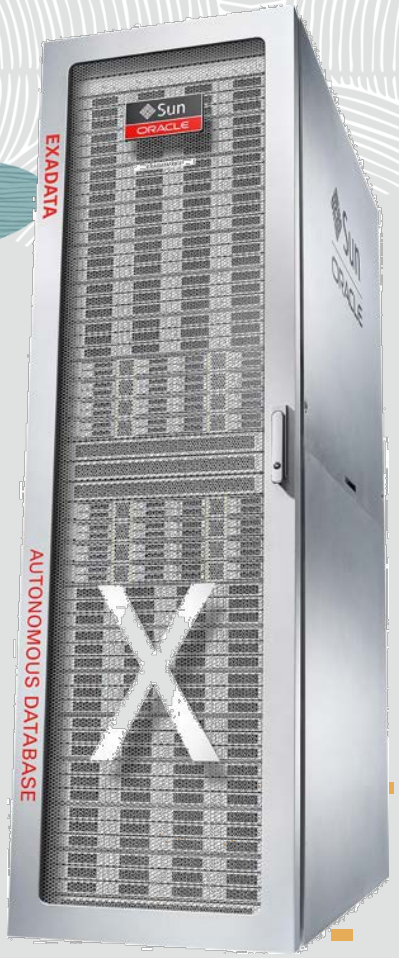
ORACLE



Oracle Exadata Database Machine

KVM Virtualization Best Practices for RoCE/PMEM-Based Systems

October 2019

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

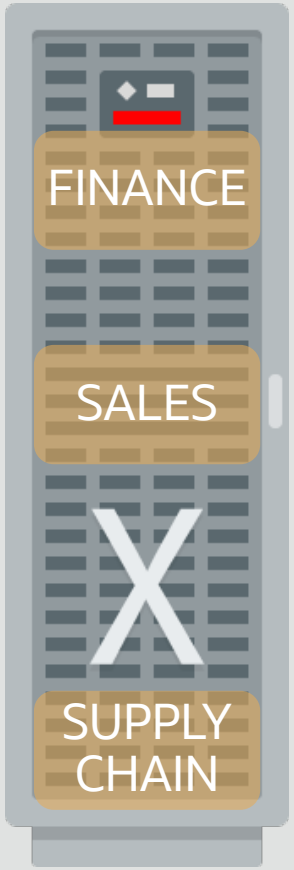
Exadata Virtualization

High-Performance Virtualized Database Platform Using KVM

X8M-2

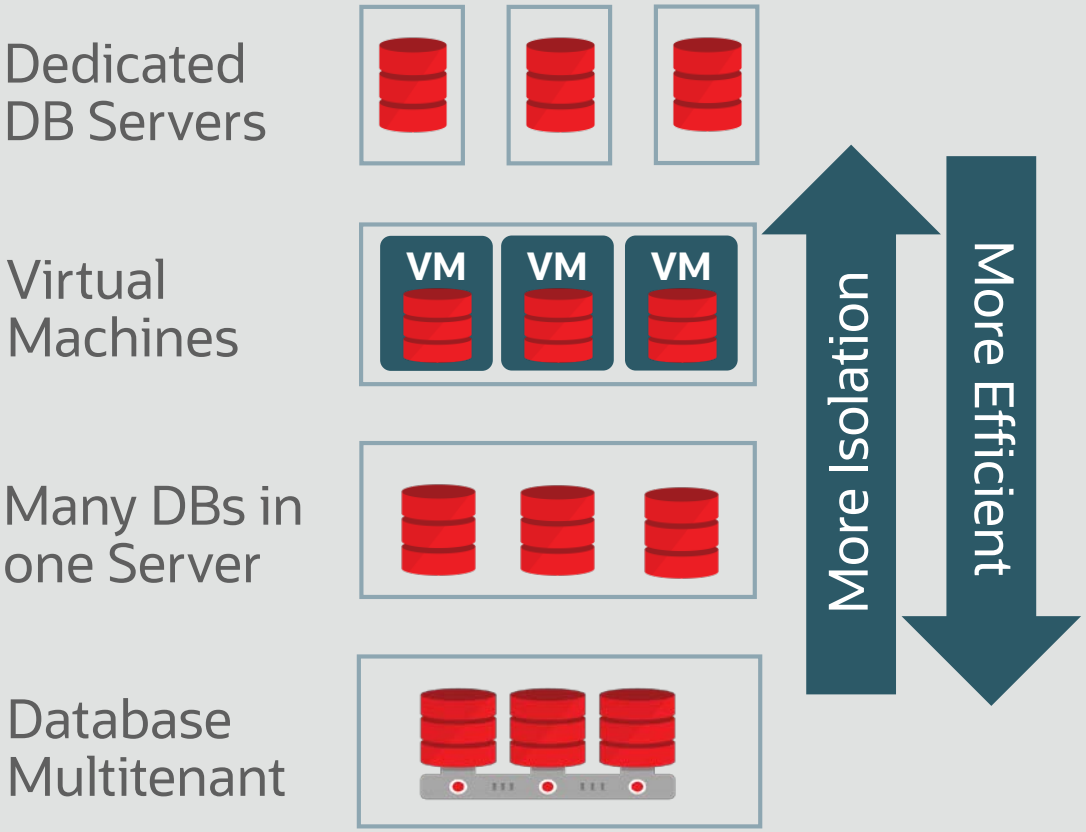
Exadata 19.3 and higher

DB 11.2.0.4 and higher
GI 12.1.0.2 and higher



- KVM hypervisor
  - Type 2 hypervisor running on a Linux kernel with improved performance
- VMs provide CPU, memory, OS, and sysadmin 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

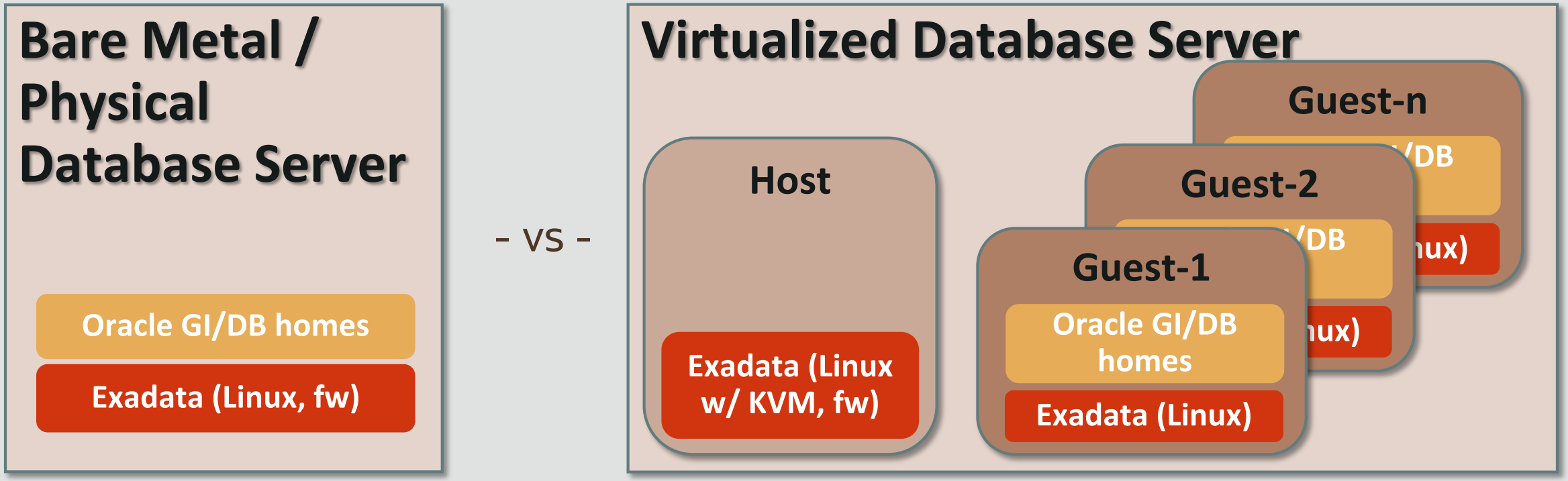
Exadata Consolidation Options



- VMs have good Isolation but poor efficiency and high management
  - 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/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

Details expanded throughout remaining slides

Topic	How Virtual differs from Physical
Hardware support	2-socket only
Cluster config	System has one or more VM clusters, each with own GI/DB install
Exadata storage config	Separate grid disks/DATA/RECO for each cluster
Dbnode disk config	Default file system sizes are small; GI/DB separate file systems
Software updates	Dbnodes require separate KVM host (Linux+fw) and guest (Linux) patchmgr updates
Exachk	Run once for KVM host/cells, run once for each VM cluster
Enterprise Manager	EM + Oracle Virtual Infrastructure plug-in + Exadata plug-in

Exadata KVM Requirements

- Hardware
  - 2-socket systems with RoCE interconnects (e.g. X8M-2)
- Software (see MOS 888828.1 for details and latest releases)
  - Exadata 19.3.0 or higher
    - Virtualization using Oracle Linux Kernel-based Virtual Machine (KVM)
    - KVM Host and guests can run different Exadata versions
  - Grid Infrastructure
    - Recommended – GI 19, using latest RU
    - Supported - GI 12.1.0.2 or higher release, using 2019-Jul RU, plus required patches
  - Database
    - Recommended - DB 19, using latest RU
    - Supported – DB 19, 18, 12.2.0.1, 12.1.0.2, or 11.2.0.4 – review MOS 888828.1 for requirements

Exadata KVM Requirements

- Interoperability between KVM/RoCE and Xen/InfiniBand
  - KVM supported only with RoCE interconnects (e.g. X8M)
  - Xen supported only with InfiniBand interconnects (e.g. X8, X7, etc.)
    - X8 and earlier upgraded to or freshly deployed with Exadata 19.3 continue to be based on Xen
  - Cannot inter-rack RoCE and InfiniBand
  - Separate KVM/RoCE and Xen/InfiniBand systems can be combined in same Data Guard / Golden Gate configuration
    - E.g. KVM-based system as primary, Xen-based system as standby
- Migration from Xen to KVM
  - Move database using Data Guard, GoldenGate, RMAN, ZDM

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
- 802.1Q VLAN Tagging for Client and Admin Ethernet Networks
  - Configured w/ OEDA during deployment (requires pre-deployment switch config) or configure manually post-deployment (MOS 2018550.1, MOS 2090345.1)
- Private network isolation
  - Server-level isolation using RoCE switch port configuration
  - Multiple VMs within same server share same VLAN
- Storage Server administration isolation through ExaCLI

Exadata KVM Sizing Recommendations

- Use Reference Architecture Sizing Tool to 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
  - Sizing tool currently does not size virtual systems

Memory Sizing Recommendations

- Can not over-provision physical memory
  - Sum of all KVM guests + KVM host reserved memory <= physical memory
- KVM host reserves 24GB + (6% total memory)
  - Not available to guests
- KVM guest memory sizing
  - Sum of all KVM guests memory <= 1390 GB (dbserver with max memory expansion)
  - Minimum 16 GB for KVM guest (to support OS, GI/ASM, starter DB, few connections)
  - Maximum 1390 GB for single KVM guest
  - Memory size on Exadata can not be changed online (guest restart required)

CPU Sizing Recommendations

- CPU over-provisioning is possible
  - But workload performance conflicts can arise if all guests become fully active
  - 1 vCPU == 1 hyper-thread; 1 core == 2 hyper-threads == 2 vCPUs
- KVM host reserves 2 cores (4 vCPUs)
  - Not available to guests
- KVM guest CPU sizing
  - Minimum CPU per KVM guest = 2 cores (4 vCPUs)
  - Maximum CPU per KVM guest = cores on system minus 2 (reserved KVM host)
    - Example: X8M-2 48 cores (96 vCPUs) – max CPU for a guest 46 cores (92 vCPUs)
  - Number of cores/vCPUs assigned to a VM can be changed online

Local Disk 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, 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 Sizing Recommendations

- X8M-2 database server – 4 x 1.2 TB local disks configured RAID5
  - Default local disk space available for VMs 1.46 TB, online resizable to 3.15 TB
- Default disk space used per KVM guest 141 GB
  - KVM guest local space can be extended after initial deployment by adding local disk images (i.e. a file in the kvmhost file system that a kvmguest sees as its local 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 db02vm01	DATAC1_CD_{00..11}_cel01 RECOC1_CD_{00..11}_cel01 DATAC1_CD_{00..11}_cel02 RECOC1_CD_{00..11}_cel02 DATAC1_CD_{00..11}_cel03 RECOC1_CD_{00..11}_cel03
clu2	db01vm02 db02vm02	DATAC2_CD_{00..11}_cel01 RECOC2_CD_{00..11}_cel01 DATAC2_CD_{00..11}_cel02 RECOC2_CD_{00..11}_cel02 DATAC2_CD_{00..11}_cel03 RECOC2_CD_{00..11}_cel03

Deployment Specifications and Limits

Category		X8M-2
VMs	Max KVM guests per database server	12
Memory	Physical per database server (default/max)	384 GB / 1.5 TB
	Min per KVM guest	16 GB
	Max per KVM guest	1390 GB
CPU/vCPU	Cores/vCPU* per database server	48/96
	Min cores/vCPU per KVM guest	2/4
	Max cores/vCPU per KVM guest	46/92
Disk	Total usable disk per database server for all KVM guests	3.15 TB
	Used disk per KVM guest at deployment	141 GB

*1 core = 1 OCPU = 2 hyper-threads = 2 vCPUs

Deployment Overview

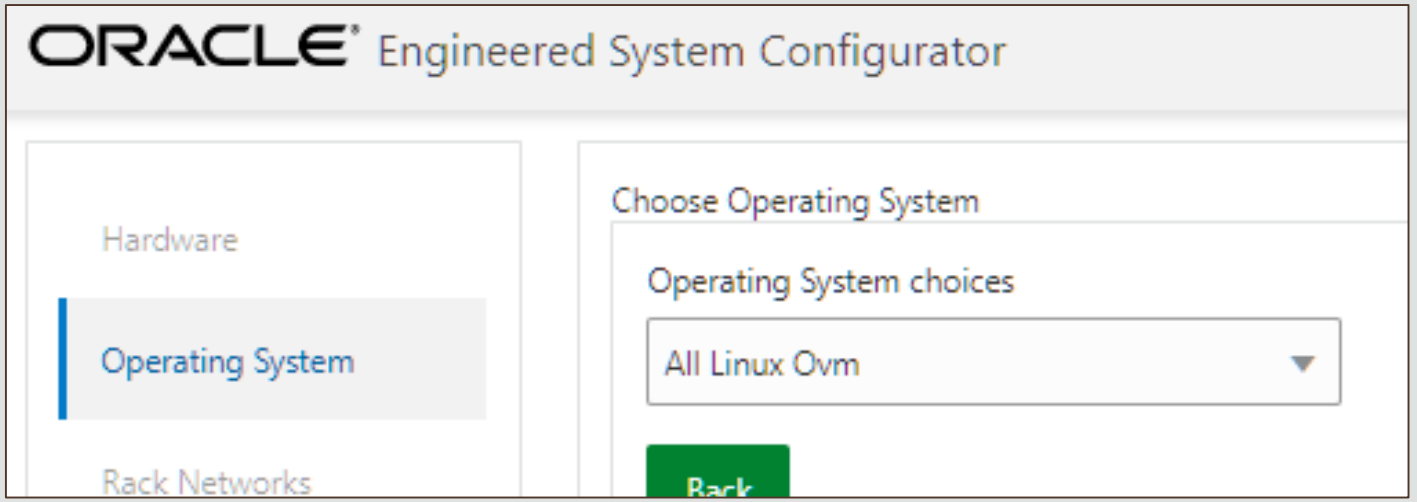
OEDA is the only tool that should be used to create VMs on Exadata

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 OVM
  - All Linux Physical
  - Custom (some OVM, some physical)

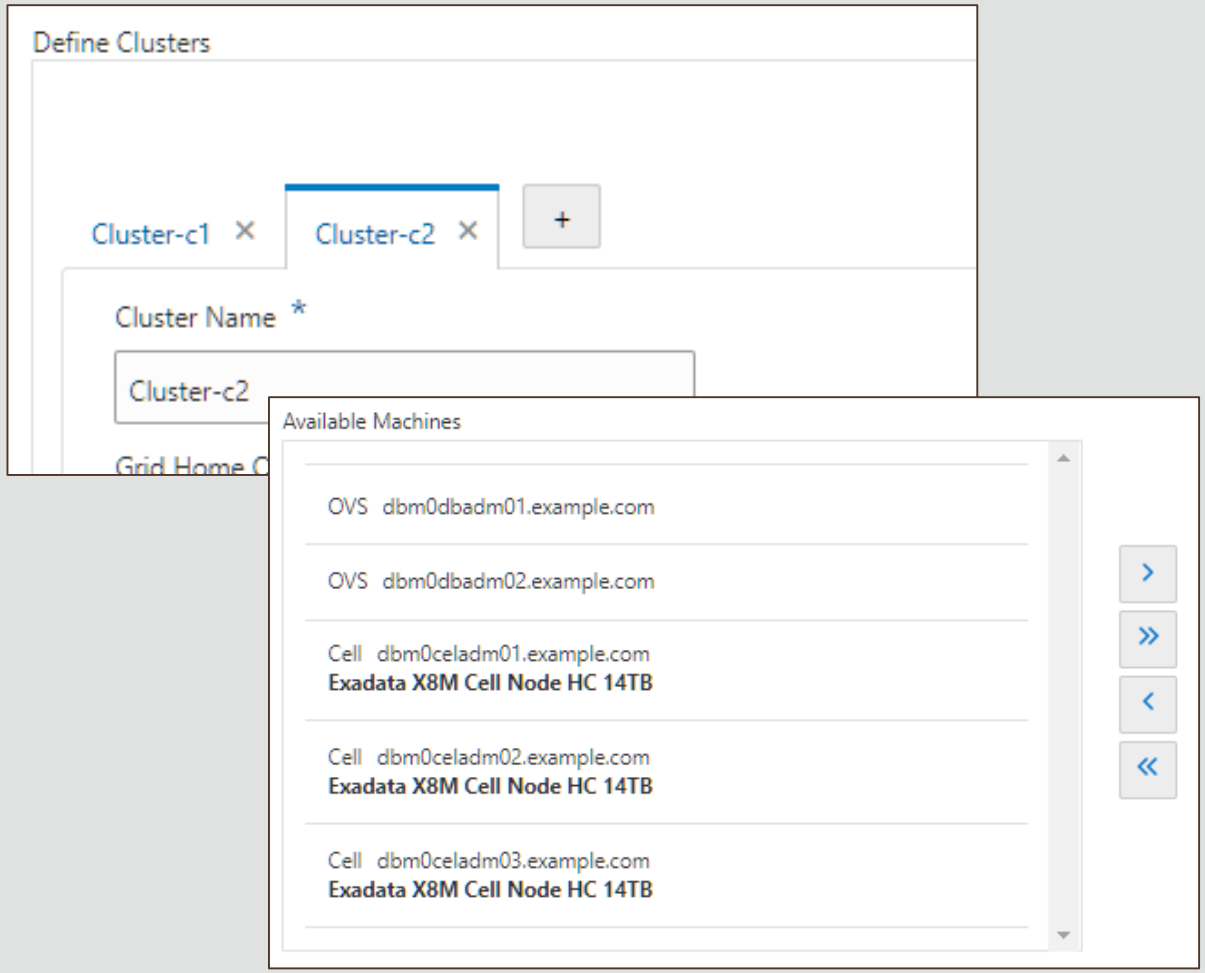


- A database server is configured either KVM or Physical

OEDA Configuration Tool

Define Clusters

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



OEDA Configuration Tool

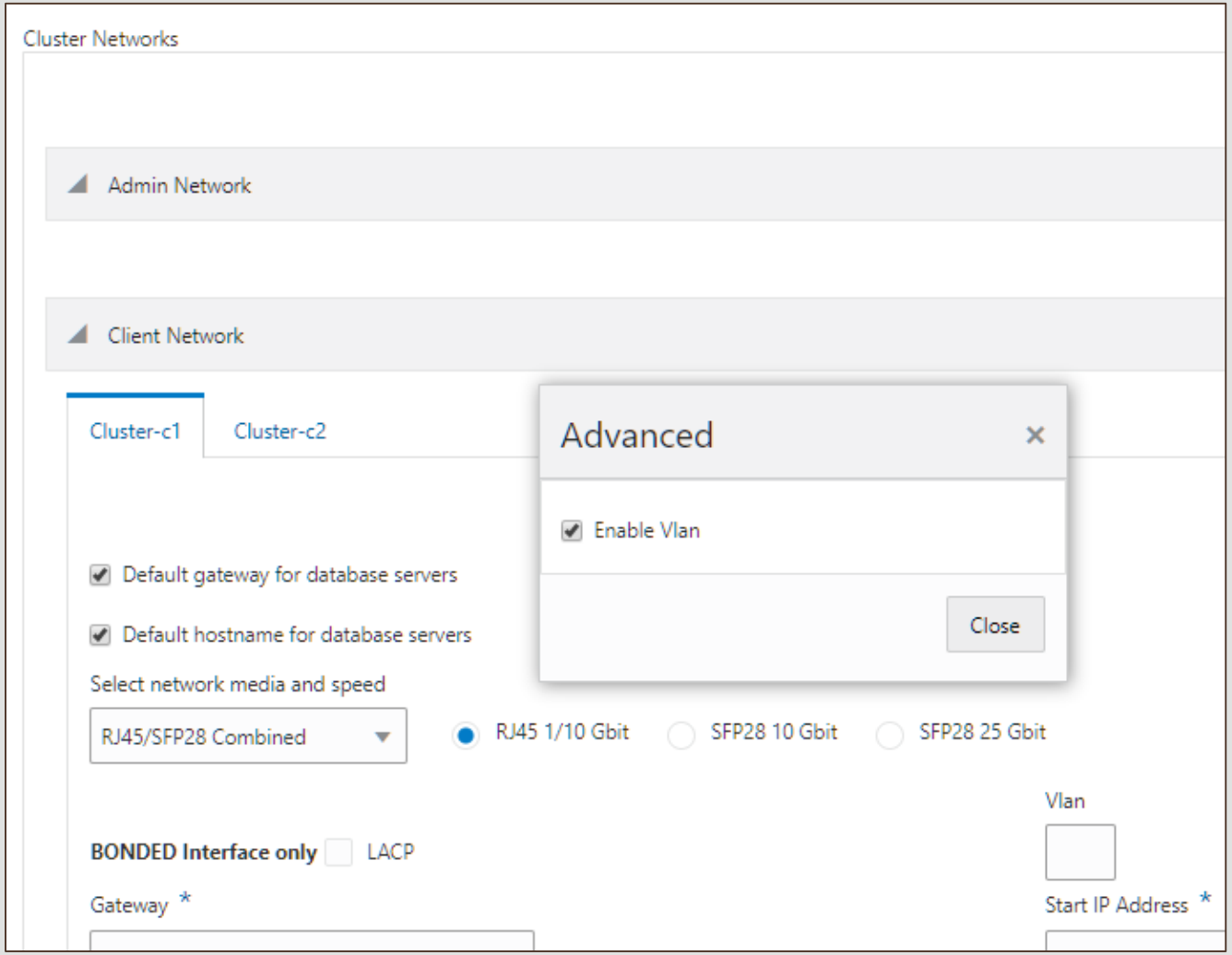
Cluster Configuration

- 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

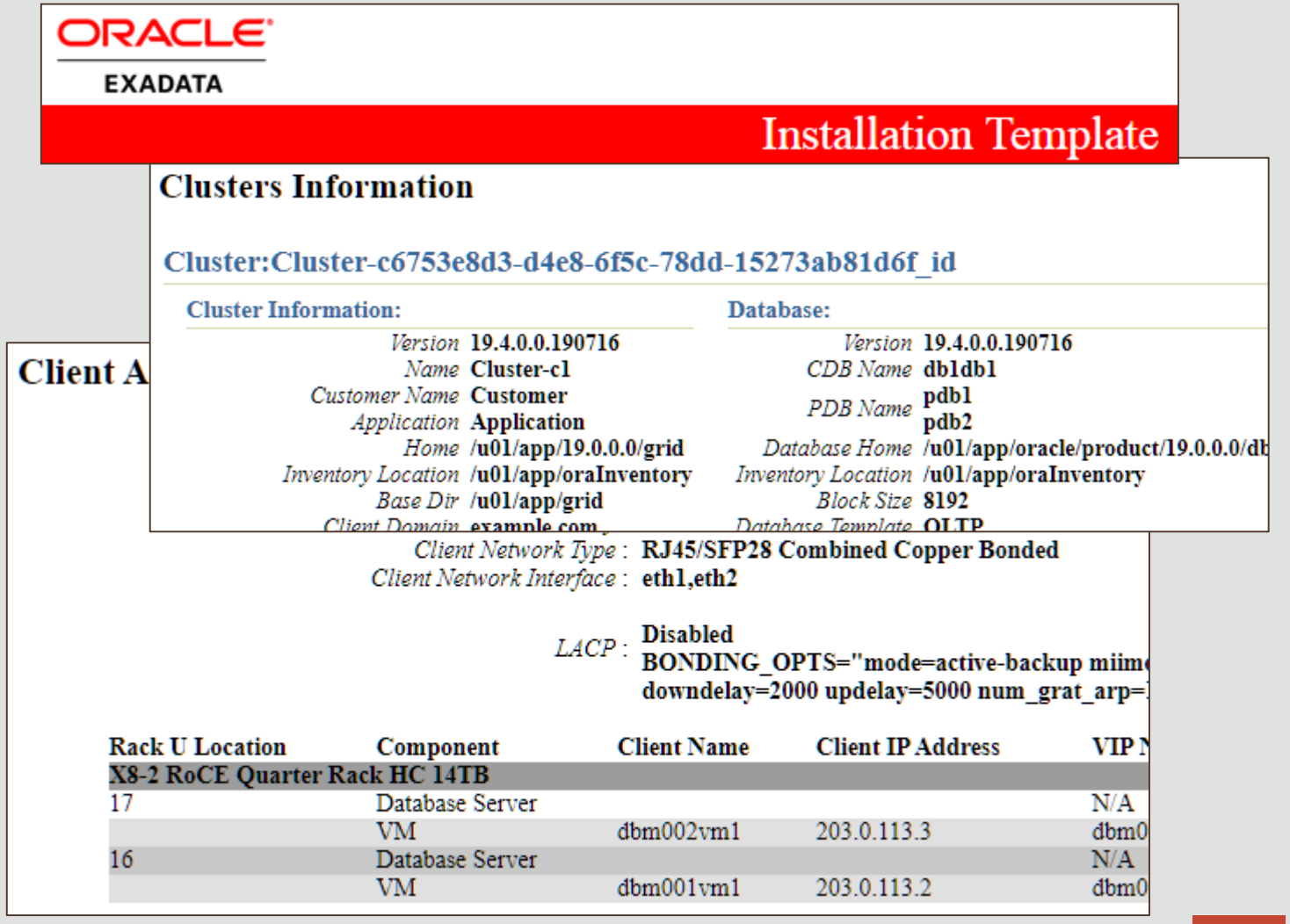
- Admin and Client Networks 802.1Q VLAN Tagging
  - To separate Admin and Client Networks traffic across VMs, use distinct VLAN ID and IP info for each cluster
  - Admin and Client Network switches must have VLAN tag configuration done before OEDA deployment



OEDA Configuration Tool

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-c6753e8d3-d4e8-6f5c-78dd-15273ab81d6f_id
Cluster Information:
Version 19.4.0.0.190716
Name Cluster-c1
Customer Name Customer
Application Application
Home /u01/app/19.0.0.0/grid
Inventory Location /u01/app/oraInventory
Base Dir /u01/app/grid
Client Domain example.com
Database:
Version 19.4.0.0.190716
CDB Name db1db1
PDB Name pdb1
pdb2
Database Home /u01/app/oracle/product/19.0.0.0/db
Inventory Location /u01/app/oraInventory
Block Size 8192
Database Template OLTP
Client A
Client Network Type : RJ45/SFP28 Combined Copper Bonded
Client Network Interface : eth1,eth2
LACP : Disabled
BONDING_OPTS="mode=active-backup miim
downdelay=2000 updelay=5000 num_grat_arp=
Rack U Location | Component | Client Name | Client IP Address | VIP
X8-2 RoCE Quarter Rack HC 14TB
17 | Database Server | | | N/A
VM | dbm002vm1 | 203.0.113.3 | dbm0
16 | Database Server | | | N/A
VM | dbm001vm1 | 203.0.113.2 | dbm0

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	dm01dbadm01**vm01**
		Client bondeth0	dm01client01**vm01**
		Client VIP	dm01client01**vm01**-vip
		Client SCAN	dm01**vm01**-scan
		Private RoCE	dm01dbadm01**vm01**-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

Exadata KVM Basic Maintenance

- Primary maintenance tools
  - vm_maker
  - OEDACLI - OEDA Command Line Interface
- Refer to Exadata Database Machine Maintenance Guide
  - Chapter 6 – "Managing Oracle VM Domains Using KVM"

Exadata KVM Migration

- Dynamic or online method to change physical to virtual
  - Data Guard or backups can be used to move databases – minimum downtime
  - Convert one node or subset of nodes to virtual at a time
- Migrating an existing physical Exadata rack to use virtual requires
  - Backing up existing databases, redeploying existing HW with OEDA and then Restoring Databases
  - Duplicating the databases to existing Exadata KVM configuration
  - If moving from source to a new target, standard Exadata migration practices still apply.
    - Refer to Best Practices for Migrating to Exadata Database Machine

Exadata KVM Migration

- Dynamic or online method to change physical to virtual using any of the procedures below
  - Migrate to OVM RAC cluster using the existing bare metal Oracle RAC cluster with zero downtime
  - Migrate to OVM RAC cluster by creating a new OVM RAC cluster with minimal downtime
  - Migrate to OVM RAC cluster using Oracle Data Guard with minimal downtime
  - Migrate to OVM RAC cluster using RMAN backup and restore with complete downtime
- For requirements and detailed steps, refer to My Oracle Support note 2099488.1: *Migration of a Bare metal RAC cluster to an OVM RAC cluster on Exadata*

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	Same as physical - run patchmgr from any server with ssh access to all cells, 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 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.

Health Checks and Monitoring

- Exachk runs in KVM host and KVM guest
  - Run in one KVM host for all KVM hosts, cells, switches
  - Run in one KVM guest of each VM cluster for 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/GI monitoring practices still apply
- Considerations
  - KVM host is not sized to accommodate EM or custom agents
  - Oracle Linux Virtualization Manager not supported on Exadata

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 DBFS on Exadata or NFS.

Exadata KVM / Xen Comparison

Category	KVM-based	Xen-based
Terminology	kvmhost, guest	dom0, domU
Hardware support	X8M-2 (using RDMA Network Fabric 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
Deployment	reclaimdisks.sh not needed	reclaimdisks.sh is manual step
File system configuration	xfs	ext4, and ocfs2 for EXAVMIMAGES

Integrated Cloud

Applications & Platform Services