

ORACLE

# Oracle Exadata Database Machine

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

**Exadata Development**

June 2023



# 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

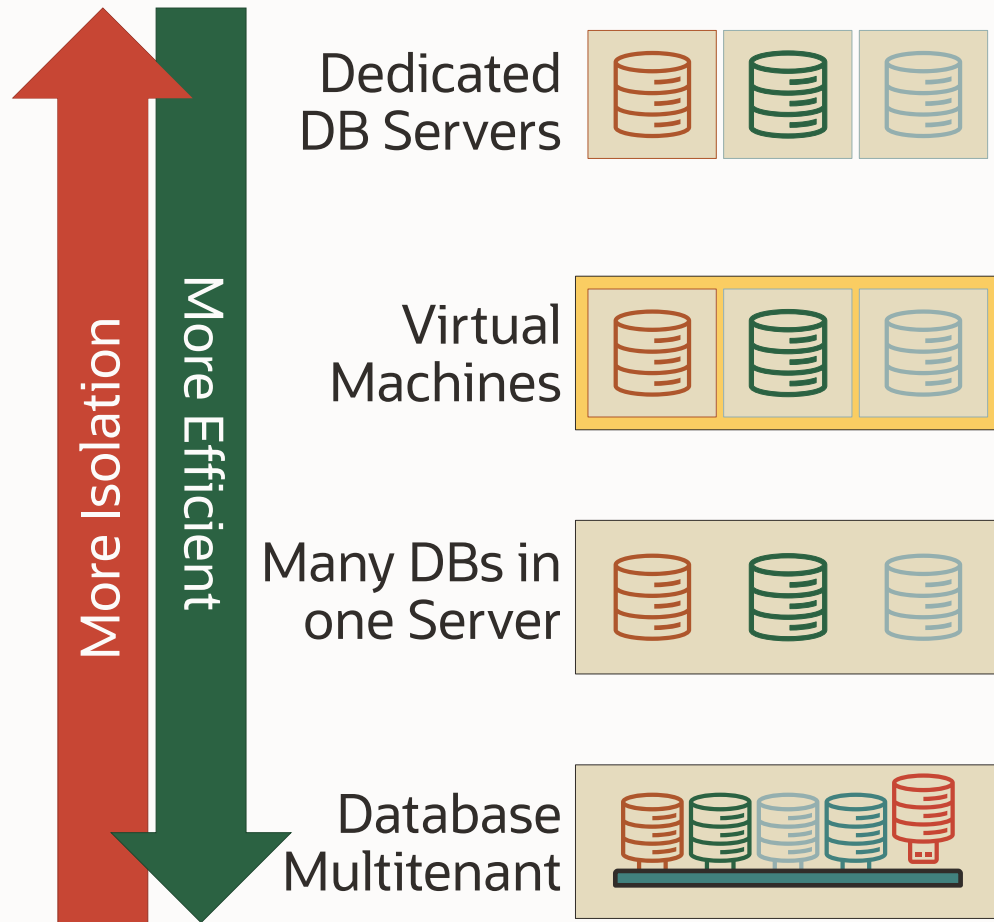
# Exadata Virtualization

## 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** (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](#)

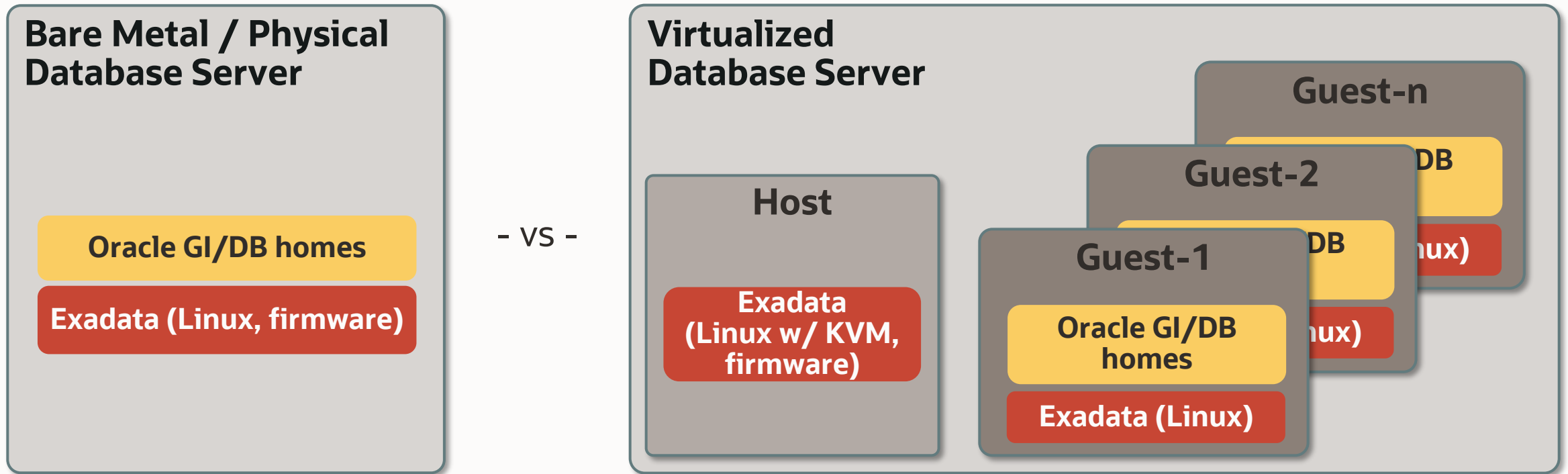
# Exadata Consolidation Options



- **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



No change to **Storage Grid, Networking,** or **Other**

# Differences Between Physical and Virtual

Topic	How Virtual differs from Physical
Reduced Licensing Option	Use Trusted Partitions to allocate OCPUs for Database License
Cluster configuration	System has one or more VM clusters, each with own Grid Infrastructure / Database software install
Network Isolation	Use Secure Fabric to isolate clusters while sharing underlying Exadata Storage
Exadata storage configuration	Separate grid disks and ASM disk groups (DATA,RECO) 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 <u>each</u> VM Cluster
Enterprise Manager	Enterprise Manager + Oracle Virtual Infrastructure plug-in + Exadata plug-in

Details expanded throughout remaining slides



# Exadata KVM Requirements



- Hardware
  - Exadata systems with RoCE interconnects (e.g., 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 KVM guests can run different Exadata database server versions
  - KVM Guests
    - Each guest runs Exadata database server software isolated from other guests
    - Each guest runs Grid Infrastructure and Database software isolated from other guests

# Exadata KVM Interoperability



- Interoperability between KVM/RoCE and Xen/InfiniBand
  - KVM supported only with RoCE interconnects
  - Xen supported only with InfiniBand interconnects (e.g., X8, X7, etc.)
    - X8 and earlier upgraded to or freshly deployed with Exadata 19.3 and later 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

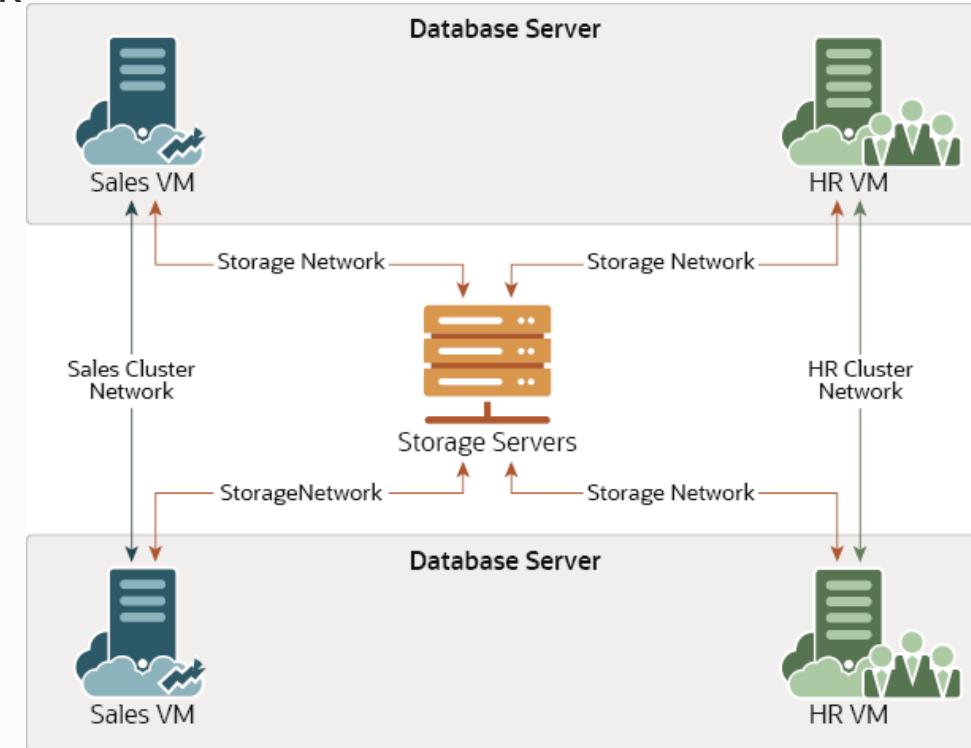


# Exadata KVM Security Isolation Recommendations

- 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>
- 802.1Q VLAN Tagging for Client and Admin Ethernet Networks
  - Configured w/ OEDA during deployment (requires pre-deployment switch config)
    - Client network manual configuration possible post-deployment (MOS 2710712.1)
- Private network isolation
  - Secure RDMA Fabric Isolation with Oracle Linux KVM
    - <https://docs.oracle.com/en/engineered-systems/exadata-database-machine/dbmmn/managing-oracle-vm-guests-kvm.html>
    - <https://docs.oracle.com/en/engineered-systems/exadata-database-machine/dbmin/exadata-network-requirements.html>
- RESTful remote access for storage server administration through ExaCLI

# Exadata 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



# Exadata KVM Sizing Recommendations



- **Maximum of 12 KVM guests per database server**
  - Eighth Rack systems support 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
  - Contact Oracle for sizing guidance

# Memory Sizing Recommendations



- **Cannot over-provision physical memory**
  - Sum of all KVM guests + KVM host reserved memory  $\leq$  installed physical memory
- **KVM Host Reserved Memory**
  - KVM host reserves portion of installed memory
  - Not available to KVM guests, enforced by `vm_maker`

# Memory Sizing Recommendations

- **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)
24 x 128 GB	X10M	3072	2800
24 x 96 GB	X10M	2304	2090
32 x 64 GB	X9M	2048	1870
24 x 64 GB	X10M, X9M, X8M	1536	1390
16 x 64 GB	X9M	1024	920
12 x 64 GB	X10M, X8M	768	660
16 x 32 GB	X10M, X9M	512	440
12 x 32 GB	X9M, X8M	384	328



# CPU Sizing Recommendations



- **CPU over-provisioning is allowed**
  - Up to 2x over-provisioning permitted with multiple VMs
    - Exceptions - No CPU over-provisioning on X10M systems:
      - with 512GB memory, or
      - 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
    - Eighth rack is allocated 2 vCPUs (1 core)



# CPU Sizing Recommendations

- **Guest CPU sizing (X10M 2x96-core CPUs)**
  - Single guest vCPU
    - Minimum 4 vCPU
    - Maximum 380 vCPU
  - Sum of all guests' vCPU
    - Max 380 vCPU if no over-provisioning
    - Max 760 vCPU if 2x over-provisioning<sup>1,2</sup>

Hardware	Min vCPU per guest	Max vCPU per guest	Max over-provision vCPU all guests
X10M	4	380	760 <sup>1,2</sup>
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



# Local Disk Space Sizing Recommendations

- KVM guest local file system disk space over-provisioning **not recommended**, but possible
- Actual allocated space initially much lower than apparent space due to sparseness and shareable reflinks (with multiple VMs), but will grow over time as shared space diverges and becomes less sparse
  - Long lived KVM guests should budget for full space allocation (assume no benefit from sparseness and shareable reflinks)
  - Over-provisioning may cause unpredictable out-of-space errors
  - Over-provisioning may prevent ability to restore disk image backup



# Local Disk Space Sizing Recommendations

- 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
- KVM guest local disk space can be extended after initial deployment by adding local disk images
- Disk space can be extended 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.

# Exadata Storage Recommendation

- Spread 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 existing disk group before adding new
- 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	DATA1_CD_{00..11}_cel01	RECOC1_CD_{00..11}_cel01
	db02vm01	DATA1_CD_{00..11}_cel02	RECOC1_CD_{00..11}_cel02
		DATA1_CD_{00..11}_cel03	RECOC1_CD_{00..11}_cel03
clu2	db01vm02	DATA2_CD_{00..11}_cel01	RECOC2_CD_{00..11}_cel01
	db02vm02	DATA2_CD_{00..11}_cel02	RECOC2_CD_{00..11}_cel02
		DATA2_CD_{00..11}_cel03	RECOC2_CD_{00..11}_cel03



# Deployment Specifications and Limits

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

- 1 – Eighth Rack systems maximum number of guests is 4
- 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

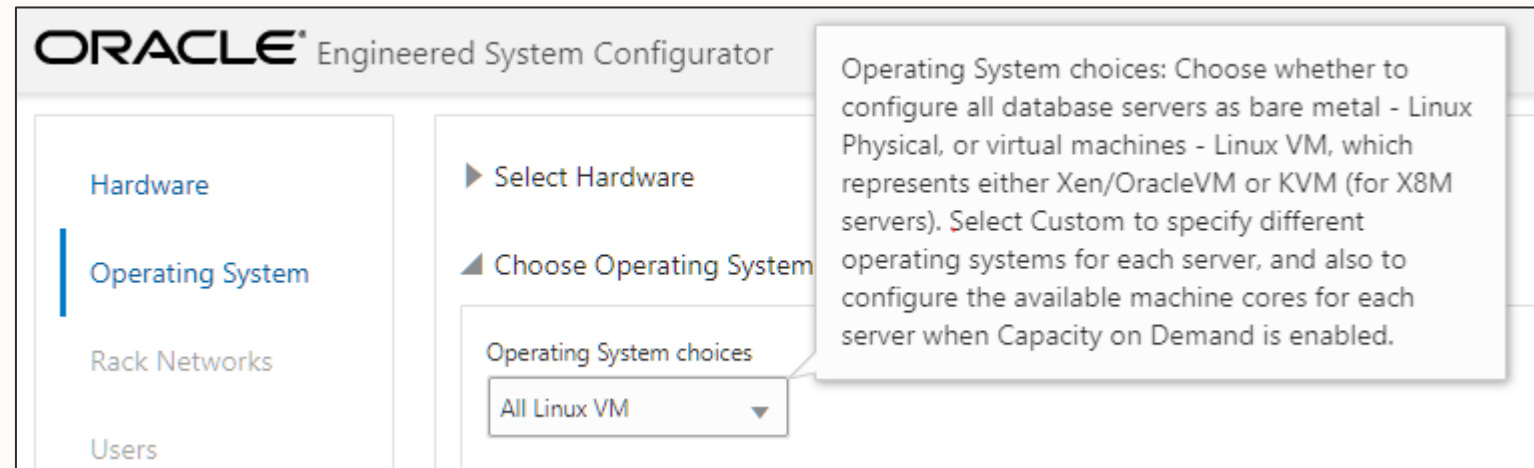
The Oracle Exadata Deployment Assistant, also known as OEDA, is the only tool to create VMs on Exadata

1. Create configuration with OEDA Configuration Tool
2. Prepare customer environment for OEDA deployment
  - Configure DNS, configure switches for VLANs (if necessary)
3. Prepare Exadata system for OEDA deployment
  - `# switch_to_ovm.sh; applyElasticConfig.sh`
4. Deploy system with OEDA Deployment Tool

# OEDA Configuration Tool

## Decide Virtual or Physical

- Section to pick KVM
  - 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

## Define Clusters

- Decide:
  1. Number of VM clusters to create
  2. 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.

Define Clusters

Cluster-c1 × Cluster-c2 × +

Cluster Name \*

Cluster-c2

Grid Home Owner

Available Machines

- OVS dbm0dbadm01.customer.nodomain
- OVS dbm0dbadm02.customer.nodomain
- Cell dbm0celadm01.customer.nodomain  
Exadata X10M Cell Node HC 22TB
- Cell dbm0celadm02.customer.nodomain  
Exadata X10M Cell Node HC 22TB
- Cell dbm0celadm03.customer.nodomain  
Exadata X10M Cell Node HC 22TB

# OEDA Configuration Tool

## Cluster Configuration

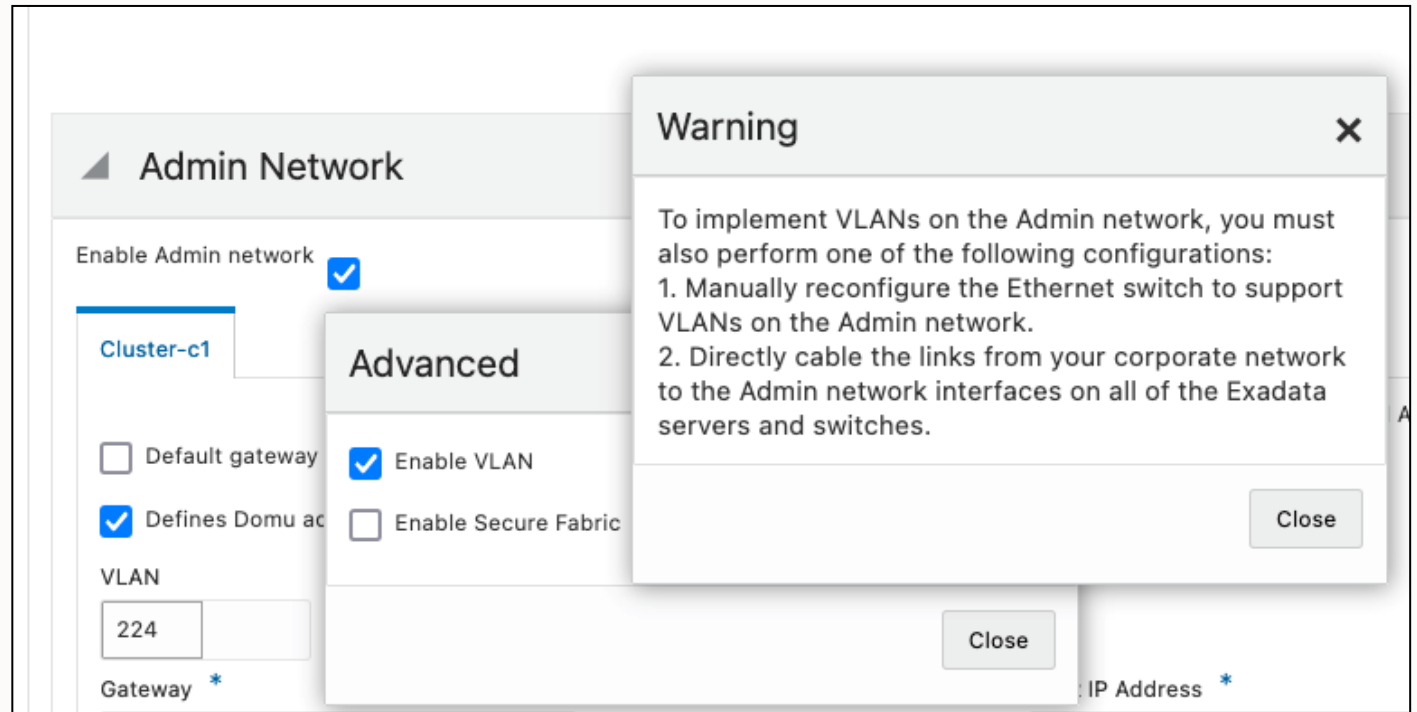


- Each VM cluster has its own configuration
  - OS users and groups
  - VM size (memory, CPU)
  - Grid infrastructure version and software location
  - Exadata software version
  - ASM disk groups (and underlying storage grid disks)
  - Database version and software location
  - Starter database configuration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

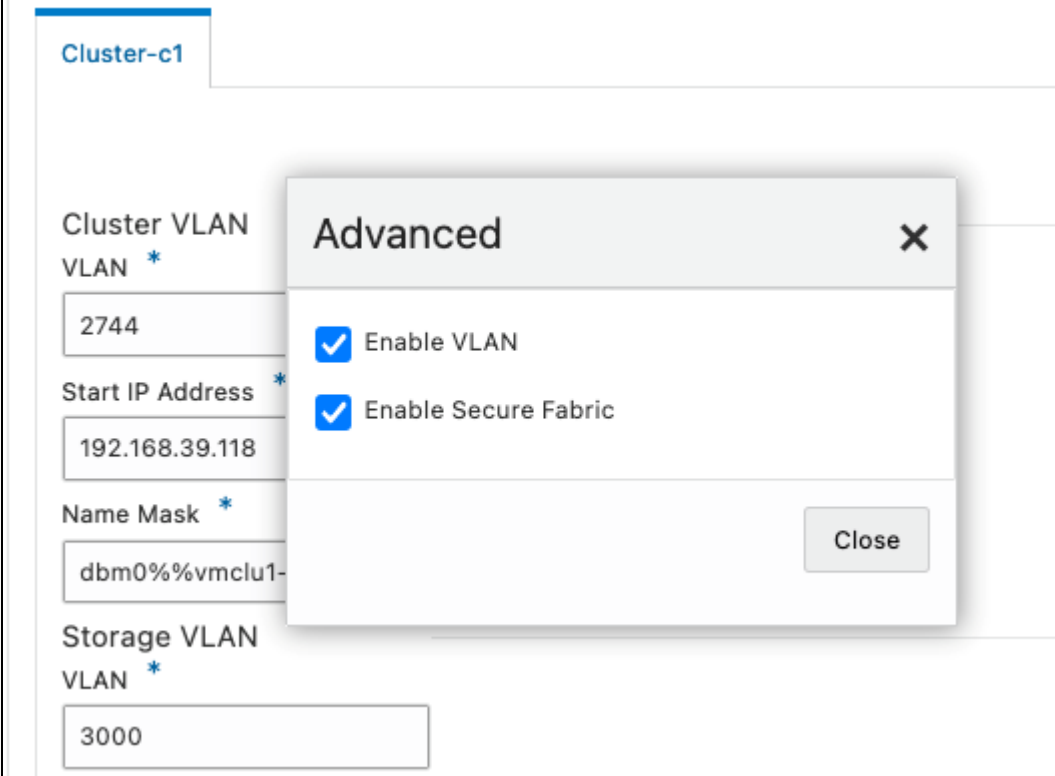




# OEDA Configuration Tool

## 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.



The screenshot displays the 'Cluster-c1' configuration page in the OEDA Configuration Tool. The page includes fields for 'Cluster VLAN', 'Start IP Address', 'Name Mask', and 'Storage VLAN'. An 'Advanced' dialog box is open, showing two checked options: 'Enable VLAN' and 'Enable Secure Fabric'. The 'Close' button is visible in the bottom right corner of the dialog box.

Cluster-c1

Cluster VLAN  
VLAN \*

2744

Start IP Address \*

192.168.39.118

Name Mask \*

dbm0%%vmclu1-

Storage VLAN  
VLAN \*

3000

Advanced

☒ Enable VLAN

☒ Enable Secure Fabric

Close

# 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 19.18.0.0.230117
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 Database
Client Domain customer.nodomain	Character Set AL32UTF8



# OEDA Configuration Tool

## Network Requirements



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
<b>TOTAL</b>	<b>228G</b>	



# 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
  - **Standard Exadata migration practices and considerations apply**
- 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 – Two Methods
  - Backup within KVM host: Snapshot the VM disk images and backup snapshot externally
  - Backup within KVM guest: Standard OS backup/restore practices apply
  - If over-provisioning local disk space - Restoring VM backup will reduce (may 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	<ul style="list-style-type: none"><li>• Same as physical - run patchmgr from any server with ssh access to all storage servers or use Storage Server Cloud Scale Software Update feature.</li></ul>
RDMA Network Fabric switches	<ul style="list-style-type: none"><li>• Same as physical - run patchmgr from any server with ssh access to all switches.</li></ul>
Database server – KVM host	<ul style="list-style-type: none"><li>• Run patchmgr from any server with ssh access to all KVM hosts.</li><li>• KVM host update upgrades database server firmware.</li><li>• KVM host reboot requires restart of all local VMs.</li><li>• KVM guest software <u>not</u> updated during KVM host update.</li><li>• KVM host/guest do not have to run same version, although specific update ordering may be required (see MOS 888828.1).</li></ul>
Database server – KVM guest	<ul style="list-style-type: none"><li>• 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.</li></ul>
Grid Infrastructure / Database	<ul style="list-style-type: none"><li>• Use Fleet Patching and Provisioning (FPP), OEDACLI, or standard upgrade and patching methods apply, maintained on a per-VM cluster scope. GI/DB homes should be mounted disk images, like initial deployment.</li></ul>



# 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 EM or custom agents

# Exadata MAA/HA



- Exadata MAA failure/repair practices still applicable.
  - Refer to [MAA Best Practices for Oracle Exadata Database Machine](#)
- Live Migration is not supported – *use RAC to move workloads between nodes*

# Resource Management



- Exadata Resource Management practices still apply
  - Exadata IO and flash resource management are all applicable and useful
- Within VMs and within a cluster, database resource management practices still apply
  - `cpu_count` still needs to be set at the database instance level for multiple databases in a VM. Recommended min `cpu_count`=2.
- No local disk resource management and prioritization
  - IO intensive workloads should not use local disks
  - For higher IO performance and bandwidth, use ACFS or NFS

# Exadata KVM / Xen Comparison

Category	KVM-based	Xen-based
Terminology	kvmhost, guest	dom0, domU
Hardware support	X8M-2 through X10M (using RoCE switches)	X2-2 through X8-2 (using InfiniBand switches)
Hypervisor	KVM (built in to UEK)	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	xf	ext4, and ocfs2 for EXAVMIMAGES



ORACLE

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

