Oracle Maximum
Availability Architecture

Oracle ASM Considerations for Exadata Deployments: On-premises and Cloud

ORACLE WHITE PAPER | JULY 2019

Disclaimer

The following is intended to outline our general product direction. It is intended for information purposes only and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. The development, release, and timing of any features or functionality described for Oracle's products remains at the sole discretion of Oracle.

Introduction

Oracle Exadata Database Machine, now in its ninth generation, is an integrated engineered system consisting of scale-out database and storage servers, utilizing database-optimized smart storage software. Exadata has been available for on-premises deployments since 2008 and is also available as an enterprise platform for deploying cloud databases, with Exadata Cloud Service (deployed at Oracle's public cloud data centers) and Exadata Cloud at Customer (deployed at customers' data centers).

With each generation, the technology within the system has component improvements designed to enhance the complete stack. One of these components is the storage technology that is the foundation for Exadata Database Machine.

The Exadata product family consists of different machine configurations with various disk, flash and memory attributes. There is a well-defined upgrade path for each configuration. There is usually a period of time between the original purchase and an upgrade, and it is likely that the upgrade components will be different from the original components. This paper explains the Exadata on-premises upgrade options, and provides ASM-related guidance and examples, for planning deployments to accommodate the generational changes in Exadata technology. This paper also discusses the nuances for utilizing ASM for database deployments with Exadata Cloud. This paper assumes a single cluster and storage expansion within that cluster.

ASM Considerations for Exadata On-premises Deployments

Available Exadata Configurations

The Exadata product family has various product configurations:

- Elastic configurations, available starting with Exadata X5, provide an extremely flexible and efficient mechanism to expand computing power and/or storage capacity of any given Exadata system. The starting configuration of an Exadata Database Machine consists of 2 database servers and 3 storage servers, which can be further elastically expanded by adding more database or storage severs as requirements grow.
- A pre-built database machine that has database servers and storage servers in eighth, quarter, half and full rack configurations in the Xn-2 family (X5-2, X6-2, X7-2, X8-2 etc.). X8-8, X7-8 and X6-8 database machines can also be expanded with an elastic configuration as mentioned above. The Xn-2 and Xn-8 products can both be configured with either Extreme Flash drives (EF) or High Capacity disks (HC).
- A pre-built storage expansion rack that has Exadata storage servers in quarter rack configurations with either EF or HC disks.
- Individual Exadata storage servers (EF or HC) that can be added to either existing racks or separate standalone racks.
- Individual Exadata X8-2 Extended storage server (also known as Exadata XT storage server) can be added to any Exadata rack that is currently supported. This storage server has 14 TB disk drives like the Exadata X8 HC storage server. It differs from an Exadata X8-2 HC storage server in 3 ways. The Exadata XT storage server does not have

flash drives included, has only one 16-core CPU and the Exadata System Software license is optional. This storage server is primarily meant to store low-use data for customers who have long-term data retention requirements for compliance purposes and high performance is not an important requirement. It can also be used for storing local database backups that are needed for immediate restore and for creating development databases where high performance is not a critical requirement.

» Individual database servers that can only be added to existing database machine racks

Exadata Storage Server Disk Types

The following table shows the disk types introduced with each Exadata generation.

Disk Type	Default Exadata Machine Model
600 GB HP disk 15K RPM	X2, X3
2 TB HC disk 7200 RPM	V2, X2
3 TB HC disk 7200 RPM	X3
4 TB HC disk 7200 RPM	X4, X5
8 TB HC disk 7200 RPM	X5, X6
1.2 TB HP disk 10K RPM	X4
1.6 TB Extreme Flash (EF) drive	X5
3.2 TB Extreme Flash (EF) drive	X6
10 TB HC disk 7200 RPM	X7
6.4 TB Extreme Flash (EF) card	X7
14 TB HC disk 7200 RPM	X8

While specific disk capacities were introduced in certain Exadata generations, the Disk Swap Service can be used to replace disks. The service allows customers to change the disks in the Exadata storage servers in their X4 and later generation Exadata and Storage Expansion Racks. They can change High Performance disks, either 600GB or 1.2TB to 14TB High Capacity disks. Swapping from High Capacity disks to High Performance disks is not part of the service. Also, swapping from 600GB High Performance disks to 1.2TB High Performance disks is not part of the service. Disk swaps are not supported on V1, V2, X2 and X3 systems. Only 14Tb High Capacity disks are available for order. More details about the disk swap service can be found in the MOS notes listed in reference (a) of this paper.

Though swapping disk drives provide a path for capacity expansion, it may not be the best solution in many cases. Replacing older Exadata servers with newer generation Exadata servers may be a better alternative in many cases given they have larger flash capacity, faster flash and more compute power in the storage servers. Also, from an HA and manageability perspective, Exadata storage servers, starting with X5, eliminate the requirement for periodic planned maintenance to replace flash card batteries. Exadata storage servers starting with X5 also automatically power cycle disk drives that appear to be failing to distinguish between a true drive failure and software issues. This allows a drive to be returned to service quickly, maintaining data protection and avoiding a service call.

There are support and cost implications too that need to be taken into account before choosing the disk swap option versus replacing with a newer generation Exadata storage server.

Exadata Upgrade Options

This section describes the various upgrade options that are available to expand an existing Exadata configuration.

Upgrade Type	Upgrade Components	Upgrade considerations
Eighth rack to quarter rack – X3, X4 and X5 generations *Note: Does not apply to X6, X7 and X8. Upgrade method for X6, X7 and X8 eighth rack machines is listed in the next row*	No additional hardware components needed for the X3, X4, X5 generations of Exadata as well as the X6 EF eighth rack An eighth rack X3, X4 and X5 is physically a quarter rack hardware machine with half the database and storage server cores disabled. EF storage servers and pre-X6 HC storage servers have half the disks/flash disabled. Adding Exadata X7-2 Eighth Rack HC storage server(s) can also be used to upgrade X3, X4 and X5 eighth rack storage servers. Details covered later in the Sample Storage Expansion Scenarios section of this document.	Upgrade both the storage and database tiers simultaneously to make it a quarter rack configuration. Database compute or storage only upgrades can also be done depending on resource requirements. Details covered later in this document
Upgrading eighth rack X6, X7 and X8 rack machines	In X6, X7 and X8 eighth racks, HC storage servers only have half the disks/flash physically present. In a X7 and X8 eighth rack, only 1 of the 2 available CPU sockets in the database nodes is populated whereas in a X6 eighth rack, the second CPU is populated but the cores are disabled. To upgrade the storage tier, add Exadata X8-2 Eighth Rack HC storage server(s) to the existing X6, X7 or X8 eighth rack and expand the ASM disk groups to the new server(s).	Minimum 1 Exadata X8-2 Eighth Rack HC storage server needs to be added, maximum determined by rack power and space limits. Can only be added to eighth rack configurations. As the X8-2 HC disks are 14 TB in size, the usable space in the newly added eighth rack storage server may not match the existing eighth rack servers. Grid disks of the same size as the existing storage servers will need to be created on the new X8-2 eighth rack storage server. Refer to Scenario 2 below for additional details. For upgrading eight rack X6 database nodes, enable the cores on the second CPU that is already present. For upgrading eighth rack X7 and X8 database servers, 2 additional CPUs need to be installed in the X7 and X8 database servers (1 per database server) and the cores enabled. Details covered later in this document.
Elastic Upgrades	Database servers or/and Storage servers. Storage servers can be either EF or HC. Exadata generation can be same as current or different, such as adding X8-2 servers to existing X4, X5, X6 or X7 machine.	Additional database or storage servers can either be added to existing cluster or to a newly-created cluster. Maximum of 22 servers in a rack. It is recommended to use the OECA tool (link in reference (b)) to validate the final configuration. If EF storage servers are being added to a machine with HC storage servers or vice versa, a new disk group must be created on the EF/HC storage servers (described in Scenario 4 in this paper). If X8-2 HC storage servers are being added to a machine that has either HC

		or HP disks then existing disk groups can be extended (described in Scenarios 2 and 3 in this paper).
Expand with new database machine	Add a new database machine configuration (EF or HC disks) to existing configuration. May need additional IB switches and cables for multi-racking. There are specific rules on what configurations can be multi-racked. The new machine can be partially licensed, such as a subset of database servers or storage servers.	The new machine can be deployed in a separate cluster or the current cluster can be expanded. If the new machine has a different drive type i.e. flash vs disk, a new disk group must be created (described in Scenario 4 in this paper). If disk-drives based storage servers are being added then existing disk groups can be extended (described in Scenarios 3 and 4 in this paper).
Expand with storage expansion rack	Add a storage expansion base rack (4 storage servers) to an existing Exadata machine. Storage expansion base rack can be either EF or HC to start with. Additional EF or HC storage servers can be added to the base rack. A subset of the storage servers can be licensed. May need additional IB switches and cables for inter-racking.	If the expansion rack has a different drive type i.e. flash vs disk, a new disk group must be created (described in Scenario 4 in this paper). If disk-drives based storage servers are being added then existing disk groups can be extended (described in Scenarios 3 and 4 in this paper). Storage expansion rack can be shared among multiple Exadata machines. Details later in this paper.
Expand with Exadata XT storage server	Add Exadata XT storage servers to any Exadata rack (including storage expansion rack) that is currently supported.	A minimum of 2 Exadata XT storage servers need to be added to start with to an existing Exadata rack. Subsequent additions can be in increments of 1 XT storage server. Given the reduced performance characteristics of this storage server, a new ASM disk group must be created for the Exadata XT storage servers to segregate low-use data. The default disk group name for XT storage servers is XTND. However, you can use a different name as required. Details are described in Scenario 5 below.
Xn-8 upgrades	Add X8-8 Database servers or/and Storage servers to existing Xn-8 base rack. If this is an older generation Xn-8 full rack with Xn-2 storage servers, then add additional database machine or storage expansion racks as described above. Add new Xn-2 rack	Disk type considerations and licensing similar to above. If Xn-2 machine added, the database servers cannot be part of the same cluster as Xn-8.

NOTE:

It is possible to have an Exadata rack with different generation machines (X2, X3, X4, X5, X6, X7), having EF, HP and/or HC disks and flash with different capacities after multiple upgrade cycles. Exadata X8-2 servers can only be added to Exadata racks that are X4 generation or higher. It is important to understand the ASM configuration possibilities to make upgrades smoother.

The same disk considerations apply to Oracle SuperCluster storage expansion because the same Exadata storage servers are used in Oracle SuperCluster. Oracle Supercluster M7 and M8 also support elastic configurations. The Exadata X8 storage servers are expected to be available for addition to a Supercluster system in the near future.

Oracle ASM: General Overview

Oracle ASM is a **volume manager** and a **file system** for Oracle Database files that supports single-instance Oracle Database and Oracle Real Application Clusters (Oracle RAC) configurations. Oracle ASM is Oracle's recommended storage management solution that provides an alternative to conventional volume managers, file systems, and raw devices.

ASM uses disk groups that consist of individual disks to store Oracle data files. To eliminate I/O hotspots, data is evenly spread across all available disks in a disk group during write operations. To get the best utilization from the ASM disk group and to avoid imbalance in data distribution, each disk in an ASM disk group must be the same size and similar performance. From a performance perspective, spreading the I/O across more disks is better as compared to fewer disks.

To provide protection against disk failures, ASM supports multiple redundancy levels. The supported levels are NORMAL, HIGH, EXTERNAL, FLEX and EXTENDED. NORMAL redundancy keeps a total of 2 copies of data (primary and mirror). HIGH redundancy keeps a total of 3 copies of data (primary and 2 mirror copies). EXTERNAL redundancy depends on an external storage array (on non-Exadata systems) to provide protection from disk failures. ASM places mirrored copies of data so that each copy is on a disk in a different failure group.

For FLEX and EXTENDED disk groups, mirroring describes the availability of the files within a disk group, not the disk group itself. For example: If a file is unprotected in a flex disk group that has five failure groups, then after one failure the disk group is still mounted, but the file becomes unavailable. To use the FLEX and EXTENDED redundancy capability, the COMPATIBLE.ASM and COMPATIBLE.RDBMS disk group attributes should be set to a minimum of 12.2.0.1 for the disk groups.

A failure group is a subset of disks in a disk group that could all fail at the same time because they share hardware such as drives in a single removable tray. If the tray is removed, all the drives in the tray would fail. It is important that primary and mirror copies are not placed in the same failure group. In the tray example, the primary and mirror copies should be on disks that are in different trays so that the simultaneous failure of all disks in a failure group does not result in data loss.

In case of a disk failure, ASM automatically starts restoring protection by re-creating the data from the failed disk onto the good disks using the data from the surviving mirror copy. The disk_repair_time (set to 3.6h on Exadata) disk group attribute determines when the failed disk is dropped and a rebalance initiated to restore data redundancy. This allows for any transient disk failures to be fixed. In effect, ASM uses the concept of spare capacity for restoring protection rather than using dedicated spare disks.

ASM on Exadata

ASM is the default and only storage management system supported on Exadata systems. Exadata systems use standard ASM, with the following stipulations:

- Each storage server is configured as a separate failure group.
- Only the NORMAL and HIGH redundancy levels are supported on Exadata.
- NORMAL redundancy provides protection against a single disk failure or an entire storage server failure. HIGH redundancy provides protection against 2 simultaneous partner disk failures from 2 distinct partner storage servers or 2 entire storage servers. HIGH redundancy provides redundancy during Exadata storage server rolling upgrades. MAA best practices recommend using HIGH redundancy.
- The ASM capacity information displayed using ASM commands may lead to mis-interpretation about the actual capacity available, due to the concept of spare capacity. Additional information is available in the MOS note listed in the reference (c) section of this paper.

ASM Disk Group Setup on Exadata

As part of a standard Exadata deployment, 2 default disk groups are created: DATA and RECO. For pre-X7 Exadata generations, the DBFS_DG disk group was also created.

- The DATA disk group is typically used for database data files (e.g. data files, log files, control files).
- The RECO disk group is used to store database recovery-related data (e.g. the Fast Recovery Area containing flashback logs, archive logs, and backups).
- The DBFS_DG, when available, is a fixed size disk group for a particular configuration and is the smallest.

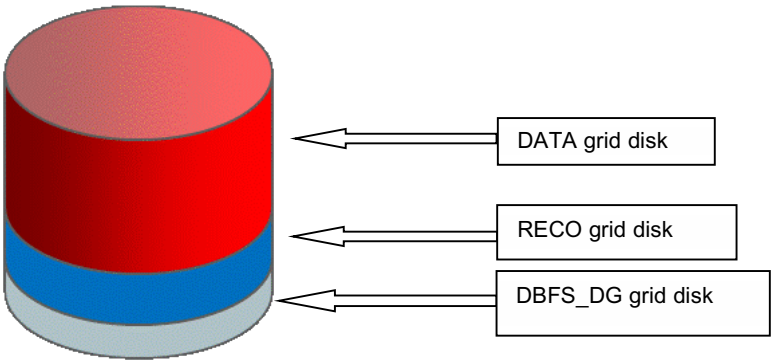
  NOTE: As noted above, the DBFS_DG disk group is not created on the X7 and X8 generation servers. In the pre-X7 generation of Exadata servers, portions of two disks at slot 0 and 1 were used as system partitions to install the operating system and Oracle Exadata System Software. The corresponding portions on disks in slots 2-11 were allocated to DBFS_DG. Starting with Exadata Database Machine X7, there are two M.2 disks dedicated for system partitions. This has removed the need for the DBFS_DG disk group on Exadata machines starting with Exadata X7. The DATA disk group can be leveraged to create the DBFS filesystem if necessary. More details on the M.2 disks can be found in the Exadata Database Machine System Overview manual.
- The ratio of DATA to RECO is configurable during deployment but is typically 80:20 or 40:60.
  - This depends on whether backups are kept on the machine or external to it.
  - Non-standard sizes can also be chosen during deployment planning phase but should be done after careful consideration.
  - The grid disk sizes must be chosen carefully as changing sizes once the system is in production can be disruptive depending on capacity utilization. Refer to the MOS note listed in reference (d) of this paper.
- The DATA and RECO disk groups can either have the same redundancy levels or different levels to meet customer-specific availability requirements.

Disks used in ASM Disk Groups on Exadata

On Exadata, each disk-based storage server has 12 physical disks except an eighth rack where only 6 physical disks can be used. An EF storage server has 8 flash cards except an eighth rack EF storage server where only 4 flash cards can be used. In order to create the 3 disk groups in pre-X7 Exadata machines, each disk is divided into 3 entities called grid disks. Each grid disk is then allocated to an ASM disk group. The following describes the default configuration using disk drives:

» Each non-eighth rack HC storage server has 12 DATA grid disks and 12 RECO grid disks. The disk group is created using the same grid disks across all the storage servers to adhere to the ASM principles of redundancy and I/O distribution. The size of these grid disks is dependent on the ratio chosen during deployment.

The following shows an example of grid disks on a physical disk in a pre-X7 Exadata storage server:



The ASM disk group is created by specifying:

» Path to storage servers and grid disk names.

» Redundancy level. If NORMAL redundancy is chosen, then grid disks from at least 2 storage servers must be specified. If HIGH redundancy is chosen, then grid disks from at least 3 storage servers must be specified. The recommended practice is to choose grid disks of the same size from all available storage servers to get the best performance and resilience.

The following example shows how to create a disk group named “data” using DATA grid disks from 3 storage servers (represented by the IP addresses (192.168.10.X)) :

```
CREATE DISKGROUP data NORMAL REDUNDANCY
DISK
'o/192.168.10.1/DATA*','o/192.168.10.3/DATA*','o/192.168.10.5/DATA*'
ATTRIBUTE 'content.type' = 'DATA',
'AU_SIZE' = '4M',
'cell.smart_scan_capable'='TRUE',
'compatible.rdbms'='12.2.0.1',
'compatible.asm'='12.2.0.1';
```

ASM Disk Group Considerations

The disks in an ASM disk group must also have similar performance characteristics in addition to having the same size. In an Exadata environment, this means that HP, HC and XT disks cannot be mixed in the same disk group. A grid disk from an HC disk must not be added to a disk group that has grid disks from HP disks. Similarly, EF drives cannot be in the same disk group that is comprised of HC or HP disks. The only exception is with X8-2/X7-2/X6-2 / X5-2 HC storage servers which can be added to HP storage servers from previous generations.

Choosing the Number of ASM Disk Groups

One deployment decision is the number of disk groups to be created. This decision should consider available capacity, isolation requirements, preferred protection levels for different data classes, management overheads, and the types of disks in the system.

It is recommended to keep the number of disk groups to the minimum specified as part of the standard deployment. The 2 main disk groups are needed to separate the database data files and the recovery area for the database, as specified by the “content.type” attribute. This enables optimal sharing of disk storage capacity as well as reducing the management effort while providing the best performance and resiliency. Exadata racks are deployed in a standard way, though the environment can be customized to match an existing environment. Oracle recommends not deviating from the deployment standards and to keep in line with global standards, which is the way the majority of customers across the globe choose to deploy Exadata machines including ASM disk groups setup.

Reasons for Additional Disk Groups

The following reasons may be why it is necessary to create additional disk groups:

- There are EF, HP and HC disks in the system. Because different drive types cannot be mixed in the same disk group, separate disk groups for the EF drives and HC disks need to be created.
- A certain database is critical and needs a HIGH redundancy disk group while most other databases can use NORMAL redundancy disk groups.
- To isolate specific databases due to operational considerations such as DATA_EBS for E-Business Suite data, and DATA_SEBL for Siebel data.
  - In this case, security mechanisms such as Exadata Database and ASM scoped security would need to be used to prevent a database from accessing the disk group in the other database.
  - More details of such an isolation scenario with security considerations can be found in the white paper listed in reference (e) of this paper.

Creating Multiple Disk Groups on Exadata

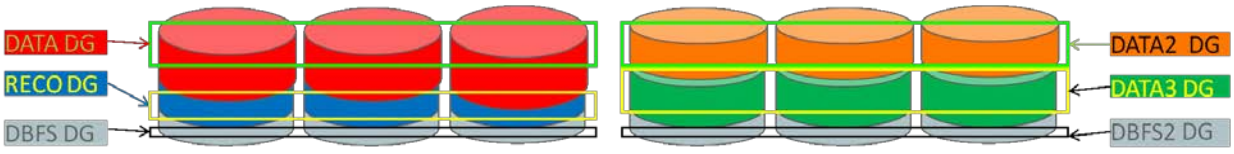
Each disk group needs its own dedicated disks. The following scenarios describe the advantages and disadvantages when using more than 3 disk groups.

Scenario 1: Distinct Storage Servers

- An Exadata rack with 6 storage servers licensed.
- Have distinct sets of storage servers for each disk group with 3 disk groups per set of storage servers.

Create the grid disks on each set of storage servers. At least 6 storage servers are needed, with each disk group being spread across the 3 storage servers. This is to provide sufficient capacity and resilience in case of failures with NORMAL redundancy with 3 failure groups per disk group.

To get the benefit of optimal I/O bandwidth, the disk groups can be spread over all 6 storage servers. There would be smaller grid disks per disk but the same overall capacity for each disk group.



- Advantages of this scenario:
  - It provides isolation at the disk group level when security is enforced.
- Disadvantages of this scenario:
  - Can only be achieved in larger Exadata configurations.
  - The I/O bandwidth is limited to a smaller number of storage servers and the amount of flash is limited those storage servers. This goes against the principle of optimal utilization of storage resources.
  - If one disk group is lightly used, then both the capacity and the performance from those servers cannot be utilized by the other databases. This defeats the purpose of consolidation efficiencies.

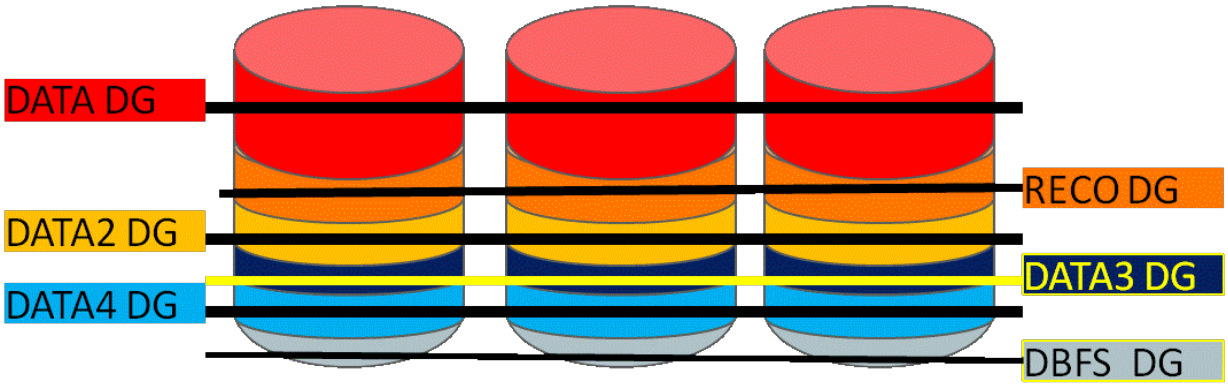
Apart from the advantages and disadvantages listed above, it is important to note that the best practice is to use all available storage servers for disk groups as described in Scenario 2 below and use Exadata Database and ASM scoped security to address the isolation requirements

Scenario 2: Shared Storage Servers

- An Exadata quarter rack with 3 storage servers
- Create additional disk groups after a default installation that share the same storage servers.

In this case, it is necessary to drop the default DATA and RECO disk groups, and then create grid disks of smaller sizes on the physical disks. Next, create disk groups on the newly-created grid disks. For example, create 6 grid disks on each physical disk instead of the standard 3 grid disks on each of the 3 storage servers. Note that in this case, since disk groups are being dropped, there must be a valid backup or a standby database corresponding to the production data, to prevent any data loss.

The number of grid disks created must be equal to the number of disk groups needed. This scenario assumes 6 disk groups (DBFS_DG + 5 additional).



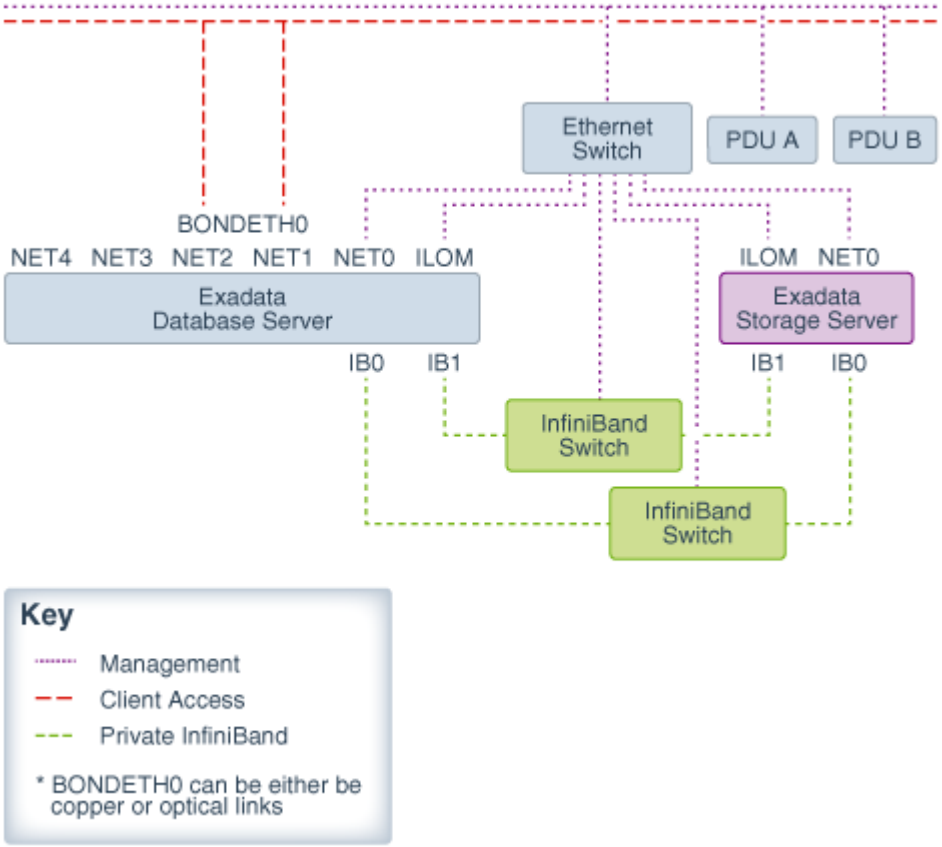
- Advantages of this scenario:
  - Ensures that all the flash is available to all the databases.
  - The I/O bandwidth is maximized.
  - It provides isolation at the disk group level when security is enforced.
- Disadvantages of this scenario are:

- Each new disk group has reduced capacity as compared to the original DATA and RECO disk groups given the smaller grid disk sizes. The spare capacity from other disk groups cannot be utilized if security isolation has been enforced.
- The time to rebalance disk groups after a disk failure increases, especially when the number of disk groups is greater than the number of ASM instances.
- Higher management overhead to create and maintain additional disk groups. It is tedious to resize grid groups when there are lots of small grid disks.

Generic Deployment Planning Considerations

- **Determine or estimate the size of the Exadata system over a period of time.**
  - Configure network subnets with the ability to host projected number of hosts.
  - Having the required number of management IP addresses in the correct subnet requires no workarounds or disruption to configure and accommodate the new nodes during the expansion process.
  - The management IP address range is provided by the customer.
  - The management network needs 2 management IP addresses for each database and storage server.
  - Management IPs are needed for InfiniBand (IB) switches, PDUs, etc.
- **The same considerations apply to the private IB network.**
  - Although a range of default addresses are chosen, choose the correct subnet mask to allow for enough IB IP addresses to accommodate expansion.
  - An example of increased IB IP address requirements is that the X4, X5, X6, X7 and X8 nodes support active-active IB bonding and need 2 IP addresses per host channel adapter (HCA).

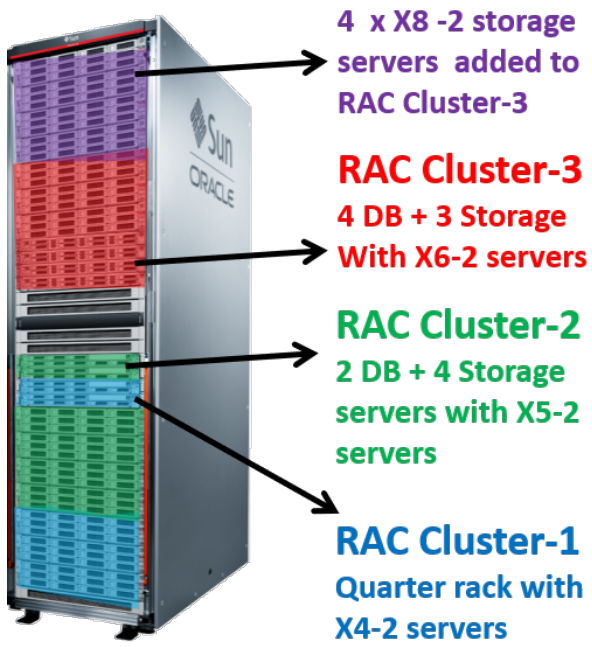
The network address requirements for Exadata are illustrated in the following graphic:



ASM on Exadata Deployment Planning Considerations

- **A single Exadata rack (at least a half rack for bare metal deployments) or a multi-rack bare metal Exadata system can be divided into multiple RAC clusters if the configuration meets the requirements below. On the other hand, using OVM, multiple clusters can also be created on a quarter or eighth rack Exadata configuration.**
  - Each cluster needs to have its own set of disk groups.
  - A minimum of 3 storage servers are required per cluster.
  - Based on the capacity requirements of the databases in each cluster, a subset of the total storage servers need to be allocated.

The following graphic shows several clusters in one rack with each cluster using a dedicated set of storage servers.



- While the new clusters could use separate storage servers as shown above, the recommended practice is to expand the existing storage pool by using all available storage servers for disk groups, as outlined in Scenario 2 in the earlier section. This would involve using the grid disk resize procedure described in the MOS notes listed in reference (d) below. Security mechanisms such as Exadata Database and ASM scoped security would need to be used to provide the necessary isolation between clusters sharing the storage servers.
- **Choose the redundancy level of the ASM disk groups appropriately.**
  - This is a CRITICAL decision because changing the redundancy level requires the disk group to be re-created. This could lead to disruption if done at a later point when the system is operational.
  - Oracle recommends HIGH redundancy for critical systems especially when there is no DR system in place. This is more practical now with the increased capacity available with 8TB, 10TB and 14TB drives for HC storage servers. The increased protection available with HIGH redundancy is a much better trade-off to the reduction in capacity in HIGH redundancy (compared to NORMAL).
  - HIGH redundancy level considerations:
    - HIGH redundancy is needed if a rolling upgrade strategy is chosen for storage servers for future software updates. This redundancy level lets the system continue running when a server is taken offline for updates or in case a failure occurs on one of the online storage servers. In such a case, there is a third copy available for the system to keep running.
    - For Exadata software releases prior to 12.1.2.3, if the cluster-related files such as the Voting disk need to be in a HIGH redundancy disk group, then a minimum of 5 storage servers is needed. In releases prior to 12.1.2.3, when Oracle Exadata systems with fewer than 5 storage servers were deployed with HIGH redundancy, the voting disk for the cluster was created on a disk group with NORMAL redundancy. If two cells go down in such a system, the data is still preserved due to HIGH redundancy, but the cluster software comes down because the voting disk is on a disk group with NORMAL redundancy. Starting with the

Exadata storage server software 12.1.2.3.0, quorum disks enable users to deploy and leverage disks on database servers to achieve highest redundancy in quarter rack or smaller configurations. Quorum disks are created on the database servers and added into the quorum failure group. This allows the use of 3 storage servers in HIGH redundancy configurations for the eighth and quarter rack Exadata configurations. Please refer to the Oracle Exadata Database Machine Maintenance Guide for software version dependencies to use this feature.

- Apart from using quorum disks on database nodes to hold the additional voting disks as described above, the best practice has been enhanced to recommend quorum disks for all high redundancy disk groups, not just those storing the vote device. Please refer to the Exadata Setup/Configuration best practices note listed in reference (f) below.

» NORMAL redundancy level considerations:

- Select when there is a Data Guard standby database corresponding to the production database and HIGH redundancy cannot be selected.
- Select when the intended use is for Dev/Test environments
- Select when you can tolerate system downtime to protect against disk failures during updates.
- The alternative to downtime would be to remove a storage server from the configuration before upgrading and then add it back. This causes multiple rebalances for each storage server upgraded. This increases patching time, operational investment, and space needed to accommodate the data on the storage server removed for patching.

» **If a DR system using Data Guard is setup for disaster protection:**

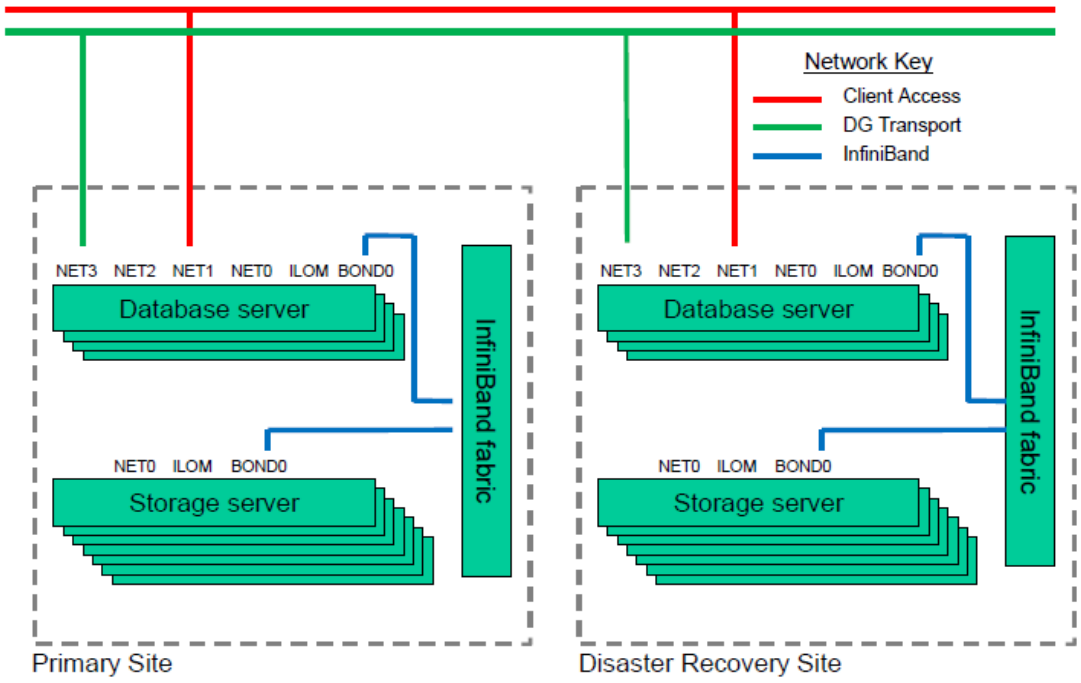
» The ASM redundancy levels for the primary and DR sites can be different, such as HIGH at production and NORMAL at the DR site.

» The disk types can be different at the 2 sites, such as EF/HP at primary site and HC at DR site.

» The Exadata DR system must not be on the same IB fabric as the primary system. The DR system must not be multi-racked on the IB network with the primary Exadata system.

» If 2 Exadata machines need to be connected for a failover capability within the same data center, then one of the Ethernet network interfaces can be used for Data Guard traffic. Alternatively, Oracle Extended Clusters on Exadata can be used as described later in this section. Additional information about using Data Guard is available in the MOS Note listed in reference (g) and is illustrated in the following graphic.

»



» **Depending on the timing of an expansion/upgrade, the software release on the existing storage servers may need to be upgraded to match the newer storage servers.** This is needed when the storage servers will be part of the same disk group.

» **The goal for deployment should be to share resources as much as possible. This does not imply that there is no isolation or compromise on SLAs.**

  » Understand the isolation requirements of the various workloads.

  » Oracle Database and Exadata have capabilities for isolating workloads, such as the Multitenant option, VLAN capability, Virtual Machines, Database Resource Management (DBRM) and Exadata I/O Resource Management (IORM) to address SLA requirements.

  » If storage isolation is needed due to perceived performance benefits, then a better option is to create a minimum number of disk groups and apply IORM policies to control I/O resource usage.

» **A storage expansion rack can be shared with different Exadata machines**:

  » All systems need to be multi-racked as one IB fabric on the same InfiniBand subnet. The MAA best practices for consolidation listed in reference (h) provide guidance on the number of Exadata machines that should be multi-racked in consolidation scenarios.

  » The individual storage servers from the expansion rack can be allocated to any of the Exadata database machines.

» **Do not attempt to design a deployment by distributing failure groups across 2 physical Exadata racks.**

  » An example of this would be deploying 2 multi-racked quarter racks instead of a half rack with an assumption that 2 racks would provide better availability.

  » As mentioned earlier, failures groups on Exadata do not have physical rack awareness so a physical rack failure would make the disk group unavailable.

  » There is nothing inside the rack that would make it fail as a unit so it should not be looked at as a point of failure.

- It is more likely that an external factor would cause a failure because all racks would be affected.
- An example of an external factor would be connecting both the PDUs in a rack to a single circuit.
- If rack level redundancy is needed, the preferred deployment would be to use Oracle Data Guard across 2 Exadata racks connected using Ethernet in the same data center.
- With Oracle Grid Infrastructure 12c release 2 (12.2), Oracle Grid Infrastructure supports the option of configuring cluster nodes in different locations as an Oracle Extended Cluster, which consists of nodes located in multiple sites. “Fire Cells” or isolated sites with independent power and cooling can created within a data center to provide rack level redundancy by deploying Oracle 12.2 Extended Clusters on Exadata. The 2 Exadata racks can be a maximum of 200m apart by placing a pair of spine switches between the 2 sites. More details of Extended Distance Clusters on Exadata can be found in the white paper listed in reference (i) of this paper.

Basic Principles and Best Practices of Exadata ASM storage Expansion

- **An ASM disk group must have disks from only 1 disk type (EF, HP or HC)**. This is true whether the new disk type is within the same Exadata rack, in a storage expansion rack or individual storage servers. The only exception is when X8 HC storage servers are being added to existing HP storage servers. In such a case, the HP storage server disk groups can be expanded to include the X8 HC storage servers as long as grid disks of the same size as the HP disks are created on the HC storage servers.
- **Options when disks of a different size, but of the same type, are being added** as follows:
  - Expand existing disk group to the new storage servers.
    - Ensure that the grid disk sizes on the new disks are exactly the same as the existing disk groups.
    - If the new disks are larger, then create grid disks of the same size as the existing (smaller) disks. Then, add the newly-created grid disks on the new storage servers to the existing disk group, and wait for the rebalance operation to complete.
    - The remaining space on the larger disks can be used to create a new disk group. This would need at least 2 storage servers for NORMAL redundancy, and 3 storage servers for HIGH redundancy.
      - **NOTE:** If only a single 14 TB disks storage server is added, then the remaining space cannot be allocated to a disk group as those disks would not have failure protection.
  - Create new disk groups on the new storage servers.
    - The new storage servers can be used to create separate disk groups. The new grid disk sizes and redundancy can be different from existing disk groups. If the disks are of a different type, then they must be in a separate disk group.
    - A minimum of 2 storage servers for NORMAL redundancy disk group and 3 for HIGH redundancy is needed. Oracle recommends 3 storage servers in a NORMAL redundancy disk group to provide extra protection in case of cell failure.
    - This should typically be done if the use case needs a separate disk group, such as a disk group to store archival data.
    - Often, a new storage server is added to the existing disk groups to improve the I/O bandwidth.

- **For older systems with HP disks: Using 1.2 TB HP disks storage servers.**
  - Even though the 1.2 TB HP disks are 10K RPM as compared to the 600 GB HP 15K RPM disks, the performance characteristics are similar due to generational advances. The disks can be mixed in the same disk group.
  - The remaining space on the new storage servers has minimal bandwidth but this is compensated by the increased amount of flash on the new storage servers.

Sample Storage Expansion Scenarios

Scenario 1: Eighth rack expansion scenario

As mentioned earlier, an older-generation eighth rack (prior to Exadata X6) is physically a quarter rack machine with half the CPU cores turned off, and half the disks/flash not configured for use. The X7 and X8 database nodes have only 1 of the 2 CPU sockets populated. The Exadata X6, X7 and X8 eighth rack HC machines have only half the number of disks and flash populated in the machine. The ASM disk groups for an eighth rack configuration are made up of 6 disks per storage server for disk-based storage servers and 4 flash drives for EF storage servers as opposed to the 12 disks/8 flash drives in quarter and larger racks.

Upgrading to a standard quarter rack (X3, X4 and X5 systems)

- Buy eighth rack to quarter rack database server hardware upgrade option and additional database software licenses
- Buy eighth rack to quarter rack storage server hardware upgrade option and the additional storage server software licenses
- The unused database CPU cores are activated.
- The disk groups are expanded to use the remaining 6 disk drives/4 flash drives that are already present in each storage server through a process of rebalancing.
- This process is described in the Exadata database machine *Oracle Exadata Database Machine Extending and Multi-Rack Cabling Guide.*

Expanding only storage capacity in an eighth rack without increasing CPU cores (X3-X8 systems)

- For the X3, X4 and X5 systems, buy the eighth rack to quarter rack storage server hardware upgrade option and the additional storage server software licenses
- The disk groups are expanded to use the remaining 6 disk drives/4 flash drives that are already present in each storage server through a process of rebalancing
- For the X4 and X5 systems, it is also valid to buy Exadata Eighth Rack Storage Server X8-2 High Capacity and the additional storage server software licenses and add the server or servers to the existing eighth rack storage servers. A minimum of one X8-2 HC eighth rack storage server needs to be bought. Exadata X8-2 storage servers cannot be added to an Exadata X3 machine
- Expand the existing disk groups to use the new eighth rack storage servers that are added.
- In the case of an Exadata X6, X7 or X8 eighth rack machine with HC storage servers, buy Exadata Eighth Rack Storage Server X8-2 High Capacity and the additional storage server software licenses. A minimum of one X8-2 HC eighth rack storage server needs to be bought. The maximum number of servers is determined by the space and cooling limits of the rack. The same approach can also be used for older generation eighth rack servers.
- Expand the disk groups to use the new eighth rack storage servers that are added.
- Alternatively, an Exadata storage expansion rack or additional storage servers can be added in the same rack or a separate rack. In this case, the new storage servers (whether of the same disk type or different) must be in a separate disk group from the existing servers. This is because the existing disk groups have only 6 disk drives or 4 flash drives per server whereas the new servers have 12 disk drives/8 flash drives enabled. Putting the new servers in the existing disk group creates an imbalance in the ASM disk group.

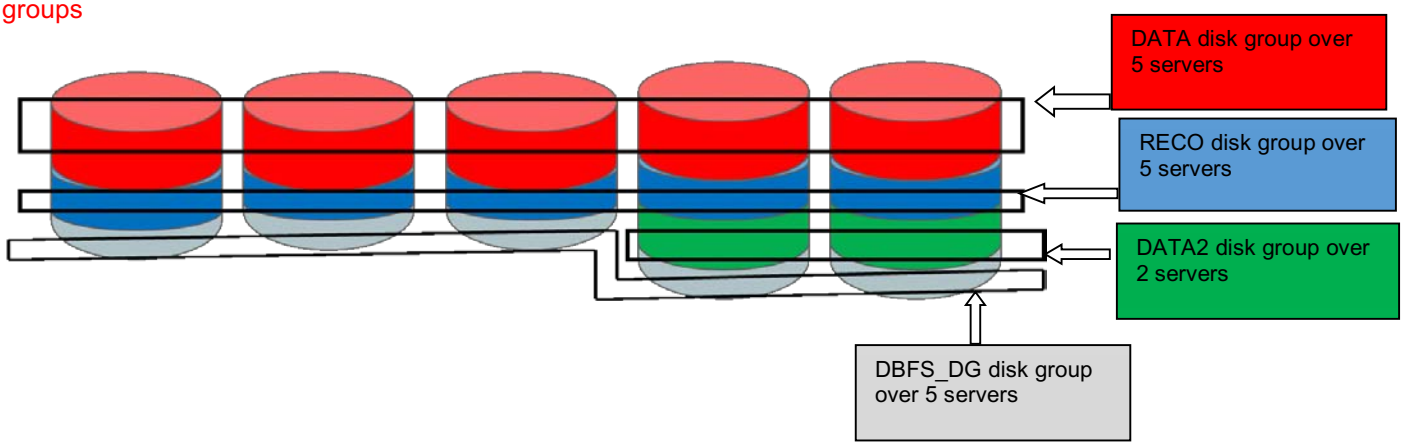
Expanding only database capacity in an eighth rack without increasing storage capacity

- Buy the eighth rack to quarter rack database server hardware upgrade option and the additional database server software licenses

» Activate the unused database CPU cores

» Alternatively, buy an additional X8-2 database server and add it to the existing cluster

Scenario 2: Add 14 TB storage servers to 4 TB or 8 TB storage servers and expand existing disk groups

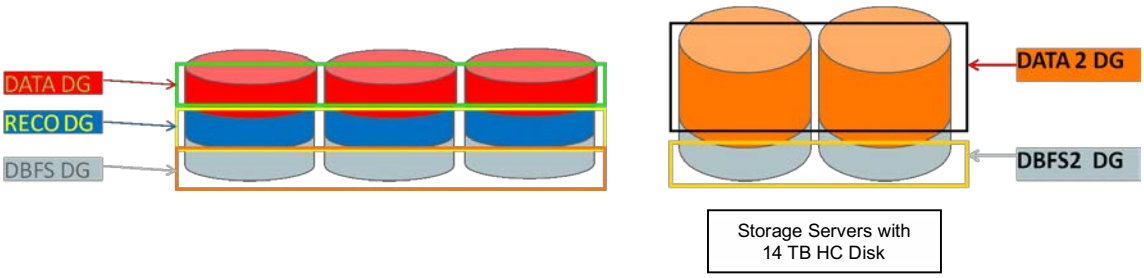


» Detailed procedure described in MOS note listed in reference (j)

Scenario 3: Add X8-2 storage servers with 14 TB HC disks to storage servers with HP disks

» This is similar to Scenario 2 above and is only valid when expanding 1.2 TB HP disk drives from X4 systems with X8-2 HC storage servers.

Scenario 4: Add 14 TB HC storage servers to X5-2 storage servers with 1.6 TB EF drives

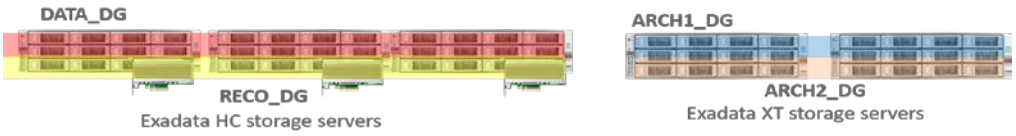


» The new 14 TB disks storage servers must be in different disk groups from the 1.6 TB EF disk groups.

» A minimum of 2 servers (recommended 3) are needed for NORMAL and 3 for HIGH redundancy.

» The new storage servers can have different grid disk configurations.

Scenario 5: Add Exadata Extended (XT) storage server to an existing Exadata rack



- The existing Exadata rack could have any type (EF or HC) and size (full or eighth rack) storage servers
- This is similar to Scenario 4 above as the performance characteristics of the Exadata XT storage server are very different from the Exadata EF or HC storage servers
- The new XT storage servers must be in a different disk group from the existing storage servers. The default disk group name for XT storage servers is XTND. However, you can use a different name as required.
- A minimum of 2 servers (recommended 3) are needed for NORMAL and 3 for HIGH redundancy.
- The new storage servers can have different grid disk configurations.

ASM Considerations for Exadata Cloud Deployments

This section of the document highlights the differences in system and storage configurations between Exadata On-premises deployments and Exadata Cloud deployments, including both Exadata Cloud Service and Exadata Cloud at Customer. For the rest of this document, the term “Exadata Cloud deployments” or “Exadata Cloud” is used to generically refer to these two deployment models.

Exadata Cloud Overview

The Oracle Exadata Cloud deployments enable you to leverage the power of Exadata in a service-based consumption model, with all Database features and options included. Each Exadata Cloud instance is based on an Exadata system configuration that contains a predefined number of compute nodes (database servers) and a predefined number of Exadata Storage Servers. Three Exadata platform versions are available with Exadata Cloud offerings currently - Exadata X5, X6 and X7 which differ in the number of compute cores available, storage capacity and IOPS provided by the storage. Note that Exadata X5 is not available with Exadata Cloud at Customer.

The following are some of the key differences between on-premises Exadata and Exadata cloud:

1. Exadata Cloud deployments currently supports the following standardized configurations:
   - **Base System**: 2 Database Servers and 3 Exadata Storage Servers.
   - **Quarter Rack**: 2 Database Servers and 3 Exadata Storage Servers.
   - **Half Rack**: 4 Database Servers and 6 Exadata Storage Servers.
   - **Full Rack**: 8 Database Servers and 12 Exadata Storage Servers
2. Each configuration above supports a minimum and maximum number of compute cores (OCPUs) and a fixed storage capacity that the customer can subscribe to. In case additional OCPUs or storage capacity beyond the maximum available in a particular configuration is needed, the cloud deployment will need to be upgraded to the next higher configuration. Detailed information about each Exadata Cloud configuration can be found in the data sheets listed in reference (k).

3. Exadata Cloud deployments currently do not support multi-racking to create large multi-rack configurations nor do they support the addition of either individual database servers, Exadata storage servers or Exadata storage expansion racks.
4. Exadata Cloud deployments do not support Exadata EF or XT storage servers.
5. Exadata Cloud offering is based on the Xn-2 generation product family. The Xn-8 product family is currently not supported with Exadata Cloud.

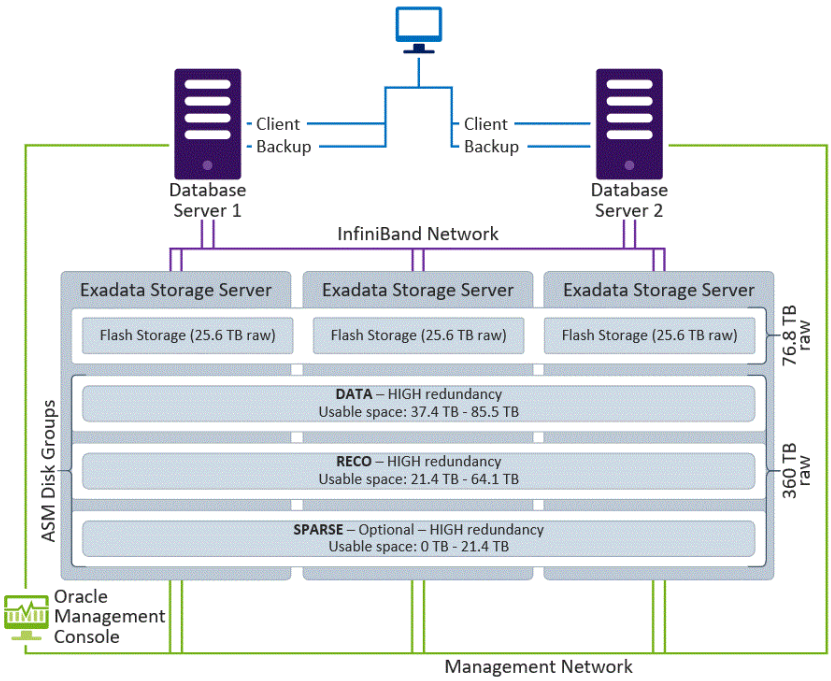
ASM Disk Group Setup in Exadata Cloud

ASM provides the storage management infrastructure in Exadata Cloud just as in on-premise Exadata systems. The ASM setup in the Exadata Cloud is different from an on-premises deployment as listed below:

- The storage in the Exadata Cloud is configured with the following default ASM disk groups: DATA and RECO.
- Only the ASM HIGH redundancy level is supported for the DATA and RECO disk groups.
- The DATA and RECO disk groups are intended for storing database files and recovery related files respectively.
- The space allocated to the DATA and RECO disk groups is dependent on the location chosen for database backups during the cloud configuration:
  - If you choose to provision for local backups on Exadata storage, approximately 40% of the available storage space is allocated to the DATA disk group and approximately 60% is allocated to the RECO disk group
  - If you choose not to provision for local backups, approximately 80% of the available storage space is allocated to the DATA disk group and approximately 20% is allocated to the RECO disk group
  - Details of the space allocated for the above 2 scenarios are provided in the documentation listed in Reference (l) below.
- For Exadata Cloud instances that are based on Oracle Exadata X5 or X6 hardware, there are additional system disk groups that support various operational purposes. These are the DBFS and ACFS disk groups.
- The DBFS and ACFS disk groups are small system disk groups that support various operational purposes. The DBFS disk group is primarily used to store the shared clusterware files (Oracle Cluster Registry and voting disks), while the ACFS disk groups are primarily used to store Oracle Database binaries e.g. cloud tooling for creating a database in the Exadata cloud service instance and to facilitate patching.
- For Exadata Cloud instances that are based on Oracle Exadata X7 hardware, there are no additional system disk groups. On such instances, a small amount of space is allocated from the DATA disk group to support the shared file systems that are used to store software binaries (and patches) and files associated with the cloud-specific tooling.
- For an Exadata Cloud deployment, a Service Request needs to be raised with Oracle to request any infrastructure changes unlike an on-premises Exadata deployment where a customer administrator can make those changes directly. Details on the process to request changes to the Exadata Cloud can be found in the MOS note listed in Reference (m) below.
- If additional staging space is needed, ACFS file systems can be created on one of the bigger disk groups for the duration they are needed and then dropped to release space. The following are the commands that can be used to create an additional ACFS file system in the RECO disk group.

```
Login as OPC and change to grid user.

[opc@svr01db01] sudo su - grid

[grid@svr01db01] asmcmd

ASMCMD> volcreate -G RECOC3 -s 1024G ingest_vol
      --this would be 1TB FS in Reco Diskgroup - change to desired size.
ASMCMD> volinfo -G RECOC3 ingest_vol
      -- **note the value listed in the volume device entry**
ASMCMD> exit

[grid@svr01db01] /sbin/mkfs -t acfs /dev/asm/ingest_vol-309
      -- use volume device from above
[grid@svr01db01] mkdir /var/opt/oracle/ingest_vol
      --location to mount ACFS filesystem
 [grid@svr01db01] /u01/app/12.1.0.2/grid/bin/srvctl add filesystem
 -d /dev/asm/ingest_vol-309 -g RECOC3 -v ingest_vol
 -m /var/opt/oracle/ingest_vol -u grid
    -- add the filesystem to be managed by Oracle clusterware
[grid@svr01db01] /u01/app/12.1.0.2/grid/bin/srvctl start filesystem
-d /dev/asm/ingest_vol-309
```

The image below shows the storage technical architecture of a generic (no specific shape) Exadata Cloud deployment. The diagram shows the layout of the Exadata storage servers from an ASM disk group perspective.



References

a. "Oracle Exadata Database Machine Disk Swap Service Process for X6 and Earlier Generation Exadata"– MOS Note 1544637.1, "Oracle Exadata Database Machine Disk Swap Service Process for X7 and Later Exadata Database Machines" – MOS Note 2531277.1

b. Oracle Exadata Configuration Assistant (OECA) (http://www.oracle.com/technetwork/database/exadata/oeca-download-2817713.html)

c. Understanding ASM capacity and Reservation of Free Space in Exadata – MOS Note 1551288.1

d. Resizing Grid disks in Exadata – MOS Notes 1465230.1 & 1684112.1

e. Oracle Exadata Database Machine Consolidation : Segregating databases and roles (http://www.oracle.com/technetwork/database/availability/maa-exadata-consolidated-roles-459605.pdf)

f. Oracle Exadata Database Machine Setup/Configuration Best Practices – MOS Note 1274318.1

g. Data Guard Transport Considerations on Exadata - MOS Note 960510.1

h. Best Practices for database consolidation on Exadata Database Machine (http://www.oracle.com/technetwork/database/features/availability/exadata-consolidation-522500.pdf)

i. Oracle Extended Clusters on Exadata Database Machine (http://www.oracle.com/technetwork/database/availability/maa-extclusters-installguide-3748227.pdf)

j. How to add Exadata storage servers with 3 TB disks to an existing database machine. The same concept applies for 4 TB and 1.2 TB disks – MOS Note 1476336.1

k. Data Sheets: Oracle Database Exadata Cloud Service (http://www.oracle.com/technetwork/database/exadata/exadataservice-ds-2574134.pdf), Oracle Database Exadata Cloud at Customer (http://www.oracle.com/technetwork/database/exadata/exacc-x7-ds-4126773.pdf)

l. Exadata Cloud Service Instance Details (http://docs.oracle.com/en/cloud/paas/exadata-cloud/csexa/service-instances.html)

m. How to Request Service Configuration for Oracle Database Cloud Exadata Service – MOS Note 2007530.1

Oracle Corporation, World Headquarters
500 Oracle Parkway
Redwood Shores, CA 94065, USA

Worldwide Inquiries
Phone: +1.650.506.7000
Fax: +1.650.506.7200

CONNECT WITH US

blogs.oracle.com/oracle

facebook.com/oracle

twitter.com/oracle

oracle.com

Integrated Cloud Applications & Platform Services