Oracle Exadata Database Machine

Oracle Exadata and OVM Best Practices

November 2019
Topics Covered

• Use Cases
• Exadata OVM Software Requirements
• Exadata Isolation Considerations
• Exadata OVM Sizing and Prerequisites
• Exadata OVM Deployment Overview
• Exadata OVM Administration and Operational Life Cycle
• Migration, HA, Backup/Restore, Upgrading/Patching
• Monitoring, Resource Management
Exadata Virtual Machines
High-Performance Virtualized Database Platform

- XEN hypervisor
- 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
  - I/Os go directly to high-speed InfiniBand bypassing hypervisor
- Combine with Exadata network and I/O prioritization to achieve unique full stack isolation
- Trusted Partitions allow licensing by virtual machine

No Additional Cost
X8-2, X7-2, X6-2, X5-2, X4-2, X3-2, X2-2
DB 11.2 and higher

Copyright © 2019 Oracle and/or its affiliates.
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.

Dedicated DB Servers

Virtual Machines

Many DBs in one Server

Database Multitenant

More Isolation

More Efficient
Software Architecture Comparison

Database Server: Bare Metal / Physical versus OVM

Bare Metal / Physical Database Server

- Oracle GI/DB homes
- Exadata (Linux, firmware)

OVM Database Server

- dom0
- Exadata (Linux, Xen, firmware)

- domU-1
- Oracle GI/DB homes
- Exadata (Linux)

- domU-2
- Oracle GI/DB homes

- domU-3

No change to **Storage Grid**, **Networking**, or **Other**
Differences Between Physical and OVM

Details expanded throughout remaining slides

<table>
<thead>
<tr>
<th>Topic</th>
<th>How OVM differs from Physical</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hardware support</td>
<td>2-socket only</td>
</tr>
<tr>
<td>Cluster config</td>
<td>System has one or more VM clusters, each with own GI/RAC/DB install</td>
</tr>
<tr>
<td>Exadata storage config</td>
<td>Separate griddisks/DATA/RECO for each VM cluster; By default no DBFS disk group</td>
</tr>
<tr>
<td>Dbnodes disk config</td>
<td>VM filesystem sizes are small; GI/DB separate filesystems</td>
</tr>
<tr>
<td>Software updates</td>
<td>Dbnodes require separate dom0 (Linux+firmware) and domU (Linux) patchmgr updates</td>
</tr>
<tr>
<td>Exachk</td>
<td>Run once for dom0/cells/ibswitches, run once for each VM cluster</td>
</tr>
<tr>
<td>Enterprise Manager</td>
<td>EM + Exadata plugin + Virtualization Infrastructure plugin</td>
</tr>
</tbody>
</table>
Exadata VM Usage

- Primary focused on consolidation and isolation
- Can only run certified Oracle Linux versions
  - Windows, RedHat, and other guest operating systems are not supported
- Can virtualize other lightweight products
  - E.g. Lightweight apps, management tools, ETL tools, security tools, etc.
- **Not** recommended for heavyweight applications
  - E.g. E-business Suite or SAP application tier
  - Instead use Private Cloud Appliance
Exadata OVM Requirements

• Hardware
  • 2-socket database servers supported (X2-2 through X8-2)

• Software
  • Recommend latest Exadata 18.x or 19.x software
    • Supplied software (update with patchmgr - see MOS 888828.1)
      • domU and dom0 run same UEK kernel as physical
      • domU runs same Oracle Linux (OL) as physical
      • dom0 runs Oracle VM Server (OVS) 3.x
  • Grid Infrastructure / Database
    • Recommend 19c with latest quarterly update
    • Supported 19c, 18c, 12.2.0.1, 12.1.0.2, or 11.2.0.4
Exadata Security Isolation Recommendations

- Each VM RAC cluster has its 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 Management Ethernet Networks
  - Dbnodes configured w/ OEDA during deployment (requires pre-deployment switch config)
  - Or configure manually post-deployment
    - Client network - MOS 2018550.1  Management network - MOS 2090345.1

- InfiniBand Partitioning with PKEYs for Exadata Private Network
  - OS and InfiniBand switches configured w/ OEDA during deployment

- Storage Server administration isolation through ExaCLI
Exadata OVM Sizing Recommendations

Use Reference Architecture Sizing Tool to determine CPUs, memory, disk space needed by each database

- Sizing evaluation should be done prior to deployment since OEDA will deploy your desired VM configuration in an automated and simple manner.
- Changes can be made post deployment, but requires many more steps
- Sizing approach does not really change except for accommodating DOM0, and additional system resources per VM
- Sizing tool currently does not size virtual systems
- Consider dom0 memory and CPU usage in sizing
Memory Sizing Recommendations

• Can not over-provision physical memory
  • Sum of all VMs + dom0 memory used cannot exceed physical memory
  • Sum of all VM memory <= 720 GB
    • X8, X7, X6 database servers support maximum 768 GB physical memory when deployed virtualized (non-virtualized systems support higher)

• dom0 memory sizing
  • 8 GB (do not change unless directed by Oracle)

• VM memory sizing
  • Initially set during OEDA configuration
  • Minimum 16 GB per VM (to support OS, GI/ASM, starter DB, few connections)
  • Maximum 720 GB for single VM
  • Memory size on Exadata can not be changed online (VM restart required)
CPU Sizing Recommendations

- CPU over-provisioning is possible
  - But workload performance conflicts can arise if all VMs become fully active
- Dom0 CPU sizing
  - Allocated 2 cores (4 vCPUs - do not change unless directed by Oracle)
- VM CPU sizing
  - Minimum per VM is 2 cores (4 vCPUs)
    - 1 vCPU == 1 hyper-thread; 1 core == 2 hyper-threads == 2 vCPUs
  - Maximum per VM per DB Server is number of cores minus 2 for dom0
    - E.g.: for X8-2, maximum per VM per DB Server is 46 cores (48 total minus 2 for dom0)
  - vCPU initially set during OEDA configuration
  - vCPU can be changed dynamically (online while VM remains up)
Local Disk Sizing Recommendations

• Total local disk space available for VMs
  • X8 - 3.2TB; X7,X6,X5 - 1.6TB, 3.7TB with disk drive expansion kit; X4 - 1.6TB
• 190GB used per VM at deployment, extendable post-deployment
• Actual allocated space for domU disk images initially much lower due to sparseness and shareable reflinks, but will grow with domU use as shared space diverges and becomes less sparse
  • Over-provisioning disk may cause unpredictable out-of-space errors inside VMs if dom0 space is exhausted
  • Restoring VM backup will reduce (may eliminate) space savings (i.e. relying on over-provisioning may prevent full VM restore)
  • Long lived / prod VMs should budget for full space allocation (assume no benefit from sparseness and shareable reflinks)
  • Short lived test/dev VMs can assume 100 GB allocation
• DomU local space can be extended after initial deployment by adding local disk images
  • Additionally, domU space can be extended with shared storage (e.g. ACFS, DBFS, external NFS) for user / app files
  • Avoid shared storage for Oracle/Linux binaries/config files. Access/network issues may cause system crash or hang.
Exadata Storage Recommendation

- DATA/RECO size for initial VM clusters should consider future VM additions
  - Using all space initially will require shrinking existing DATA/RECO before adding new
- Spread DATA/RECO for each VM cluster across all disks on all cells
  - By default no DBFS disk group
- Enable ASM-Scoped Security to limit grid disk access

<table>
<thead>
<tr>
<th>VM Cluster</th>
<th>Cluster nodes</th>
<th>Grid disks (DATA/RECO for all clusters on all disks in all cells)</th>
</tr>
</thead>
<tbody>
<tr>
<td>clu1</td>
<td>db01vm01, db02vm01</td>
<td>DATAC1_CD_{00..11}<em>cel01, RECOC1_CD</em>{00..11}<em>cel01, DATAC1_CD</em>{00..11}<em>cel02, RECOC1_CD</em>{00..11}<em>cel02, DATAC1_CD</em>{00..11}<em>cel03, RECOC1_CD</em>{00..11}_cel03</td>
</tr>
<tr>
<td>clu2</td>
<td>db01vm02, db02vm02</td>
<td>DATAC2_CD_{00..11}<em>cel01, RECOC2_CD</em>{00..11}<em>cel01, DATAC2_CD</em>{00..11}<em>cel02, RECOC2_CD</em>{00..11}<em>cel02, DATAC2_CD</em>{00..11}<em>cel03, RECOC2_CD</em>{00..11}_cel03</td>
</tr>
</tbody>
</table>
## Deployment Specifications and Limits

<table>
<thead>
<tr>
<th>Hardware</th>
<th>X3-2</th>
<th>X4-2</th>
<th>X5-2</th>
<th>X6-2</th>
<th>X7-2</th>
<th>X8-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max VMs per database server</td>
<td></td>
<td></td>
<td></td>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physical per database server (default/max)</td>
<td>256 GB</td>
<td>256 GB</td>
<td>256 GB</td>
<td>256 GB</td>
<td>384 GB</td>
<td>384 GB</td>
</tr>
<tr>
<td></td>
<td>512 GB</td>
<td>512 GB</td>
<td>768 GB^2</td>
<td>768 GB</td>
<td>768 GB^2</td>
<td>768 GB^2</td>
</tr>
<tr>
<td>Min per VM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>16 GB</td>
<td></td>
</tr>
<tr>
<td>Max per VM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>464 GB</td>
<td>720 GB</td>
</tr>
<tr>
<td>Default setting</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Initially set during OEDA configuration</td>
<td></td>
</tr>
<tr>
<td>Cores/vCPU^1 per database server</td>
<td>16</td>
<td>24</td>
<td>36</td>
<td>44</td>
<td>48</td>
<td></td>
</tr>
<tr>
<td>Min cores/vCPU per VM</td>
<td></td>
<td></td>
<td></td>
<td>2 core (4 vCPUs)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Max cores/vCPU per VM</td>
<td></td>
<td></td>
<td></td>
<td>Cores minus 2 (dom0 assigned 2 cores/4vCPUs)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Default setting</td>
<td></td>
<td></td>
<td></td>
<td>Initially set during OEDA configuration</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total usable disk per dbserver for all VMs</td>
<td>700 GB</td>
<td>1.6 TB</td>
<td>1.6 TB (3.7 TB w/ DB Storage Expansion Kit)</td>
<td>3.2 TB</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Used disk per VM at deployment</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Footnotes: 1) 1 core = 1 OCPU = 2 hyper-threads = 2 vCPUs; 2) Systems deployed non-virtual support higher physical memory.

---

Copyright © 2019 Oracle and/or its affiliates.
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; reclaimdisks.sh; applyElasticConfig.sh
4. Deploy system with OEDA Deployment Tool
Note: OS VLAN config can be done by OEDA or post deployment (MOS 2018550.1)
OEDA Configuration Tool

Configuring OVM

- Screen to decide OVM or Physical
  - All OVM
  - All Physical
  - Some OVM, some physical
OEDA Configuration Tool

Define Clusters

• Decide
  • Number of VM clusters to create
  • Dbnodes and Cells that will make up those VM clusters
    • Recommend using all cells
• What is a “VM cluster?”
  • 1 or more user domains on different database servers running Oracle GI/RAC, each accessing the same shared Exadata storage managed by ASM.
Each VM cluster has its own configuration

- VM size (memory, CPU)
- Exadata software version
- Networking config
- OS users and groups
- GI/DB version and location
- Starter database config
- ASM disk group config
Grid infrastructure installed in each VM (grid disks “owned” by a VM cluster)

- Cluster 1 - DATAC1 / RECOC1 across all cells
- Cluster 2 - DATAC2 / RECOC2 across all cells
- Consider future clusters when sizing
- DBFS not configured
- ASM-Scoped Security permits a cluster to access only its own grid disks. Available with Advanced button.
• Ethernet VLAN ID and IP details
  • To separate Ethernet traffic across VMs, use distinct VLAN ID and IP info for each cluster
  • Ethernet switches (customer and Cisco) must have VLAN tag configuration done before OEDA deployment

• InfiniBand PKEY and IP details
  • Typically just use OEDA defaults
  • **Compute Cluster network** for dbnode-to-dbnode RAC traffic. Separates IB traffic by using distinct Cluster PKEY and IP subnet for each cluster.
  • **Storage network** for dbnode-to-cell or cell-to-cell traffic - same PKEY/subnet for all clusters
OEDA Configuration Tool

Installation Template

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

### Network Requirements

<table>
<thead>
<tr>
<th>Component</th>
<th>Domain</th>
<th>Network</th>
<th>Example hostname</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Database servers</strong></td>
<td><strong>dom0</strong> (one per database server)</td>
<td>Mgmt eth0</td>
<td>dm01dbadm01</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mgmt ILOM</td>
<td>dm01dbadm01-ilom</td>
</tr>
<tr>
<td></td>
<td><strong>domU</strong> (one or more per database server)</td>
<td>Mgmt eth0</td>
<td>dm01dbadm01<strong>vm01</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Client bondeth0</td>
<td>dm01client01<strong>vm01</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Client VIP</td>
<td>dm01client01<strong>vm01</strong>-vip</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Client SCAN</td>
<td>dm01<strong>vm01</strong>-scan</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Private ib</td>
<td>dm01dbadm01<strong>vm01</strong>-priv1</td>
</tr>
<tr>
<td><strong>Storage servers (same as physical)</strong></td>
<td></td>
<td>Mgmt eth0</td>
<td>dm01celadm01</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mgmt ILOM</td>
<td>dm01celadm01-ilom</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Private ib</td>
<td>dm01celadm01-priv1</td>
</tr>
<tr>
<td><strong>Switches (same as physical)</strong></td>
<td></td>
<td>Mgmt eth0</td>
<td>dm01sw-*</td>
</tr>
</tbody>
</table>
Exadata OVM Basic Maintenance

Refer to Exadata Database Maintenance Guide: Managing Oracle VM Domains on Oracle Exadata Database Machine

- Show Running Domains, Monitoring, Startup, Shutdown
- Disabling User Domain Automatic Start
- Modify Memory, CPU, local disk space in a user domain
- Remove/Create RAC VM Cluster
- Expand Oracle RAC VM cluster
- Create User Domain without Grid Infrastructure (e.g. App VM)
- Moving a User Domain to a Different Database Server
- Deleting a User Domain from an Oracle RAC VM Cluster
- Running exachk
Exadata OVM Basic Maintenance

- Backing Up and Restoring Oracle Databases on Oracle VM User Domains
- Creating Oracle VM Oracle RAC Clusters
- Creating Oracle VM without GI and Database for Apps
- Add or Drop Oracle RAC nodes in Oracle VM
- Expanding /EXAVMIMAGES on User Domains after Database Server Disk Expansion
- Implementing Tagged VLAN Interfaces
- Implementing InfiniBand Partitioning across OVM RAC Clusters on Oracle Exadata
- Backing up the Management Domain (dom0) and User Domains (domU) in an Oracle Virtual Server Deployment
- Migrating a Bare Metal Oracle RAC Cluster to an OVM RAC Cluster
OEDACLI to Perform Maintenance Operations

- OEDA Command Line Interface
  - Orchestrate Exadata life cycle management tasks
- Supported post-deployment operations with VMs – Examples:
  - Add/Remove VM cluster
  - Add/Remove node
  - Add/Remove database
  - Add/Remove database home
  - Add/Remove storage cell
  - Resize ASM disk group
  - Upgrade Clusterware
Exadata OVM 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 OVM 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

Copyright © 2019 Oracle and/or its affiliates.
Exadata OVM 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

- **Dom0**
  - Standard backup/restore practices to external location

- **DomU – Two Methods**
  - Backup within Dom0: Snapshot the VM image and backup snapshot externally
  - Backup within DomU: 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 backups/restore**
  - Use standard Exadata MAA practices with RMAN, ZDLRA, and Cloud Storage

- Refer to [Exadata Database Machine Maintenance Guide](#)
### Updating Software

<table>
<thead>
<tr>
<th>Component to update</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Storage servers</td>
<td>Same as physical - run patchmgr from any server with ssh access to all cells, or use Storage Server Cloud Scale Software Update feature (starting in 18.1).</td>
</tr>
<tr>
<td>InfiniBand switches</td>
<td>Same as physical - run patchmgr from dom0 with ssh access to all switches.</td>
</tr>
<tr>
<td>Database server – dom0</td>
<td>Run patchmgr from any server with ssh access to all dom0s. Dom0 update upgrades database server firmware. Dom0 reboot requires restart of all local domUs. DomU software <strong>not</strong> updated during dom0 update. Dom0/domU do not have to run same version, although specific update ordering may be required (see 888828.1).</td>
</tr>
<tr>
<td>Database server – domU</td>
<td>Run patchmgr from any server with ssh access to all domUs. Typically done on a per-VM cluster basis (e.g. vm01 on all nodes, then vm02, etc.), or update all VMs on a server before moving to next.</td>
</tr>
<tr>
<td>Grid Infrastructure / Database</td>
<td>Standard upgrade and patching methods apply, maintained on a per-VM cluster scope. GI/DB homes should be mounted disk images, like initial deployment. 12.2 upgrade MOS 2111010.1.</td>
</tr>
</tbody>
</table>
Health Checks and Monitoring

Exachk runs in Dom0 and DomU (cells and IB switches checks run with Dom0)
   Run in one dom0 for all dom0s, cells, switches
   Run in one domU of each VM cluster for all domUs, GI/DB of that cluster

EM Monitoring support (MOS 1967701.1)

Exawatcher runs in Dom0 and DomU

Database/GI monitoring practices still apply

Considerations
   Dom0-specific utilities (xmtop)
   Dom0 is not sized to accommodate EM or custom agents
   Oracle VM Manager not supported on Exadata
EM Support for Exadata Virtualization Provisioning

• Includes VMs/DB/GI/AS M

Create / delete RAC Cluster

Include DB/GI/ASM

Scale up / down RAC Cluster by adding or removing VMs

• Includes DB/GI/ASM

Increase Operational Efficiency by Deploying RAC Cluster Faster on Virtualized Exadata

Copyright © 2019 Oracle and/or its affiliates.
Exadata MAA/HA

- Exadata MAA failure/repair practices still applicable. Refer to [MAA Best Practices for Oracle Exadata Database Machine](#)
- OVM 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 = 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.
Integrated Cloud
Applications & Platform Services