ORACLE

Oracle Exadata Database Machine

KVM Virtualization Overview and Best Practices for On-Premises RoCE-Based Systems

Exadata X11M and Exadata System Software 25.1 Update

Exadata Development

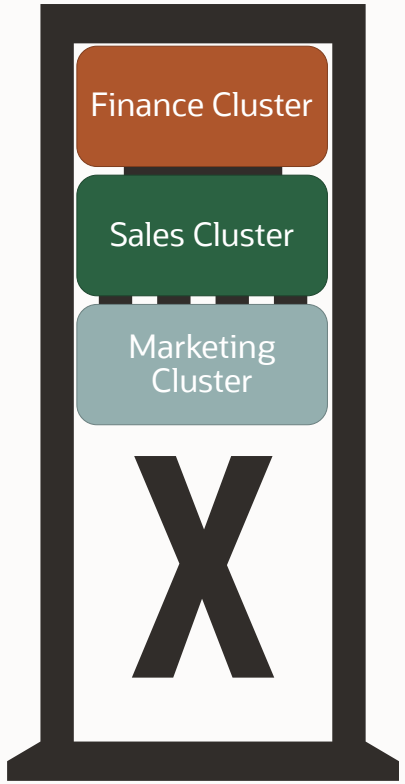
January 2025

Topics Covered

- Use Cases
- Exadata Virtualization Software Requirements
- Exadata Isolation Considerations
- Exadata KVM Sizing and Prerequisites
- Exadata KVM Deployment Overview
- Exadata KVM Administration and Operational Life Cycle
- Migration, HA, Backup/Restore, Upgrading/Patching
- Monitoring, Resource Management

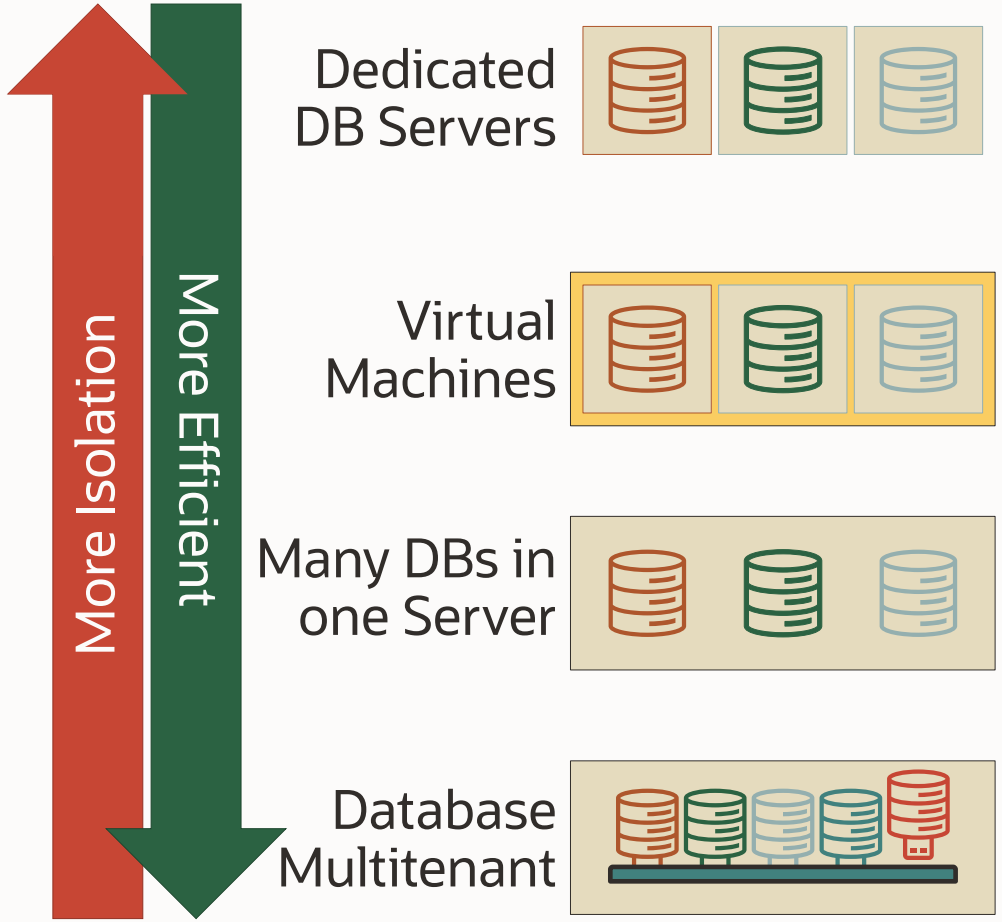
Exadata Virtualization

High-Performance Virtualized Database Platform Using KVM



- Kernel-based Virtual Machine (KVM) hypervisor
  - Linux kernel-based type 2 hypervisor with improved performance
  - Exadata RoCE based systems ***only*** (X11M, X10M, X9M-2, X8M-2)
- VMs provide CPU, memory, OS, and system admin isolation for consolidated workloads
  - Hosting, cloud, cross department consolidation, test/dev, non-database or third-party applications
- Exadata VMs deliver near **raw hardware performance**
  - Database I/Os go directly to high-speed RDMA Network Fabric bypassing hypervisor
- Combine with Exadata network and I/O prioritization to achieve unique full stack isolation
- Trusted Partitions allow licensing by virtual machine
  - See Oracle Exadata Database Machine Licensing Information User's Guide

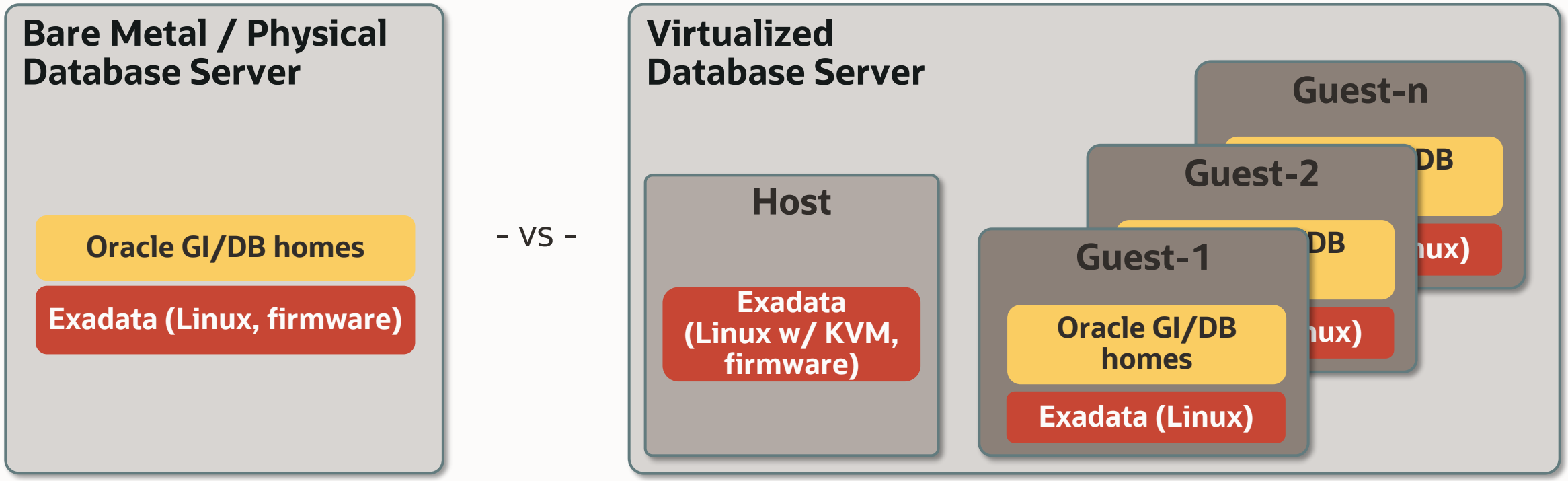
Exadata Consolidation Options



- **Dedicated** Database Servers provide the **best isolation**
- Virtualization has good isolation but requires more management overhead and resource usage
  - VMs have separate OS, memory, CPUs, and patching
  - Isolation without need to trust DBA, System Admin
- Database consolidation in a single OS is **highly efficient** but less isolated
  - DB Resource manager isolation adds no overhead
  - Resources can be shared much more dynamically
  - But must trust admins to configure systems correctly
- **Best strategy is to combine VMs with database native consolidation**
  - Multiple trusted DBs or Pluggable DBs in a VM
  - Few VMs per server to limit overhead of fragmenting CPUs/memory/software updates/patching etc.

Software Architecture Comparison

Database Server: Bare Metal / Physical versus Virtualized



Differences Between Physical and Virtual Deployments

Topic	How Virtual differs from Physical
Reduced licensing option	Use Trusted Partitions to license database options within specific VM Clusters
Cluster configuration	System has one or more VM clusters, each with own Grid Infrastructure / Database software installations
Network isolation	Use Secure Fabric to isolate cluster infrastructure network while sharing underlying Exadata storage
Exadata storage configuration	Separate grid disks / ASM disk groups or Exascale vaults for each cluster
Database server disk configuration	Default file system sizes are small Grid Infrastructure and Database software homes attached as separate file systems
Software updates	Database servers require separate KVM host (Linux, firmware) and KVM Guest (Linux) updates
EXAchk	Run once for KVM host + storage servers + switches, run once for each VM Cluster
Enterprise Manager	Enterprise Manager + Oracle Virtual Infrastructure plug-in + Exadata plug-in

Requirements

- Hardware
  - Exadata systems with RDMA over Converged Ethernet (RoCE) networks – X11M, X10M, X9M-2, X8M-2

- Software
  - Review MOS 888828.1 for recommended and minimum required versions
  - KVM Host
    - Virtualization using Oracle Linux Kernel-based Virtual Machine (KVM)
    - KVM host and guests can run different Exadata System Software versions
  - KVM Guests
    - Each guest runs Exadata System Software isolated from other guests
    - Each guest runs Grid Infrastructure and Database software isolated from other guests

Interoperability

- Interoperability between KVM/RoCE and Xen/InfiniBand
  - KVM supported only with RoCE interconnects
  - Xen supported only with InfiniBand interconnects
    - X8 and earlier upgraded to or freshly deployed with latest Exadata software continue to be based on Xen
  - Cannot inter-rack RoCE and InfiniBand
  - Separate KVM/RoCE and Xen/InfiniBand systems can be used in same Data Guard / GoldenGate configuration
    - e.g., KVM-based system as primary, separate Xen-based system as standby

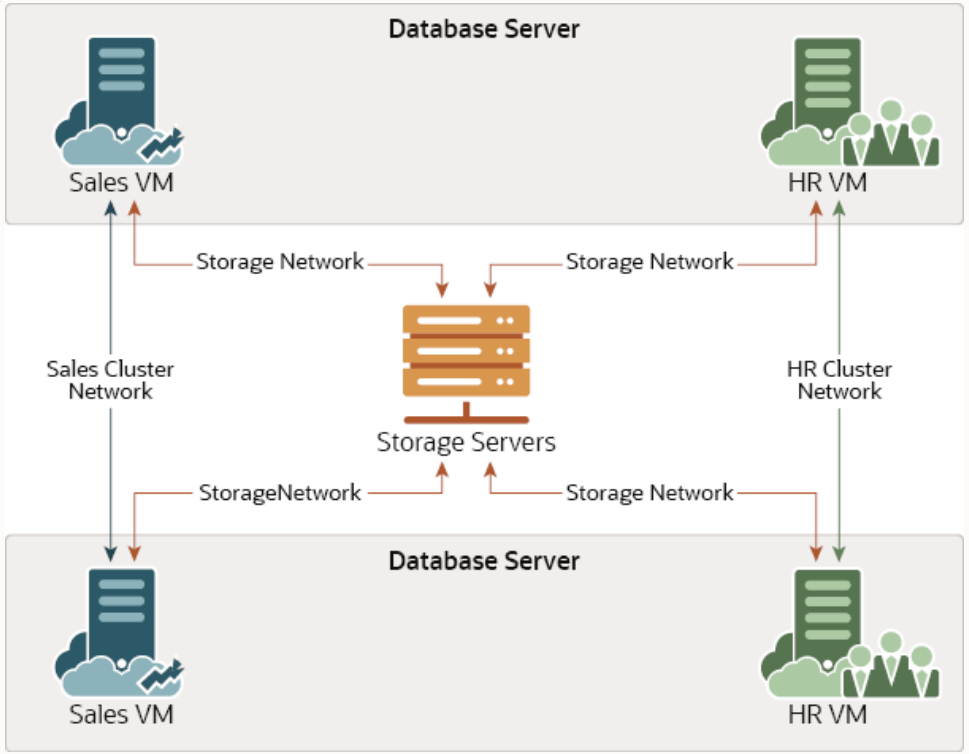
- Migration from Xen to KVM
  - Move database using Data Guard, GoldenGate, RMAN, ZDM

Security

- Storage Isolation - Each VM RAC cluster has own Exadata grid disks and ASM Disk Groups
  - Setting Up Oracle ASM-Scoped Security on Oracle Exadata Storage Servers
    - https://docs.oracle.com/en/engineered-systems/exadata-database-machine/dbmsq/exadata-security-practices.html
- Network Isolation - 802.1Q VLAN Tagging for Client and Admin Ethernet Networks
  - Configured w/ OEDA during deployment (requires pre-deployment switch config)
  - https://docs.oracle.com/en/engineered-systems/exadata-database-machine/dbmmn/managing-oracle-vm-guests-kvm.html
- RoCE Network Isolation - Secure RDMA Fabric Isolation with Oracle Linux KVM
  - https://docs.oracle.com/en/engineered-systems/exadata-database-machine/dbmin/exadata-network-requirements.html
- KVM Guest Secure Boot with UEFI (Exadata 24.1)
  - https://docs.oracle.com/en/engineered-systems/exadata-database-machine/dbmsq/exadata-security-features.html
- RESTful remote access for storage server administration through ExaCLI
  - https://docs.oracle.com/en/engineered-systems/exadata-database-machine/dbmmn/exacli.html

Secure RDMA Fabric Isolation for RoCE

- Exadata **Secure Fabric** for RoCE systems implements network isolation for Virtual Machines while allowing access to common Exadata Storage Servers
  - Each VM cluster is assigned a private network
  - VM clusters cannot communicate with each other
  - All VMs can communicate to the shared storage infrastructure
  - Security cannot be bypassed
    - Enforcement done by the network card on every packet
    - Rules programmed by hypervisor automatically
- Enabled by default in OEDA (Exadata 25.1) for new on-premises deployment



System Sizing and VM Resource Allocation

- **Maximum of 50 KVM guests per database server (X11M, X10M with Exadata 24.1 and later)**
  - X10M with Exadata 23.1, X9M, X8M – Maximum of 12 KVM guests per database server
  - X11M-Z and Eighth Rack systems – Maximum of 4 KVM guests per database server

- Determine peak CPU, memory, disk space needed by each database
  - Perform sizing evaluation prior to deployment, configure in OEDA accordingly
  - Consider KVM host reserved memory
  - Consider KVM host reserved CPU
  - Consider KVM guest long-term local disk file system growth
    - Long lived KVM guests should budget for full space allocation (assume no benefit from sparseness and shareable reflinks)
  - Each VM cluster has its own grid disks and disk groups and/or Exascale vaults
  - Contact Oracle for sizing guidance

VM Memory Sizing

- **Cannot over-provision physical memory**
  - Sum of all KVM guests + KVM host reserved memory <= installed physical memory

- **KVM Host Reserved Memory**
  - KVM host reserves portion of installed memory
  - Not available to KVM guests, enforced by vm_maker

VM Memory Sizing

- **KVM Guest memory sizing**
  - Total VM Memory Available
    - Allocate to single guest or divide among multiple guests
- Minimum 16 GB memory for a guest
  - To support OS, GI/ASM, starter DB, few connections
- VM Memory size can not be changed online
  - Guest restart required

Memory Config	Supported Platforms				Installed Memory (GB)	VM Memory (GB)
	X11M	X10M	X9M	X8M		
24 x 128 GB	●	●			3072	2800
24 x 96 GB	●	●			2304	2090
32 x 64 GB			●		2048	1870
24 x 64 GB	●	●	●	●	1536	1390
12 x 96 GB	●	●			1152	1010
16 x 64 GB			●		1024	920
12 x 64 GB	●	●		●	768	660
16 x 32 GB	●	●	●		512	440
12 x 32 GB		●	●	●	384	328

VM CPU Sizing

- **CPU over-provisioning is allowed**
  - Up to 2x over-provisioning permitted with multiple VMs
    - Exceptions – No CPU over-provisioning on X10M systems:
      - with 512 GB memory
      - when Capacity-On-Demand is used
  - Large increase in cores with X10M
    - CPU over-provisioning use cases decrease significantly compared to previous Exadata hardware
  - Performance degradation may occur if all guests become fully active when over-provisioning

- **Number of vCPUs assigned to a VM can be changed online**

- **KVM Host Reserved CPU**
  - Host is allocated 4 vCPUs (2 cores) – Not available to guests
    - X11M-Z and Eighth rack is allocated 2 vCPUs (1 core)

1 vCPU == 1 hyper-thread; 1 core == 2 hyper-threads == 2 vCPUs

VM CPU Sizing

- **Guest CPU sizing (X11M 2x96-core CPUs)**
  - Single guest vCPU
    - Minimum 4 vCPU
    - Maximum 380 vCPU
    - X11M-Z (1x32-core CPU)
      - Maximum 62 vCPU
  - Sum of all guest's vCPUs
    - Max 380 vCPU if no over-provisioning
    - Max 760 vCPU if 2x over-provisioning[1,2]
    - X11M-Z (1x32-core CPU)
      - Max 62 vCPU if no over-provisioning
      - Max 124 vCPU if 2x over-provisioning[1]

Hardware	Min vCPU per guest	Max vCPU per guest	Max over-provision vCPU all guests
X11M, X10M	4	380	760[1,2]
X11M-Z, X10M Eighth	4	62	124[1]
X9M-2	4	124	248
X9M-2 Eighth	4	62	124
X8M-2	4	92	184
X8M-2 Eighth	4	46	92

1 – No CPU over-provisioning when Capacity-on-Demand is used
2 – No CPU over-provisioning on systems with 512GB memory

1 vCPU == 1 hyper-thread; 1 core == 2 hyper-threads == 2 vCPUs

VM Storage Considerations

Database files and Grid Infrastructure shared files

- Default – store in Exadata storage managed by ASM
- Optional (Exadata 24.1) – store in Exadata storage managed by Exascale
  - Native Exascale support in Database 23ai
  - Database 19c may use Exascale Direct Volume (EDV) devices in VM guest as ASM grid disks in ASM disk groups

VM Guest disk images

- Default – store in KVM host local disk, limited by local disk size (3.84TB)
- Optional (Exadata 24.1) – store in Exadata storage managed by Exascale (EDV devices in KVM host)

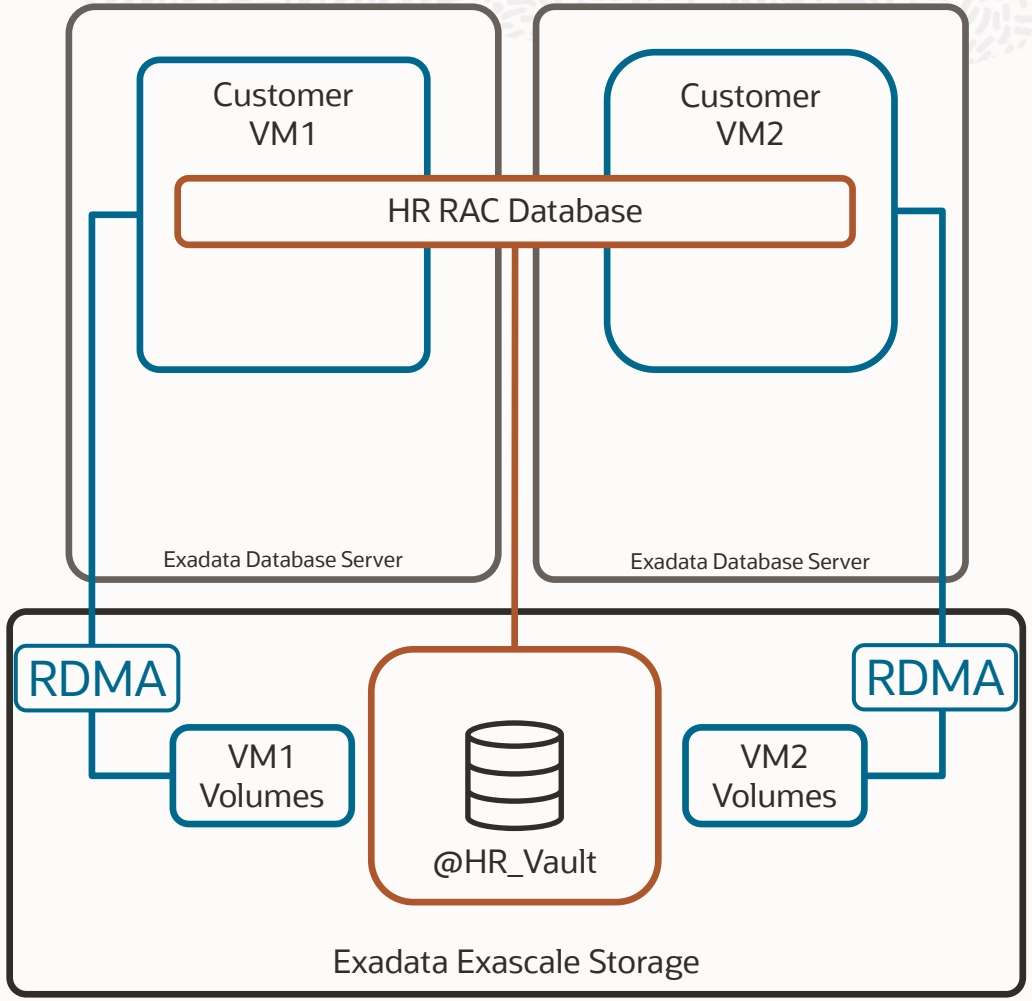
Storage Type	Storage Options
Database files (data, control, online/archived redo, etc.) Grid Infrastructure shared files (OCR, voting files)	ASM (Automatic Storage Management)
	Exascale native (GI/DB 23ai only)
	Exascale EDV devices in VM guest + ASM (GI/DB 19c and 23ai)
VM guest disk images	KVM host local disk
	Exascale EDV devices in KVM host

Storage for VM Guest Disk Images – KVM Host Local Disk

- X11M, X10M database server – 2 x 3.84TB NVME drives configured RAID1
  - Default local disk space available for VMs 1.46 TiB, online resizable to 3.4 TiB
  - Option to add 2 x 3.84TB NVME drives RAID1, increase local disk space to 6.9 TiB
- Default disk space used per KVM guest 228 GiB
  - Extend after initial deployment by allocating new disk images in KVM host local disk
  - Extend with shared storage (e.g., ACFS, DBFS, external NFS, OCI File Storage) for user files
    - Do not use shared storage for Oracle/Linux binaries/configuration/diagnostic files. Access/network issues may cause system crash or hang.
- KVM guest local file system disk space over-provisioning **not recommended**, but possible
  - Allocated space initially much lower than apparent space due to sparseness and shareable reflinks (with multiple VMs), but will grow over time (shared space diverges and becomes less sparse)
    - Long lived KVM guests should budget for full space allocation (assume no sparse/reflink benefit)
    - Over-provisioning may cause unpredictable out-of-space errors or prevent ability to restore disk image backup

Storage for VM Guest Disk Images – Exascale

- Enables VM image files to be located on shared Exascale storage as Exascale Volumes
- Increases the available storage capacity for VM images significantly
- Requires Exascale to be deployed in the environment and the VM cluster must be created with OEDA specifying Exascale will be used for VM storage
- Exascale Volumes can also be used to provide extended shared storage using ACFS or other Linux filesystems



Storage for Database and Grid Infrastructure – ASM

- Storage on cells may be shared between ASM and Exascale
- Spread ASM disk groups for each VM cluster across all disks on all cells
  - Every VM cluster has its own grid disks
  - Disk group size for initial VM clusters should consider future VM additions
    - Using all space initially will require shrinking the existing disk group before adding a new disk group

- Enable ASM-Scoped Security to limit grid disk access

VM Cluster	Cluster Nodes	Grid Disks (DATA/RECO for all clusters on all disks in all cells)	
clu1	db01vm01 db02vm01	DATAC1_CD_{00..11}_cel01 DATAC1_CD_{00..11}_cel02 DATAC1_CD_{00..11}_cel03	RECOC1_CD_{00..11}_cel01 RECOC1_CD_{00..11}_cel02 RECOC1_CD_{00..11}_cel03
clu2	db01vm02 db02vm02	DATAC2_CD_{00..11}_cel01 DATAC2_CD_{00..11}_cel02 DATAC2_CD_{00..11}_cel03	RECOC2_CD_{00..11}_cel01 RECOC2_CD_{00..11}_cel02 RECOC2_CD_{00..11}_cel03

Storage for Database and Grid Infrastructure – Exascale

- Cell disks may be shared between Exascale and ASM
- Cell disk partitions can be configured as pool disks (Exascale) or grid disks (ASM)
- One storage pool spans all pool disks of the same type (HC or EF) on all cells
- Each database uses Exascale vaults or ASM disk groups for storage, not both
- Databases within a VM cluster can be a mixture of Exascale and ASM
- Recommend one vault per VM cluster

VM Cluster	Cluster Nodes	Exascale storage pool	Exascale vault
Cluster-c1	db01vm01 db02vm01	hcpool1	Cluster-c1vault
Cluster-c2	db01vm02 db02vm02		Cluster-c2vault

Deployment Specifications and Limits

Category		X8M-2	X9M-2	X10M	X11M
VMs	Max guests per database server	12	12	50 (12[1])	50
	Max guests per Eighth Rack and X11M-Z database servers	4	4	4	4
Memory	Min GB per guest	16	16	16	16
	Max GB per guest / all guests	1390[2]	1870[2]	2800[2]	2800[2]
CPU/vCPU	Min vCPU per guest	4	4	4	4
	Max vCPU per guest	92	124	380	380
	Max over-provisioned vCPU all guests	184	248	760[3,4]	760[3,4]
Disk space	Usable TiB per DB server for all guests	3.15	3.40 / 6.97[5]	3.40 / 6.97[5]	3.40 / 6.97[5]
	Used GiB per guest at deployment	141	228	228	228

1 – X10M supports 50 guests w/ Exadata 24.1 and later, Exadata 23.1 and earlier supports 12
2 – Using maximum memory configuration
3 – No CPU over-provisioning when Capacity-on-Demand is used
4 – No CPU over-provisioning on systems with 512GB memory
5 – When local disk expanded to 4 drives

Deployment Overview

Oracle Exadata Deployment Assistant

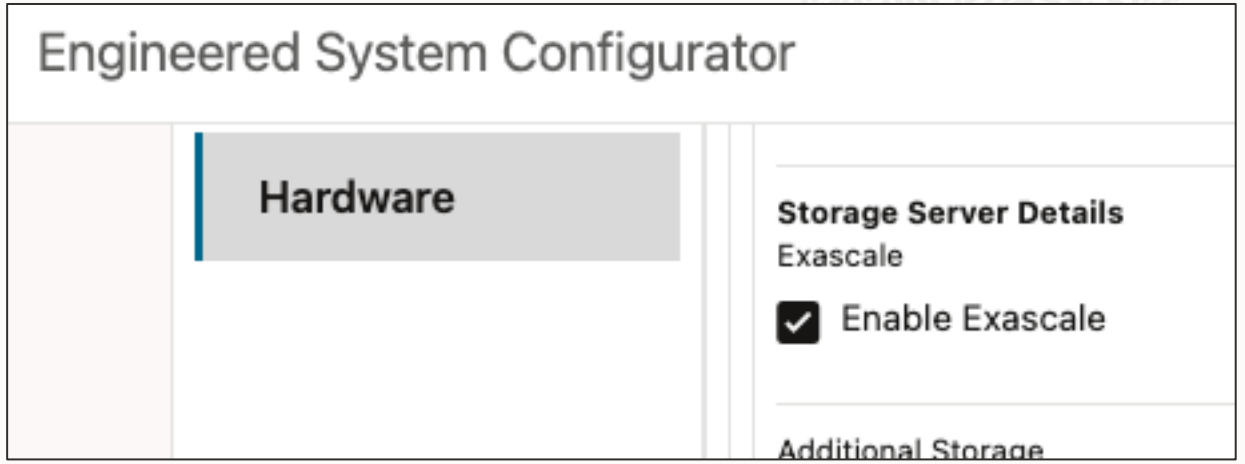
The Oracle Exadata Deployment Assistant, also known as OEDA, is the only tool to create VMs on Exadata

1. Create configuration with OEDA Web User Interface
2. Prepare customer environment for OEDA deployment
   - Configure DNS, switches for VLANs, and Secure Fabric
3. Prepare Exadata system for OEDA deployment
   - # switch_to_ovm.sh, applyElasticConfig.sh
4. Deploy system with OEDA Deployment Tool
   - # install.sh

OEDA Configuration Tool

Choose to deploy Exascale

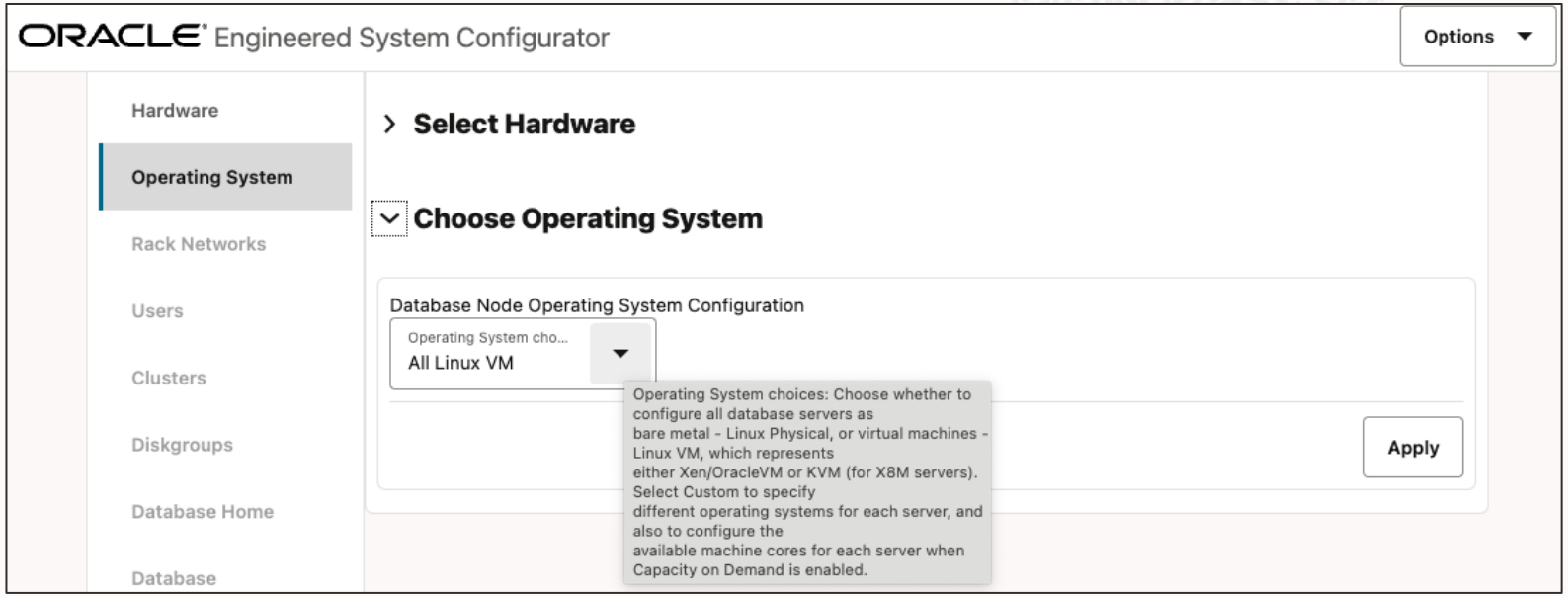
- Initial configuration choice is whether to enable Exascale
- Storage pool size and VM use of Exascale in later OEDA screens
- Exascale can also be deployed into existing environments



OEDA Configuration Tool

Choose Virtual or Physical Database Server Configuration

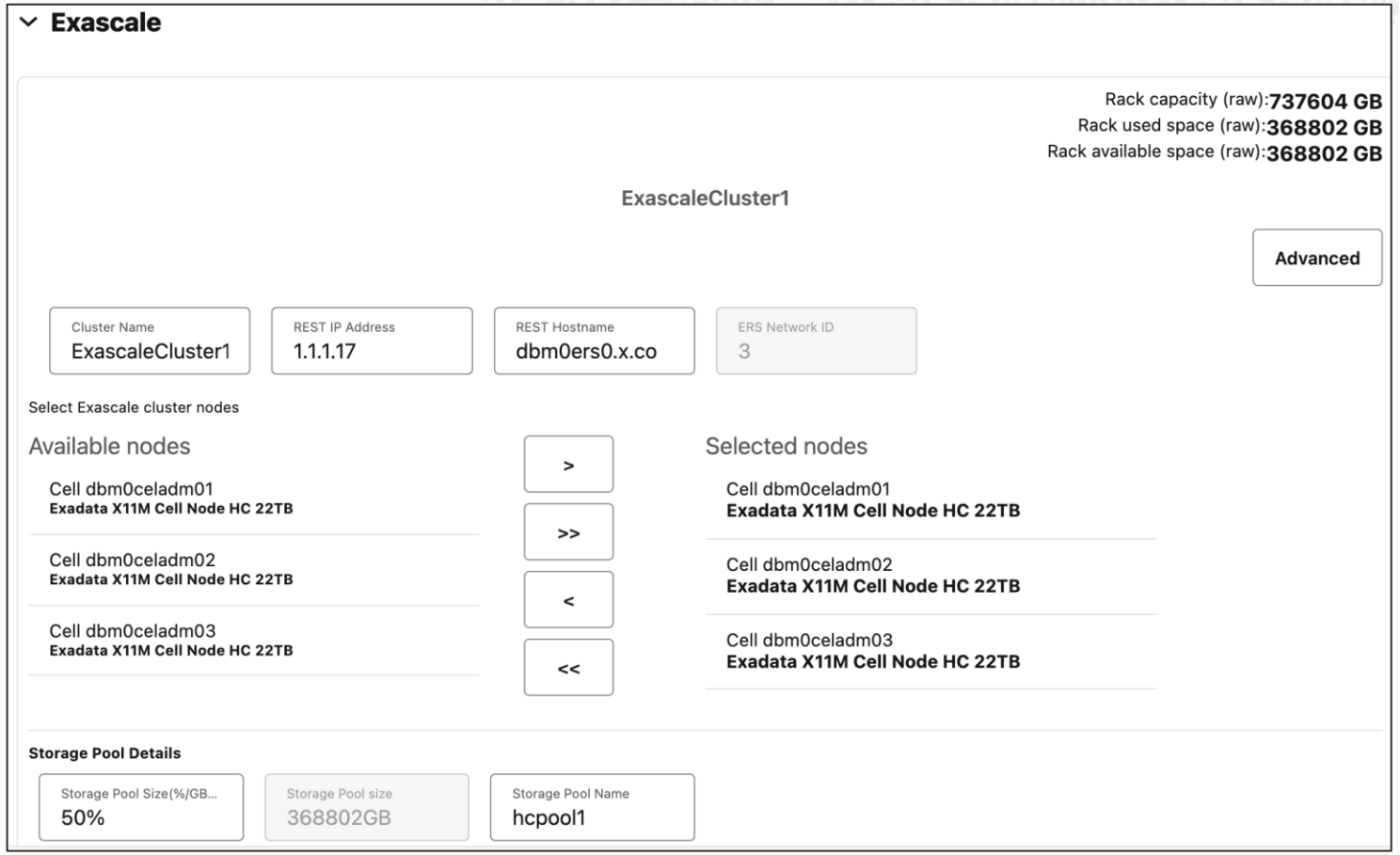
- Virtual or physical configuration
  - All Linux VM
  - All Linux Physical
  - Custom (some servers VM, some servers physical)
- An individual database server is configured either VM or Physical



OEDA Configuration Tool

Configure Exascale Storage Pool

- Select cells for Exascale storage pool
  - Recommend selecting all available cells
- Configure storage pool size
  - Space allocated to storage pool used for Exascale
  - Unallocated space used for future Exascale expansion or ASM storage



OEDA Configuration Tool

Define Clusters

- Decide
  - Number of VM clusters to create
  - Database servers and Cells that will make up those VM clusters
    - Recommend using all cells for each cluster

- What is a VM cluster?
  - One or more guests on different database servers running Oracle GI/RAC, each accessing the same shared Exadata storage managed by ASM or Exascale.

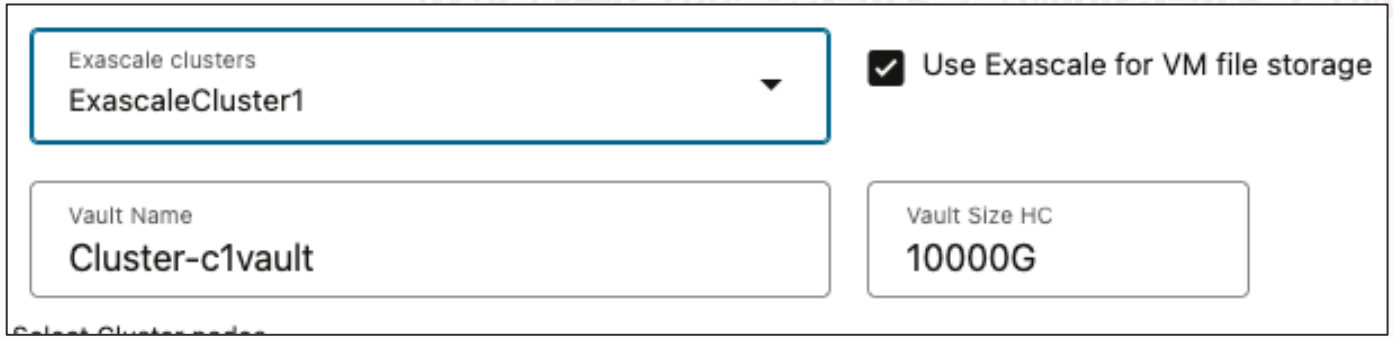
OEDA Configuration Tool

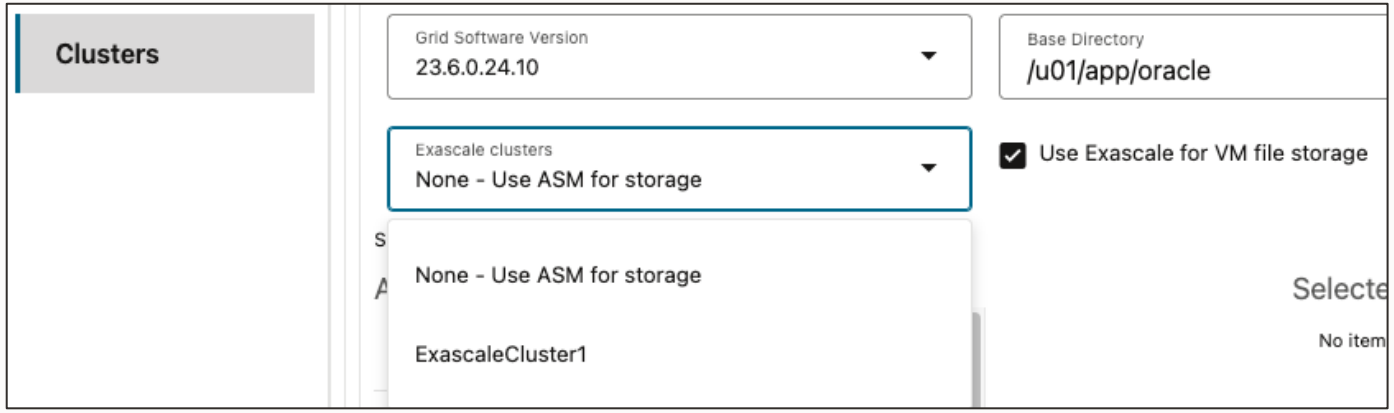
Configure VM Use of Exascale

- Decide
  - Vault size
  - Store VM guest disk images in Exascale vault(s) or local disks
  - Store GI/DB files in Exascale vault(s) (23ai only) or ASM disk groups

Note:

- A VM may use Exascale for VM image files and ASM for database file storage
- A 23ai database may use only Exascale or ASM for file storage





OEDA Configuration Tool

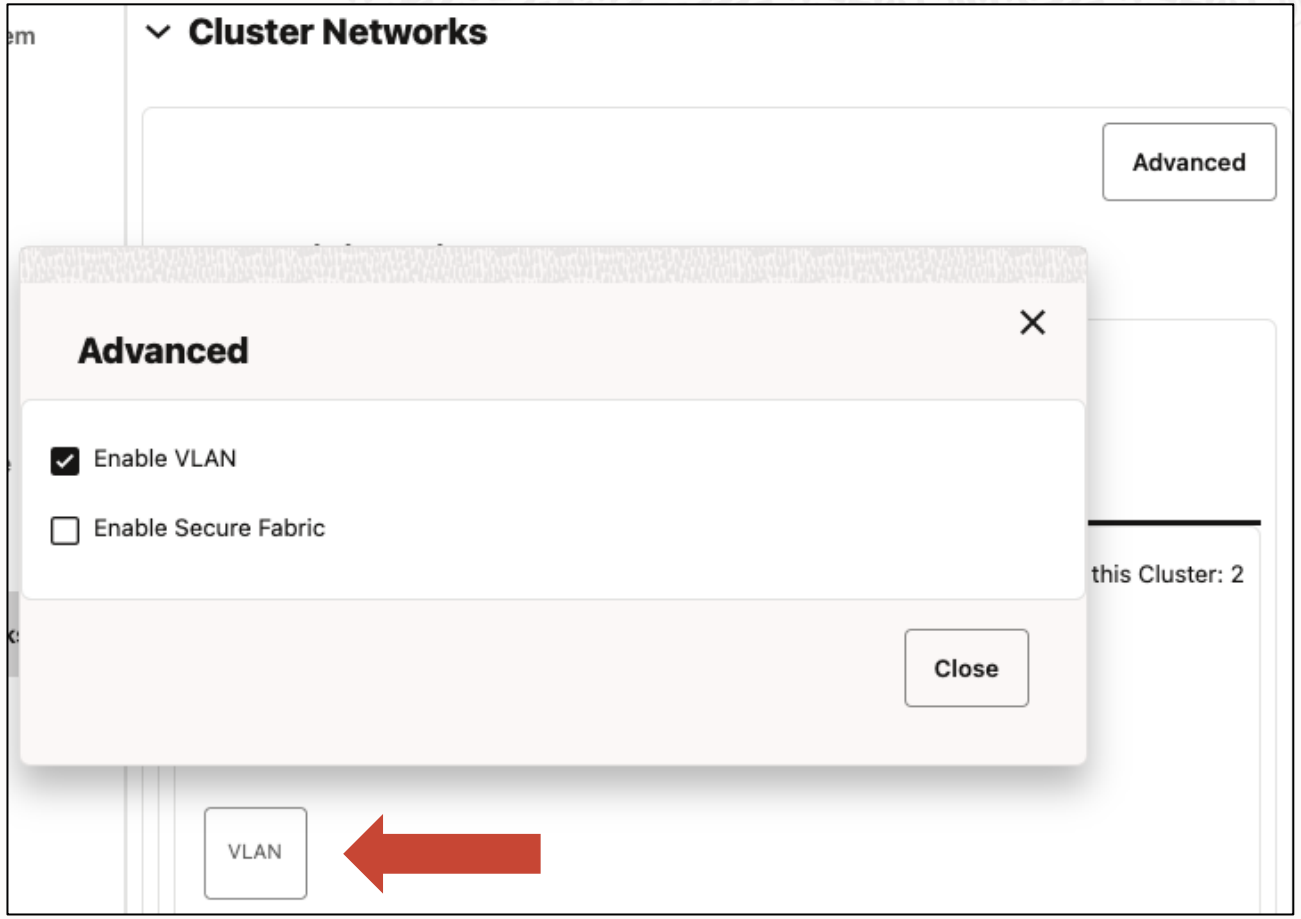
Cluster Configuration

- Each VM cluster has its own configuration
  - OS users and groups
  - VM size (memory, CPU)
  - Grid infrastructure version and software location
  - Exadata software version
  - ASM disk groups (and underlying storage grid disks)
  - Database version(s) and software location(s)
  - Starter database configuration
  - Client, Backup, and Admin networking configuration

OEDA Configuration Tool

Advanced Network Configuration

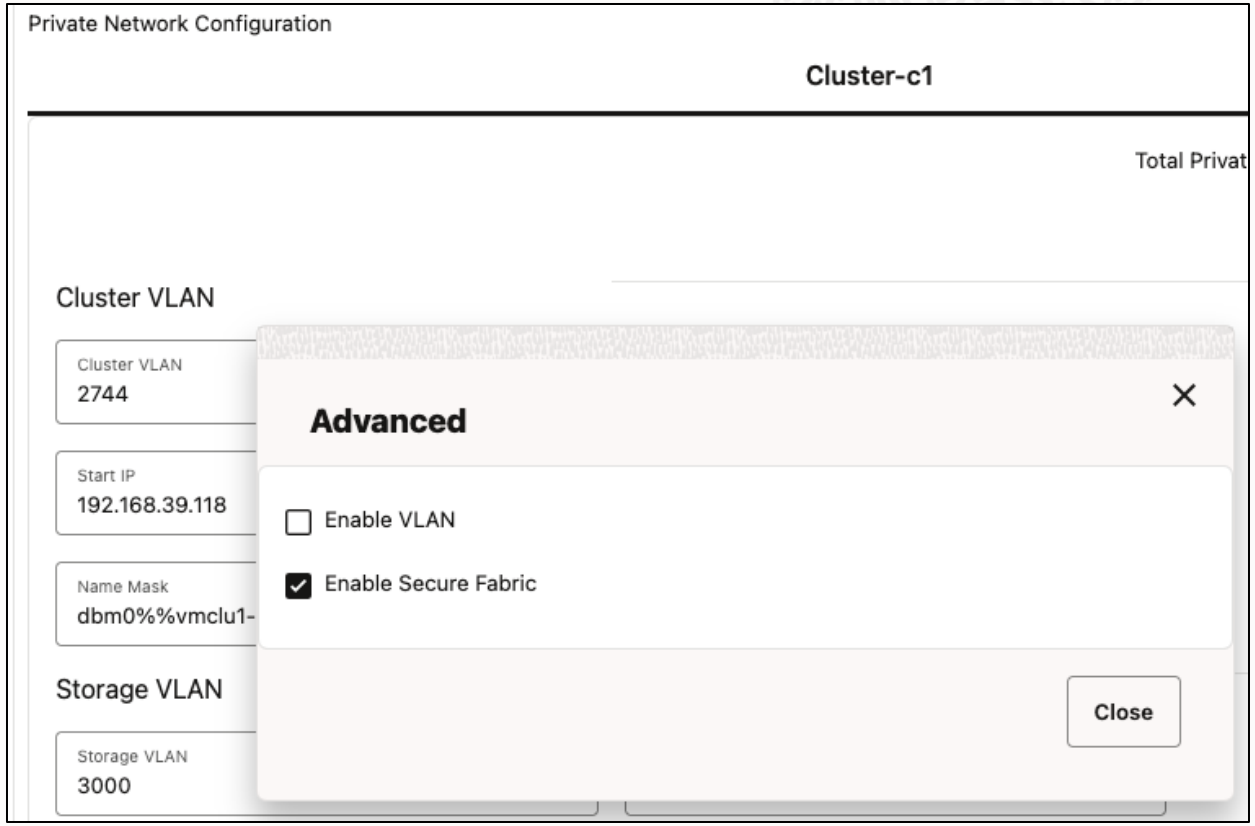
- Admin and Client Networks 802.1Q VLAN Tagging
  - To separate Admin and Client Networks traffic across VMs, use distinct VLAN ID and IP info for each cluster
  - Admin and Client Network switches must have VLAN tag configuration done before OEDA deployment



OEDA Configuration Tool

Advanced Network Configuration

- Private Network Secure Fabric
  - Secure RDMA Fabric Isolation uses RoCE VLANs to enable strict network isolation for Oracle RAC clusters
  - Multiple VM clusters share storage server resources but cannot communicate with each other
  - Secure Fabric is chosen by default for new on-premises deployments starting with the Oct 2024 OEDA release



OEDA Configuration Tool

Installation Template

- Verify proper settings for all VM clusters in Installation Template so the environment can be properly configured before deployment (DNS, switches, VLANs, etc.)


ORACLE
EXADATA
Installation Template

Clusters Information

Cluster:Cluster-cf3413015-352d-4443-0932-9f4ce4ca0314_id

Cluster Information:		Database:	
Version	**19.18.0.0.230117**	*Version*	**19.18.0.0.230**
Name	**Cluster-cl**	*Name*	**db1db1**
Customer Name	**Customer**	*Database Home*	**/u01/app/ora**
Application	**Application**	*Inventory Location*	**/u01/app/ora**
Home	**/u01/app/19.0.0.0/grid**	*Block Size*	**8192**
Inventory Location	**/u01/app/oraInventory**	*Database Template*	**OLTP**
Base Dir	**/u01/app/oracle**	*Database Type*	**RAC Databa**
Client Domain	**customer.nodomain**	*Character Sets*	**AL32UTF8**

OEDA Configuration Tool

Network Requirements for VM Deployment

Component	Domain	Network	Example hostname
Database servers	KVM host (one per database server)	Mgmt eth0	dm01dbadm01
		Mgmt ILOM	dm01dbadm01-ilom
	KVM guest (one or more per database server)	Mgmt eth0	dm01dbadm01vm01
		Client bondeth0	dm01client01vm01
		Client VIP	dm01client01vm01-vip
		Client SCAN	dm01vm01-scan
		Private RoCE	dm01dbadm01vm01-priv
Storage servers (same as physical)		Mgmt eth0	dm01celadm01
		Mgmt ILOM	dm01celadm01-ilom
		Private RoCE	dm01celadm01-priv
Switches (same as physical)		Mgmt and Private	dm01sw-adm, dm01sw-roce

Guest Disk Layout

File system	Size	Use
/ (root)	15G	Root file system
/u01	20G	Oracle BASE
/u01/app/<ver>/grid	50G	Grid infrastructure software home
/u01/app/oracle/product/<ver>/dbhome_1	50G	Database software home
/tmp	3G	/tmp
/home	4G	User home directories
/var	2G	/var
/var/log	18G	System logs
/var/log/audit	1G	System audit logs
/crashfiles	20G	System kdump kernel crash vmcore
/boot	512M	System boot
Other LVM space	44G	LVDbSwap1, LVDbSys2, LVDbVar2, LVDoNotRemoveOrUse
TOTAL	**228G**	

Exadata KVM Basic Maintenance

- Primary maintenance tools
  - OEDACLI – OEDA Command Line Interface
  - vm_maker

- Refer to Exadata Database Machine Maintenance Guide
  - Managing Oracle Linux KVM Guests
    - https://docs.oracle.com/en/engineered-systems/exadata-database-machine/dbmmn/managing-oracle-vm-guests-kvm.html

Exadata KVM Migration

- Migrate databases on existing system to new Exadata KVM system
  - Methods
    - Create Data Guard standby on new Exadata KVM system, switchover (minimal downtime)
    - Duplicate existing databases to new Exadata KVM system
    - Back up existing databases, restore databases on new Exadata KVM system

- Convert existing ROCE-based Exadata system deployed bare metal/physical to KVM
  - Methods
    - Back up existing databases, redeploy system to KVM, restore databases
    - Convert one or subset of database servers at a time to KVM

Backup/Restore of Virtualized Environment

- KVM host
  - Standard backup/restore practices to external location
- KVM guest
  - Backup within KVM host: Snapshot the VM disk images and back up snapshot externally
  - Backup within KVM guest: Standard OS backup/restore practices apply
  - If over-provisioning local disk space – Restoring VM backup will reduce or eliminate space savings (i.e., relying on over-provisioning may prevent full VM restore)
- Database backup/restore
  - Use standard Exadata MAA practices with RMAN, ZDLRA, and Cloud Storage
- Refer to Exadata Database Machine Maintenance Guide

Updating Software

Component to update	Method
Storage servers	• Same as physical – run patchmgr from any server with ssh access to all storage servers or use Storage Server Cloud Scale Software Update feature.
RDMA Network Fabric switches	• Same as physical – run patchmgr from any server with ssh access to all switches.
Database server – KVM host	• Run patchmgr from any server with ssh access to all KVM hosts. • KVM host update upgrades database server firmware. • KVM host reboot requires restart of all local VMs. • KVM guest software not updated during KVM host update. • KVM host/guest do not have to run same version, although specific update ordering may be required (see MOS 888828.1).
Database server – KVM guest	• Run patchmgr from any server with ssh access to all KVM guests. Typically done on a per-VM cluster basis (e.g., vm01 on all nodes, then vm02, etc.), or update all VMs on a KVM host before moving to next.
Grid Infrastructure / Database	• Use Fleet Patching and Provisioning (FPP), OEDACLI, or standard upgrade and patching methods, maintained on a per-VM cluster scope. GI/DB homes should be mounted disk images, like initial deployment.

Health Checks and Monitoring

- Exachk (AHF) runs in KVM host and KVM guest
  - Run in one KVM host – evaluates all KVM hosts, cells, switches
  - Run in one KVM guest of each VM cluster – evaluates all KVM guests, GI/DB of that cluster
- Exadata Storage Software Versions Supported by the Oracle Enterprise Manager Exadata Plug-in (MOS 1626579.1)
- Exawatcher runs in KVM host and KVM guest
- Database and Grid Infrastructure monitoring practices still apply
- Considerations
  - KVM host is not sized to accommodate Enterprise Manager or custom agents

Exadata MAA/HA

- Exadata MAA failure/repair practices
  - Refer to MAA Best Practices for Oracle Exadata Database Machine

- Live Migration is not supported – *use RAC to move workloads between nodes*

Resource Management

- Manage Exadata resources using Exadata I/O Resource Management (IORM)

- Manage resources within VMs and within a cluster using Database Resource Management (DBRM)
  - cpu_count set at the database instance level for multiple databases in a VM
  - Recommended minimum cpu_count=2

- No local disk resource management and prioritization
  - I/O intensive workloads should not use local disks
  - For higher I/O performance and bandwidth, use ACFS or NFS

Exadata KVM / Xen Comparison

Category	KVM-based	Xen-based
Terminology	host, guest	dom0, domU
Hardware support	X8M-2 through X11M (using RoCE network)	X2-2 through X8-2 (using InfiniBand network)
Hypervisor	KVM	Xen
VM management	vm_maker, OEDACLI	xm, OEDACLI, domu_maker
Database server software update	patchmgr using same ISO/yum repo for KVM host and guests	patchmgr using different ISO/yum repo for dom0 and domUs
File system configuration	xfs	ext4, and ocfs2 for EXAVMIMAGES


ORACLE

Our mission is to help people see data in new ways, discover insights, unlock endless possibilities.