Oracle 12c RAC Flex
Architecture and Oracle 12c RAC Support for Oracle Multitenant Architecture

OTN MENA TOUR 2014, May 27, 2014

Kai Yu
Senior Principal Architect
Oracle Solutions Engineering, Dell Inc.
Agenda

• Oracle 12c Grid Infrastructure and RAC overview
• Oracle Flex Clusters
• Flex ASM Architecture
• Oracle RAC Support for Pluggable Databases
• Oracle RAC Troubleshooting and Health Check
About Author

- Kai Yu, Senior Principal Architect, Dell Database Engineering
  - 19 years Oracle DBA/Apps DBA and Solutions Engineering
  - Specializing in Oracle RAC, Oracle VM and Oracle EBS
  - Oracle ACE Director, Oracle papers author/presenter
  - 2011 OAUG Innovator of Year, 2012 Oracle Excellence Award: Technologist of the Year: Cloud Architect by Oracle Magazine
  - Co-author Apress Book “Expert Oracle RAC 12c”
What I do

Providing solutions on the whole stack from the ground up

http://www.delltechcenter.com/oracle

- Solutions Deliverable List
- Validated integration
- Best practices

- Virtualization
- Oracle EM12c
- Oracle Applications
- Performance Study
**Oracle 12c Grid Infrastructure and RAC**

- **Oracle Real Application Clusters: active-active cluster database**
  - Protect database availability against up to N-1 server failure
  - Reduce planned downtime for hardware, OS, software upgrade
  - Add node or remove node based on demand of capacity
  - Application load balancing
  - Provide high availability, scalability and flexibility

![Oracle 12c Grid Infrastructure and RAC Diagram]
Oracle 12c Grid Infrastructure and RAC

- **Database Clients connect to RAC Database**
  - Use Virtual IP (VIP) instead of RAC node Host Name/IP
  - Virtual IP (VIP) automatic failover by Oracle Clusterware without waiting for TCP/IP timeout
  - Application connections failed over to surviving nodes
    - DML will be rolled back and started over after reconnecting.

  **Transparent Application Failover (TAF):** Client side failover:
  - specify how to failover query
  - **Oracle Notification Services (ONS)** for notifying down event
  - **Fast Connect Failover (FCF)** for fast connection failover: database clients registered with **Fast Application Notification (FAN)**, database clients get notified the up and down event and react accordingly.
  - **Application Continuity (AC)** of Oracle 12c:
    - During the instance outage, automatically replay the transaction on another instance without the need for end-users and applications resubmitting the transaction.
Oracle 12c Grid Infrastructure and RAC

- The Oracle RAC Stack:
  - Oracle Grid Infrastructure: Clusterware + ASM
  - Oracle RAC coordinates and synchronize multiple DB instances through Cache Fusion technology
Oracle 12c Grid Infrastructure and RAC

Oracle 12c Clusterware
- Enable the communication between the cluster server
- Managing resources: ASM instances, DB instances, Virtual IPs, SCAN, etc
- Foundation for RAC database and HA features
  - Manage failover of Virtual IP to other node
  - Restarts failed Oracle processes
  - Manages node memberships & prevent the split-brain syndrome
- Installed into the same Grid Infrastructure home with ASM

Oracle 12c Clusterware Components
- Required shared storage and private interconnects between cluster nodes
- Technology stack
Oracle 12c Grid Infrastructure and RAC

- Clusterware/ASM Startup Sequences:
  - Oracle clusterware is started up automatically when OS:
    On Linux: OS init process $\rightarrow$ init.ohasd $\rightarrow$ multiple levels:

```
Level 0
- cssdmonitor
  - OHASD oraagent
    - cssdagent
  - OHASD oraclerootagent
  - cssdagent

Level 1
- ASM
  - mDNSD
  - GIPCD
  - EVMD
  - GPNPD
  - CRSD
    - CTSSD
      - Diskmon
  - OHASD oraclerootagent

Level 2
- CRSD orarootagent

Level 3
- CRSD orarootagent
- Network sources
- SCANIP
- Node VIP
- ACF Registry
- GNSVIP
- ASM instance
- Diskgroup
- DB Resource
- SCAN
- Listener
- Listener
- Services
- eONS
- ONS
- GNS
- GSD
```

Process on the High Availability Stack
Process on the Cluster Ready Service Stack
Resource managed by Cluster Ready Service
Oracle 12c Grid Infrastructure and RAC

- **Clusterware OCR and Voting disks**
  - Voting disk: stores cluster membership used by CSS
  - OCR stores information about clusterware resources
  - Multiplexed OCR, Odd numbers of Voting disks
  - Preferably to be stored in ASM

- **Oracle 12c RAC New Features Overview**
  - New features and enhancements focus on business continuity, high availability, scalability, agility, cost-effective workload management:
    - Oracle Flex Clusters for cluster scalability
    - Oracle Flex ASM for high availability scalability
    - Oracle RAC support for Oracle 12c Pluggable databases
    - What-If Command Evaluation
    - Public Networks for RAC: IPv6 Support Added
    - Application Continuity:
      - Cluster Health Monitor (CHM) Enhancements
      - Shared ASM Password File stored in ASM diskgroup
Oracle 12cR1 Flex Clusters

Oracle Flex Clusters Architecture

- Scalability limitation of the standard cluster
  - All nodes tightly-connected: $N \times (N-1)/2$ interconnect paths
  - All nodes directly connected to storage: total $N$ storage paths
  - Preventing the cluster to go beyond 100 nodes.
- Two two-layered hub-and-spoke topology:
  - Hub Node: interconnected and directly connected to the shared storage
  - Leaf Node: connected a Hub node, no storage connection is required
  - Scalability of Oracle 12cR1 RAC
    - 64 Hub Nodes
    - up to 2000 Hub nodes + Leaf Nodes
  - Each Leaf node has its dedicated Hub node to connect
Oracle 12cR1 Flex Clusters

- A Leaf node with access to shared storage can be changed to a Hub node
- A Standard cluster → A Flex Cluster, but can’t change back without a reconfiguration of a cluster
- All Hub nodes function in same way as the standard cluster node using Flex ASM.
- When you design the cluster initially, you may choose either one. If you are not sure, choose the standard Cluster as it can converted to Flex Cluster.

**Configuring Flex Clusters**
- Configuring Flex Cluster during the OUI : Advanced Installation

![Configuring Flex Clusters during the OUI](image-url)
Oracle 12cR1 Flex Clusters

- Configuring Flex Clusters
  - Requires a fixed GNS VIP address for the Flex Cluster

Single Client Access Name (SCAN) allows clients to use one name in connection strings to connect to the cluster as a whole. Client connect requests to the SCAN name can be handled by any cluster node.

- Cluster Name: knewrac
- SCAN Name: knewrac-scan.kcloud.dblab.com
- SCAN Port: 1521
- Configure GNS
  - Configure nodes Virtual IPs as assigned by the Dynamic Networks
  - Create a new GNS
    - GNS VIP Address: 172.16.150.9
    - GNS Sub Domain: kcloud.dblab.com
After Grid Infrastructure installation, Hub node status:

```bash
[grid@knewracn1 ~]$ crsctl get node role status
Node 'knewracn1' active role is 'hub'
[grid@knewracn1 ~]$ crsctl check crs
CRS-4638: Oracle High Availability Services is online
CRS-4537: Cluster Ready Services is online
CRS-4529: Cluster Synchronization Services is online
CRS-4533: Event Manager is online
[grid@knewracn1 ~]$ crsctl get node role status -all
Node 'knewracn1' active role is 'hub'
Node 'knewracn2' active role is 'hub'
Node 'knewracn3' active role is 'hub'
Node 'knewracn4' active role is 'hub'
Node 'knewracn6' active role is 'leaf'
Node 'knewracn5' active role is 'leaf'
Node 'knewracn7' active role is 'leaf'
Node 'knewracn8' active role is 'leaf'
```
Oracle 12cR1 Flex Clusters

After Grid Infrastructure installation, Leaf node:

```
$ crsctl get node role status
Node 'knewracn5' active role is 'leaf'
```

- **Changing a standard Cluster to Flex Cluster:**
  - **Prerequisites:** Add GNS service with a Fixed Virtual IP
    ```
    # srvctl add gns --vip <VIP_address> --domain <domain_name>
    ```
  - Enable Flex ASM (see next slide)
  - Convert to Flex Cluster
    ```
    # crsctl set cluster mode flex
    # crsctl stop crs
    # crsctl start crs --wait
    ```
  - Add Leaf listener: `$srvctl add listener --leaflistener --skip`
Oracle 12cR1 Flex ASM

- **Oracle Flex ASM Architecture**
  - Limitation of the standard ASM
    - Each node has an ASM instance which costs CPU/memory
    - Local ASM instance failure will cause DB instance failure

- Flex ASM is an option on Oracle 12c: enabled or disabled.
  - A small # of ASM instance. (default 3, specified by admin)
  - DB instances connects to any ASM instance (local/remote)
Oracle 12cR1 Flex ASM

- Two kinds of Oracle ASM configurations:
  - Local ASM clients connect to local ASM instance
  - Flex ASM clients connect to a remote ASM instance

- ASM network
  - Added for Flex ASM for communication between ASM clients and ASM
  - On Oracle 12cR1, share the network with cluster private interconnect

- Database instances to access ASM servers on different nodes:
  - Flex ASM uses password file authentication
  - ASM password is shared and stored in ASM Disk group

```
ASM> ls -1 orapwasm
Type Redund Striped Time  Sys Name
PASSW UNPROT COARSE SEP 11 10:00:00 M  orapwasm => +DATA1/ASM/PASSWORD/pwdasm.256.825849121
```
Oracle 12cR1 Flex ASM

- **Flex ASM and Flex Clusters**
  - Flex ASM is enabled if you choose Flex Cluster. ASM instance runs on Hub nodes only as Hub nodes have access to the shared storage.
  - Flex ASM can be enabled for standard Cluster. Only a subset of nodes run ASM instance.

- **Configure Flex ASM**
  - On GI OUI, select Flex ASM option or select Flex Cluster.
  - Need to specify ASM network.

- **Convert to standard ASM to Flex ASM**
  - Setup ASM network.
Oracle 12cR1 Flex ASM

- Convert ASM to Flex ASM with asmca tool such as
  $asmca -silent -convertToFlexASM -asmNetworks eth1/192.168.141.0 -asmListenerPort 1521
- Run converttoFlexASM.sh as the root on all nodes, one at a time

### Managing Flex ASM

- No specific tasks for Flex ASM management
- Use asmcmd and SRVCTL commands:

```
[grid@knewracn1 ~]$ asmcmd showclustermode
ASM cluster: Flex mode enabled
[grid@knewracn1 ~]$ srvctl status asm -detail
ASM is running on knewracn1,knewracn4,knewracn3
ASM is enabled.
[grid@knewracn1 ~]$ srvctl config asm
ASM home: /u01/app/12.1.0/grid
Password file: +DATA1/orapwASM
ASM listener: LISTENER
ASM instance count: 3
Cluster ASM listener: ASMNET1LSNR_ASM
```

- **Provide a better HA of database instances**
  - Database instance can connect to remote ASM instance if case the local ASM instance fails
Deploy Oracle Flex ASM: Architecture Options

- Oracle Flex ASM for new Oracle 12c Database Deployment:
Oracle 12cR1 Flex ASM

- Oracle Flex ASM for mixing Oracle pre-12c DBs +12c DBs Deployment:
  - Run ASM instance on every node. 12c DBs can failover to other ASM

- Run pre-12c DB on ASM nodes only, 12c DBs can failover to other ASM
Oracle RAC Support for Pluggable Databases

- **Pluggable Databases Architecture Overview**
  - A pluggable database (PDB) is a self-contained collection of schemas.
  - A container database (CDB) which is a superset of the pluggable databases, the root $CDB$ROOT, the seed named $PDB$SEED, Zero or more PDBs.

```sql
SQL> SELECT NAME, CON_ID, DBID, CON_UID FROM V$CONTAINERS ORDER BY CON_ID;
```

<table>
<thead>
<tr>
<th>NAME</th>
<th>CON_ID</th>
<th>DBID</th>
<th>CON_UID</th>
</tr>
</thead>
<tbody>
<tr>
<td>$CDB$ROOT</td>
<td>1</td>
<td>1776735948</td>
<td>1</td>
</tr>
<tr>
<td>$PDB$SEED</td>
<td>2</td>
<td>4057735188</td>
<td>4057735188</td>
</tr>
<tr>
<td>PDB1</td>
<td>3</td>
<td>3322827582</td>
<td>3322827582</td>
</tr>
<tr>
<td>PDB2</td>
<td>4</td>
<td>3881372142</td>
<td>3881372142</td>
</tr>
<tr>
<td>PDB3</td>
<td>5</td>
<td>3931156879</td>
<td>3931156879</td>
</tr>
</tbody>
</table>
Oracle RAC Support for Pluggable Databases

- How 12c Pluggable Database works on Oracle 12c RAC
  - Different open modes of a PDB: Mounted; Read only; Read and Write
    check the Open modes of all the PDBs on this RAC instance when you connect to the CDB root:

```
SELECT NAME, OPEN_MODE, RESTRICTED FROM V$PDBS;
NAME     OPEN_MODE       RESTRICTED
--------- -------------- --------------
PDB$SEED  READ ONLY     NO
PDB1      READ WRITE     NO
PDB2      READ WRITE     NO
PDB3      Mounted        NO
```

- Startup a PDB in sqlplus

```
STARTUP OPEN
STARTUP OPEN READ ONLY
STARTUP RESTRICT OPEN READ ONLY
SHUTDOWN IMMEDIATE
ALTER PLUGGABLE_DATABASE OPEN READ ONLY
```
Oracle RAC Support for Pluggable Databases

- How 12c Pluggable Database works on Oracle 12c RAC
Consolidate Database using PDB

- **How 12c Pluggable Database works on Oracle 12c RAC**
  - Create a PDB
  - Create *dynamic* database service for PDB

```
$ srvctl add service -db cdb -service hr1 -pdb pdb1 -preferred host1  -available host2
```

- Connect PDB through service

```
HR_PDB1 =
(DESCRIPTION =
 (ADDRESS = (PROTOCOL = TCP)(HOST = knewracscan.kcloud.dblab.com)
  (PORT = 1521))
 (CONNECT_DATA =
  (SERVER = DEDICATED)
  (SERVICE_NAME = hr1.kcloud.dblab.com))
)
```
Oracle RAC Troubleshooting and Health Check

- Clusterware health check and troubleshooting
  - Clusterware utility: `crsctl` for check `crs` status and start/stop
    `crsctl check cluster --all; crsctl stat --ret -t`
  - Log files: `$GRID_HOME/log/<host>/alert<host>.log` and
    `$GRID_HOME/log/<host>/<process>/log`
- **Clusterware health Verification Utility: CLUFY**
  - Verifies Clusterware, RAC best practices, mandatory requirements: 
    $./cluvfy$ comp healthcheck –collect cluster –bestpractice –html

### Detailed report for Best Practices checks

#### Summary of environment

<table>
<thead>
<tr>
<th>Date (mm/dd/yyyy)</th>
<th>10/14/2013</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time (hh:mm:ss)</td>
<td>04:34:34</td>
</tr>
<tr>
<td>Cluster name</td>
<td>knwrac</td>
</tr>
<tr>
<td>Clusterware version</td>
<td>12.1.0.0.0</td>
</tr>
<tr>
<td>Grid home</td>
<td>/u01/app/12.1.0.0/grid</td>
</tr>
<tr>
<td>Grid User</td>
<td>grid</td>
</tr>
<tr>
<td>Operating system</td>
<td>Linux2.6.39-400.17.1.el6uek.x86_64</td>
</tr>
</tbody>
</table>

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

1. System recommendations
2. Clusterware recommendations

### System recommendations

<table>
<thead>
<tr>
<th>Verification Check</th>
<th>Verification Result</th>
<th>Verification Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hardware Clock synchronization at shutdown</td>
<td>PASSED</td>
<td>Checks whether Hardware Clock is synchronized with the system clock during system shutdown</td>
</tr>
<tr>
<td>availability of port 8888</td>
<td>PASSED</td>
<td>availability of port 8888</td>
</tr>
<tr>
<td>Reverse path filter setting</td>
<td>PASSED</td>
<td>Checks if reverse path filter setting for all private interconnect network interfaces is correct... details</td>
</tr>
</tbody>
</table>

### Clusterware recommendations

<table>
<thead>
<tr>
<th>Verification Check</th>
<th>Verification Result</th>
<th>Verification Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSS misscount parameter</td>
<td>PASSED</td>
<td>Checks if the CSS misscount is set correctly on the system... details</td>
</tr>
<tr>
<td>CSS reboottime parameter</td>
<td>PASSED</td>
<td>Checks if the CSS reboottime is set correctly on the system... details</td>
</tr>
<tr>
<td>CSS disktimeout parameter</td>
<td>PASSED</td>
<td>Checks if the CSS disktimeout is set correctly on the system... details</td>
</tr>
</tbody>
</table>
Oracle RAC Troubleshooting and health Check

- **Oracle RACcheck**: A RAC Configuration Audit tool
  - audit the configurations settings for RAC, Clusterware and ASM
  - MOS note ID 1268927.1, download the tool
  - To invoke: ./raccheck; it will produce an auditing report

- **CHM**: detect/analyze OS and Cluster’s resource related degradations and failure.
  - A set of tools tracing OS resource consumptions
  - Enhanced in Oracle 12cR1 and consists three components:
    - **osysmond**: System Monitor Service process on each node monitor and collects real time OS metric data and send to Ologgerd
    - **ologgerd**: cluster logger service, one for each 32 node
    - Grid Infrastructure Management Repository (the CHM repository): central repository to store metrics data

- **Grid Infrastructure Management Repository (GIMR)**
  - A single instance Oracle database run by grid user
  - Installed on one of the cluster nodes
  - Need to select Advanced Installation option
  - run on the same node that runs the ologgerd service to reduce the traffic
Oracle RAC Troubleshooting and health Check

– by default the database is stored in same location of OCR/Voting disks

$ oclumon manage -get repsize reppath allllogger -details

CHM Repository Path = +DATA1/_MGMTDB/DATAFILE/sysmgmtdata.260.807876429
CHM Repository Size = 38940
Logger = knewracn1
Nodes = knewracn1,knewracn2,knewracn4,knewracn7,knewracn5,knewracn8,knewracn6

– Use OCLUMON to manage the size and retention of the repository

- **diagcollecton.pl** to collect the CHM data
  - Get the master node: $ oclumon manage -get master
  - login to the master node as root
  - run the command: diagcollection.pl -collect -crshome <CRS_HOME>
  - It produces four .gz files which include various log files for diagnosis.

- **use OCLUMON query the CHM repository for node specific data**
  $ oclumon dumpnodeview -allnodes -v -s "begin_timestamp" -e "end_timestamp"
Node Eviction:
Cluster split brain condition: a node failure partitions the cluster into multiple sub-clusters without knowledge of the existence of others.

Possible causes:
- not responding network heartbeat, disk heartbeat, a hung node or hung ocssd.bin process
- Consequence: data collision and corruption
- IO fencing: fencing the failed node off from all the IOs: STOMITH (Shoot The Other Machine In The Head) algorithm
- Node eviction: pick a cluster node as victim to reboot.
  Always keep the largest cluster possible up, evicted other nodes two nodes: keep the lowest number node up and evict other
- Two CSS heartbeats and misscounts to detect node eviction
  1. Network HeartBeat (NHB) over private interconnect to check node membership; misscount: 30 secs
  2. Disk heartbeat: between the cluster node and voting disk: misscount: 200 secs
Oracle RAC Troubleshooting and health Check

- Troubleshooting node eviction
  - Common causes for OCSSD eviction:
    - network failure latency exceeds CSS miscount 30 seconds
    - access disk issue: CSS miscount 200 sec OCSSD failure,
  - Common causes of : CSSDAGENT OR CSSDMONITOR eviction: OS scheduler problem caused by OS locked up in driver or hardware or the heavy loads; thread of CSS demon hung
  - Review the log files, refer to metalink note [1050693.1]

- Node Eviction Diagnosis Examples
  - Case 1 : Node 2 was rebooted in a 2-node cluster on Linux: OCSSD log:
    $CRS_HOME/log/<hostname>/cssd/ocssd.log file in Node1:

```
root@k4r815n1:/opt/app/11.2.0/grid/log/k4r815n1/cssd
2010-11-23 17:11:55.221: [ CSSD][1342572864]clssnmPollingThread: node k4r815n2 (2) at 75% heartbeat failed, removal in 7.500 seconds
2010-11-23 17:11:59.231: [ CSSD][1353062720]clssnmSendingThread: sending status msg to all nodes
2010-11-23 17:11:59.231: [ CSSD][1353062720]clssnmSendingThread: sent 5 status msgs to all nodes
2010-11-23 17:12:00.232: [ CSSD][1342572864]clssnmPollingThread: node k4r815n2 (2) at 90% heartbeat failed, removal in 2.490 seconds, seedhbimpd 1
2010-11-23 17:12:02.718: [ CSSD][1342572864]clssnmPollingThread: Removal started for node k4r815n2 (2), flags 0x3040c, state 3, wt4c 0
2010-11-23 17:12:02.718: [ CSSD][1342572864]clssnmDiscHelper: k4r815n2, node(2) connection failed, endp (0x264), probe (0x10000000), ninf->endp 0x264
2010-11-23 17:12:02.718: [ CSSD][1342572864]clssnmDiscHelper: node 2 clean up, endp 0x264, init state 5, cur state 5
```
Oracle RAC Troubleshooting and health Check

- Case 2: node 1 reboot: $CRS_HOME/log/<hostname>/cssd/ocssd.log file

```
[cssd(16288)] CRS-1604: CSSD voting file is offline. ORCL:OCR2; details at ("/CSSNM00058:"") in /opt/app/11.2.0/grid/log/k4r815n1/cssd/ocssd.log.
2010-11-23 18:10:15.898
[cssd(16288)] CRS-1606: The number of voting files available, 0, is less than the minimum number of voting files required, 3, resulting in CSSD termination to ensure data integrity; details at ("/CSSNM00018:") in /opt/app/11.2.0/grid/log/k4r815n1/cssd/ocssd.log
2010-11-23 18:23:52.704
[ohasd(9966)] CRS-2112: The OLR service started on node k4r815n1.
2010-11-23 18:23:53.424
[ohasd(9966)] CRS-8017: location: /etc/oracle/lastgasp has 56 reboot advisory log files, 0 were announced and 0 errors occurred
```

- Case 3: One node rebooted once a month in 11 nodes cluster:

```
------- /var/log/message:
Jul 23 11:15:23 racdb7 logger: Oracle CRS failure. Rebooting for cluster integrity
------- OCSSD log: $CRS_HOME/log/<hostname>/cssd/ocssd.log file
    clssnmPollingThread: node racdb7 (7) at 90% heartbeat fatal, eviction in 0.550 seconds
...
    clssnmDoSyncUpdate: Terminating node 7, racdb7, misstime(60200) state(3)
```
Contact me at kai_yu@dell.com or visit my Oracle Blog at http://kyuoracleblog.wordpress.com/

I am a member of

Presenting at the First Oracle Technology Network (OTN) MENA Conference Tour  Leave a comment

I will give an Oracle technology presentation at the first OTN MENA (Middle East & North Africa) tour May 27, 2014. This presentation will be given online which can be attended remotely anywhere in the world. The following is detailed information about this presentation:

Title: Oracle 12c RAC Flex Architecture and support for RAC 12c Multitenant Architecture
Speaker: Kai Yu
Time: May 27th 2:35pm – 3:25pm Tunis time, which is 8:35-m-9:35 am US central time.
Web link to attend the presentation remotely: https://www1.gotomeeting.com/register/439651800
Abstract:
Oracle 12c RAC Flex Architecture and support for RAC 12c Multitenant Architecture
High Availability and Scalability become increasingly important for IT infrastructure as applications and database consolidation has become a key initiative in the data center. The Latest Oracle 12c RAC provide the right solution to achieve these goals with Oracle12c RAC and the multitenant architecture. Come to this session to learn these new features from the author of the Oracle 12c New feature chapter of the Apress book Expert Oracle RAC 12c.