

ORACLE®



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.



ORACLE®

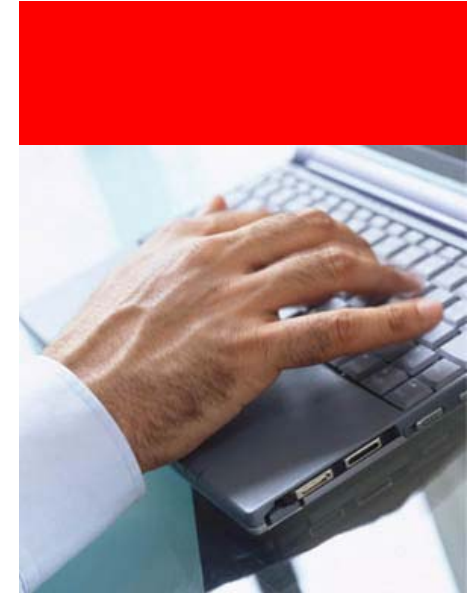


Oracle Real Application Clusters Installation and Configuration Best Practices

Duane Smith & Su Tang
RACPACK – Server Technologies
March 6, 2008

Agenda

- Installation Overview
- Cluster Verification Utility (CVU ; cluvfy)
- Cluster Interconnect
- Clusterware
- Storage / MPIO
- ASM / Shared filesystem
- Patching
- Monitoring / Tuning
- Tips & Troubleshooting





Oracle RAC Installation Overview

- Validate and prepare Hardware & OS
 - Consult Oracle Validated Configurations on OTN (Linux only)
 - <http://www.oracle.com/technology/tech/linux/validated-configurations/index.html>
- Determine cluster interconnect
- Determine storage methodology
- Install and configure the Oracle Clusterware Software
- Install the Oracle RDBMS RAC software
 - Can install ASM and create a database automatically

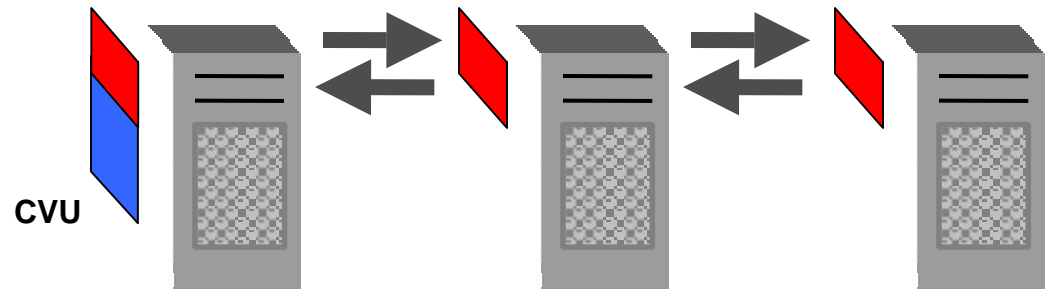


Cluster Verification Utility (CVU, cluvfy)

- Allows customers to verify cluster during various stages of its deployment from hardware setup, Clusterware Install, RDBMS install, storage, etc.
- Extensible framework
- Command Line only
 - `$./cluvfy comp peer -n node1,node2 | more`
- Does not take any corrective action following the failure of a verification task
 - Non-intrusive verification

Deployment of cluvfy

- **Install only on local node. Tool deploys itself on remote nodes during execution, as required.**
 - User installs on local node
 - Issues verification command for multiple nodes
 - Tool copies the required bits to the remote nodes
 - Executes verification tasks on all nodes and generates report





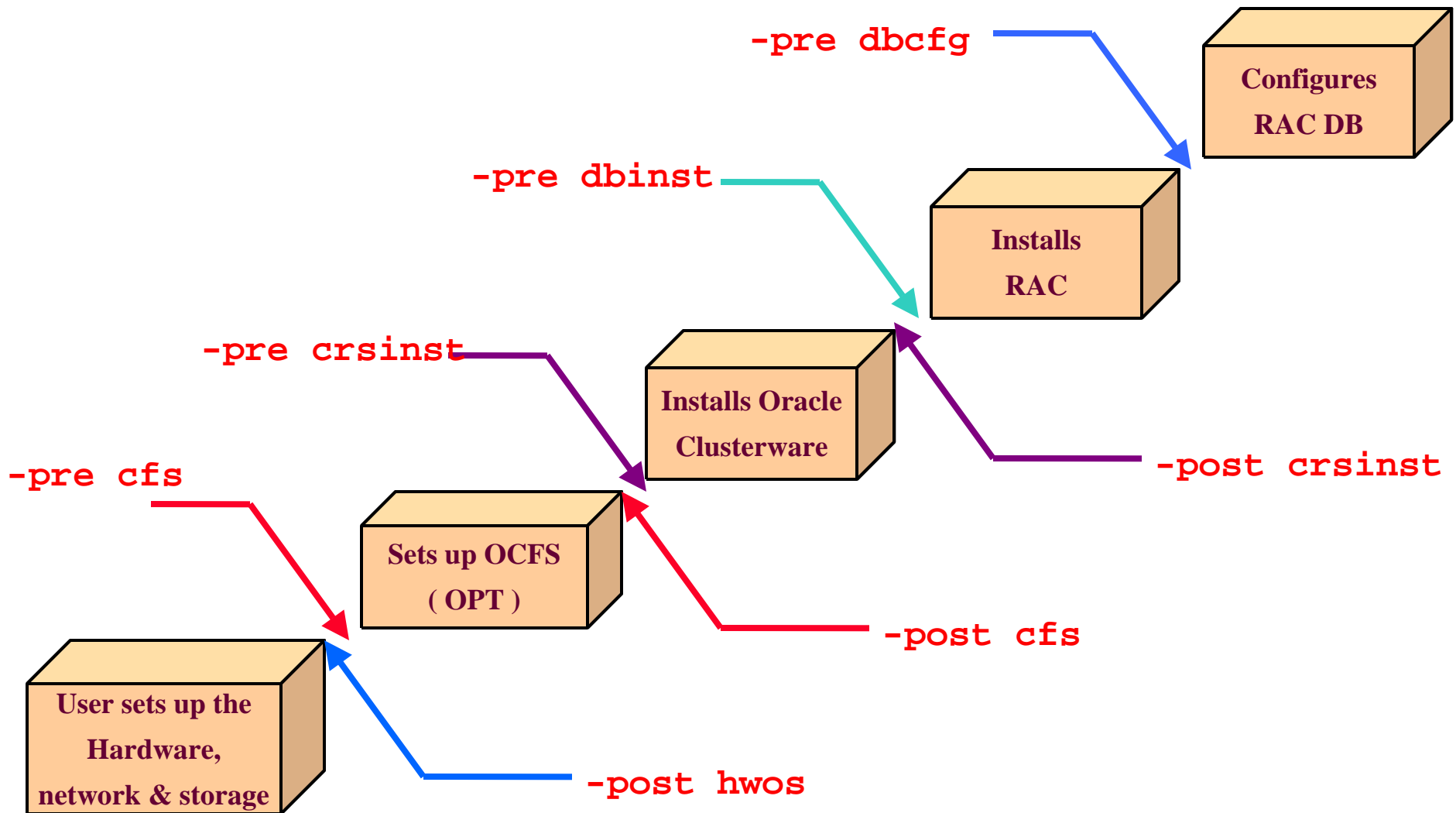
cluvfy Stage List

- Valid stage options and stage names are:

```
$ ./cluvfy stage -list
```

```
-post hwos      : post-check for hardware & operating system  
-pre  cfs       : pre-check for CFS setup  
-post cfs       : post-check for CFS setup  
-pre  crsinst   : pre-check for Clusterware installation  
-post crsinst   : post-check for Clusterware installation  
-pre  dbinst    : pre-check for database installation  
-pre  dbcfg     : pre-check for database configuration
```

cluvfy Stage List - Graphical





cluvfy Component List

- Valid components are:

```
$ ./cluvfy comp -list
```

nodereach	: checks reachability between nodes
nodecon	: checks node connectivity
cfs	: checks CFS integrity
ssa	: checks shared storage accessibility
space	: checks space availability
sys	: checks minimum system requirements
clu	: checks cluster integrity
clumgr	: checks cluster manager integrity
ocr	: checks OCR integrity
crs	: checks CRS integrity
nodeapp	: checks node applications existence
admprv	: checks administrative privileges
peer	: compares properties with peers



cluvfy Stage or Component

- Use **Stage** checks during installation of Oracle Clusterware and RAC.

- Use the appropriate `–pre` and `–post` check for the stages, e.g:

```
$ ./cluvfy stage –pre crsinst –n node1,node2 –verbose
```

- To verify a particular component while the stack is running or to isolate a cluster subsystem for diagnosis, use appropriate **Component** checks.



CVU locations

- Pre-Installation:
 - Cluster Verification Utility on OTN:
 - http://www.oracle.com/technology/products/database/clustering/cvu/cvu_download_homepage.html
 - Oracle DVD
 - `clusterware/cluvfy/runcluvfy.sh`
- Clusterware Home
 - `<crs_home>/bin/cluvfy`
- Oracle Home
 - `$ORACLE_HOME/bin/cluvfy`



Cluster Interconnect Best Practices

- Use UDP over Gigabit Ethernet
 - TCP if Windows, RDG on Tru64 (See [Note 278132.1](#))
- Use OS Bonding/teaming to “virtualize” interconnect
 - Failover ; Load-balancing ; Improved bandwidth
- Set UDP send/receive buffers high enough
 - Platform dependant – typically 256K is adequate
 - Linux: `net.core.rmem_max`, `net.core.wmem_max`,
`net.core.rmem_default`, `net.core.wmem_default`
- Use a private dedicated non-routable Switch or VLAN
 - crossover cables are **not** supported
- Eliminate any Transmission Problems
 - Packet errors/drops can manifest into more serious outages

Cluster Interconnect (Cont')

- Use same interconnect for both Clusterware and DB (GCS/DLM/PQ) communications
 - Clusterware uses: `oifcfg getif`
 - DB uses: `select * from v$cluster_interconnects;`
 - DB may override Clusterware via the `cluster_interconnects` init.ora on a per database basis
 - Might be needed with many databases on the same cluster
- Private NICs and public NICs should be kept the same name/order



Clusterware Misscount

- Oracle Clusterware has two heartbeats
 - Network: **misscount**, defaults to 30sec (Linux 10g: 60sec)
 - Disk (IOT): function of **misscount** varies by release
 - Problematic approach to tie the two together (not granular enough)
- Patch 4896338 decouples above timeouts (disk/network)
 - New css **disktimeout** parameter; defaults to 200 seconds
 - Reconfiguration/reboot **only** if **misscount** exceeded for network or **disktimeout** exceeded for voting disk
- Included in 11g; 10.2.0.2 patchset & 10.1.0.4 cumulative Clusterware patch
- Note 294430.1: Misscount definitions
- Note 284752.1: Change misscount/reboottime/disktimeout
- Best Practice: Do NOT change misscount or disktimeout unless on the recommendation of Support



VIP IP in RAC

- Used to mitigate TCP/IP timeout delays on client connections
- When configuring (VIPCA) choose only the **public** interfaces
 - Watch out the default subnet is 255.255.255.0, correct it if needed
 - On SLES10 / RHEL5 / OEL5: VIPCA fails during root.sh, see [Note:414163.1](#) (10g)
- The VIP must be a DNS known IP address
 - Clients connect to VIP address from tnsnames connect description
 - Listeners listen on VIP for client connections
- Use ifconfig (on most platforms) to verify VIP interface is configured **after** Clusterware is running
 - IP address on the new VIP interface eg: eth0:1, should respond to pings
- The VIP is stored within the OCR
 - To modify the VIP IP see [Note:276434.1](#)

Mirror OCR/Voting disk

- Oracle Cluster Repository (OCR), and split brain resolution mechanism (Voting Disk)
- Storage Options: Block, RAW, CFS or certified NFS
- 10gR2: Oracle mirroring is recommended at install time
 - Configure 3 Voting disks and 2 OCR devices
 - Post install:
 - # **crsctl add css votedisk path**
 - # **ocrconfig -replace ocrmirror destination file/disk**
- 10gR1: Limited to hw RAID and OS LVM
- Split brain resolution requires majority of disks to allow sub-cluster to continue
 - Stretched clusters may place voting on 3rd location over NFS (Linux, AIX, Solaris)
- Auto OCR Backups: # **ocrconfig -showbackup**
 - New in 11g: # **ocrconfig -manualbackup**



Determine Storage Methodology

- ASM (Automatic Storage Management) -
RECOMMENDED
- Clustered Filesystems:
 - Oracle Cluster Filesystem (OCFS1 for 2.4 kernels; OCFS2 for 2.6 kernels) [FREE]
 - Certified 3rd party clustered filesystems
- NFS (Certified Network Filesystem, see OTN for supported list)
- Raw Devices (**Avoid!**, limited to 255; needed only for OCR/Voting in 10g, not needed in 11g)
- iSCSI provides block devices (for use with ASM, OCFS2, etc.)



IO Multipathing

- Device driver automatically or manually combines multiple paths to the same device
 - Two HBAs become one virtual HBA (Host Bus Adapter)
 - Failover, Bandwidth aggregation, path rediscovery
- On Linux (Open Source, FREE)
 - 2.6 kernels: **Device Mapper (DM)** (decent)
 - Fixes all the lacks of MD and then some
 - 2.4 kernels: **Multipath Device (MD)** (mdadm) **AVOID**
 - Long timeout (90 seconds) for failover to kick in
 - Manual configuration of path, No path rediscovery
 - **Use 3rd Party instead**
- Third-party (HP, EMC, IBM, Sun, HDS, Veritas, Qlogic) multipathing on Linux
 - No Unbreakable support from Oracle



3rd Party IO Multipathing

- HP – Secure Path, Auto Path XP
 - Only HP storage
- IBM (Varies by storage/OS)
 - MPIO Driver (Multi-Path Input Output): AIX
 - SDD (Subsystem Device Driver): AIX, Linux, HP-UX, Solaris, Windows
 - RDAC (Redundant Disk Array Controller):
 - AIX, Linux, Windows
- Sun – StorEdge Traffic Manager (Sun Storage only)
- Microsoft – MPIO Software dev kit (not AIX's MPIO)
- EMC – Power Path
 - Compatible with many storage arrays
- Qlogic – Must use Qlogic HBAs
- Symantec/Veritas – Dynamic Multipathing (DMP, VxVM)
 - Must use Veritas LVM, create logical volumes for ASM to use



When NOT to use Multipathing

- MP is transparent to ASM/ASMLIB, avoid these cases:
 - When MP requires root access to MP device
 - When MP requires an LVM in the path
 - When 3rd party vendor does not certify



ASM Recommendations

- Install ASM on a separate Oracle Home
 - NEW in 11g – ASM rolling upgrades are possible, 11g onwards
- Set INIT.ORA on ASM and DB as per recommendations
- Remove ASM dependency on VIP (bug 4865736)
 - If VIP fails ASM instance remains operational
 - Fixed in 11g and 10.2.0.3 patchset; download fix for 10.2.0.2
- If mirroring is done in the storage array, set REDUNDANCY=EXTERNAL for the diskgroup
- On Linux use ASMLib
 - Protects against device name changes across reboots without compromising security ([Note 394959.1](#))
 - Fewer kernel resources, no configs to modify as disks are added
 - Global Open/Close for ASM devices



Shared Oracle Home

- Shared Oracle Home requires a shared filesystem
 - OCFS2, Certified NFS device, etc.
- Only one copy of the software to maintain & faster installation, however with following drawbacks:
 - Can not perform rolling upgrade of patches/sets
 - Binaries have local dependencies
 - Requires cross-node OS compatibility
 - Single point of failure
- Avoid using Shared Home
 - Especially for the Oracle Clusterware Home

http://www.oracle.com/technology/products/database/clustering/pdf/oh_rac.pdf



Summary: Install Oracle RAC 10g / 11g on Linux

- Consult Oracle Validated Configurations on OTN
- OS at latest revision + set kernel parameters correctly
- Run Cluster Verification Utility (CVU) at various stages
- Install and configure the Oracle Clusterware Software
- Install the Oracle RDBMS RAC software
 - Can install ASM and create a database automatically

Patching/SW Maintenance

- Stay current with:
 - CPU's (Critical Patch Update)
 - RDBMS Patchsets ; Clusterware bundle patches
- Use latest Opatch; download from Metalink
 - 10.2 placeholder bug 4898608; 10.1 placeholder bug 2617419
- Review Support/Metalink (e.g. 10.2.0.2 see Notes: [359415.1](#), [391116.1](#)) recommended 1-off patches from Support/Metalink
 - Ensure Support has specific platform info
 - `$ opatch lsinventory -detail` to ensure no patch conflicts
- Read individual patch readme's carefully
 - Not all patches install exactly the same way
- Confirm patch successfully applied to all nodes
 - `$ opatch lsinventory -oh <home location>`
- Patch first in test/QA environment
- NEW in Opatch 10.2.0.3: Apply/remove N patches at once
- NEW in 11g: Online Patching; some patches can be applied to running code



Patchsets (10.1.0.4, 10.2.0.3, etc.)

- Consist of two portions (Clusterware & RDBMS/ASM)
- Install using Oracle Universal Installer (OUI)
- Latest patchset (10.2.0.3) is always advised
- Oracle Clusterware must be newer or equal version of any RDBMS or ASM installed
- Oracle Clusterware portion can **always be installed** in a rolling upgrade fashion
- New in 11g: ASM can be upgraded as a rolling upgrade, 11g onwards
- RDBMS portion can only be installed in a rolling upgrade fashion if a logical stand-by exists



Patching Mixed Oracle Homes

- Mixed Oracle Home is when Clusterware and RDBMS/ASM versions are not identical
 - Fully supported; Clusterware always higher version
 - Patching is slightly different as follows
- Clusterware patches consist of two portions
 - One applied to Clusterware Home
 - Second applied to ASM or RDBMS Home
 - ASM & RDBMS treated equally for this purpose
- Attempt to install a 10.2 patch in a 10.1 RDBMS will fail
 - Patches must always be applied to exact version



Patching Mixed Oracle Homes (Cont')

Refer to Metalink [Note 363254.1](#) for full details

- You may skip the RDBMS portion
 - Bug may still be visible on that RDBMS Home
- Or; Request a one-off for the needed older RDBMS version
- Remember:
 - Never force a patch to be installed into incorrect version home
 - A single one-off zip will always contain exact versions for both portions of the patch (Clusterware,RDBMS)



Tuning Philosophy

- Philosophies differ
 - Tuning for new or existing database
 - Tend to start with things we know
 - Perception of a problem may sway your philosophy
- Here's mine...
 - Go for the best bang for your buck
 - Translation: Go after the big things first



Monitoring: General

- Vital to have good baseline to compare with
- Correlate I/O timing reported by Oracle to I/O timing reported by OS utilities & Hardware
 - For example: Database says I/O takes 60ms but hardware says 10ms, investigate why.
- OS and database statistics should be collected at the same time periods to have a meaningful comparison
 - Run **OSWatcher** and **statspack** continuously

Monitoring Tools: Linux Specific

- Overall tools
`sar , vmstat`
- CPU
`/proc/cpuinfo , mpstat , top`
- Memory
`/proc/meminfo , /proc/slabinfo`
- Disk I/O
`iostat, sar`
- Network
`iptraf, netstat, mii-tool`
- Individual process debugging
`strace , ltrace, lsof`



Monitoring Tools: RAC

- Oracle Enterprise Manager (recommended)
 - DB Control or Grid Control
 - With Diagnostics Pack license provides Automatic Database Diagnostic Monitor (ADDM) and Automatic Workload Management (AWR)
 - Comprehensive & concrete tuning recommendations
- Statspack
 - Manual snapshot/reporting, similar to AWR reports
 - No recommendations, user must conclude based on report
 - Metalink [Note: 94224.1](#)
- OS Watcher
 - Continuous collection of OS metrics automatically
 - Metalink [Note: 301137.1](#)



RAC Performance Recommendations

- Good SQL
- Reduce Hot Spots
 - Same as you would for a single instance
 - Set sequence cache to 1000 or more
- Scalable I/O sub-system
 - Implement multipathing
- Confirm Interconnect is actually being used
- Use Automatic Segment Space Management
 - “SEGMENT SPACE MANAGEMENT AUTO” in create tablespace
 - Remove PCTUSED, FREELIST & FREELISTS GROUPS



Tips & Troubleshooting

- Installing on a cluster with many nodes?
 - Use cluster configuration file (Text file with node names)
 - Metalink [Note 336912.1](#)
- Want silent OUI installs?
 - Try the **-record** flag to generate a response file
- Use REMOTE_NODES & CLUSTER_NODES options in OUI to install/manipulate a subset of nodes
- Setup SSH equivalency using **runSSHSetup.sh** on install directory of Clusterware CD/DVD.
- Use **pdsh** (Public Domain Shell) runs commands on all nodes



Tips & Troubleshooting (Cont')

- Correctly mount clustered filesystems
 - OCFS2: “**datavolume**” for database mount points ([Note 428356.1](#))
 - NFS: Correct NFS mount options ([Note 359515.1](#))
- Clusterware needs storage and network UP
 - Verify host startup sequence of network & I/O drivers, iSCSI
- Please use NTP (Network Time Protocol)
 - Easier debugging/diagnostics as time is in sync
 - Some issues may exist for Clusterware & DBMS_SCHEDULER if time drifts wildly
 - Jobs get scheduled incorrectly
 - May reboot nodes as **misscount** calculations will be incorrect
 - Use **-x** (if available) to prevent time from moving backwards



Tips & Troubleshooting (Cont')

- Ensure IO Storage scalability for multiple nodes early on
 - As nodes are added more storage bandwidth should be added
 - ORION: Oracle tool on OTN (Linux, Windows)
 - IOzone: Freeware on Internet (Cross platform)
- If OS stack size set too high (e.g. 200MB), Oracle Clusterware fails to start
 - Each thread consumes stack-size (200MB!!)
 - Leave at port-specific defaults
- Work with hardware vendor to confirm latest/certified firmware & drivers on all equipment, including network switches
- Enterprise Linux users should use Oracle Validated Configurations to fulfill all installation pre-requirements

<http://www.oracle.com/technology/tech/linux/validated-configurations/index.html>

Tips & Troubleshooting (Cont')

- On Windows: Disable Media Sense ([Note: 243549.1](#))
- Increase SYS.AUDSES\$ sequence cache ([Note: 395314.1](#)):
`alter sequence sys.audses$ cache 10000;`
 - Affects 9i up to and including 10.2.0.2
- Clusterware relies on OS authentication, if using LDAP ensure it's at High Availability standards or decouple the RAC nodes from LDAP
- Oracle RAC 11g on Linux uses `oprocd` to detect hangs
 - `hangcheck-timer` can still pickup lower level (device driver) hangs in 10g



Tips & Troubleshooting (Cont')

- Collect RAC traces/diagnostics
 - Remote Diagnostic Agent (RDA) 4.2 or above: [Note:359395.1](#)
 - RAC Diagnostic Data Tool (RAC-DDT): [Note 360926.1](#)
`<CRS_home>/bin/diagcollection.pl`
- Cluster Deconfig/Deinstall tool on OTN (10g: Linux x86)
 - Helps deinstall RAC 10g software for a clean reinstallation

<http://download.oracle.com/otndocs/products/clustering/deinstall/clusterdeconfig.zip>

ORACLE®



ORACLE IS THE INFORMATION COMPANY

A large, stylized, black 'QA' logo is centered on the page. The 'Q' is a thick, rounded letter, and the 'A' is a tall, narrow letter with a thick stroke. The letters are slightly overlapping.

QUESTIONS
ANSWERS