ACHIEVING MAINFRAME RELIABILITY USING ORACLE RAC AND COMMODITY HARDWARE

Steele G. Arbeeny, PhD
GL Associates
arbeeny@glassoc.com

Agenda

- Introduction to clustering
- Components of a RAC system
- In-depth review of each component
- Real-world case study showing reduced TCO
- Questions

What is a cluster ?

- A cluster is a interconnected set of computers that can function as one.
- Even though the cluster is made up of distinct computers, it typically appears to external systems as a single computer.
- The computers are connected to each other using a private cluster interconnect.

Types of Clusters

- **Clustering provides the following benefits:**
  - ***High Availability (HA)***
    - Important services are replicated on multiple machines
    - Single points of failure should be eliminated on all cluster components.
    - If one component fails, the redundant components take over seamlessly.
    - 2 of each cluster component is required (at a minimum) to provide HA.

Types of Clusters

- **Clustering provides the following benefits:**
  - ***Load Balancing***
    - Incoming requests can be routed to different machines in the cluster based on a load balancing algorithm
    - Load balancing algorithms can be simple such as round robin or complex, and balance based on server load or request "state".
    - This type of cluster will typically also be highly available.
    - Allows for horizontal scaling where many nodes can be added to increase capacity.

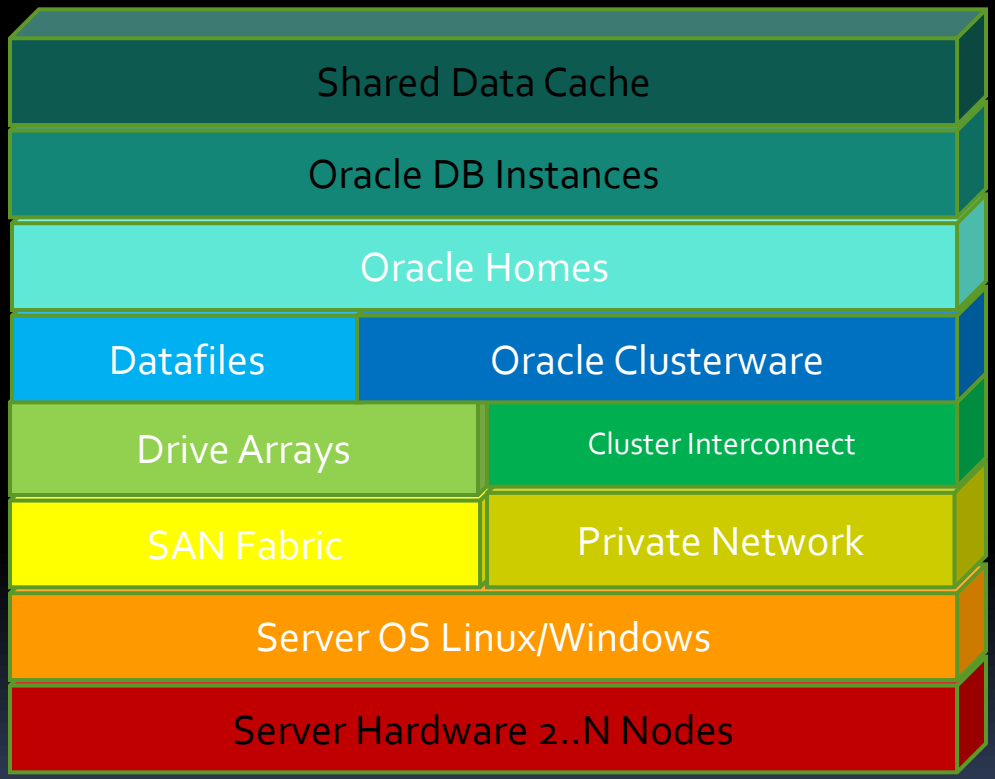
Types of Clusters

- **Clustering provides the following benefits:**
  - ***Compute Clusters***
    - Designed for processor intensive operations such as scientific applications
    - The program is usually tightly coupled to the architecture.
    - Could employ any or all of the following architectures:
      - SIMD – Single Instruction Multiple Data
      - MIMD – Multiple Instruction Multiple Data
      - MISD – Multiple Instruction Single Data
    - May require high bandwidth cluster interconnects to achieve unified memory. Ie NUMA, InfiniBand

Oracle RAC

- ***Just so you know…***
  - RAC is designed to provide HA and horizontal scaling
  - Clients connect to a VIP which is balanced to a real node
  - From that point on, the user is connected to a single cluster node and a single instance of Oracle.
  - Queries executed on that connection run on a single server as if RAC was not present.

RAC Cluster Components



- It takes many layers to bake this cake !
- Seems like we do it all for the frosting on the top !

Let's look layer by layer from the bottom up.

RAC Server Hardware

- Reliability starts with node hardware selection. Choose a enterprise class hardware vendor and machines with a high level of redundancy built in.

Redundant Hot-Swappable Power Supplies	Multiple Hot-Swappable Drive Bays
Lots of Memory and Processor Slots	On-board RAID1 for the OS disk and Oracle SW
Adequate hot swappable cooling	Redundant teaming NICs and Storage Adapters

- *Tip:* Always use the same hardware, accessories and software versions on each cluster node. It will make your support much easier.
- Eliminate single points of failure.

The RAC OS

- ***Get ready for a fight !***
  - **Windows and Linux both work well**
  - **For large deployments use only 64bit**
  - **I have done both, but personally prefer Linux for several reasons**
    - **Cheaper Licensing**
    - **Easier to Automatically Administer**
    - **More "customizable" and "tunable" at the OS level**
  - **To be fair, there are a lot more mistakes to make though.**
  - **Choose what you are comfortable with. Both do the job well.**

RAC Storage

- Choose a storage system that itself is highly available and have multiple redundant paths to the storage. Ie: HBAs, Switches and RAID.

- Tip: Install the OS, swap file, Clusterware and Oracle DB HOMES on locally attached RAID1 storage instead of out on the SAN. At least you'll be able to boot your nodes if you have SAN or fabric issues.

RAC Storage Options

- Clusterware requires an OCR and Voting disk
- These can be stored on:
  - NFS (Network File System) (Not DNFS)
  - OCFS (Oracle Cluster File System)
- Other obscure options are available but why make your life miserable.
- Pick OCFS if you are using a FC SAN for the data, pick NFS if you will access the data via NFS
- The storage required is very small, just a few GB. Use external redundancy.

RAC Storage Options

- Oracle Datafiles can be stored several ways too.
- ***ASM*** – Automatic Storage Management – This is your only option if you use Oracle DB SE.
- ***NFS/DNFS*** – Requires EE. All data and log storage is on an IP SAN.

  (There are some other options but they are out on the fringe)
- ***Lets look at each one in a little more detail.***

ASM

- ASM has been around a long time
- Uses a stripped down instance of Oracle DB to manage the storage
- The ASM instance must start before Oracle can access the data files.
- OCR and Voting disks cannot be on ASM because for ASM to start, Clusterware must already be started.
- External redundancy is recommended but RAID groups can be created in ASM if needed. Its simpler to manage disk failure on the SAN only.
- If you are using a FC SAN or SE choose ASM

ASM

- Create large RAID disk aggregates on your SAN
- Carve them up into smaller LUNS that you can present to the hosts
- Popular SAN vendors publish a RAC best practice guide to detail how to configure your LUNS and SAN for RAC
- The optimal size for an ASM disk is usually stated in this document

ASM

Raw vs Cooked File Systems

- ASM Requires RAW devices to be enrolled as disks.
- You should experiment with partitioned vs unpartitioned disks or look at your SAN vendor's recommendation.
- Disks are then lumped together into diskgroups
- Data files are placed on the diskgroups
- Create separate DGs for DATA, ARCH, REDO and CONTROL
- Disks can be added or removed from the diskgroups and ASM will rebalance the files across the remaining disks.
- You can use the EM GUI for all of this.

A Note About Multipath

- Since our target architecture contains multiple HBAs that provide redundant paths to the SAN, some form of multipathing is required.
- There are both Linux and Windows versions of multipath, as well as versions from the SAN vendor.
- There are a lot of multipath configuration options that are unique to each SAN/HBA/OS combination.
- Typically the SAN vendor publishes a guide on how to setup multipath. Get It. Use It.
- Tip: Leave extra time in your project to experiment with different multipath setups to see which gets the best throughput.

NFS

- All modern SANs offer an NFS option
- SAN controllers can have multiple high bandwidth GiG-E ports for NFS access
- The SAN controller manages the storage and how the files are laid out
- This can lead to an all-IP solution.
- Stay away from server-hosted NFS because it reduces reliability unless you want to do another clustering project over there.
- Still remember to eliminate single points of failure with multiple NICs, Switches and RAID.

NFS

- NFS drives can be mounted on the nodes OR you can use Oracle DNFS where a NFS client is included inside Oracle DB.
- This bypasses the OS kernel for Oracle I/O
- This can give a 20 – 30% improvement over K-NFS when using multiple bonded NICs. No need for multipath !
- While ASM currently has the most installs, DNFS is winning out in many NEW installs.

The Cluster Interconnect

- This is a private network that connects the nodes to each other.
- Put this on its own switch with no other traffic or its own VLAN if that is not possible
- Heartbeat messages are sent across this link
- It is also used to provide a single data cache to all cluster nodes using GCF – Global Cache Fusion.

The Cluster Interconnect

- When a data block is requested, Oracle will first check to see if it is in the local DB cache, it will then check the other nodes to see if it is in their local DB cache.
- If it is found in the cache of another node, it is pulled over the interconnect
- Only if it is not found on any node will it be read from disk.

The Cluster Interconnect

- We have just exposed the weakest link in RAC
- When reading a large table on node 1, if it is in the cache on node 2, it will be pulled over the interconnect. If the interconnect is several times slower than the disk system, performance will be terrible.
- Tip: This is why RAC is not ideally suited for data warehouse applications

The Cluster Interconnect

- Tip: Use at a bare minimum switched Gig-E !
- Tip: 10 Gig-E is even better
- Tip: For clusters with many nodes or data warehouse applications consider Infiniband.
- Tip: Redundant NICs and switches are critical here also.
- If the interconnect goes down, the whole cluster goes down.
- Our target architecture uses as much RAM as possible, this means the interconnect will get most of the data traffic.

Software Installation

- Tip: OS, Oracle program files and drivers should all be the same version and patch levels and stored on the local disk.
- Tip: Only data files should go on the SAN
- You can have multiple instances and even multiple Oracle HOMEs on a single clusterware/ASM install.
- Not all DBs must be running on all cluster nodes. Single node instances are also possible.

Software Install

To spfile or not to spfile…that is the question…

- **Either way is OK. Use what you prefer.**
- **I prefer pfiles on the local storage of each node when I am doing initial setup and tuning.**
- **Once the config stabilizes I move to a single spfile out on ASM or NFS**

Archived Redo Logs

- These can be placed on ASM and/or other storage
- Each instance writes its own archived log.
- Write them initially to your primary cluster storage, either ASM or NFS.
- Tip: Set up alternate locations if the primary fills up. You can also send drop copies to NAS.

Backup

- RMAN still works the same way as always.
- You can do hot or cold backups on your normal schedule.
- Also get the archived logs and back them up and clear out the log destinations.
- SAN vendors also make Oracle specific backup solutions.

Backup

- SAN Vendor Specific Backup Solutions
  - These allow duplicate instances to be setup and only the differences between the actual data and the copy are stored.
  - This is also great for table level restore and test copies of the DB
  - They can do snap-backups of the data files in a lot less time than it takes RMAN to copy the files.
  - They also have a GUI for managing the snapshots.

Common Objections

- **My staff is already trained in our current platform…this change will be too big.**
  - When compared with the cost of a hardware upgrade and continued maintenance costs, the cost of training is about 1% of that
  - Only sys admins need to be retrained
  - Developers will not notice a difference
  - Cost of iSeries/pSeries upgrade $1,000,000
  - Linux Training $10,000 per employee

Common Objections

- **You cant compare the performance and reliability of our current platform with PCs.**
  - These are not your home PCs ! They are 64-Bit systems which can have > 10, > 20, > 30 processor cores with hundreds of GB of RAM
  - Individual nodes can be clustered to increase performance and reliability beyond that possible with any single machine.
  - Cost per MFLOP is about 5% that of the "big iron".
  - SANs, snap backup and DR are much cheaper too

Case Study: Faster Performance on Low Cost hardware

Client Profile

- Three iSeries JD Edwards database instances
- Large data volumes
- JD Edwards a 7.3 and 8.12

GLA Solution

- Two 24-Core Intel Nehalem Servers, 256GB RAM
- Linux OS & Oracle 11g - All 64-bit
- One JD Edwards 9.0 instance

Client's Bottom Line

70% faster screen refresh, package builds from 5hrs to 1hr, 8hr batch jobs complete in under 1 hr

Saved $100's thousands on hardware upgrade

High availability, rolling upgrades, > 50% less power and cooling usage

Hardware Comparison

IBM Power 550 8204-E8A
4 x 4.3 GHz Power 6 Dual Core
256GB RAM
DB2
AIX or OS/400
DS3400 SAN

TPC-C Throughput 629,159 tpm

Total Cost $ 1,226,021

Response Time:
New Order
Average: 0.89s
90%: 1.35s
Payment
Average:0.89s
90%: 1.37s

HP Proliant DL580 G5
4 x 2.7GHx Hex Core
256 GB RAM
Oracle11g
OEL5
HP StorageWorks SAN

TPC-C Throughput 639,253 tpm

Total Cost $80,000 (6% !)

Response Time:
New Order
Average: 0.392s
90%: 1.25s
Payment
Average: 0.378s
90%: 1.236s

About GL Associates …

What differentiates GL Associates?

- → 30 years experience in Financial Systems improvement
- → Experts in your ERP system's technical architecture
- → Management reporting experts
- → Automated COA conversion & Change Management tools – Proven on large databases
- → Automated Consolidation Tool
- → Financial Data Integrity Services

About GL Associates

- → Global client base
- → Jersey City, NJ and Regional Offices
- → ERP Consulting: Oracle, PeopleSoft, JD Edwards
- → Financial Reporting Process / Chart of Accounts
- → Strategic Business Technologies
- → Data Management

GL Associates – Industries Served

- Automotive Parts
- Chemicals
- Consumer Packaged Goods
- Engineering
- Financial Services
- Government
- Homebuilding
- Industrial Products
- Media & Entertainment
- Medical Devices
- Mining
- Non-profit
- Pharmaceuticals
- Real Estate Management
- Transportation
- Utilities

The only true drag-and-drop reporting software!

- *Integrated reporting for end-users*
- *Performance Management*
- *Budgeting and Planning*
- *Database query application*

www.cetova.com

Cetova is a GL Associates Company

GL Associates

How can GL Associates help?

ERP Consolidation questionnaire?

Questions on material?

Tools demo?

Copy of the presentation?

Request GL Associates white papers?

GL Associates Whitepapers

- COA Design
- COA Conversions

Questions?