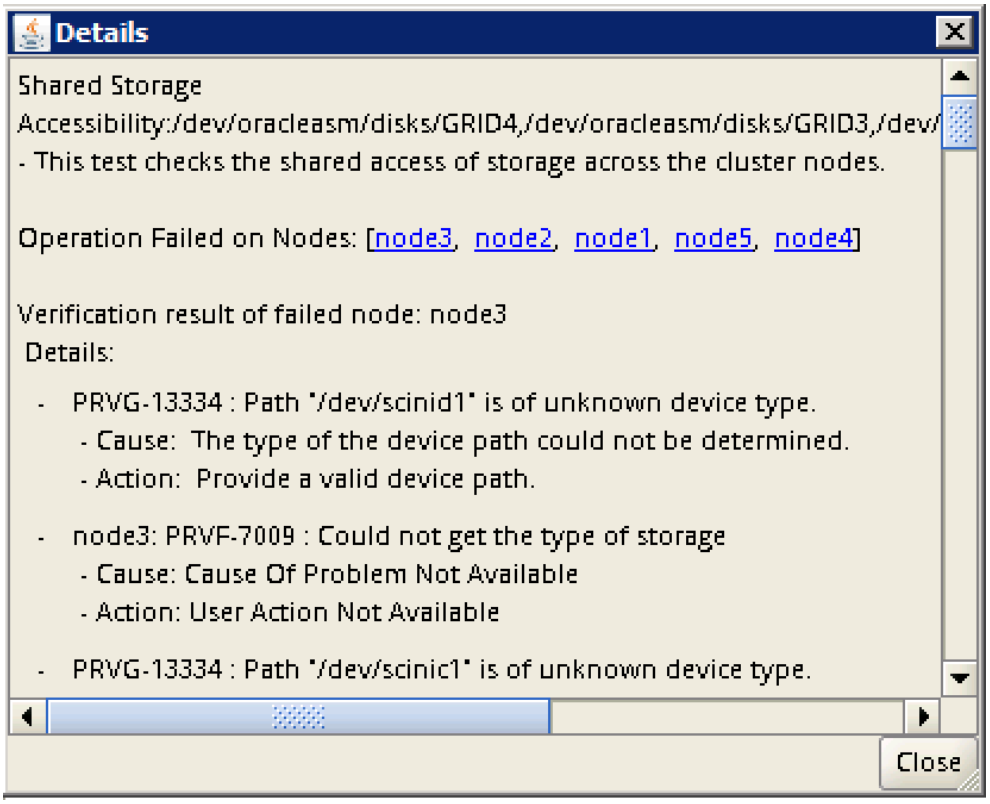

Figure 9 Pre-Install Checks detail

As of the release of this document, Oracle is working on a patch. Thus far, the error will not appear in an Oracle Linux/UEK environment using asmlib (assuming patch 25784424 is applied.) When using UDEV rules for devices, the errors still appear, and regardless of the OS and disk management method, any device checks using the cluster verification utility will produce the same errors and warnings. These errors and warnings CAN BE SAFELY IGNORED. Keep an eye on the above patch by logging onto support.oracle.com and searching the patch number. This issue will be resolved in the future. There is additional information on applying the patch later in this document.

4.7 Notes on Oracle Patch 25784424

Oracle has provided a patch to 12cR2 which can be used to address VxFlex OS device related warnings and errors when installing and configuring using the ASMLIB option. The patch can be applied during installation.

The following is an example of installing using the Grid Home method (using the installation file of the form: linuxx64_12201_grid_home.zip.)

1. Unzip the file to the desired $GRID_HOME.
2. Create a patch directory in your $GRID_HOME.
3. Change the working directory to that location.
4. Unzip the contents of the 25784424 to that location.
5. Launch the installer with the patch option:

```
$GRID_HOME/gridSetup.sh -applyPSU $GRID_HOME/patch/25784424
```

This will patch the $GRID_HOME before launching the installer. The installer pre-check will no longer complain about the scini devices.

Be aware that the patch does not fix the cluster verification utility, cluvfy. That will continue to report bad device types. (This is being further pursued with Oracle, and you can refer to the latest notes on the patch for current information.)

If you wish to use the conventional installation method, using files of the form: v46096.zip, refer to Oracle Support Document 1410202.1 for specific patching information.

There is no point in patching existing installations. Since the existing environments actually function properly and the database is running, and given the patch does not fix the cluster verification utility, there is no benefit to applying the patch in that case.

4.8 Updated 18c patch

Oracle has provided a patch to 18cR1 which can be used to address VxFlex OS device related warnings and errors when installing and configuring using the ASMLIB option. The patch can be applied during installation. This differs from the 12cR2 patch in that it also addresses the cluvfy bug:

```
[grid@node1 grid]$ ls -l /dev/scini*
crw------- 1 root root     238,  0 Dec 12 15:10 /dev/scini
brw-rw---- 1 grid asmadmin 253, 16 Dec 12 15:10 /dev/scinib
brw-rw---- 1 grid asmadmin 253, 32 Dec 12 15:10 /dev/scinic
brw-rw---- 1 grid asmadmin 253, 48 Dec 12 15:10 /dev/scinid
brw-rw---- 1 grid asmadmin 253, 64 Dec 12 15:10 /dev/scinie
brw-rw---- 1 root disk     253, 80 Dec 18 21:52 /dev/scinif
[grid@node1 grid]$ ./runcluvfy.sh comp ssa -asm -asmdev /dev/scinib

Verifying Shared Storage Discovery ...
  Disk                                  Sharing Nodes (1 in count)
  ------------------------------------  ------------------------
  /dev/scinib                           node1
Verifying Shared Storage Discovery ...PASSED

Verification of shared storage accessibility was successful.

CVU operation performed:      shared storage accessibility
Date:                         Dec 20, 2018 3:49:37 PM
CVU home:                     /u01/app/18.0.0/grid/
User:                         grid
```

Follow the procedure outlined in section 4.7 to install the patch, but use the 18c grid home file and patch.

5 Data Protection

5.1 Backup Utilities

VxFlexOS can be protected with any backup utility, just like any other block-based storage solution. In the case of Oracle, RMAN is recommended. There is a RecoverPoint Appliance solution available as well. Contact your sales team for more information.

5.2 Snapshots

VxFlex OS also offers volume based writeable snapshots. Snapshots can be initiated via consistency groups as seen here:

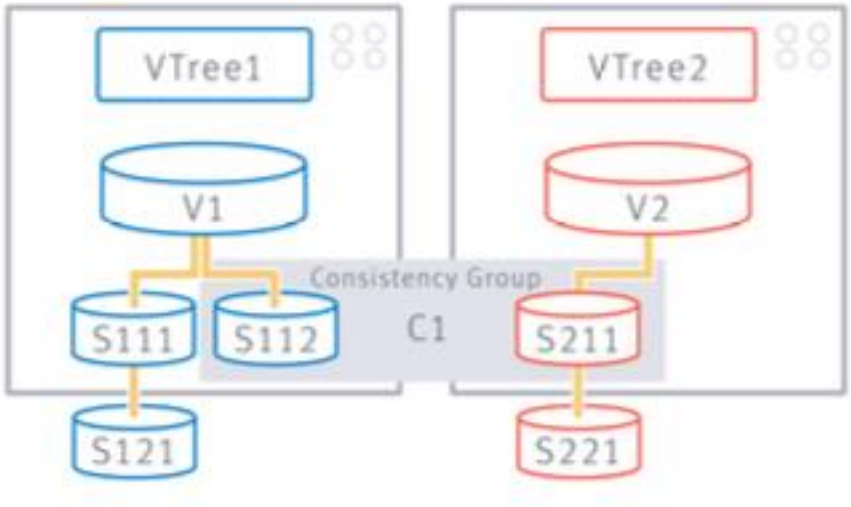

Figure 10 VxFlex OS Snapshots

As seen above, snapshots can cascade, providing the ability to create gold clones of your database.

5.2.1 Volume Layout for Snapshot-based Backups and Database Cloning

In order to facilitate snapshot-based backups, you must create a volume layout that separates data file, redo log, and archive log/flash recovery content (FRA), so two or three volumes must be used. When using multiple storage volumes for each ASM disk group, all the volumes in a given disk group must be part of the consistency group.

5.2.2 Snapshot Backup Workflow

For crash-consistent snapshot-based backups, initiate a consistency group snapshot of all database content. However, for point-in-time recoverable clones, one must follow traditional backup processing, treating the snapshots like any other backup media:

1. Quiesce: “Alter database begin backup;”
2. Initiate the consistency group snapshot of the volumes containing data files.
3. Un-Quiesce: “Alter database end backup;”
4. Preserve your control file into your archive logging location: “alter database backup controlfile to …”
5. Initiate a snapshot of your archive logging location

Conclusion

In retrospect, when comparing database layout and configuration, there is very little difference between building your database on traditional SAN verses VxFlex OS storage.
The real difference is in the flexibility of the deployment of your storage. You'll have more options to select a storage configuration that best balances business needs and cost. For environments requiring non-production activities, storage system node count can be reduced, and for production environments, the increased node count offers greater system performance and resiliency. Also, as you learn more about the VxFlex family of solutions, you'll discover how easy it is to maintain software and firmware certification. You'll also learn that in working with VxFlex OS, the days of the forklift upgrade are over.

There are only a few steps in this guide that are unique to using VxFlex OS, and most of those are related to performance. Therefore, it is clear that deploying Oracle databases on VxFlex OS is something that you can enter into confidently and with complete peace of mind.