Oracle OpenWorld 2019

SAN FRANCISCO

Safe Harbor

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, timing, and pricing of any features or functionality described for Oracle's products may change and remains at the sole discretion of Oracle Corporation.

Statements in this presentation relating to Oracle's future plans, expectations, beliefs, intentions and prospects are "forward-looking statements" and are subject to material risks and uncertainties. A detailed discussion of these factors and other risks that affect our business is contained in Oracle's Securities and Exchange Commission (SEC) filings, including our most recent reports on Form 10-K and Form 10-Q under the heading "Risk Factors." These filings are available on the SEC's website or on Oracle's website at http://www.oracle.com/investor. All information in this presentation is current as of September 2019 and Oracle undertakes no duty to update any statement in light of new information or future events.

ORACLE

Exadata Maximum Availability Architecture
Best Practices and Recommendations

Michael Nowak

MAA Solutions Architect
Oracle Product Development

Christopher Guillaume

Senior Database Administrator
CME Group

Our Goals

- To briefly review *Exadata MAA*
- To understand *MAA characteristics of three new Exadata X8M components*
- To take home and consider *three new Exadata MAA best practices*
- To see *Exadata MAA in action at the Chicago Mercantile Exchange*

Exadata MAA

A brief review

Impact of Database Downtime

$350K

Average cost of downtime per hour

$10M

Average cost of unplanned data center outage or disaster

87 hours

Average amount of downtime per year

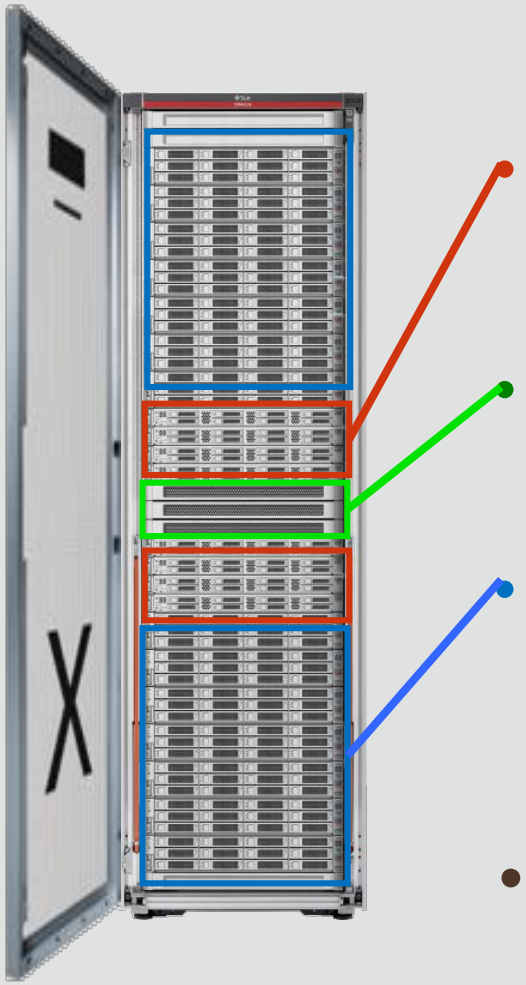
91%

Percentage of companies have experienced an unplanned data center outage in the last 24 months

Source: Gartner, Data Center Knowledge, IT Process Institute, Forrester Research

Oracle Maximum Availability Architecture(MAA) Solution Options

Exadata X8M (changes from X8 in red)



- Scale-Out 2 or 8 Socket Database Servers
  - Latest 24 core Intel Cascade Lake
- 100Gb RDMA over Converged Ethernet (RoCE) Internal Fabric
- Scale-Out Intelligent 2-Socket Storage Servers
  - 1.5 TB Persistent Memory per storage server
  - Three tiers of storage: PMEM, NVMe, HDD
- Enhanced consolidation using Linux KVM

MAA Characteristics of Three New Exadata X8M Components

KVM, Persistent Memory (PMEM), and RDMA Network Fabric

KVM
MAA Characteristics

- Full set of Exadata KVM best practices will be available here: https://www.oracle.com/database/technologies/high-availability/exadata-maa-best-practices.html
- Some MAA notables:
  - The number of guests supported on KVM is 12 (8 on Xen)
  - Prior generations of Exadata can be connected via Data Guard or Golden Gate
  - Standard backup procedures apply
    - The KVM host can optionally shapshot VM disk images and store them externally
  - Update core Exadata infrastructure with patchmgr
  - Update Grid Infrastructure and Database ORACLE_HOMEs with oedacli
  - Perform lifecycle operations with vm_maker and oedacli

“

This demo will show the online addition of CPU capacity to an Exadata X8M with oedacli, stabilizing a CPU bound workload.”

—

Exadata MAA Team

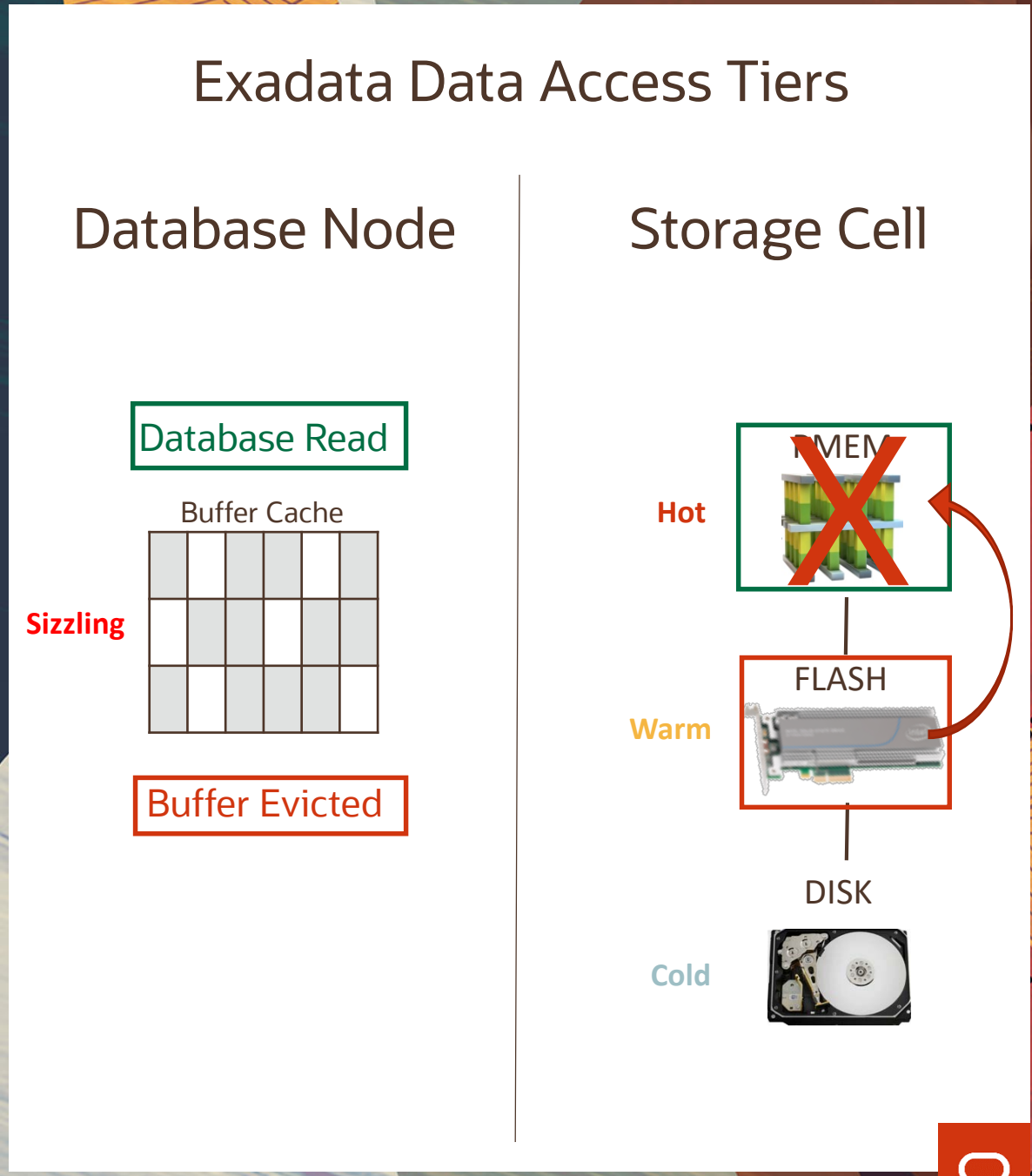
PMEM
MAA Characteristics

- Not drawn to scale ☺
- Primary copy of data placed in PMEM cache on a read miss
- Secondary copy of data placed in flash cache on buffer eviction

If a pmem fails in Writethrough mode, no redundancy restoration is required

If a pmem fails in Writeback mode, a resilver operation is run to restore redundancy

Low latency flash reads will repopulate super low latency pmem



“

This demo will show a Persistent Memory (PMEM) replacement on an Exadata X8M cell with no application service level impact.”

Exadata MAA Team

RDMA Network Fabric
MAA Characteristics

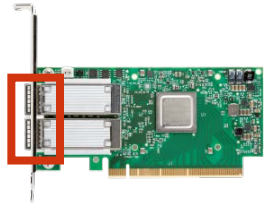
2 Active-Active ports in every RDMA Network Fabric Adapter

2 RDMA Network Fabric Switches in every Exadata single rack

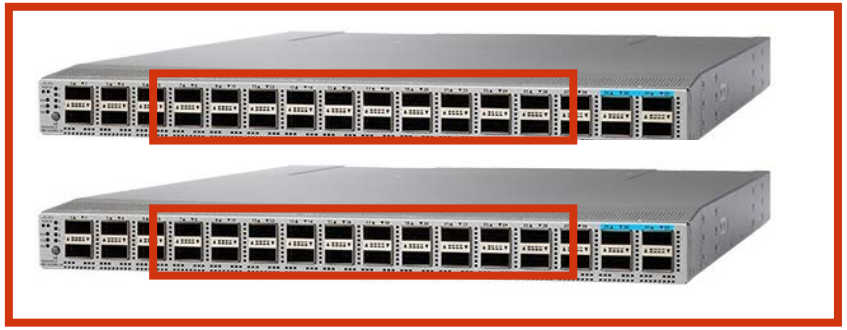
22 Ports per switch used for internal cluster network, cabled ensuring no single point of failure exists

RDMA over Converged Ethernet (RoCE)

RDMA Network Fabric Adapter



RDMA Network Fabric Switch

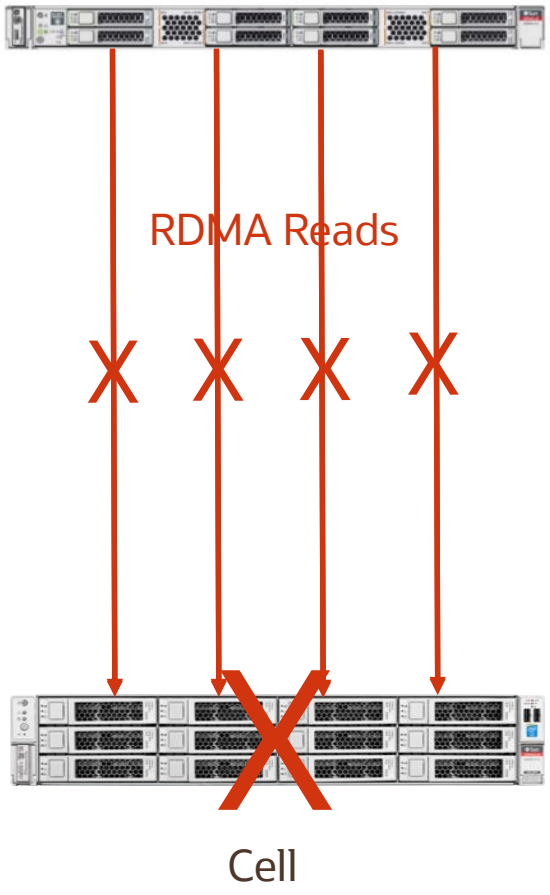


Wait, in the past you have told me about how Exadata Fast Node Death Detection (FNDD) uses the InfiniBand Subnet Manager, but Exadata X8M does not have InfiniBand switches. How does FNDD work?

4 RDMA paths exist between database nodes and cells to monitor cell liveliness.

If all four are unavailable after a short timeout expires, the cell is evicted

<1 Second to complete cell eviction, maintaining SLA



“

This demo will show Fast Node Death Detection (FNDD) on an Exadata X8M with minimal application service level brownout.”

—

Exadata MAA Team

Copyright © 2019 Oracle and/or its affiliates.

RDMA Network Fabric
MAA Characteristics

Network Fabric Switch Software Updates

- Same tool, patchmgr
- Separate software update package
- Optimized, built-in handling of port down/up events
- -verify-config and –roceswitch-precheck options available to check state ahead of time

```
$./patchmgr --roceswitches /tmp/switch_group --upgrade -log_dir /tmp/roce_switches

 2019-09-05 09:46:19 -0700        :Working  Initiate upgrade of 2 RoCE switches to 7.0(3)I7(6) Expect up to 15 minutes for each switch

Thu Sep  5 09:46:19 PDT 2019 1 of 2 :Updating switch switch1

Thu Sep  5 09:46:21 PDT 2019:      [INFO     ] Switch switch1 will be upgraded from nxos.7.0.3.I7.5.bin to nxos.7.0.3.I7.6.bin
Thu Sep  5 09:46:21 PDT 2019:      [INFO     ] Checking for free disk space on switch
Thu Sep  5 09:46:22 PDT 2019:      [INFO     ] disk is 97.00% free,  available: 236578451456 bytes
Thu Sep  5 09:46:22 PDT 2019:      [SUCCESS  ] There is enough disk space to proceed
Thu Sep  5 09:46:22 PDT 2019:      [INFO     ] Found  nxos.7.0.3.I7.6.bin on switch, skipping download
Thu Sep  5 09:46:22 PDT 2019:      [INFO     ] Verifying sha256sum of bin file on switch
Thu Sep  5 09:46:38 PDT 2019:      [SUCCESS  ] sha256sum matches: 70003934d6669da969bad0617aa3657acb5f6e5418bfcddaa987ff577293f531
Thu Sep  5 09:46:38 PDT 2019:      [INFO     ] Performing FW install pre-check of nxos.7.0.3.I7.6.bin (eta: 2-3 minutes)
Thu Sep  5 09:48:38 PDT 2019:      [SUCCESS  ] FW install pre-check completed successfully
Thu Sep  5 09:48:38 PDT 2019:      [INFO     ] Performing FW install of nxos.7.0.3.I7.6.bin on switch1 (eta: 3-7 minutes)
Thu Sep  5 09:51:42 PDT 2019:      [SUCCESS  ] FW install completed
Thu Sep  5 09:51:42 PDT 2019:      [INFO     ] Waiting for switch to come back online (eta: 6-8 minutes)
Thu Sep  5 09:59:42 PDT 2019:      [INFO     ] Verifying if FW install is successful
Thu Sep  5 09:59:46 PDT 2019:      [SUCCESS  ] switch1 has been successfully  upgraded to nxos.7.0.3.I7.6.bin!

Thu Sep  5 09:59:46 PDT 2019 2 of 2 :Updating switch switch2

Thu Sep  5 09:59:48 PDT 2019:      [INFO     ] Switch switch2 will be upgraded from nxos.7.0.3.I7.5.bin to nxos.7.0.3.I7.6.bin
Thu Sep  5 09:59:48 PDT 2019:      [INFO     ] Checking for free disk space on switch
Thu Sep  5 09:59:48 PDT 2019:      [INFO     ] disk is 97.00% free,  available: 237016682496 bytes
Thu Sep  5 09:59:48 PDT 2019:      [SUCCESS  ] There is enough disk space to proceed
Thu Sep  5 09:59:49 PDT 2019:      [INFO     ] Found  nxos.7.0.3.I7.6.bin on switch, skipping download
Thu Sep  5 09:59:49 PDT 2019:      [INFO     ] Verifying sha256sum of bin file on switch
Thu Sep  5 10:00:05 PDT 2019:      [SUCCESS  ] sha256sum matches: 70003934d6669da969bad0617aa3657acb5f6e5418bfcddaa987ff577293f531
Thu Sep  5 10:00:05 PDT 2019:      [INFO     ] Performing FW install pre-check of nxos.7.0.3.I7.6.bin (eta: 2-3 minutes)
Thu Sep  5 10:02:06 PDT 2019:      [SUCCESS  ] FW install pre-check completed successfully
Thu Sep  5 10:02:06 PDT 2019:      [INFO     ] Checking if previous switch switch1 is fully up before proceeding (attempt 1 of 3)
Thu Sep  5 10:02:07 PDT 2019:      [SUCCESS  ] switch1 switch is fully up and running
Thu Sep  5 10:02:07 PDT 2019:      [INFO     ] Performing FW install of nxos.7.0.3.I7.6.bin on switch2 (eta: 3-7 minutes)
Thu Sep  5 10:05:11 PDT 2019:      [SUCCESS  ] FW install completed
Thu Sep  5 10:05:11 PDT 2019:      [INFO     ] Waiting for switch to come back online (eta: 6-8 minutes)
Thu Sep  5 10:13:11 PDT 2019:      [INFO     ] Verifying if FW install is successful
Thu Sep  5 10:13:15 PDT 2019:      [SUCCESS  ] switch2 has been successfully  upgraded to nxos.7.0.3.I7.6.bin!

Thu Sep  5 10:13:16 PDT 2019 1 of 2 :Verifying config on switch switch1

Thu Sep  5 10:13:16 PDT 2019:      [INFO     ] Dumping current running config locally as file: /tmp/roce_switches/run.switch1.cfg
Thu Sep  5 10:13:18 PDT 2019:      [SUCCESS  ] Backed up switch config successfully
Thu Sep  5 10:13:18 PDT 2019:      [INFO     ] Validating running config against template [1/3]: /u01/software/PATCH_ROCE_SWITCHES/patch_switch_19.3.0.0.0.190903/roce_switch_templates/roce_leaf_switch.cfg
Thu Sep  5 10:13:18 PDT 2019:      [INFO     ] Config matches template: /u01/software/PATCH_ROCE_SWITCHES/patch_switch_19.3.0.0.0.190903/roce_switch_templates/roce_leaf_switch.cfg
Thu Sep  5 10:13:18 PDT 2019:      [SUCCESS  ] Config validation successful!

Thu Sep  5 10:13:18 PDT 2019 2 of 2 :Verifying config on switch switch2

Thu Sep  5 10:13:18 PDT 2019:      [INFO     ] Dumping current running config locally as file: /tmp/roce_switches/run.switch2.cfg
Thu Sep  5 10:13:19 PDT 2019:      [SUCCESS  ] Backed up switch config successfully
Thu Sep  5 10:13:19 PDT 2019:      [INFO     ] Validating running config against template [1/3]: /u01/software/PATCH_ROCE_SWITCHES/patch_switch_19.3.0.0.0.190903/roce_switch_templates/roce_leaf_switch.cfg
Thu Sep  5 10:13:19 PDT 2019:      [INFO     ] Config matches template: /u01/software/PATCH_ROCE_SWITCHES/patch_switch_19.3.0.0.0.190903/roce_switch_templates/roce_leaf_switch.cfg
Thu Sep  5 10:13:19 PDT 2019:      [SUCCESS  ] Config validation successful!
2019-09-05 10:13:19 -0700 ++++++++++++++++++ Logs so far end ++++++++++
2019-09-05 10:13:19 -0700        :SUCCESS: Config check on RoCE switch(es)
2019-09-05 10:13:19 -0700 ++++++++++++++++++ Logs so far begin ++++++++++
2019-09-05 10:13:19 -0700 ++++++++++++++++++ Logs so far end ++++++++++
2019-09-05 10:13:19 -0700        :SUCCESS: upgrade 2 RoCE switch(es) to 7.0(3)I7(6)
2019-09-05 10:13:19 -0700        :SUCCESS: Completed run of command: ./patchmgr --roceswitches /tmp/switch_group --upgrade -log_dir /tmp/roce_switches
2019-09-05 10:13:19 -0700        :INFO   : upgrade attempted on nodes in file /tmp/switch_group: [switch1 switch2]
2019-09-05 10:13:19 -0700        :INFO   : For details, check the following files in /tmp/roce_switches:
2019-09-05 10:13:19 -0700        :INFO   :  - updateRoceSwitch.log
2019-09-05 10:13:19 -0700        :INFO   :  - updateRoceSwitch.trc
2019-09-05 10:13:19 -0700        :INFO   :  - patchmgr.stdout
2019-09-05 10:13:19 -0700        :INFO   :  - patchmgr.stderr
2019-09-05 10:13:19 -0700        :INFO   :  - patchmgr.log
2019-09-05 10:13:19 -0700        :INFO   :  - patchmgr.trc
2019-09-05 10:13:19 -0700        :INFO   : Exit status:0
2019-09-05 10:13:19 -0700        :INFO   : Exiting.
================ PatchMgr run ended 2019-09-05 10:13:19 -0700 ================
```

Three New Exadata MAA Best Practices

Exachk critical issue repair, Cell drop/failure, XT storage cells

Repair of Exadata Critical Issues

```
# ./exachk -repaircheck -h

      -repaircheck <file|checkids|all>
            Repair check(s).
            Options:
               file       : file containing check ids which need to be repaired
               checkids   : comma separated check ids which need to be repaired
               all        : repair all checks(for which command to repair is available)
            Example:
               ./exachk -repaircheck <check_id>,[<check_id>,<check_id>..]
               ./exachk -repaircheck <file>
               ./exachk -repaircheck all

# ./exachk -showrepair -h

      -showrepair <checkid>
            Display check repair command
            Options:
               checkid    : Show repair command for given check id.
            Example:
               ./exachk -showrepair <check_id>.
```

Database Server

Status	Type	Message	Status On	Details
REPAIR-PASS	OS Check	System is not exposed to Exadata critical issue EX50	All Database Servers	Hide

Exadata Critical Issue EX50

Recommendation	Benefit / Impact: A system exposed to a critical issue may experience system-wide impact to performance or availability Risk: Kernel service systemd-tmpfiles-clean.service may remove required socket files in /var/tmp/.oracle, which may cause database startup or connections to fail, or clusterware connection to fail on Exadata database servers running Oracle Linux 7 (i.e. Exadata 19.1). Action / Repair: See EX50 in below document 1270094.1 for additional details
Links	1. Note: 1270094.1 - Exadata Critical Issues

Cell Drop/Failure Best Practices

- Quick review on disk failure coverage

Grid Infrastructure Version	Number of Failgroups	Required % Free of Diskgroup Capacity
12.1.0	Any	15
12.2, 18.x	less than 5	15
12.2, 18.x	5 or more	9
19.x with high redundancy diskgroups (smart rebalance)	Any	0

- Cell failure coverage:
  - Cell failures are extremely rare but some conservative customers like to be prepared for them
  - The failgroup_repair_time diskgroup attribute defines the amount of time disks are left offline before dropped, and defaults to 24 hours. Reducing the time to repair/replace a failed cell is preferable to increasing the failgroup_repair_time diskgroup attribute.
  - If a cell must be dropped, each diskgroup should have FREE_MB greater than a cell's worth of the total diskgroup space plus an additional 5% of that space. In more technical ASM terms, this means FREE_MB > DG TOTAL_MB/num_of_cells * 1.05
- Refer to MOS note 1551288.1 for more details.

Extended (XT) storage cells

Best Practices

- Do use XT storage cells for their intended purposes – ex: historical infrequently accessed data, development databases, local backups
- Do understand the OEDA default name for diskgroups on XT storage is XTND
- Do use standard database node quorum devices when implementing a normal redundancy diskgroup with two XT cells. You will have four devices available for ASM metadata and that is OK ☺

Worst Practices

- Do not mix Exadata storage cell types in the same diskgroup
- Do not create a normal redundancy diskgroup using 2 HC cells or 2 EF cells

Oeda / oedacli implements all best practices automatically (and prevents worst practices)

Sneak peek into Exadata MAA future

CME Group Overview

CME Group is the world's leading and most diverse derivatives marketplace bringing together those who need to manage risk or those that want to profit by accepting it.

- Operating Multiple Exchanges – CME, CBOT, Nymex and COMEX
- Trade hundreds of products across the globe on a single platform
- Average daily volume of 15.6 million contracts

- CME Clearing – matches and settles all trades and guarantees the creditworthiness of every transaction
- Cleared more than 4.9 billion contracts with a value exceeding $1 quadrillion
- Highest Volume Day – 51.9 million contracts

Who am I

- Christopher Guillaume, Senior Oracle DBA @ CME Group
- Oracle DBA since over 20 years
- Former Certified Oracle Instructor
- Working in Financial Industry since 2004, at CME Group since 2012
- Lead Exadata Response Team

Exadata

What we currently have:

- Originally started on X2
- Currently on X6 and an X7
- Datawarehouse utilizes a 1 ½ rack
- OLTP utilizes 3 separate ½ racks
- DR Matches Prod (minus Local Standby)

CME HA/DR Requirements :

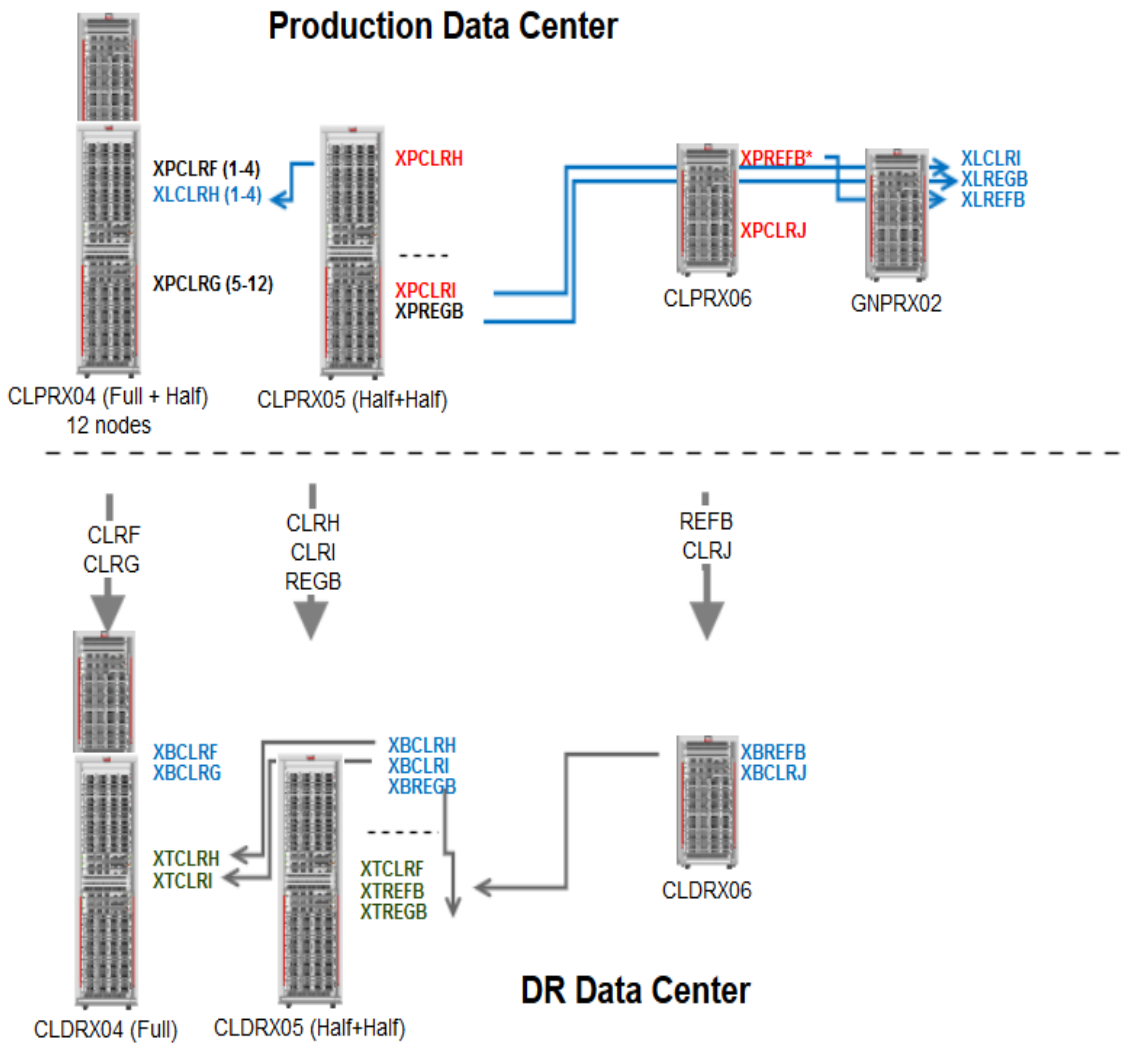
- An Exadata Failure Cannot Cause a DR Event
- Allow Mid-Week End-to-End testing in DR
- Provide safe means to test applications in Production (Saturday Test)
- Continuous DB availability through planned maintenance for critical applications
- Allow for customers to retrieve critical data while Recovery is happening.

Disaster Recovery:

- **Critical DB's – 10 second outage**
- Recovery Point Objective (RPO) – 30 seconds
- Recovery Time Objective (RTO) – 2 hours for system
- RTO – 5 minutes for Databases

CME HA Architecture

- Multiple Databases running on
  - Single Servers
  - RAC
  - Exadata (Shown)
- Each Prod Database is replicated locally and remotely
  - BLUE Local (Fast Sync)
  - Gray – Async
- Dedicated Local DG Recipient
- Active DG in DR
- Multiple Complete Exadata Failures need to occur in order for DR event to happen
- Running over 100 apps and more than 200 services



Session Survey

Help us make the content even better. Please complete the session survey in the Mobile App.