
ORACLE®

Oracle Real Application Clusters (RAC) 12c Rel. 2 – Operational Best Practices

Markus Michalewicz
Senior Director of
Product Management,
Oracle RAC Development

Markus.Michalewicz@oracle.com
@OracleRACpm
http://www.linkedin.com/in/markusmichalewicz
http://www.slideshare.net/MarkusMichalewicz

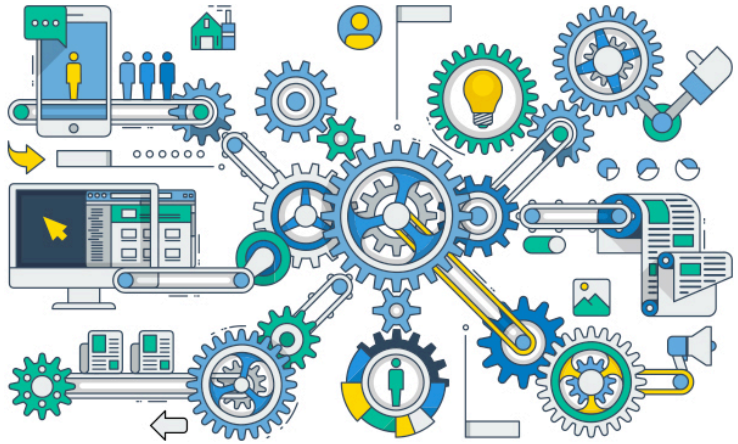
December 6, 2017

Safe Harbor Statement

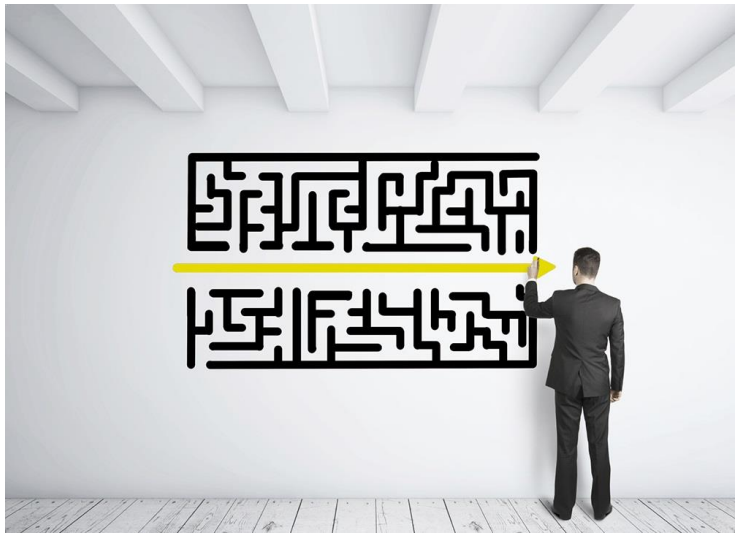
The following is intended to outline our general product direction. It is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. The development, release, and timing of any features or functionality described for Oracle's products remains at the sole discretion of Oracle.

Best Practice Recommendations Go with the Flow

Fact



Way:
Automation



Goal:
Simplification

RAC is Now Part of The World's First Autonomous Database

Self-Driving - User defines service levels, database makes them happen

Self-Tuning - Continuous adaptive performance tuning

Self-Scaling - Instantly resize compute and storage without downtime

Self-Securing - Protection from both external attacks and internal users

Self-Repairing - Automated protection from all downtime

Less Labor, Lower Cost, Fewer Errors, More Secure, More Reliable

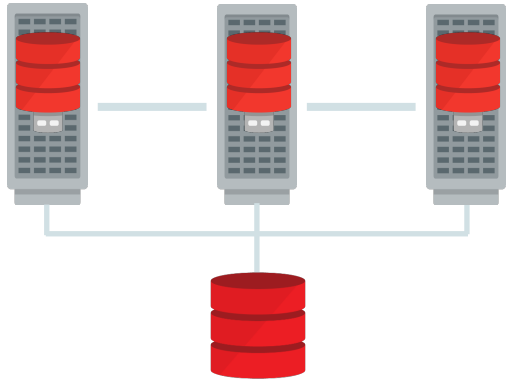
Program Agenda

1. Fundamentals To Be Considered First
2. Architecture Choices
3. Applying Operational Best Practices
4. Using Smart Features Automatically
5. Summary

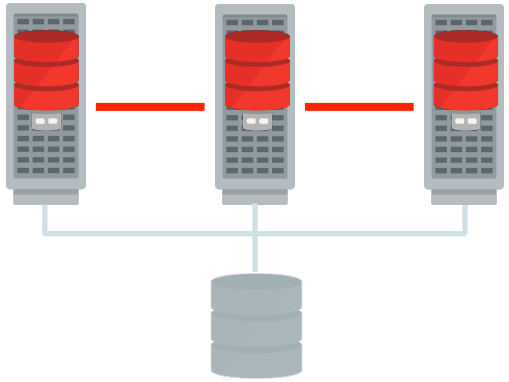
Program Agenda

1. Fundamentals To Be Considered First
2. Architecture Choices
3. Applying Operational Best Practices
4. Using Smart Features Automatically
5. Summary

Fundamentals To Be Considered First



Shared Storage

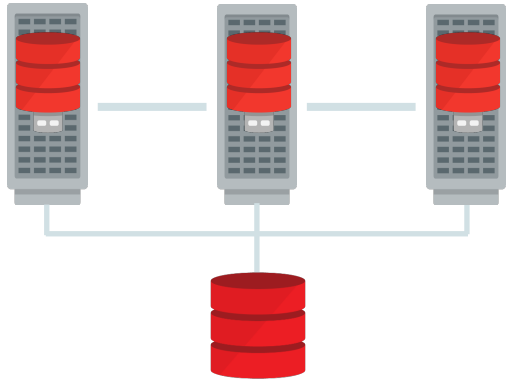


Network / Interconnect

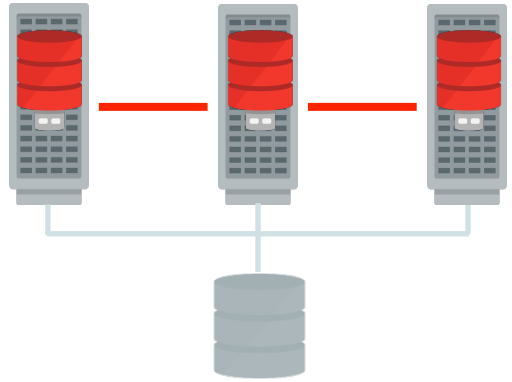
Virtualization / Container

Cloud

Fundamentals To Be Considered First



Shared Storage



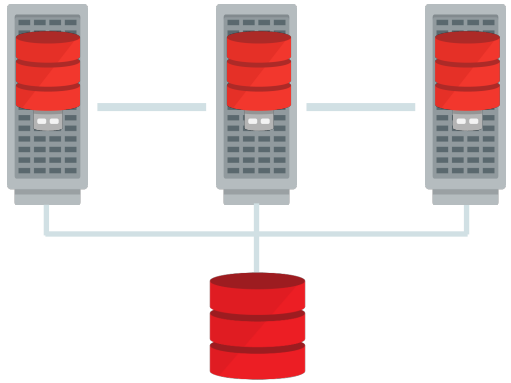
Network / Interconnect

Virtualization / Container

Cloud

More information on those topics:
https://www.slideshare.net/MarkusMichalewicz/oracle-real-application-clusters-rac-12c-rel-2-whats-next

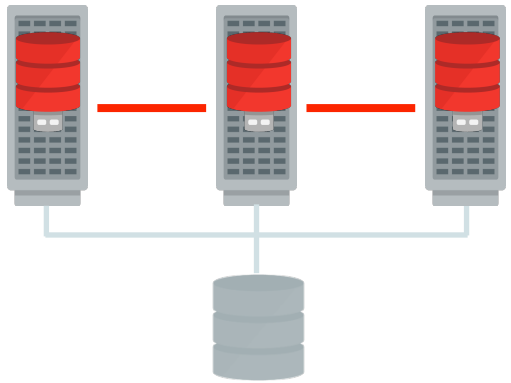
Shared Storage Fundamentals



- Shared storage is the top requirement for Oracle RAC.
- Block storage that is supported as shared storage for RAC by the storage vendor *and* the server vendor is supported for RAC.

  https://www.slideshare.net/MarkusMichalewicz/how-to-use-oracle-rac-in-a-cloud-a-support-question
- Exceptions prove the rule.
  - Shared File System require certification.
  - Storage solutions may be subject to “special ruling”.
- Currently, No 3rd Party Public Cloud supports shared storage for RAC.
- **Oracle ASM and ACFS are by far the preferred storage management solutions for Oracle RAC.**
- Caution: no Oracle provided tool can fully reliably check the appropriate configuration of shared storage beforehand.

Network / Interconnect Fundamentals



- The interconnect is crucial to the operation of Oracle RAC.
  - Think of it as a RDMA access path rather than a network.
  - 1 GigE and redundancy should be the bare minimum.
  - 10 GigE and redundancy are strongly recommended.
- Converged networks and switches are supported. Routing is not.
  - I.E. combining public / private network communication.
    - In a future release, communication on interconnect is planned to be secured by default.
  - VLANs are a logical separation, not a physical one.
    - E.g. they do not protect from interference on the same physical network.
  - When network and storage communication are combined, storage communication should get prioritized.
- Caution: no Oracle provided tool can fully reliably check the appropriate configuration of the interconnect beforehand.

Program Agenda

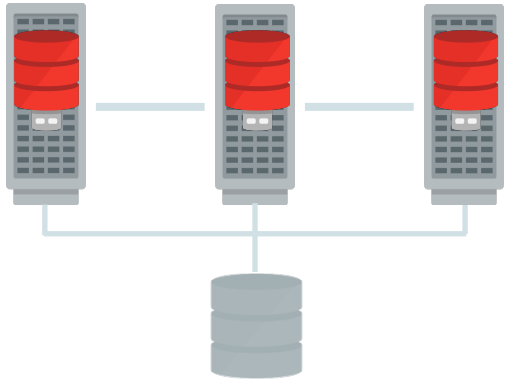
1. Fundamentals To Be Considered First
2. Architecture Choices
3. Applying Operational Best Practices
4. Using Smart Features Automatically
5. Summary

Architecture Choices

Oracle Restart

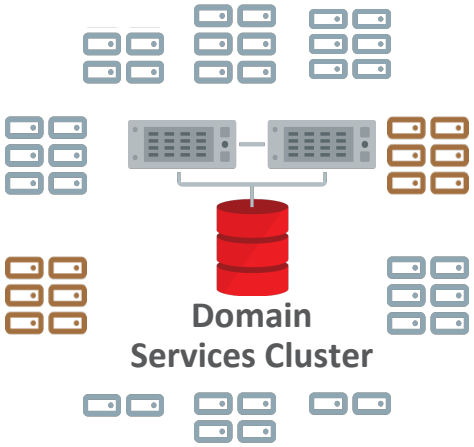
Not a cluster,
but back for good!

See updates in
MOS Note 1584742.1



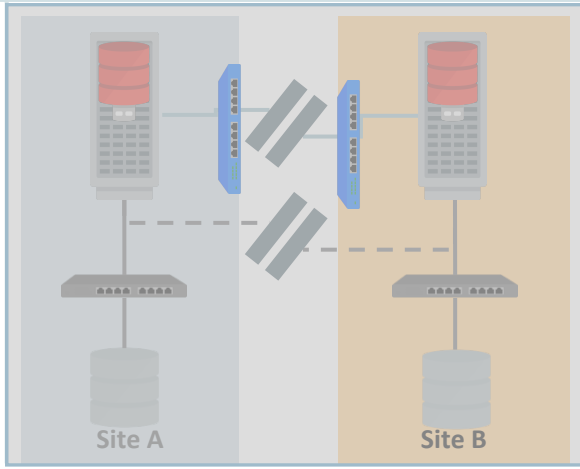
Standalone Cluster

Dedicated cluster for OLTP,
DWH or mixed-WL systems.



Cluster Domain

Domain Services Cluster (DSC) &
Member Clusters for large cluster
estates and storage consolidation.

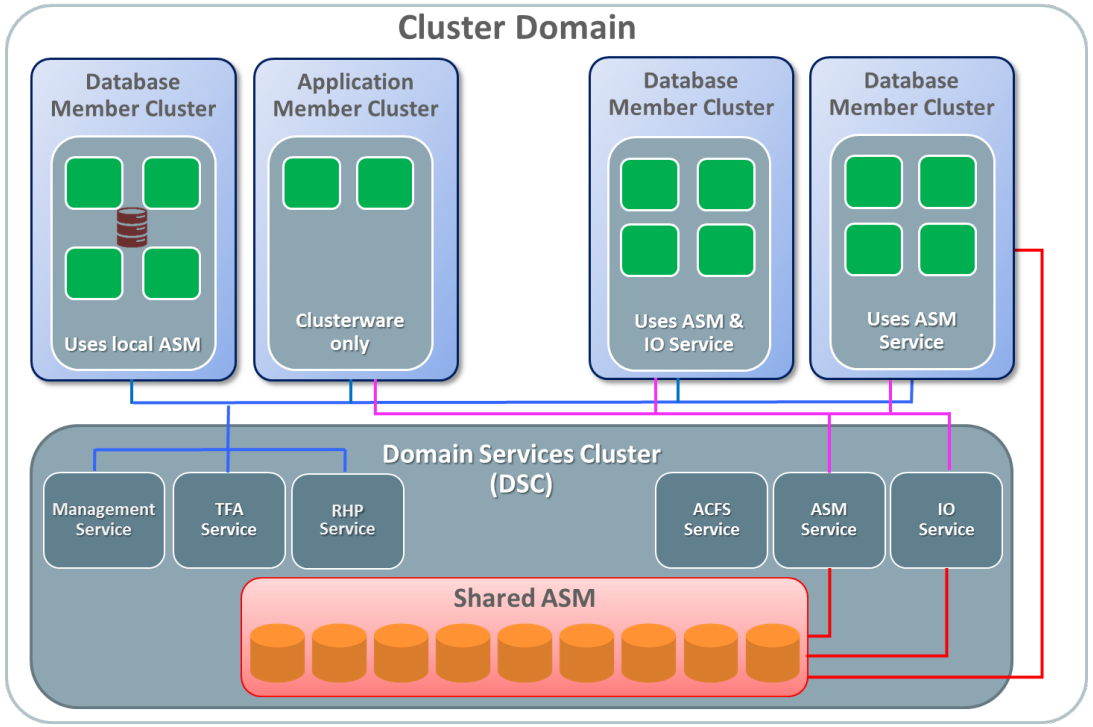


Extended Cluster

Standalone Cluster or DSC
maintaining two independent
storage arrays (over distance).

Oracle RAC 12*c* Rel. 2 Cluster Domain

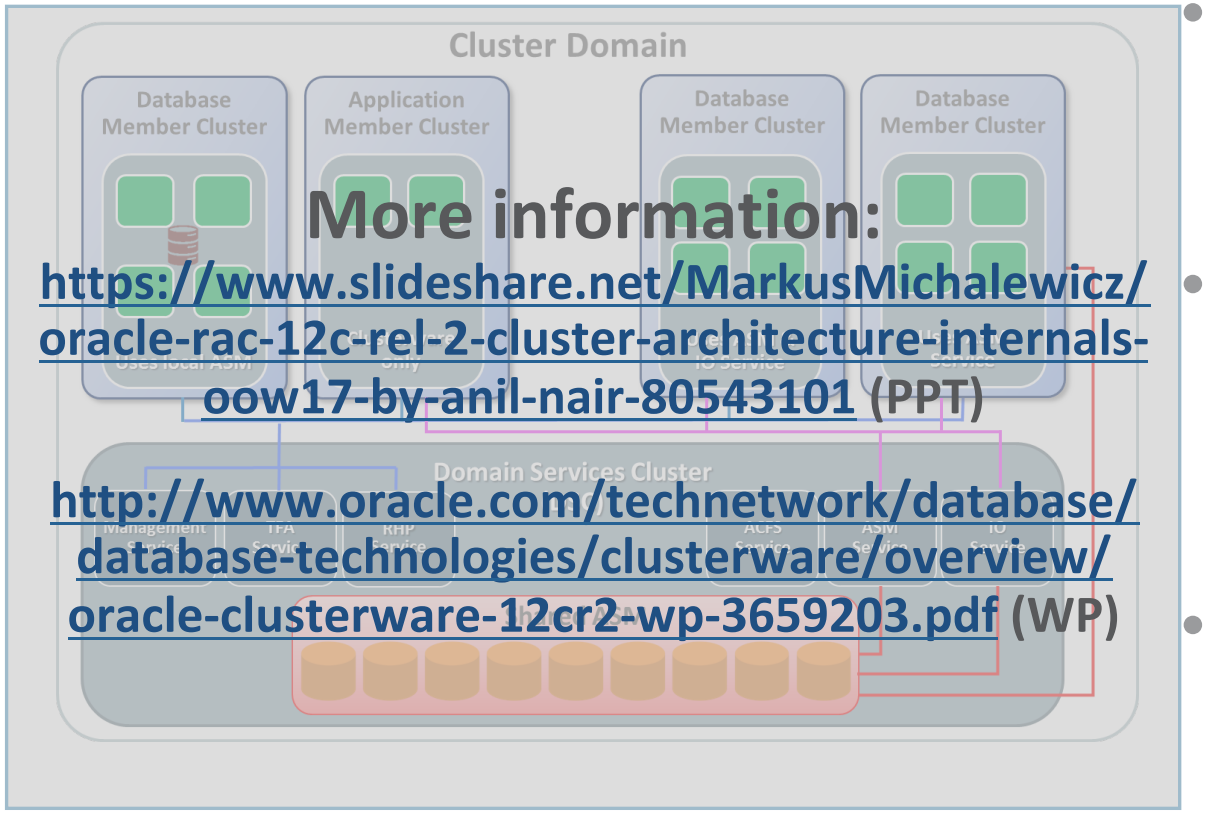
Centralized Management for Cluster Estates "too big to manage" otherwise



- Simplified Management
  - Fleet Management for installation, update, patching and maintenance
- Reduced Local Overhead
  - Member Clusters benefit from the consolidation of common services on the Domain Services Cluster
- Improved IO Performance
  - Utilizing consolidated shared storage

Oracle RAC 12*c* Rel. 2 Cluster Domain

Centralized Management for Cluster Estates “too big to manage” otherwise



More information:

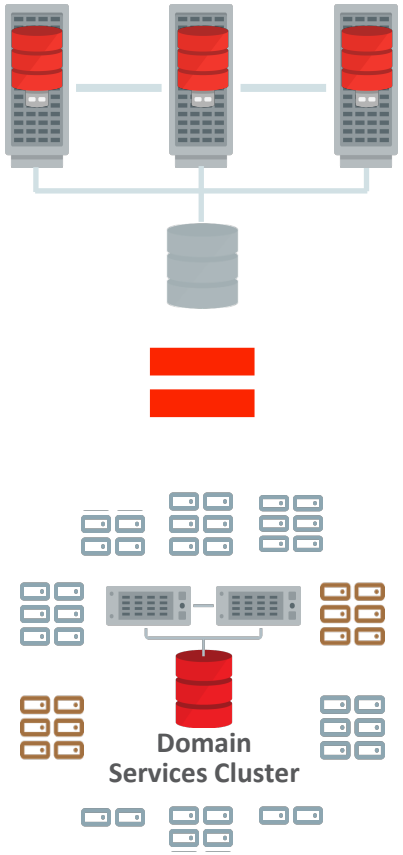
https://www.slideshare.net/MarkusMichalewicz/oracle-rac-12c-rel-2-cluster-architecture-internals-oow17-by-anil-nair-80543101 (PPT)

http://www.oracle.com/technetwork/database/database-technologies/clusterware/overview/oracle-clusterware-12cr2-wp-3659203.pdf (WP)

- Simplified Management
  - Fleet Management for installation, update, patching and maintenance
- Reduced Local Overhead
  - Member Clusters benefit from the consolidation of common services on the Domain Services Cluster
- Improved IO Performance
  - Utilizing consolidated shared storage

Best Practices Apply Across Architectures

The simplified Best Practices (BP) equation

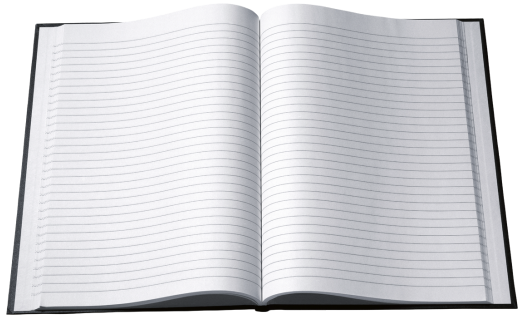


- **BP are designed for Standalone Cluster deployments**
- A DSC is a Standalone Cluster
  - that offers additional services to Member Clusters.
  - **BP for Standalone Clusters apply to a DSC**
- Member Clusters come with a simplified deployment
  - **Basic BP for Standalone Clusters apply to Member Clusters**
- Extended Clusters for RAC come with additional BP:
  - https://www.slideshare.net/MarkusMichalewicz/oracle-extended-clusters-for-oracle-rac

Program Agenda

1. Fundamentals To Be Considered First
2. Architecture Choices
3. Applying Operational Best Practices
4. Using Smart Features Automatically
5. Summary

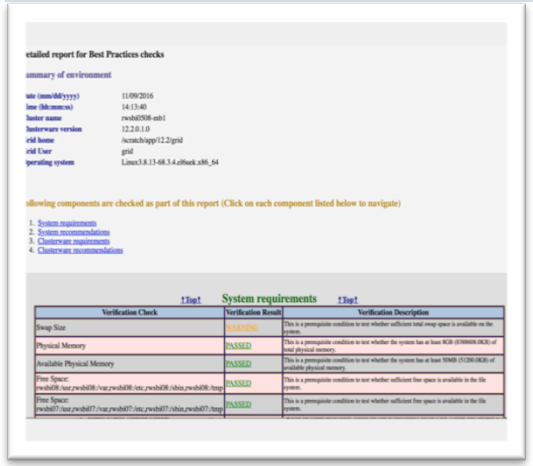
How To Obtain and Apply Best Practices for Oracle RAC?



Read & implement Oracle Documentation and My Oracle Support (MOS) Notes

Read & implement Blogs & Forum Entries

Not always vetted by Oracle.



etailed report for Best Practices checks

ummary of environment

ate (mm/dd/yyyy)	11/09/2016
ime (hh:mm:ss)	14:13:40
luster name	rwsbl0508-nb1
lusterware version	12.2.0.1.0
rid home	/scratch/app/12.2/grid
rid User	grid
perating system	Linux3.8.13-68.3.4.el6uek.x86_64

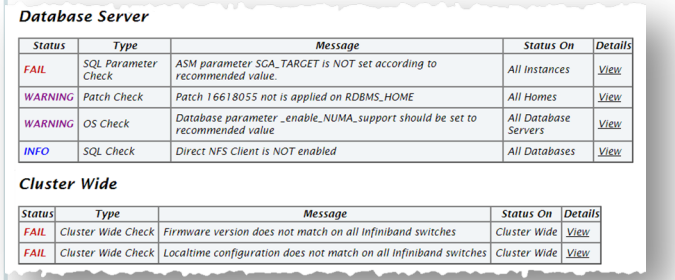
llowing components are checked as part of this report (Click on each component listed below to navigate)

1. System requirements
2. System recommendations
3. Clusterware requirements
4. Clusterware recommendations

↑Top↑ System requirements ↑Top↑

Verification Check	Verification Result	Verification Description
Swap Size	WARNING	This is a prerequisite condition to test whether sufficient total swap space is available on the system.
Physical Memory	PASSED	This is a prerequisite condition to test whether the system has at least 8GB (8388608.0KB) of total physical memory.
Available Physical Memory	PASSED	This is a prerequisite condition to test whether the system has at least 50MB (51200.0KB) of available physical memory.
Free Space: rwsbl0508:/usr,rwsbl0508:/var,rwsbl0508:/etc,rwsbl0508:/sbin,rwsbl0508:/tmp	PASSED	This is a prerequisite condition to test whether sufficient free space is available in the file system.
Free Space: rwsbl0507:/usr,rwsbl0507:/var,rwsbl0507:/etc,rwsbl0507:/sbin,rwsbl0507:/tmp	PASSED	This is a prerequisite condition to test whether sufficient free space is available in the file system.

Run & apply Cluster Verification Utility (CVU) & Oracle Universal Installer (OUI)



Database Server

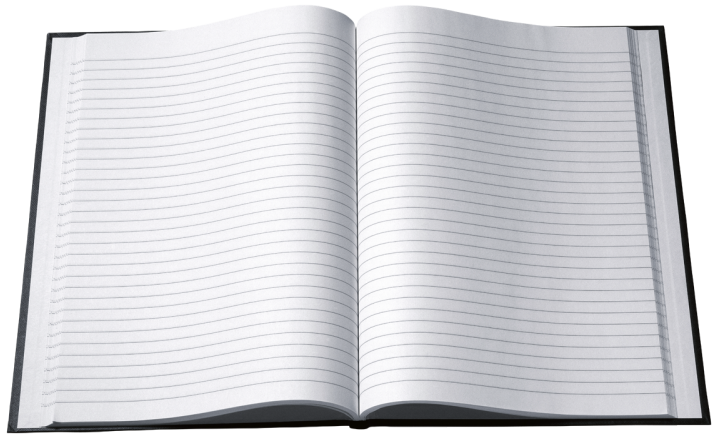
Status	Type	Message	Status On	Details
FAIL	SQL Parameter Check	ASM parameter SGA_TARGET is NOT set according to recommended value.	All Instances	View
WARNING	Patch Check	Patch 16618055 not is applied on RDBMS_HOME	All Homes	View
WARNING	OS Check	Database parameter _enable_NUMA_support should be set to recommended value	All Database Servers	View
INFO	SQL Check	Direct NFS Client is NOT enabled	All Databases	View

Cluster Wide

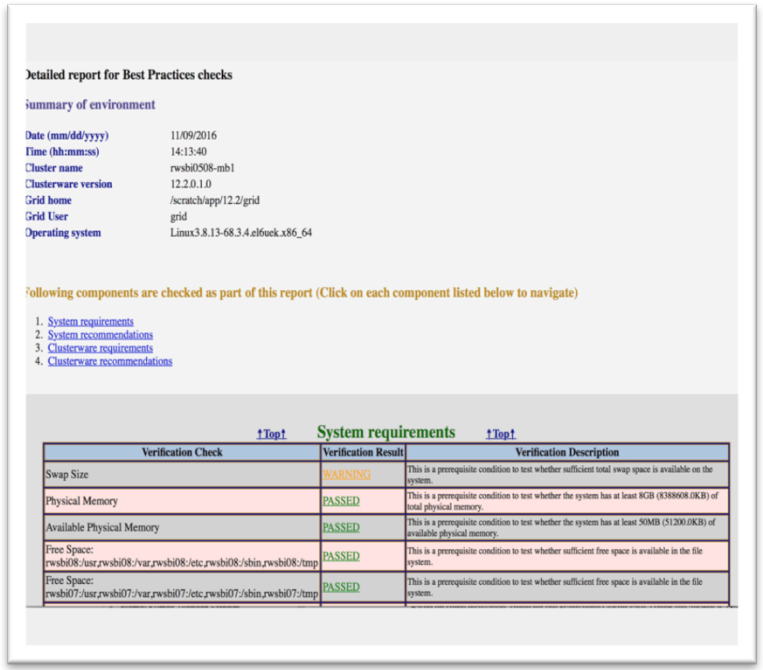
Status	Type	Message	Status On	Details
FAIL	Cluster Wide Check	Firmware version does not match on all Infiniband switches	Cluster Wide	View
FAIL	Cluster Wide Check	Localtime configuration does not match on all Infiniband switches	Cluster Wide	View

Run & follow ORAchk

Each “Tool” has a Different Purpose



Establish the base system;
e.g.: hardware and OS
minimum requirements
(fundamentals)



Detailed report for Best Practices checks

Summary of environment

Date (mm/dd/yyyy)	11/09/2016
Time (hh:mm:ss)	14:13:40
Cluster name	rwsbi0508-mb1
Clusterware version	12.2.0.1.0
Grid home	/scratch/app/12.2/grid
Grid User	grid
Operating system	Linux3.8.13-68.3.4.el6uek.x86_64

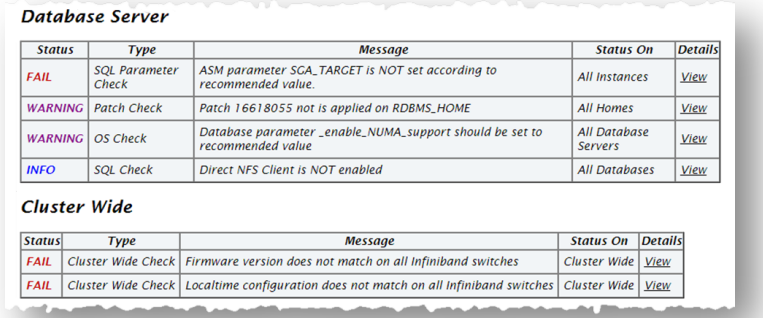
Following components are checked as part of this report (Click on each component listed below to navigate)

1. System requirements
2. System recommendations
3. Clusterware requirements
4. Clusterware recommendations

↑Top↑ System requirements ↑Top↑

Verification Check	Verification Result	Verification Description
Swap Size	WARNING	This is a prerequisite condition to test whether sufficient total swap space is available on the system.
Physical Memory	PASSED	This is a prerequisite condition to test whether the system has at least 8GB (8388608.0KB) of total physical memory.
Available Physical Memory	PASSED	This is a prerequisite condition to test whether the system has at least 50MB (51200.0KB) of available physical memory.
Free Space: rwsbi08:/usr,rwsbi08:/var,rwsbi08:/etc,rwsbi08:/sbin,rwsbi08:/tmp	PASSED	This is a prerequisite condition to test whether sufficient free space is available in the file system.
Free Space: rwsbi07:/usr,rwsbi07:/var,rwsbi07:/etc,rwsbi07:/sbin,rwsbi07:/tmp	PASSED	This is a prerequisite condition to test whether sufficient free space is available in the file system.

Check and confirm baseline;
e.g. check for and apply OS
minimum parameter requirements;
supports diff comparison



Database Server

Status	Type	Message	Status On	Details
FAIL	SQL Parameter Check	ASM parameter SGA_TARGET is NOT set according to recommended value.	All Instances	View
WARNING	Patch Check	Patch 16618055 not is applied on RDBMS_HOME	All Homes	View
WARNING	OS Check	Database parameter _enable_NUMA_support should be set to recommended value	All Database Servers	View
INFO	SQL Check	Direct NFS Client is NOT enabled	All Databases	View

Cluster Wide

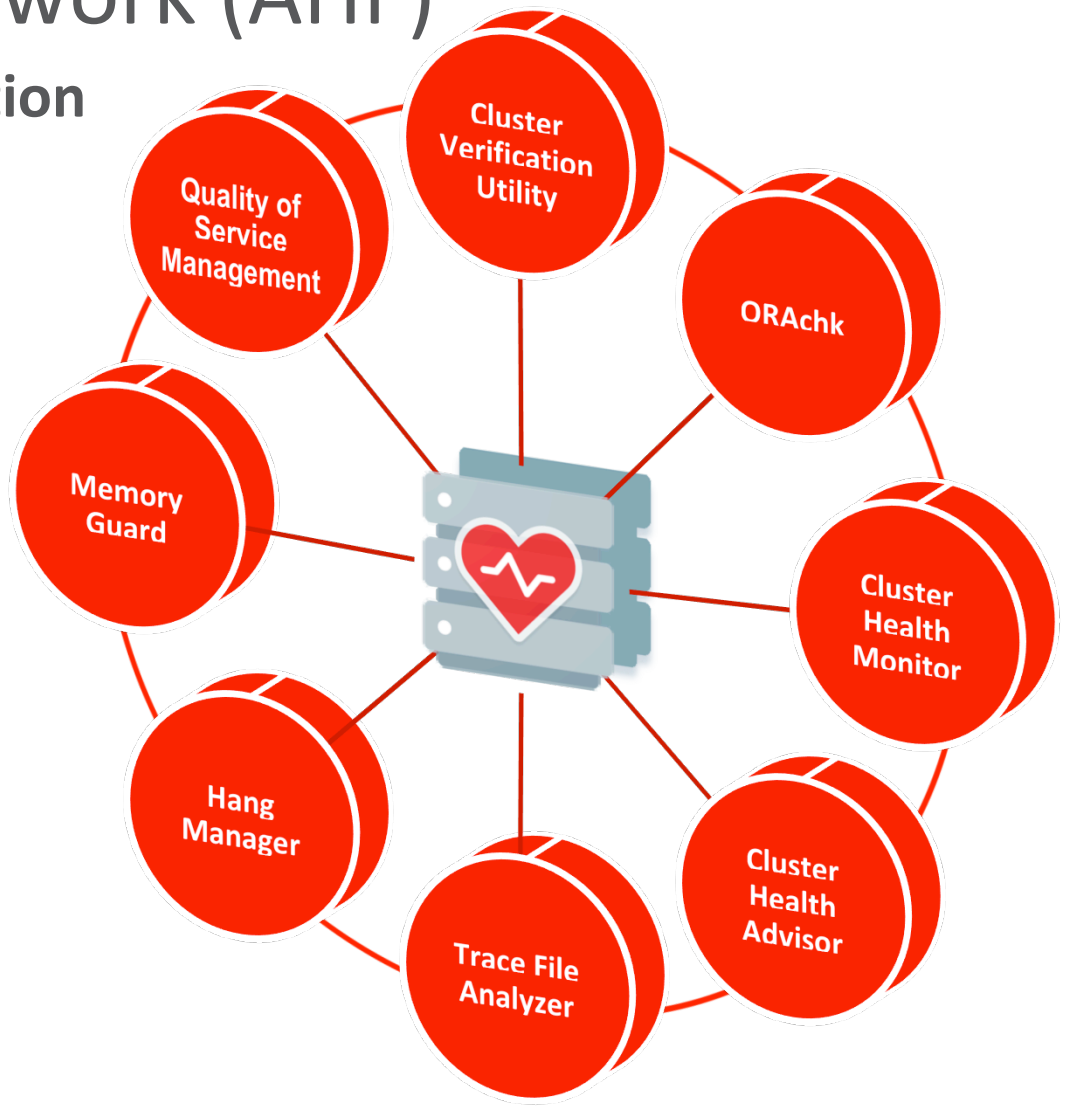
Status	Type	Message	Status On	Details
FAIL	Cluster Wide Check	Firmware version does not match on all Infiniband switches	Cluster Wide	View
FAIL	Cluster Wide Check	Localtime configuration does not match on all Infiniband switches	Cluster Wide	View

Check for
recommended best practices
and implement manually

Oracle Autonomous Health Framework (AHF)

For automated and simpler best practices application

Oracle AHF integrates next generation tools running as components - 24/7



Oracle Cluster Verification Utility (CVU)

Establishes Baseline and Maintains Best Practice Configurations

- Use during install and operation
- Checks O/S, GI and DB mandatory compliance and best practices with healthcheck option
- **Runs as daemon every 6 hours**
- ASM Best Practices Check
- **Creates Baseline collections**
- *Add user-specified/disable problem checks*
- *New user-friendly report format*

Detailed report for Best Practices checks

Summary of environment

Date (mm/dd/yyyy)	11/09/2016
Time (hh:mm:ss)	14:13:40
Cluster name	rwsbi0508-mb1
Clusterware version	12.2.0.1.0
Grid home	/scratch/app/12.2/grid
Grid User	grid
Operating system	Linux3.8.13-68.3.4.el6uek.x86_64

Following components are checked as part of this report (Click on each component listed below to navigate)

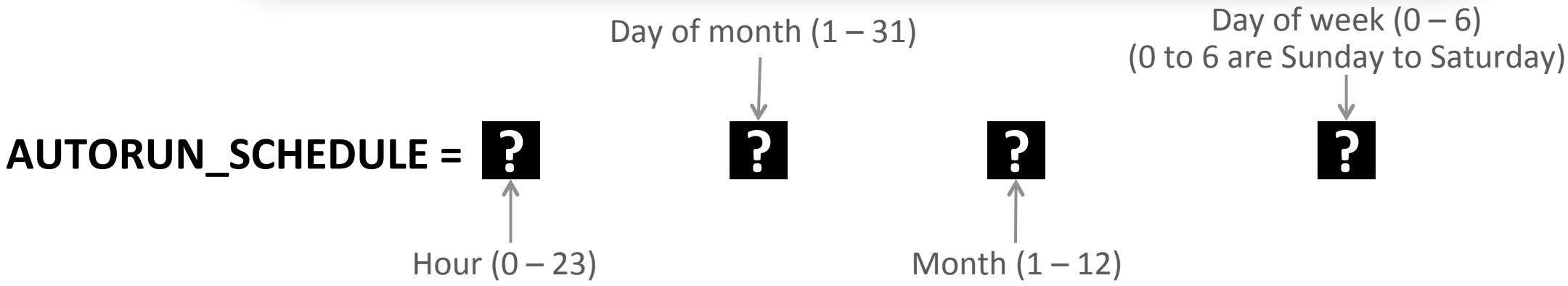
1. System requirements
2. System recommendations
3. Clusterware requirements
4. Clusterware recommendations

↑Top↑ **System requirements** ↑Top↑

Verification Check	Verification Result	Verification Description
Swap Size	WARNING	This is a prerequisite condition to test whether sufficient total swap space is available on the system.
Physical Memory	PASSED	This is a prerequisite condition to test whether the system has at least 8GB (8388608.0KB) of total physical memory.
Available Physical Memory	PASSED	This is a prerequisite condition to test whether the system has at least 50MB (51200.0KB) of available physical memory.
Free Space: rwsbi08:/usr,rwsbi08:/var,rwsbi08:/etc,rwsbi08:/sbin,rwsbi08:/tmp	PASSED	This is a prerequisite condition to test whether sufficient free space is available in the file system.
Free Space: rwsbi07:/usr,rwsbi07:/var,rwsbi07:/etc,rwsbi07:/sbin,rwsbi07:/tmp	PASSED	This is a prerequisite condition to test whether sufficient free space is available in the file system.

Schedule ORAchk

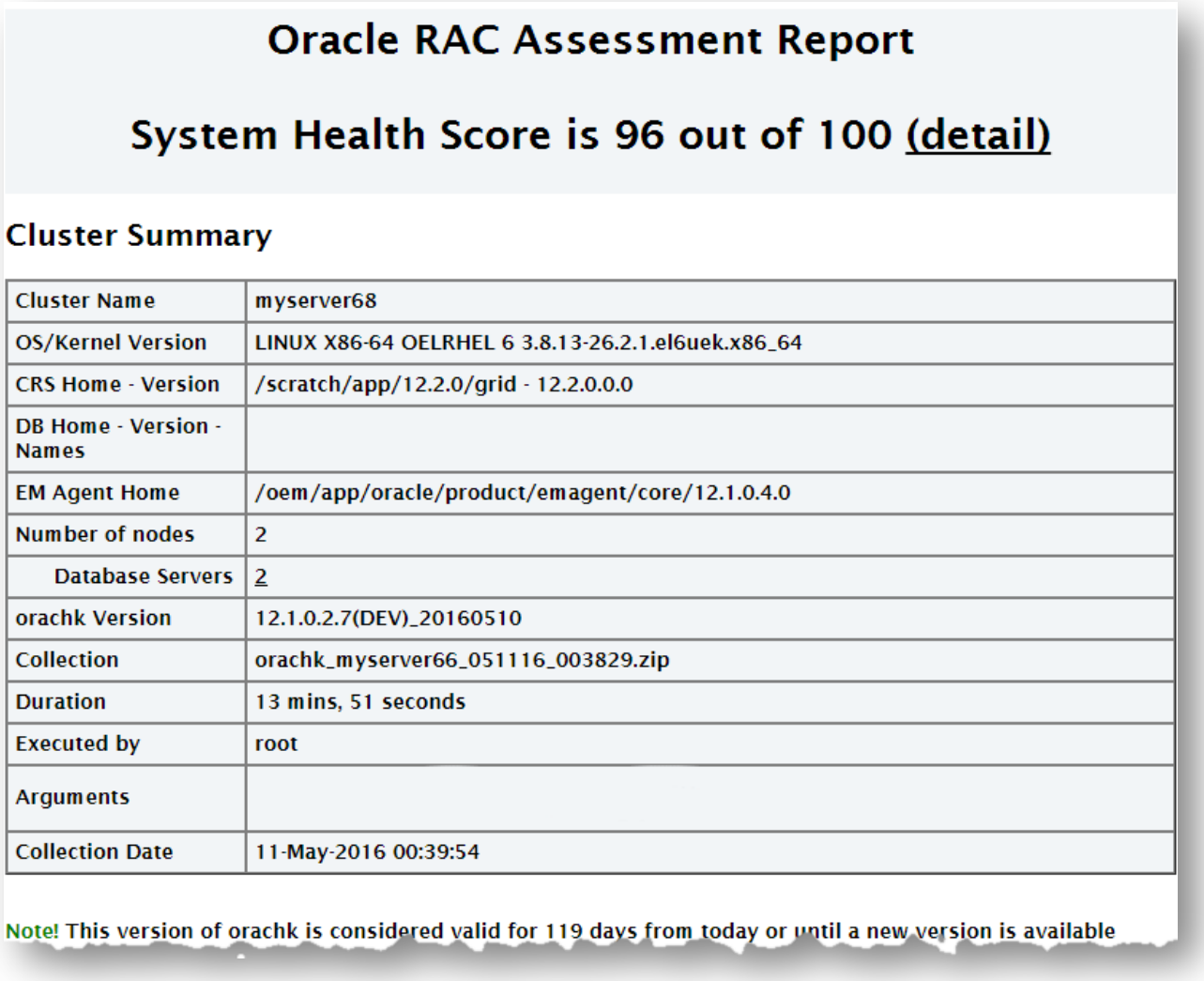
```
$./orachk -set "AUTORUN_SCHEDULE=3 * * 0 ;NOTIFICATION_EMAIL=some.body@company.com"

Updated autorun_schedule for ID[orachk.default]

Updated notification_email for ID[orachk.default]

$./orachk -d start
```



example: orachk -set 'AUTORUN_SCHEDULE=8,20 * * 2,5' will schedule runs on Tuesday and Friday at 08:00 & 20:00

View ORAchk Report

- Health score
- Summary of ORAchk run
- Table of content
- Controls for report features
- ***Findings***
- ***Recommendations***



Oracle RAC Assessment Report

System Health Score is 96 out of 100 (detail)

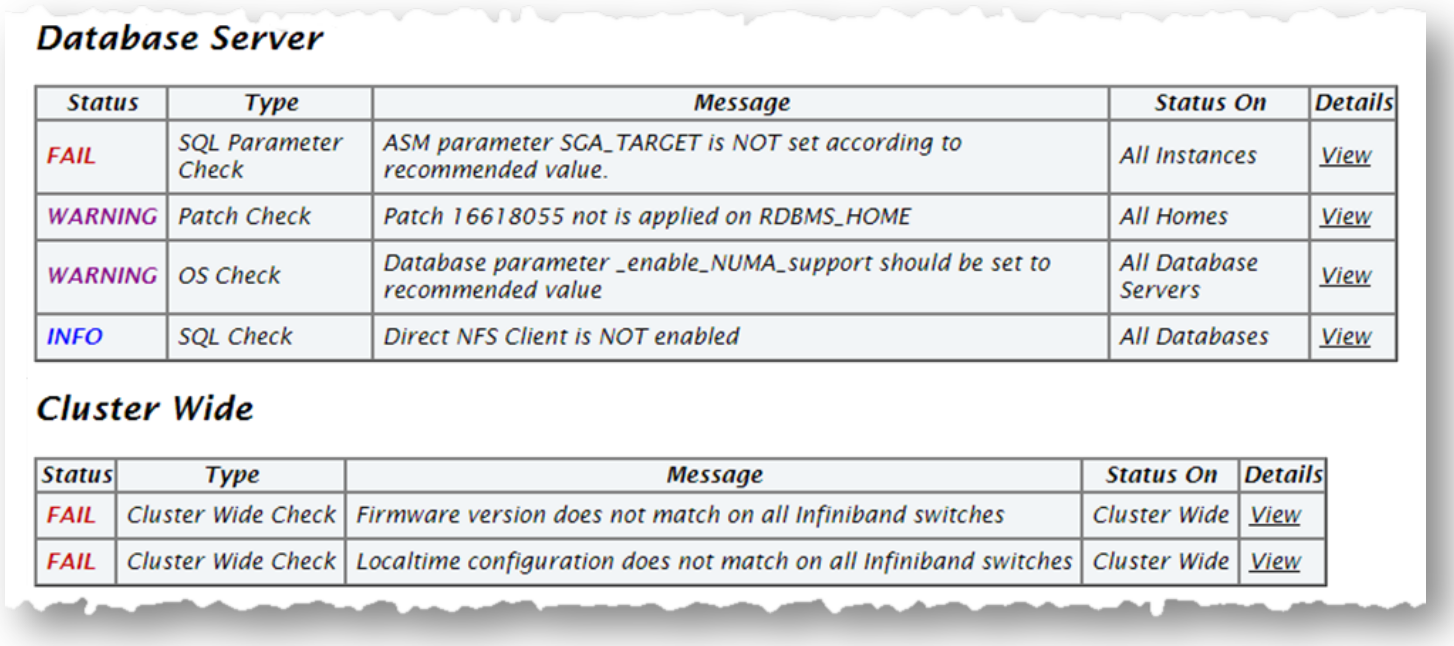
Cluster Summary

Cluster Name	myserver68
OS/Kernel Version	LINUX X86-64 OELRHEL 6 3.8.13-26.2.1.el6uek.x86_64
CRS Home - Version	/scratch/app/12.2.0/grid - 12.2.0.0.0
DB Home - Version - Names	
EM Agent Home	/oem/app/oracle/product/emagent/core/12.1.0.4.0
Number of nodes	2
Database Servers	2
orachk Version	12.1.0.2.7(DEV)_20160510
Collection	orachk_myserver66_051116_003829.zip
Duration	13 mins, 51 seconds
Executed by	root
Arguments	
Collection Date	11-May-2016 00:39:54

Note! This version of orachk is considered valid for 119 days from today or until a new version is available

View ORAchk Findings

- Check status
- Type of Check
- Check Message
- Where the check was run
- Link to expand details



Database Server

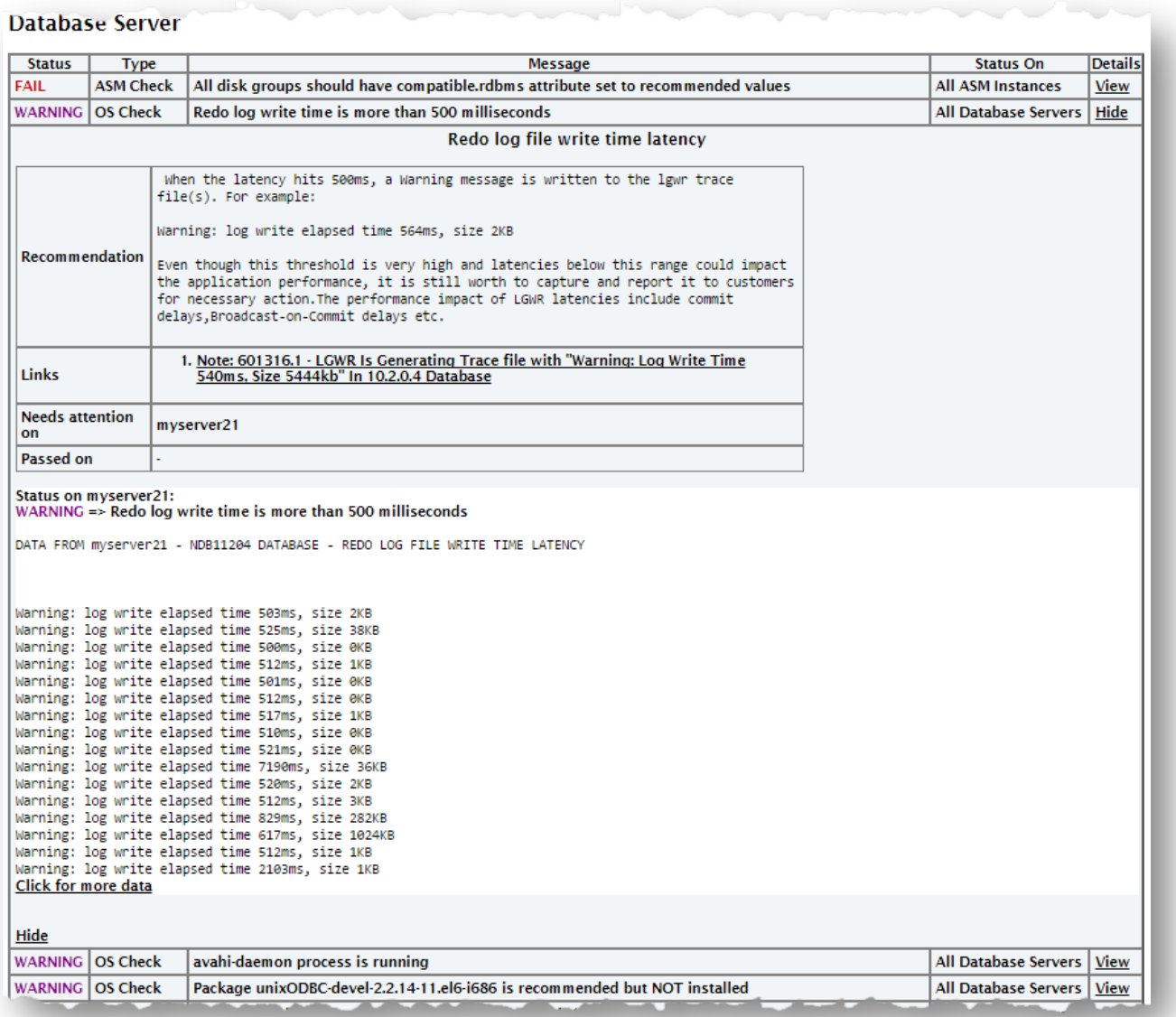
Status	Type	Message	Status On	Details
FAIL	SQL Parameter Check	ASM parameter SGA_TARGET is NOT set according to recommended value.	All Instances	View
WARNING	Patch Check	Patch 16618055 not is applied on RDBMS_HOME	All Homes	View
WARNING	OS Check	Database parameter _enable_NUMA_support should be set to recommended value	All Database Servers	View
INFO	SQL Check	Direct NFS Client is NOT enabled	All Databases	View

Cluster Wide

Status	Type	Message	Status On	Details
FAIL	Cluster Wide Check	Firmware version does not match on all Infiniband switches	Cluster Wide	View
FAIL	Cluster Wide Check	Localtime configuration does not match on all Infiniband switches	Cluster Wide	View

Act on Recommendations

- What to do to solve the problem
- Links to relevant Knowledge docs
- Where recommendation applies
- Where problem doesn't apply
- Example of data the recommendation is based on



Database Server

Status	Type	Message	Status On	Details
FAIL	ASM Check	All disk groups should have compatible.rdbms attribute set to recommended values	All ASM Instances	View
WARNING	OS Check	Redo log write time is more than 500 milliseconds	All Database Servers	Hide

Redo log file write time latency

Recommendation	When the latency hits 500ms, a Warning message is written to the lgwr trace file(s). For example: Warning: log write elapsed time 564ms, size 2KB Even though this threshold is very high and latencies below this range could impact the application performance, it is still worth to capture and report it to customers for necessary action.The performance impact of LGWR latencies include commit delays,Broadcast-on-Commit delays etc.
Links	1. Note: 601316.1 - LGWR Is Generating Trace file with "Warning: Log Write Time 540ms, Size 5444kb" In 10.2.0.4 Database
Needs attention on	myserver21
Passed on	-

Status on myserver21:
WARNING => Redo log write time is more than 500 milliseconds

```
DATA FROM myserver21 - NDB11204 DATABASE - REDO LOG FILE WRITE TIME LATENCY



Warning: log write elapsed time 503ms, size 2KB
Warning: log write elapsed time 525ms, size 38KB
Warning: log write elapsed time 500ms, size 0KB
Warning: log write elapsed time 512ms, size 1KB
Warning: log write elapsed time 501ms, size 0KB
Warning: log write elapsed time 512ms, size 0KB
Warning: log write elapsed time 517ms, size 1KB
Warning: log write elapsed time 510ms, size 0KB
Warning: log write elapsed time 521ms, size 0KB
Warning: log write elapsed time 7190ms, size 36KB
Warning: log write elapsed time 520ms, size 2KB
Warning: log write elapsed time 512ms, size 3KB
Warning: log write elapsed time 829ms, size 282KB
Warning: log write elapsed time 617ms, size 1024KB
Warning: log write elapsed time 512ms, size 1KB
Warning: log write elapsed time 2103ms, size 1KB
```

Click for more data

Hide

WARNING	OS Check	avahi-daemon process is running	All Database Servers	View
WARNING	OS Check	Package unixODBC-devel-2.2.14-11.el6-i686 is recommended but NOT installed	All Database Servers	View

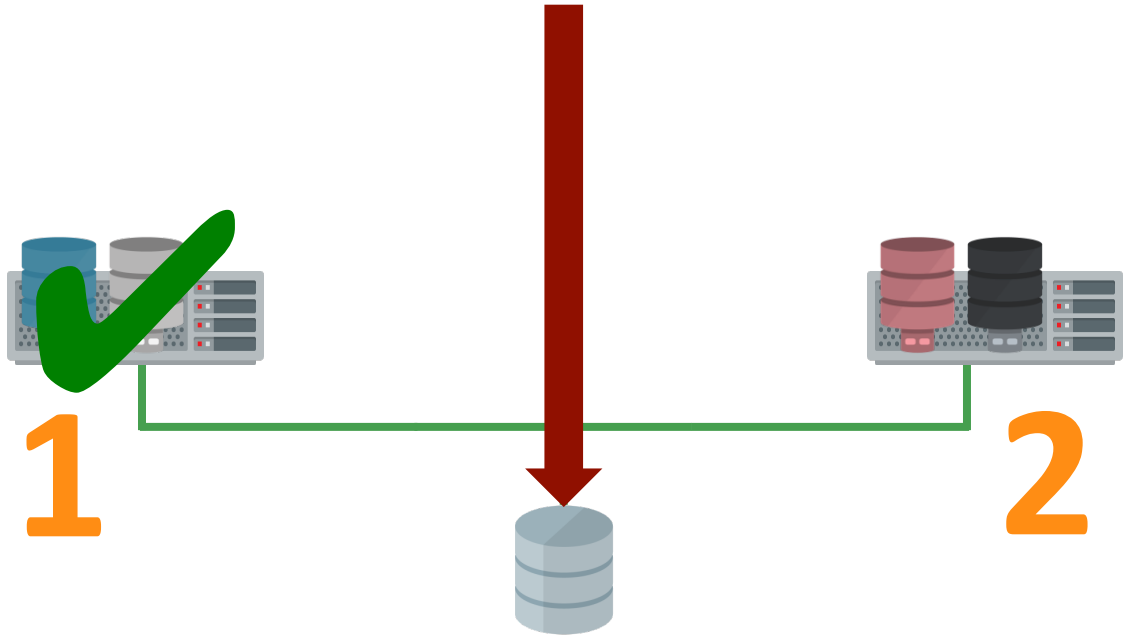
Program Agenda

1. Fundamentals To Be Considered First
2. Architecture Choices
3. Applying Operational Best Practices
4. Using Smart Features Automatically
5. Summary

Smart Fencing

Node Eviction Basics

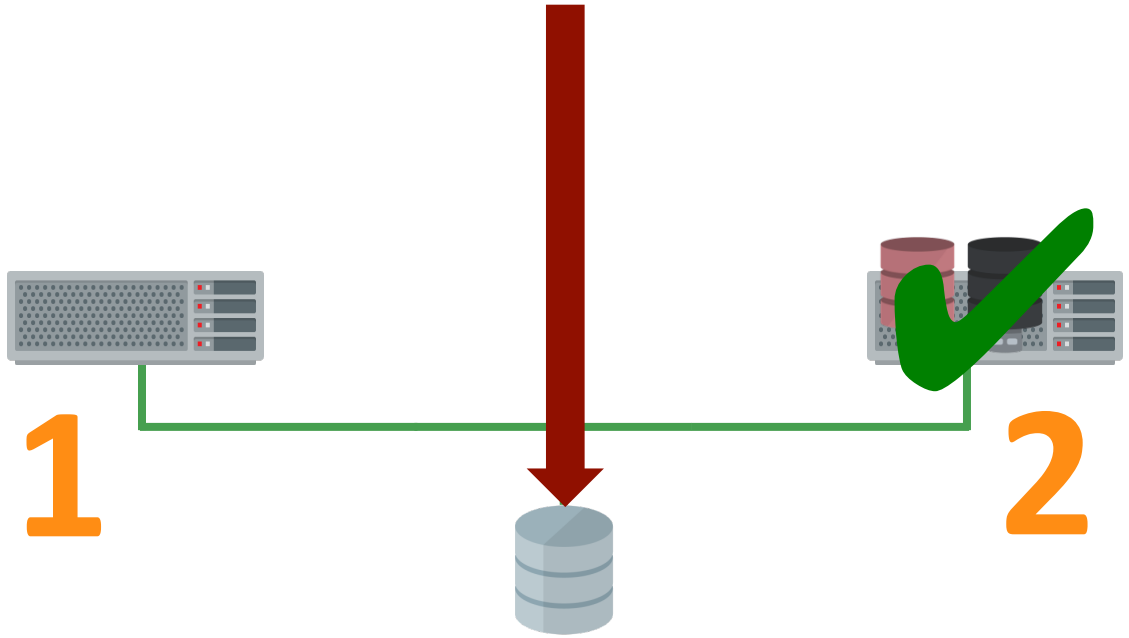
http://www.slideshare.net/MarkusMichalewicz/oracle-clusterware-node-management-and-voting-disks



- Pre-12.2, node eviction follows a rather “ignorant” pattern
  - Example in a 2-node cluster: The node with the lowest node number survives.
- Customers must not base their application logic on which node survives the split brain.
  - As this may(!) change in future releases

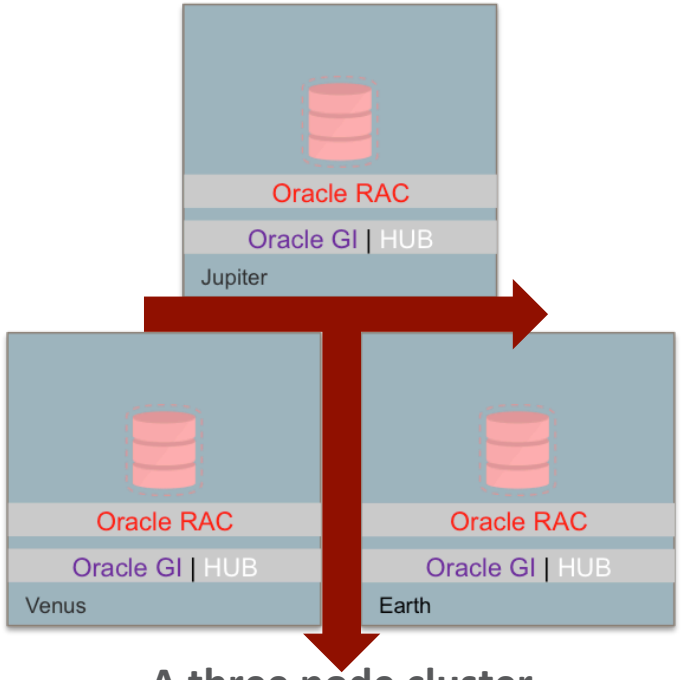
Node Weighting in Oracle RAC 12*c* Release 2

Idea: *Everything equal*, let the majority of work survive

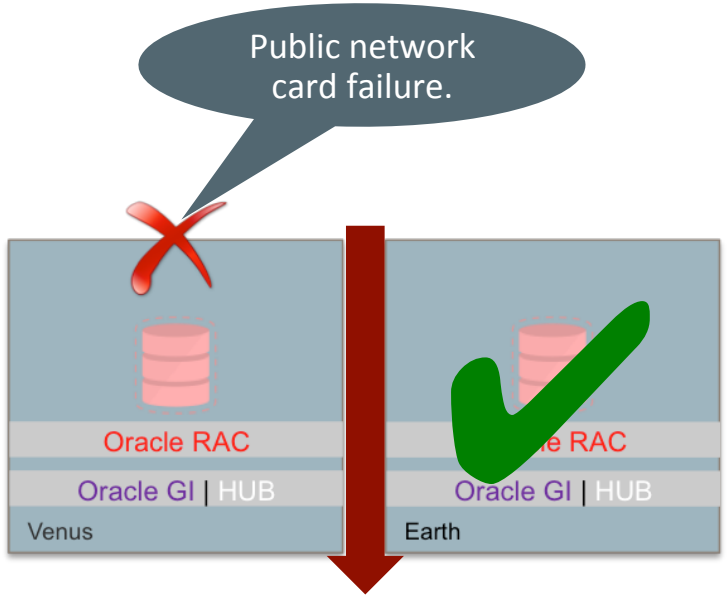


- Node Weighting is a new feature that considers the workload hosted in the cluster during fencing
- The idea is to let the majority of work survive, if *everything else is equal*
  - Example: In a 2-node cluster, the node hosting the majority of services (at fencing time) is meant to survive

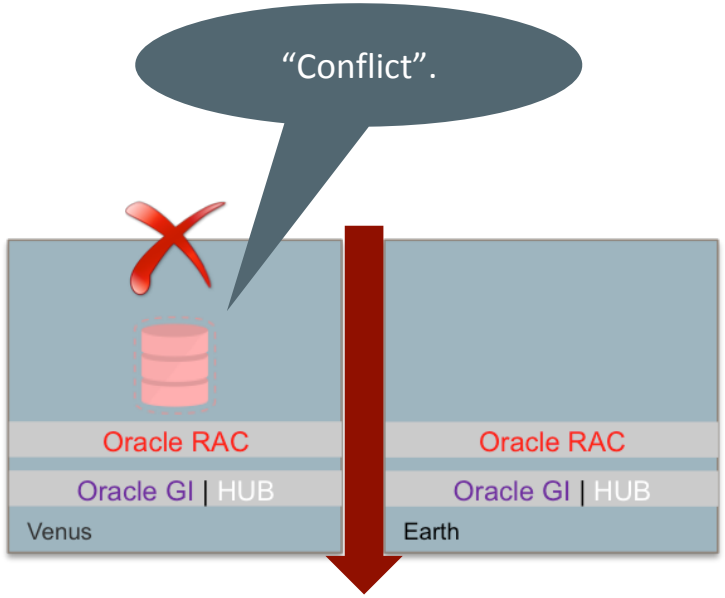
Let’s Define “Equal”



A three node cluster will benefit from “Node Weighting”, if three equally sized sub-clusters are built as s result of the failure, since two differently sized sub-clusters are not equal.

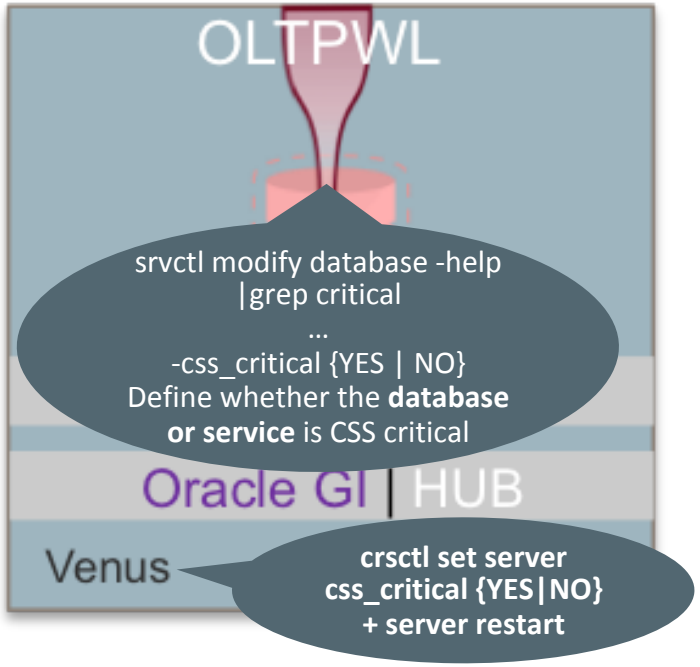


Secondary failure consideration can influence which node survives. Secondary failure consideration will be enhanced successively.

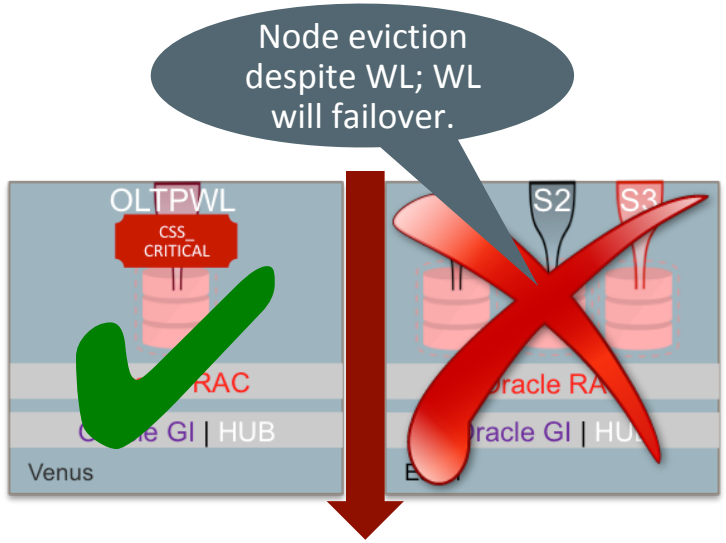


A fallback scheme is applied if considerations do not lead to an actionable outcome.

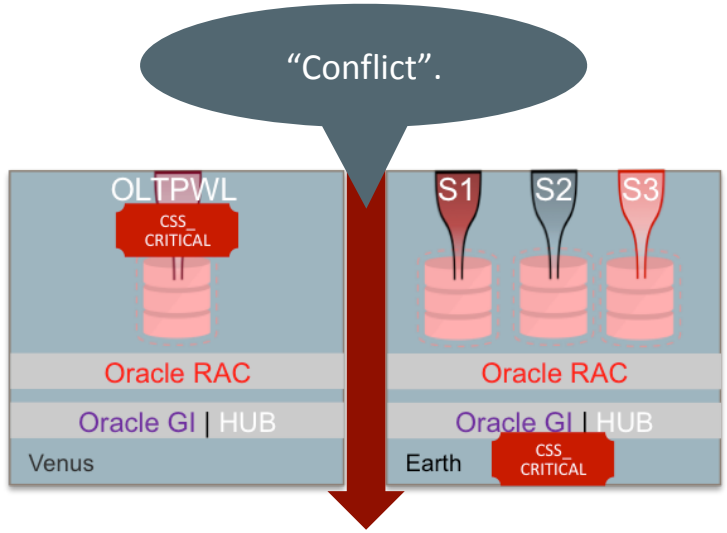
CSS_CRITICAL – Fencing with Manual Override



CSS_CRITICAL
can be set on various levels / components to mark them as "critical" so that the cluster will try to preserve them in case of a failure.



CSS_CRITICAL will be honored
if no other technical reason prohibits survival of the node which has at least one critical component at the time of failure.

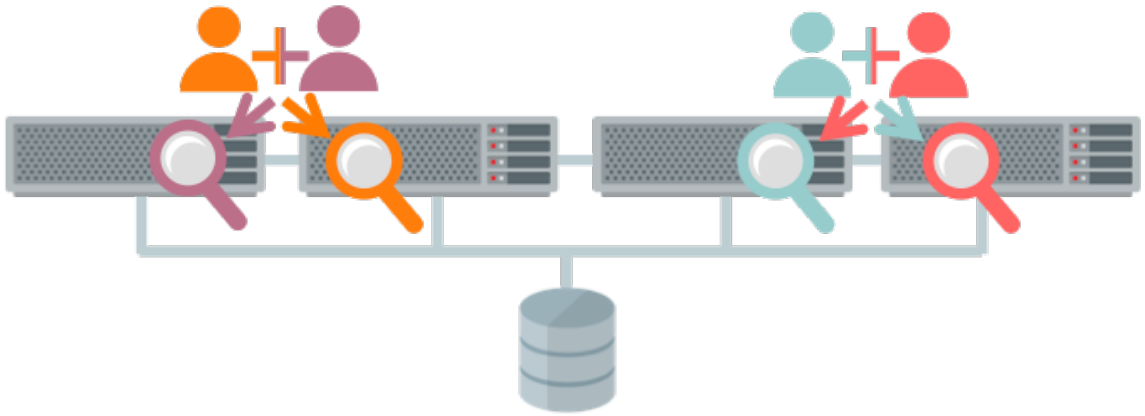


A fallback scheme is applied if CSS_CRITICAL settings do not lead to an actionable outcome.

Recovery Buddies

Near Zero Reconfiguration Time with Recovery Buddies

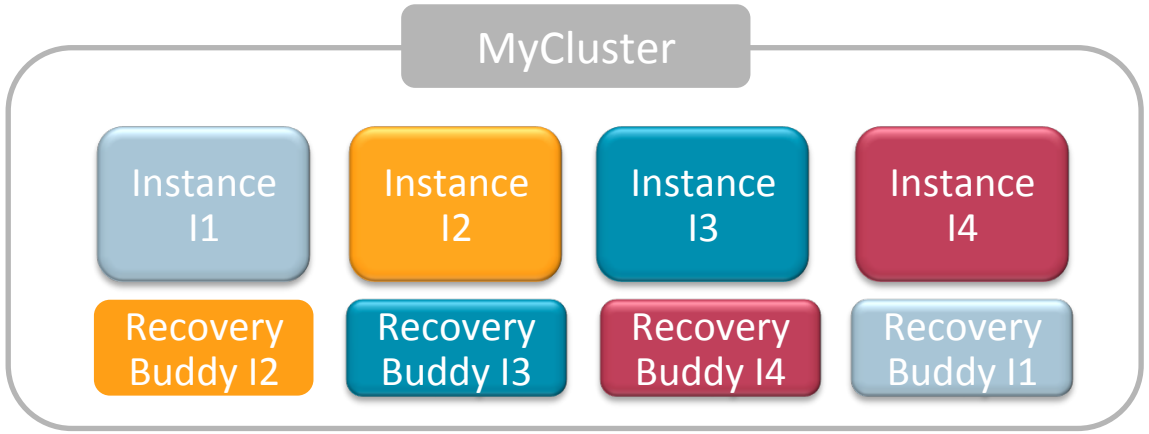
A.k.a. Buddy Instances



- Recovery Buddies
  - Track block changes on buddy instance
  - Quickly identify blocks requiring recovery during reconfiguration
  - Allow rapid processing of transactions after failures

Near Zero Reconfiguration Time with Recovery Buddies

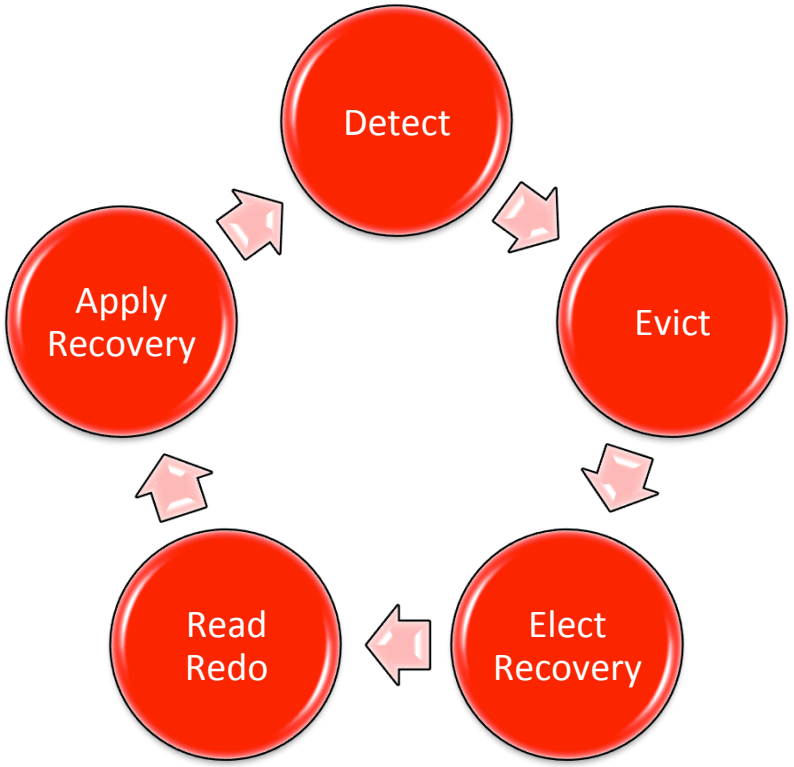
How it works under the hood



- Buddy Instance mapping is simple (random)
  - e.g. I1 → I2, I2 → I3, I3 → I4, I4 → I1
- Recovery buddies are assigned during startup
- RMS0 on each recovery buddy instance maintains an in-memory area for redo log change
- An in-memory area is used during recovery
  - Eliminates the need to physically read the redo

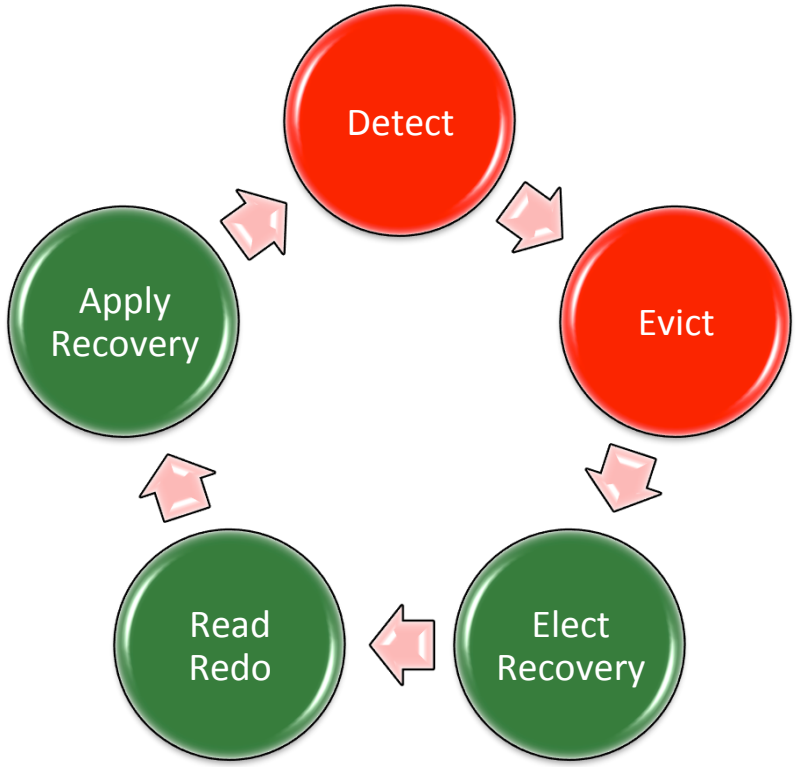
How Recovery Buddies Help Reducing Recovery Time

Without Recovery Buddies

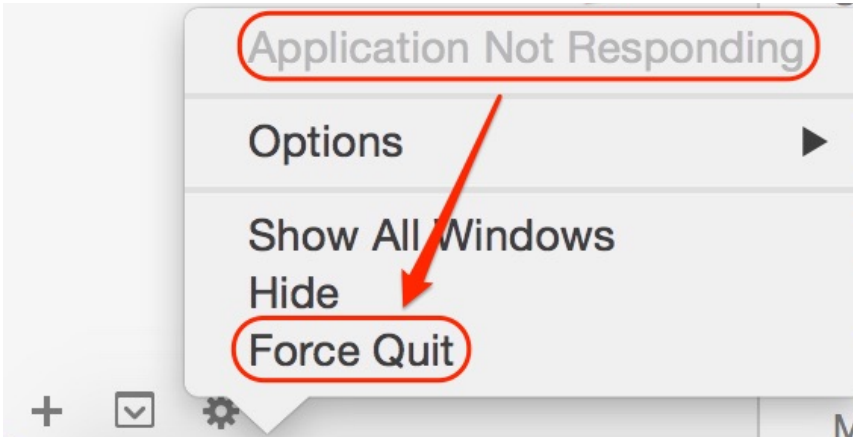


Up to 4x faster

With Recovery Buddies

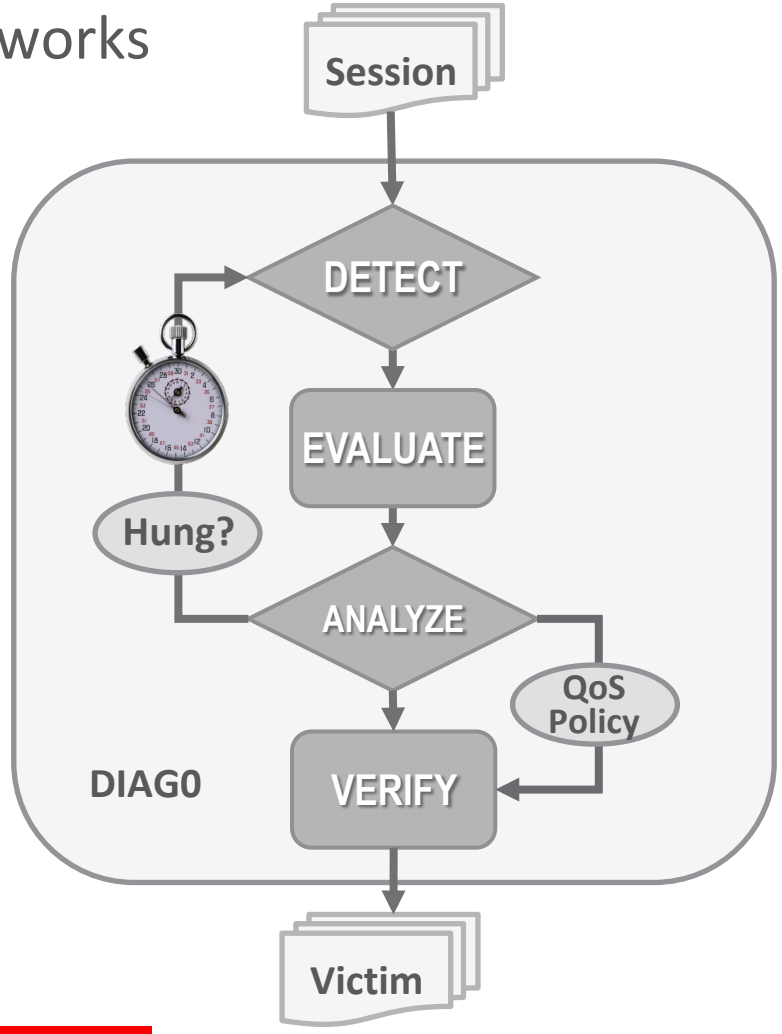


Database Hang Manager



Introduction to Hang Manager

How it works



- Always on, as enabled by default
- Reliably detects database hangs
- Automatically resolves hangs
- Considers QoS policies for hang resolution
- Logs all detected hangs & their resolutions

Hang Manager Optimizations with Oracle RAC 12c (Rel. 2)

Tuning under the hood

```
Hang Statistics

            current number of active sessions 7
              current number of hung sessions 6
  instance health (in terms of hung sessions) 15.00%
```

- Hang Manager auto-tunes itself by periodically collecting instance-and cluster-wide hang statistics
- Metrics like cluster health/instance health is tracked over a moving average
- This moving average is considered during resolution
- Holders waiting on SQL*Net break/reset are fast tracked

DBMS_HANG_MANAGER.Sensitivity

A new SQL interface to set Hang Manager sensitivity

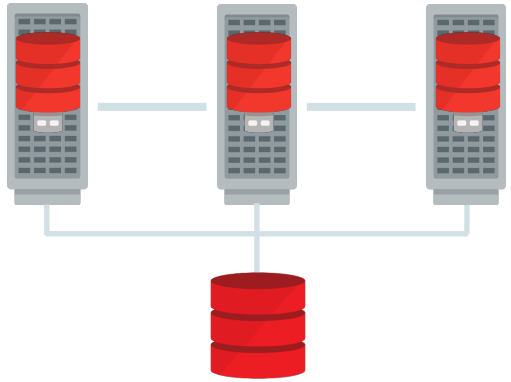
Hang Sensitivity Level	Description	Note
NORMAL	Hang Manager uses its default internal operating parameters to try to meet typical requirements for any environments.	Default
HIGH	Hang Manager is more alert to sessions waiting in a chain than when sensitivity is in NORMAL level.	

- Early warning exposed via (V$ view)
- Sensitivity can be set higher
  - If the default level is too conservative
- Hang Manager considers QoS policies and data during the validation process

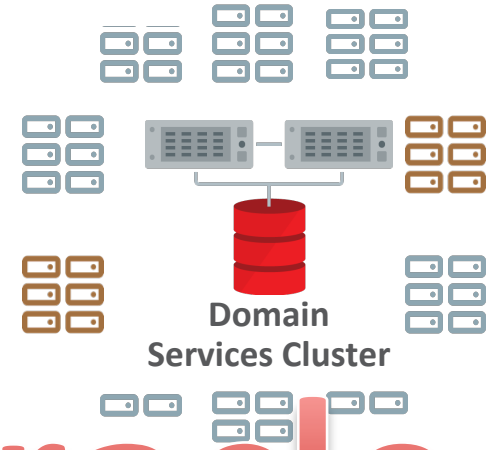
Program Agenda

1. Fundamentals To Be Considered First
2. Architecture Choices
3. Applying Operational Best Practices
4. Using Smart Features Automatically
5. Summary

Summary



Standalone Cluster



Cluster Domain

It's all Oracle RAC!

Virtualized environments

Cloud

Integrated Cloud

Applications & Platform Services