High Availability Scenarios for Oracle Databases on IBM z Systems

Sam Amsavelu
Oracle on z Architect
IBM ISV & Channels Technical Sales – Oracle
samvelu@us.ibm.com

29th Annual International Oracle on IBM z Systems™ SIG Conference
Trademarks

The following are trademarks of the International Business Machines Corporation in the United States and/or other countries.

- **AIX**
- **FICON**
- **IBM**
- **IBM (logo)**
- **GDPS**
- **Geographically Dispersed Parallel Sysplex**
- **HiperSockets**
- **HyperSwap**
- **IMS**
- **InfiniBand**
- **Lotus**
- **MQSeries**
- **Parallel Sysplex**
- **POWER**
- **POWER4**
- **Rational**
- **Redbooks**
- **PowerHA**
- **Power Systems**
- **PowerVM**
- **PR/SM**
- **PureSystems**
- **RMF**
- **System x**
- **System z**
- **System z10**
- **Tivoli**
- **WebSphere**
- **zEnterprise**
- **z/VSE**

* Registered trademarks of IBM Corporation

The following are trademarks or registered trademarks of other companies.

Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.

IT Infrastructure Library is a registered trademark of the Central Computer and Telecommunications Agency which is now part of the Office of Government Commerce.

Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel Speed Step, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

Windows Server and the Windows logo are trademarks of the Microsoft group of companies.

ITIL, is a registered trademark, and a registered community trademark of the Office of Government Commerce, and is registered in the U.S. Patent and Trademark Office.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Java and all Java based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

Oracle Java are registered trademarks of Oracle and/or its affiliates.

Cell Broadband Engine is a trademark of Sony Computer Entertainment, Inc. in the United States, other countries, or both and is used under license therefrom.

Linear Tape-Open, LTO, the LTO Logo, Ultrium, and the Ultrium logo are trademarks of HP, IBM Corp. and Quantum in the U.S. and other countries.

* Other product and service names might be trademarks of IBM or other companies.

Notes:

Performance is in Internal Throughput Rate (ITR) ratio based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput improvements equivalent to the performance ratios stated here.

IBM hardware products are manufactured from new parts, or new and serviceable used parts. Regardless, our warranty terms apply.

All customer examples cited or described in this presentation are presented as illustrations of the manner in which some customers have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics will vary depending on individual customer configurations and conditions.

This publication was produced in the United States. IBM may not offer the products, services or features discussed in this document in other countries, and the information may be subject to change without notice. Consult your local IBM business contact for information on the product or services available in your area.

All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

Information about non-IBM products is obtained from the manufacturers of those products or their published announcements. IBM has not tested those products and cannot confirm the performance, compatibility, or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

Prices subject to change without notice. Contact your IBM representative or Business Partner for the most current pricing in your geography.
Notice Regarding Specialty Engines (e.g., zIIPs, zAAPs and IFLs):

- Any information contained in this document regarding Specialty Engines ("SEs") and SE eligible workloads provides only general descriptions of the types and portions of workloads that are eligible for execution on Specialty Engines (e.g., zIIPs, zAAPs, and IFLs). IBM authorizes customers to use IBM SE only to execute the processing of Eligible Workloads of specific Programs expressly authorized by IBM as specified in the “Authorized Use Table for IBM Machines” provided at www.ibm.com/systems/support/machine_warranties/machine_code/aut.html (“AUT”).

- No other workload processing is authorized for execution on an SE.

- IBM offers SEs at a lower price than General Processors/Central Processors because customers are authorized to use SEs only to process certain types and/or amounts of workloads as specified by IBM in the AUT.
Session Objective

- Discuss about the various scenarios for Oracle Databases on IBM z Systems
High Availability - primer

- What would cost my business if my systems are not available for a long period of time?
  - Customer satisfaction
  - Loosing customers
  - Loss of $$$

- So how long I can afford to loose? - Downtime (RTO)

- Next question is how much Data I can afford to loose? (RPO)

- Those two define the SLA and once the SLA is defined next question is
  - To meet that SLA how much money can I spend?

- That defines your High Availability

- Be Realistic
High Availability is not Continuous Availability and is not DR

- Continuous Availability masks from end users both planned and unplanned outages
- High Availability masks the end users from unplanned outages and should meet the agreed SLA
- Many times customers expect high availability to perform like Continuous Availability.
- I had one customer with one z System and wanted to have zero downtime for the databases. (SPOF)
- Disaster Recovery
  - DR is different from HA and CA
High Availability is an iterative process

- Planning
- Development
- Validation
- Implementation and updating
Customer choices for Oracle Databases on z Systems

- Stand alone databases
- Active / Passive implementations RAC One
- RAC
- Databases with Data Guard
- Databases with GoldenGate IBM CDC
- Databases for custom (OLTP, DW), Weblogic, WebSphere, PeopleSoft etc., applications
Typical z Systems Oracle Databases implementation
Exposures

- Unplanned outages
  - Oracle instance
  - Linux instance
  - zVM
  - Network
  - Storage
  - In rare occasions z Systems hardware

- Planned outages
  - System HW refresh
  - POR
  - Patches (for SW components)
Single instance Oracle Database Server

- Can be implemented as
  - Standalone
  - Multitenant (12c)
- Major protection is data BACKUP
- RTO can be from minutes to days
- Oracle FLASHBACK technologies may come in hand for
  - Error investigation
  - Error correction
  - Quick Point-in-Time recovery compared to Backup-Restore
- Bottom level at High Availability
- There are many SPOF components
Oracle Database Server Best practices for High Availability

- Frequently **updated and validated** local and remote backups constitute the foundation of an overall HA strategy.
- Multiplex Redo logs
- Use spfile
- Create two or more control files
- Enable Archive mode and use Flash Area
- Avoid NOLOGGING
- Take advantages of Flashback Database
- Enable Block checking
- Use Temp tablespace with temp files and use temp tablespaces groups
- Usage of ASM will be a good choice
  - for mirroring
  - data distribution
  - Add / removing luns
z/VM High Availability features

- z/VM is developed to provide the highest quality of service and availability to the virtual machines it hosts.
  - CP, the hypervisor component in a z/VM system, includes transparent error handling and recovery, whenever possible.
  - There are several built-in transparent network failover techniques available, when using dedicated OSA cards or virtual switches.

- These techniques try to guarantee the availability of a z/VM Logical Partition (LPAR).

- However, they do not prevent unrecoverable hardware errors or any planned outages for z/VM.

- To improve the availability of virtual machines in case of a hardware maintenance or failure:
  - Single System Image (SSI)
  - Live Guest Relocation (LGR)
**z/VM Single System Image**

- **Member 1**
  - Non-shared volumes

- **Member 2**
  - Multiple CTCs for ISFC-based SSI communications

- **Member 3**
  - Shared volumes

- **Member 4**
  - Common LAN for guest IP communications (optionally, with shared SAN for guest FCP connections)
Linux high availability features

- Linux provides HA by clustering and there are four options:
  - Linux-HA
  - SUSE Linux Enterprise High Availability Extension
  - Red Hat Enterprise Linux Add-on
  - IBM Tivoli System Automation
Linux-HA

- Open HA project hosted at [http://www.linux-ha.org](http://www.linux-ha.org)

- Set of building blocks for high availability clusters
  - Heartbeat (messaging layer)
  - resource-agents
    - Scripts that start/stop clustered services
    - Templates and scripts for many applications
  - Pacemaker
    - cluster resource manager (CRM)
  - Cluster-glue (plumbing library)
    - Everything that is not messaging layer and not resource manager

- The tools provided by the Linux-HA project are multi platform. However, there are no Linux-HA packages available in the standard Linux on System z distribution medias. They have to be hand-compiled to be used on Linux on System z.

- Achieving High Availability on Linux for System z with Linux-HA Release 2, SG24-7711
SUSE Linux Enterprise High Availability Extension (SLE HAE)

- Bundled with base SUSE Linux Enterprise Server at no additional charge
  - Support level inherited by base SUSE Linux Enterprise Server

- Provides the packaged tools to create clusters
  - Corosync, (messaging layer)
  - Pacemaker
  - OCF resource agents
  - Integrated management tools (YaST, Pacemaker GUI)
  - Oracle Cluster File System 2 (OCFS2) and Clustered Logical Volume Manager (cLVM)
  - Distributed Replicated Block Device (DRBD)

- Reference: (IBM RedGuide)

- Keep Your Systems and Services Available with IBM z Systems and SUSE Linux Enterprise Server for System z (REDP-5249-00)

- Mike Friesenegger (mikef@suse.com)
SUSE Linux Enterprise High Availability Extension (SLE HAE)

- Source: Implementing the SUSE Linux Enterprise High Availability Extension on System z: Mike Friesenegger
Red Hat Enterprise Linux High Availability Add-on

- Refer to Filipe charts (Filipe Miranda fmiranda@redhat.com)
- The Red Hat Enterprise Linux High Availability Add-on is currently not supported by Red Hat on System z. However, it has been ported to System z by an IBM Business Partner, Sine Nomine, which also provides support for this add-on.
- For more information about this product, see this website:
  - http://www.sinenomine.net/products/linux/hao
IBM Tivoli System Automation

- Family of products to help reduce the frequency and duration of service disruptions with advanced policy-based automation
  - IBM Tivoli System Automation Application Manager (TSA AM)
  - IBM Tivoli System Automation for Multiplatforms (SA MP)
- Policy-based and Goal-driven automation
Reference

End-to-end Automation with IBM Tivoli System Automation for Multiplatforms

- Achieve proactive high availability of heterogeneous environments
- Covers multiplatforms, Linux, AIX, and z/OS
- Includes real world case study scenarios

Edson Manoel
Desmond Krishna
Randy R. Watson
Creighton Hicks

ibm.com/redbooks
Oracle Grid Infrastructure (Active / Passive)

- Allows clustering of independent servers so that they cooperate as a single system
  - If a clustered server fails, then any managed application can be restarted on the surviving servers.
  - The managed applications can be like Siebel, GoldenGate, WebSphere®, including Oracle databases
- The applications are protected in active / passive environment
  - Built-in agents to start at the primary node or at other nodes
  - Monitoring frequency, starting, and stopping of the applications and the application dependencies - all can be configured
Oracle Grid Infrastructure (Active / Passive)

- HADR applicability
  - Economical
  - Downtime can be from seconds to minutes (restart time)
  - Protection from Computer hardware failures
  - Protection from OS (Linux / zVM) failures
  - Protection from Oracle instance failures
  - Active / Passive implementation, so recovery is not instantaneous
Oracle RAC One Node (Active / Passive)

- A single Oracle instance is active
- If the Oracle instance or the node on which it is running fails, RAC One Node automatically starts the instance on the second node
  - Applications are notified and automatically connect to the new instance
- Uses ‘omotion’ technology to relocate the instance without any downtime and does not need manual intervention
  - During the short period of time when the instance is moved from one node to another, both instances are active. Once all the connections are migrated the first instance goes down.
- Built-in cluster failover for high availability
- Easy path to RAC
Oracle RAC One Node (Active / Passive)

- HADR applicability
  - Protection from Computer hardware failures
  - Protection from OS (Linux / zVM) failures
  - Protection from Oracle instance failures
  - Protection from storage failures (when ASM is used)
  - In planned outages it is possible to have continuous availability of Oracle instances
  - Rolling database patches
  - Rolling OS upgrades
Oracle RAC (Active / Active)

- Oracle RAC technology allows multiple Oracle instances running across multiple nodes to access the same database providing a single logical instance view

- With Oracle RAC all nodes are active and it enables the continuous availability of Oracle instance

- Scalability – Add or remove node quickly

- Oracle Extended RAC is an architecture where the nodes in the cluster are separated into different data centers
Oracle RAC (Active / Active)

- HADR applicability
  - Protection from Computer hardware failures
  - Protection from OS (Linux / zVM) failures
  - Protection from Oracle instance failures
  - Protection from storage failures (when ASM is used)
  - Active / Active configuration and hence continuous availability
  - Fast application notification (FAN) with integrated Oracle client failover
  - Scalability
  - Complex, expensive solution

- Reference:
  Graceful Application Switchover in RAC with No Application Interruption (Doc ID 1593712.1)
Oracle HA implementation can span across two System z servers

Comments:
- Physically separate System z machines
RAC - Gotchas

- RAC is a complex solution
- RAC is expensive
- Node evictions in RAC environment are more prevalent than outages
  - when that happens many times Oracle instances hang
- Protects only Server, Linux and Oracle instance failures
  - Voting disks, switches, network are still issues
- The switchover from one node to another takes 30 to 60 seconds
- Applications have to be RAC aware
- Only the SELECT statements continue in the failover node
- Not all the patch updates are not rolling
  - In some cases cluster has to be down
High Availability for Oracle on z Systems – multiple options

- Guards against Linux failure, LPAR failure, z/VM failure, Oracle instance failure, LPAR maintenance
- Can be: Active/active, active/passive
- Not limited to two nodes

Oracle provided HA
- RAC
- Data Guard
- Flashback
- CRS
- Grid Control

Operating system HA
- Linux
  - Linux Clustering
- z/VM
  - Mature Hypervisor
  - Hardware assist
  - z/VM SSI/LGR

Hardware provided HA
- Spare CPUs
- N+1 power supplies
- Chip sparing in memory
- Concurrent maintenance
- 50 years MTBF (system fail.)
DR – Oracle Data Guard

- When disaster happens or fore site failures
  - complete restore of a multi-terabyte backup can take for ever
  - the backups may not contain the most up to date versions of data
- You need to maintain one or more synchronized replicas of the production database in separate data center
- Oracle MAA recommends Oracle Data Guard and Active Data Guard
- Maintains a synchronized physical replica (standby) of a production database (primary)
- Administrators can choose either manual or automatic failover to a standby database if the primary database is unavailable
- Client connections can quickly and automatically failover to the standby and resume service.
Oracle Data Guard concept diagram
DR – Oracle Data Guard

- Highest level of data protection
  - Oracle-aware data validation
  - System and software defects, data corruption, and administrator errors that affect a primary database are not mirrored to the standby

- Asynchronous protection
  - simple to deploy, with no performance impact to the primary, regardless of the distance that separates primary and standby databases

- Synchronous protection (zero data loss)
  - will affect performance and thus imposes a practical limit to the distance between primary and standby database.
  - Performance is affected because the primary database does not proceed with the next transaction until the standby acknowledges the commit.
  - The time spent waiting for acknowledgement increases as the distance between primary and standby increases, directly affecting application response time and throughput
Oracle 12c Data Guard enhancements

- **Fast Sync**
  - Allows a standby to acknowledge the primary database as soon as it receives redo in memory, without waiting for disk I/O to a standby redo log file.
  - This reduces the impact of synchronous transport on primary database performance by shortening the total round-trip time between primary and standby.

- **Far Sync**
  - Maintains a synchronized standby database located at a short distance from the primary location.
  - This database instance does not have data, so minimal CPU and memory requirements. This database receives the changes redo synchronously from the primary and then it sends to secondary asynchronously.
  - New York to Sydney is an example with zero data loss.
DR – Oracle GoldenGate

- Logical replication
- Multi-master replication
- Hub and spoke deployment
- Subset replication and data transformation
- Replication between a broad range of heterogeneous hardware platforms and database management systems (DB2, SQL Server)

- Possible solution for customers with single z System

- Bidirectional is a challenge

- IBM Infosphere CDC is another solution for logical replication
DR – Storage replication

- Oracle Data Guard database-centric
- Gets complicated when thousands of databases and maintaining their standby databases
- Datafiles invalidation with NOLOGGING at primary databases
- Additional hardware / software
- Maintenance (upgrades, patches complications)
DR – Storage replication

- IBM Storage replication products

- **Metro Mirror**
  - Metro mirroring is a function of the IBM System Storage® Server.
  - System Storage Peer-to-Peer Remote Copy (PPRC) function provides replication of the independent disk pool to another System Storage Server.
  - You also have the ability to combine this solution with other System Storage-based copy services functions, including IBM FlashCopy

- **Global Mirror**
  - Global Mirror uses the same base technology as Metro Mirror except the transmission of data is done in an asynchronous manner and FlashCopy to a third set of disks is required to maintain data consistency.

- **Metro Global Mirror**
  - Combination of Metro and Global Mirror Technologies
Metro Mirror Two Site replication (with Global Mirror Asynchronous)
Three site replication (Metro Mirror / Global Metro Mirror)
## Disaster Recovery – Sample Tiered Guidelines

<table>
<thead>
<tr>
<th>Tier</th>
<th>Criticality</th>
<th>RTO</th>
<th>RPO</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Mission Critical</td>
<td>&lt;5 Hours</td>
<td>PoF</td>
</tr>
<tr>
<td>2</td>
<td>Business Critical</td>
<td>5 to 24 Hours</td>
<td>PoF</td>
</tr>
<tr>
<td>3</td>
<td>Required</td>
<td>&gt;24 Hours</td>
<td>Intra-Day or LC</td>
</tr>
</tbody>
</table>

- **RTO** – Recovery Time Objective
- **RPO** – Recovery Point Objective
- **PoF** – Point of Failure
- **Intra-Day** – Intra-Day Data Capture Point
- **LC** – Last Capture Point (Backups)
IBM Geographically Dispersed Parallel Sysplex (GDPS)
DR – Typical Scenario

- Different LPAR names between primary and backup data centers
- Guests see the LPAR name at boot up and loads its customized config file based on the name of LPAR
- SAN target WWPN and LUN differ between primary and backup data centers
- Guests automatically configure WWPN and LUN based on the LPAR name it is getting at boot up
- Hipersocket address DO NOT have to be changed at backup data center
- External network address DO change at backup data center and the applications have to point to them
Flexible DR

- Oracle Data Guard (MAA)
- GDPS (or Mirroring)
  - Usable space closer to raw disk capacity
  - Smaller HW configuration with CBU (lower cost dark processors)
  - Multiple options for cold standby
  - Minimized SW license cost
Additional References

- IBM Redbooks
  - End-to-End High Availability Solution for System z from a Linux perspective SG24-8233
  - IBM System z Connectivity Handbook, SG24-5444
  - Achieving High Availability on Linux for System z with Linux-HA Release 2, SG24-7711
  - End to End Automation with IBM Tivoli System Automation for Multiplatforms, SG24-7117
  - GDPS Family an Introduction to Concepts and Facilities, SG24-6374

- OTN HA Portal: http://www.oracle.com/goto/availability

- Maximum Availability Architecture: http://www.oracle.com/goto/maa

- Maximize Availability with Oracle Database 12c – Oracle White paper
THANK YOU

- THANKS