Database as a Service (DBaaS) Cloud: HA Architecture and Consolidation Methods
About Author

Kai Yu, Senior Principal Architect, Dell Database Engineering

- 20 years Oracle DBA/Apps DBAS and Solutions Engineering
- Specializing in Oracle RAC/DB, 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”
My Work: Dell Oracle Solutions Engineering

- Flash SSD based high performance storage solution for database
  - Remove storage bottlenecks for Oracle Databases.
  - Shared storage for Oracle RAC database
My Work: Dell Oracle Solutions Engineering

- Preconfigured Ready Infrastructure with all flash storage for Oracle Database and DBaaS.
Agenda

- Database Consolidation for Database Cloud
- High Availability for Database Cloud
- Database Cloud Architecture Design
- High Availability Infrastructure Configuration
- Achieving HA through Oracle RAC
- Oracle RAC Troubleshooting and Health Check
Database Consolidation

• The Journey to Database Cloud
  – Challenges to the traditional computing architecture
  – Consolidate multiples databases: multitenant architecture
  – Integrate all the resources to allow provisioning on demand: dynamically provisioning to meet the workload needs
  – Provide Database as a Service (DBaaS)/ Platform as a Service (PaaS)
  – Provide Measuring and Charge back service
Database Cloud Consolidation Methods

- **Database Consolidation Models**
  - Multitenant Architecture for Database Consolidations consolidating multiple databases on a shared infrastructure
  - Multiple Levels of consolidation and abstraction

- **Multiple schemas in a database**
- **Multiple PDBs in a container database**
- **Multiple database instance running on an OS**
- **Multiple VMs on a physical server (virtualization)**
- **Multiple databases/apps shared a storage**
Considerations for selecting consolidation models.

- Level of isolation between the databases: security, HA, manageability, etc
- Efficiency: Resource utilizations and admin cost
- Different models work together
Database Consolidation Models

- **Example: consolidating multiple databases on a cluster**
  - Multiple databases (Single/RAC/ RAC one node) on a single cluster
  - Possible multiple versions Oracle Homes on 11g R2 Clusterware
  - A example that consolidates 100 Oracle EBS Databases
Database Consolidation Models

- **Example: consolidate database with Oracle VM**
  - Consolation of multiple RAC databases in fewer physical servers
  - Each database instance runs on its own VM independently
  - One database instance node eviction will not impact other databases
  - Less impact of downtime during OS and Oracle software upgrade
  - Easy provisioning with Oracle VM templates
  - Only need to pay for the licenses based on virtual CPUs
  - Overhead on OS, database Instances and extra maintenance costs
Database Consolidation Models

- Multitenant Architecture with Oracle 12c Pluggable Databases

Diagram showing the multitenant architecture with Oracle 12c Pluggable Databases, featuring CDBs, PDBs, and shared storage for multiple applications.
Database Consolidation Models

- Architecture Comparison: Isolation vs Efficiency

Oracle EBS APPs tier
16 Nodes RAC
Each RAC database serves an EBS Instance

Oracle ASM diskgroups for RAC Databases

EqualLogic PS6000XV Storage Arrays

Oracle 12c R1 Clusterware/ASM
Multiple Apps
Oracle ASM diskgroups for RAC Databases

Oracle 12c RAC
Single Database instance per Node

Shared Storage
Oracle Public Cloud DBaaS Offerings:

- **Database Schema as a Service**
  - 1 schema on Oracle Database 11g: 5/10/50GB size
  - SQL & PL/SQL, APEX for development, RESTful web service
  - no SQL*Net, no DBA access, fully managed by Oracle,

- **Database as a Service**
  - Dedicated Virtual machine for running Oracle instance 11g/12c
  - Preinstalled Oracle software, wizard to create oracle database
  - OS root and SYSDBA access, full SQL*Net access
  - Automated backup and point-in-time Recovery

- **Exadata Service**
  - Full featured 11.2.0.4 and Oracle 12.1.0.2 on Exadata
  - Choose Quarter/Half/Full RACK Exadata
  - Offer root /SYSDBA, but Server, storage, network managed by Oracle
  - Support OLTP/Analytic/mixed workload at any scale.
  - Secure network access, customer database run on separate VM
  - Backup & Recovery to Exadata or Oracle Database backup service
  - Server, storage, network managed by Oracle
  - Data encrypted by default in your tablespace
High Availability for Database Cloud

- **Database High Availability Requirement**
  - Defined by Service Level Agreement (SLAs):
  - HA goal is to meet SLA requirement
  - Balance between the availability and implementation cost
  - SLA: for example, 99.95%, annual 4 hrs 22 minutes downtime
    Downtime window: first Saturday: 8pm-10pm every quarter

- **High Availability Challenges in Cloud Environment**
  - Consolidating many databases in a single cloud infrastructure
  - Great business impact due to the infrastructure downtime
  - Databases may have different SLAs for different business:
    • Different business requirements
    • Different time zones
    • Infrastructure downtime means downtime for all the databases
    • Very difficult to find the downtime for maintenance that meets every SLA
High Availability for Database Cloud

- **Causes of Impacting System Availability**
  - Service outage by unplanned downtime:
    - hardware or software failure, human error, nature disaster, etc.
  - Service disruption by planned downtime:
    - hardware/software upgrade, patching and migration from old system to new system
  - Service performance degradation: violate performance SLA for example, 99% transactions finished in a 2 seconds window

- **Architect a High Availability Cloud Infrastructure**
  - Architect a highly available cloud architecture
  - Configure HA infrastructure to reduce unplanned outage
  - Implement methods/options to minimize planned downtime
  - Use configuration and implementation best practices for HA
  - Administration and troubleshooting tips for High Availability
  - Establish the pre-active real time monitoring system
Database Cloud Architecture Design

- **Oracle Real Application Clusters: active-active cluster database**
  - Based on share everything clustering architecture
  - 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 and scalability architecture
Database Cloud Architecture Design

- **Basic Software Components:**
  - Grid Infrastructure: Oracle Clusterware and ASM
  - Oracle RAC: Cache Fusion technology

- **Hardware requirements:**
  - Shared storage
  - Private Interconnect network for cluster heartbeat and RAC Cache
Database Cloud Architecture Design

**Oracle 12c 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 of 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
  - 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
Database Cloud Architecture Design

- **Oracle RAC one node database: active-passive database**
  - Single node database on Clusterware; no load balancing
  - Require the same system architecture as RAC: network/storage
  - Install Oracle Grid Infrastructure and RAC on all nodes
  - Specify RAC One node Database during the database creation
Database Cloud Architecture Design

- Protect database against server failure with VIP automatic failover
- Reduce planned downtime: Online relocating database to another node for hardware, OS, database software upgrade
  
  ```
srvctl relocate database -d <dbname> -n <nodename> -w -15
  ```

- RAC and RAC one node can be switched back and forth
  
  Convert to RAC from RAC one node
  
  ```
srvctl convert database -d <dbname> -c RAC -n <nodename>
srvctl add instance -d <dbname> -i <instname> -n <nodename>
  ```

  Convert to RAC one node from RAC
  
  ```
srvctl remove instance -d <dbname> -i <instname>
srvctl convert database -d <dbname> -c RAC -n <hostname>
  ```

- Why RAC one node:
  
  Low cost alternative to RAC: $10k/p vs $23k/p (EE)
  
  two nodes 2 X8 core servers, $80K vs. $512K(EE)
Database Cloud Architecture Design

- Oracle VM provides HA against physical server failure
- Databases run on virtual machine
- High isolation: each database instance runs on a dedicated OS
- Flexible for maintenance windows
- Lower efficiency: extra overhead on OS, database instance
- Virtual machines run on a pool of VM servers (VM server pool)
- Virtual machines can be failed over or migrate to different physical server to reduce planned/unplanned downtime
High Availability Infrastructure Configuration

- **High Availability Storage Infrastructure**
  - Storage HA plays a key role in the infrastructure HA
  - Redundant Storage Controllers for high availability
  - Oracle ASM diskgroup redundancy settings
  - Redundant IO paths from servers to storage array
High Availability Infrastructure Configuration

- Redundant IO paths from servers to storage array

  Multiple physical paths: (DB-HBA#, Switch#, Controller-HBA#, Controller#),
  DB-HBA#: =1,2; Switch# = 1,2; Controller-HBA#=1,2,3,4, Controller# =1,2
  A storage access should be able to fail over to another path

```
[root@ibdbnode1 ~]# multipath -ll
ibond_v_d2 (26363646436333335) dm-13 FUSIONIO,ION LUN
size=1.5T features='3 queue_if_no_path pg_init_retries 50' hwhandler='1 alua' wp=rw
|-- policy='queue-length 0' prio=100 status=active
  |-- 45:0:0:7 sdbm 68:0  active ready running
  |-- 46:0:0:7 sdcc 69:0  active ready running
  |-- 48:0:0:7 sdck 69:128 active ready running
  |-- 49:0:0:7 sdcz 70:112 active ready running
  |-- 47:0:0:7 sdda 70:128 active ready running
  |-- 50:0:0:7 sddi 71:0  active ready running
  |-- 51:0:0:7 sddq 71:128 active ready running
  `-- 52:0:0:7 sdy 128:0  active ready running
    `-- policy='queue-length 0' prio=1 status=enabled
       |-- 39:0:0:7 sdx 65:128 active ready running
       |-- 37:0:0:7 sdi 65:128 active ready running
       |-- 38:0:0:7 sdx 65:112 active ready running
       |-- 41:0:0:7 sdao 66:128 active ready running
       |-- 77:0:0:7 sdan 66:112 active ready running
       |-- 78:0:0:7 sdaw 67:0  active ready running
       |-- 79:0:0:7 sdbf 67:128 active ready running
          `-- 80:0:0:7 sdbu 68:128 active ready running
```
Software multipathing: to group the multipath

Linux Device Mapper (DM) or from storage vendors

Group them together to alias (ibion1_v_d2) using multipathing software

Linux Configuration: /etc/multipath.conf

```bash
multipath {
    wwid 2636364643633335
    alias ibion1_v_d2
}
```

#service multipathd restart

– Oracle ASM diskgroup redundancy settings
– SAN Disk Array RAID for Redundancy: Raid 10/5 Configuration
High Availability Infrastructure Configuration

- Network High Availability Configuration
  - fully redundant interconnects for cluster configuration
  - Network bonding vs Oracle Highly Available Virtual IP (HAIP)
  - Dedicated switches for private interconnects

- Redundant Hardware Infrastructure for Cluster Database
High Availability Infrastructure Configuration

VM Server Configuration

RAC nodes on VMs:
Achieving HA through Oracle RAC
Achieving HA through Oracle RAC

- **Preventing Unplanned downtime through RAC**
  - 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. Failover works for most database clients: JDBC, OCI, UCP, etc

- **Application Continuity (AC) of Oracle 12:**
  During the instance outage, automatically replay the transaction on another instance without the need for end-users and applications resubmitting the transaction.
Achieving HA through Oracle RAC

- **Reducing planned downtime for upgrade**
  - Online patches can be applied to a running database. The patches contain a single shared library, and do not require shutting down the instance or relinking the Oracle binary to apply the patches.
  - $ opatch query -all online

..... "Patch is a rolling patch: true."

Rollback all the online patch and reapply the offline version at next downtime.

Refer MOS note: 761111.1

- Rolling upgrade: one instance is shutdown down for upgrade while other are functioning:
  - rolling upgrade of server hardware, firmware, BIOS, OS
  - rolling upgrade of Oracle Clusterware and ASM
  - rolling upgrade of RAC: database patch, PSU, CPU patches

- Rolling upgrade process in RAC:
  - Relocate the database connections to the other instance
  - Shutdown (DB/OS/Server), upgrade, restart
  - Repeat on other nodes.
Achieving HA through Oracle RAC

– Rolling upgrade requirements Oracle RAC
  . Local Oracle Home; Rolling upgradable patch:
    $opatch query --all <Patch_location> | grep rolling
    .... “Patch is a rolling patch: true.”
– Rolling database Upgrade to reduce downtime:
  • Use Data Guard SQL apply and transient Logical Standby feature
  • MOS note #949322.1

1. Start with physical standby
2. Convert to logical standby and queued all redo data
3. Upgrade the logical standby
4. Apply all redo data to Standby
5. Switch over roles
6. Upgrade the old production (the new standby)
7. Apply all redo data to the new standby (old production)
8. Switch over roles
9. Convert back to physical standby
Oracle RAC Troubleshooting
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: $GIRD_Home/log/<host>/alert<host>.log and $GIRD_Home/log/<host>/<process>/log

- 12c CRS log: Log files: $GIRD_Home/diag/crs/<host>/crs/trace/alert.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>12.1.0.1.0</td>
</tr>
<tr>
<td>Grid home</td>
<td>/u01/app/12.1.0/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

| Verification Check               | Verification Result | Verification Description                                                                 |
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 alllogger -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"
Oracle RAC Troubleshooting and health Check

- **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 misscount 30 seconds
  - access disk issue: CSS misscount 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:

```
2010-11-23 17:11:55.221: [ CSSD] [1342572864] clssmmPollingThread: node k4r815n2 (2) at 75% heartbeat failed, removal in 7.500 seconds
2010-11-23 17:11:59.231: [ CSSD] [1353062720] clssmmSendingThread: sending status msg to all nodes
2010-11-23 17:11:59.231: [ CSSD] [1353062720] clssmmSendingThread: sent 5 status msgs to all nodes
2010-11-23 17:12:00.232: [ CSSD] [1342572864] clssmmPollingThread: node k4r815n2 (2) at 90% heartbeat failed, removal in 2.490 seconds, seedhbinmpd 1
2010-11-23 17:12:02.718: [ CSSD] [1342572864] clssmmPollingThread: Removal started for node k4r815n2 (2), flags 0x3040c, state 3, wt4c 0
2010-11-23 17:12:02.718: [ CSSD] [1342572864] clssmnDiscHelper: k4r815n2, node(2) connection failed, endp (0x264), probe(0xa0000000), ninf->endp 0x264
2010-11-23 17:12:02.718: [ CSSD] [1342572864] clssmnDiscHelper: 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

- 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)
Thank You and QA
Contact me at kai_yu@dell.com or visit my Oracle Blog at http://kyuoracleblog.wordpress.com/