Mission-Critical Databases in the Cloud. Oracle RAC in Microsoft Azure Enabled by FlashGrid® Software.

White Paper
rev. 2018-06-06
Abstract

Ensuring high availability of backend relational databases is a critical part of the cloud migration strategy. Many IT organizations have been using Oracle RAC as a trusted high-availability solution for running mission-critical databases and would like to continue using it in the cloud. Additionally, many of those organizations that have not been using Oracle RAC in their on-premise environments see Oracle RAC as the technology enabling the required level of database high availability in the cloud. However, Oracle RAC has the following infrastructure requirements that are not directly available in public clouds:

- Shared high-performance storage accessible from all nodes in the cluster
- Multicast enabled network between all nodes in the cluster
- Separate networks for different types of traffic: client, cluster interconnect, and storage

FlashGrid Storage Fabric and FlashGrid Cloud Area Network™ technologies address these requirements and enable mission-critical Oracle RAC clusters in public clouds. This paper provides architectural overview of the solution that should help with planning and designing Oracle RAC deployments in Microsoft Azure public cloud.

Introduction to FlashGrid Software

High-speed shared storage is critical for seamless database infrastructure failure handling with zero downtime and zero data loss. FlashGrid Storage Fabric software enables high-speed shared storage in a variety of infrastructure environments including bare-metal servers, virtual machines, or extended distance clusters, without the use of proprietary storage arrays. FlashGrid Storage Fabric adds shared access required by Oracle RAC to the standard storage capabilities of Azure compute cloud.

FlashGrid Cloud Area Network™ software enables migration of mission-critical applications to Azure cloud by bridging the gap between the standard network capabilities of the Azure virtual networks and the networking requirements of Oracle RAC.

Why Oracle RAC in Azure Cloud

Oracle RAC provides an advanced technology for database high availability. Many organizations use Oracle RAC for running their mission-critical applications, including most financial institutions and telecom operators where high-availability and data integrity are of paramount importance.

Oracle RAC is an active-active distributed architecture with shared database storage. The shared storage plays a central role in enabling automatic failover, zero data loss, 100% data consistency, and in preventing application downtime. These HA capabilities minimize outages due to unexpected failures, as well as during planned maintenance.

Oracle RAC technology is available for both large scale and entry level deployments. Oracle RAC Standard Edition 2 provides a very cost-efficient alternative to open-source databases, while ensuring the same level of high availability that the Enterprise Edition customers enjoy.

FlashGrid software brings the superior economics, flexibility, and agility of Azure to a broad range of Oracle RAC customers. It enables existing enterprise Oracle RAC customers to realize the full benefits of migrating their entire IT infrastructure to Azure. It also lowers entry barriers for new customers starting with small scale database deployments.
Supported Cluster Configurations

The FlashGrid architecture enables variety of RAC cluster configurations in Azure. Two or three node clusters are recommended in most cases. Clusters with four or more nodes can be used for extra HA or performance. It is possible to have Grid Infrastructure clusters with 4+ nodes containing several 2- or 3-node database clusters. It is also possible to use Grid Infrastructure for running single-instance databases with automatic fail-over.

Nodes of a cluster can be spread across availability zones to ensure the highest degree of fault isolation. Availability sets with fault domains can be used in the regions that do not currently support availability zones.

Configurations with two RAC nodes

Configurations with two RAC nodes have 2-way data mirroring using Normal Redundancy ASM disk groups. An additional VM is required to host quorum disks. Such cluster can tolerate loss of any one node without database downtime.

Configurations with three RAC nodes

Configurations with three RAC nodes have 3-way data mirroring using high redundancy ASM disk groups. Two additional EC2 instances are required to host quorum disks. Such a cluster can tolerate the loss of any two nodes without database downtime.
Four or more RAC nodes across availability zones

It is possible to configure clusters with 4 or more nodes across availability zones with 2 or more database nodes per availability zone. The database nodes are spread across two availability zones. The third availability zone is used for the quorum node. Such cluster can tolerate loss of an entire availability zone. But in addition, it allows HA within each availability zone, which helps with application HA design.

How It Works

Architecture Highlights

- FlashGrid Cloud Area Network™ enables high-speed overlay networks with multicast, virtual IP addresses, and bandwidth allocations, leveraging Azure Advanced Network capabilities with speeds up to 30 Gbps
- FlashGrid Storage Fabric turns local drives (e.g. Premium SSDs) into shared drives accessible from all nodes in the cluster
- FlashGrid Read-Local™ Technology minimizes network overhead by serving reads from local drives
- 2-way or 3-way mirroring of data across separate nodes or Availability Zones
- Oracle ASM and Clusterware provide data protection and availability

Network

FlashGrid Cloud Area Network™ (CLAN) enables running high-speed clustered applications in public clouds or multi-datacenter environments with the efficiency and control of a Local Area Network.

The network connecting Azure VMs is effectively a single IP network with a fixed amount of network bandwidth allocated per VM for all types of network traffic. However, the Oracle RAC architecture requires separate networks for client connectivity and for the private cluster interconnect between the cluster nodes. There are two main reasons for that: 1) the cluster interconnect must have low latency and sufficient bandwidth to ensure adequate performance of the inter-node locking and Cache Fusion, 2) the cluster interconnect is used for transmitting raw data and for security reasons must be accessible by the database nodes only. Also, Oracle RAC requires network with multicast capability, which is not available in Azure.
FlashGrid CLAN addresses the limitations described above by creating a set of high-speed virtual LAN networks and ensuring QoS between them.

**Figure 4. FlashGrid CLAN virtual subnets**

Network capabilities enabled by FlashGrid CLAN for Oracle RAC in Azure:

- Each type of traffic has its own virtual LAN with a separate virtual NIC, e.g. `fg-pub`, `fg-priv`, `fg-storage`
- Negligible performance overhead compared to the raw network
- Minimum guaranteed bandwidth allocation for each traffic type while accommodating traffic bursts
- Low latency of the cluster interconnect in the presence of large volumes of traffic of other types
- Transparent virtual IP failover between nodes
- Multicast support

**Shared Storage**

FlashGrid Storage Fabric turns local drives into shared drives accessible from all nodes in the cluster. The local drives shared with FlashGrid Storage Fabric can be block devices of any type including Premium SSD disks or LVM volumes. The sharing is done at the block level with concurrent access from all nodes.

**Figure 5. FlashGrid Storage Fabric with FlashGrid Read-Local Technology**

Each database node has a full copy of user data stored on Azure Premium SSD disks attached to that database node. The FlashGrid Read-Local™ Technology allows serving all read I/O from the locally attached disks and increases both read and write I/O performance. Read requests avoid the extra network hop, thus reducing the
latency and the amount of the network traffic. As a result, more network bandwidth is available for the write I/O traffic.

The FlashGrid software maintains persistent disk names and sets the required disk permissions. There is no need to configure ASMlib or UDEV rules.

**ASM Disk Group Structure and Data Mirroring**

FlashGrid software leverages proven Oracle ASM capabilities for disk group management, data mirroring, and high availability. In Normal Redundancy mode each block of data has two mirrored copies. In High Redundancy mode each block of data has three mirrored copies. Each ASM disk group is divided into failure groups – one failure group per node. Each disk is configured to be a part of a failure group that corresponds to the node where the disk is located. ASM stores mirrored copies of each block in different failure groups.

![ASM disk group with Normal Redundancy](image)

*Figure 6. Example of a Normal Redundancy disk group in a 2-node RAC cluster*

A typical Oracle RAC setup in Azure will have three Oracle ASM disk groups: GRID, DATA, FRA.

In a 2-node RAC cluster all disk groups must have Normal Redundancy. The GRID disk group containing voting files is required to have a quorum disk for storing a third copy of the voting files. Other disk groups also benefit from having quorum disks as they store a third copy of ASM metadata and improve failure handling.

In a 3-node cluster all disk groups, except the GRID disk group, must have High Redundancy in order to enable full Read-Local capability. In a 3-node RAC cluster the GRID disk group would typically have Normal Redundancy. Note that in such 3-node RAC cluster loss of no more than one node is tolerated without causing downtime. If a 3-node RAC cluster must tolerate simultaneous failure of two nodes without causing downtime then the GRID disk group must have High Redundancy and additional two quorum nodes must be provisioned to accommodate five copies of voting files. Details of such configuration are not covered in this paper. To learn more, contact FlashGrid. Our contact information is at the end of this document.
High Availability Considerations

FlashGrid Storage Fabric and FlashGrid Cloud Area Network™ have a fully distributed architecture with no single point of failure. The architecture leverages HA capabilities built in Oracle Clusterware, ASM, and Database.

Availability Sets and Availability Zones

Azure offers two features that allow protecting a cluster from two VMs going offline simultaneously: Availability Sets and Availability Zones.

Configuring an Availability Set allows placing cluster nodes in separate Update Domains and separate Fault Domains. Placing VMs in separate Update Domains ensures that those VMs will not be rebooted simultaneously during a planned update of the underlying Azure infrastructure. Placing VMs in separate Fault Domains ensures that those VMs have separate power sources and network switches. Thus, failure of a power source or a network switch will be localized to a single Fault Domain and will not affect VMs in other Fault Domains. Note that for using separate Fault Domains the region must support three Fault Domains. It is still possible to deploy 2-node clusters in the regions that provide only two Fault Domains by placing the quorum VM in a different region. Details of such configuration are beyond the scope of this white paper.

Availability Zones offer better degree of failure isolation by having independent power, cooling, and networking in physically separate data centers. FlashGrid recommends spreading the cluster nodes across Availability Zones in regions where Availability Zones are supported.

Because all instances are virtual, failure of a physical host causes only a short outage for the affected node. The node VM will automatically restart on another physical host. This significantly reduces the risk of double failures.

Data Availability

A Premium SSD disk in Azure provides persistent storage that survives a failure of the node VM. After the failed VM restarts on a new physical node all its volumes are attached with no data loss.

Premium SSD disks have built-in redundancy that protects data from failures of the underlying physical media. The mirroring by ASM is done on top of the built-in protection of Premium SSD disks. Together Premium SSD disks plus ASM mirroring provide durable storage with two layers of data protection, which exceeds the typical level of data protection in on-premises deployments.
Performance Considerations

Supported VM Types and Sizes
Database node VMs must have 2+ CPU cores, 8+ GB of memory, and Premium storage support. The following VM types are recommended for database nodes:

- E8s_v3, E16s_v3, E32s_v3, E64s_v3
- M64s, M128s, M64ms, M128ms
- DS11_V2, DS12_V2, DS13_V2, DS14_V2, DS15_V2
- GS1, GS2, GS3, GS4, GS5

DS1_V2 (1 core, 3.5GB memory) type is recommended for use as a quorum node. Note that there is no Oracle Database software installed on the quorum node.

Supported Disk Types
Currently only Premium SSD disks are supported.

Each disk provides up to 7,500 IOPS and 250 MB/s depending on its size. The maximum performance of 7,500 IOPS is available for 2048 GB and 4096 GB disks. For databases that require high performance, but smaller capacity, use of multiple 512 GB or 1024 GB disks may be optimal to maximize the total IOPS and MB/s. Note that maximum number of IOPS per VM is also capped and depends on the VM size - 80,000 for E64s_v3.

Reference Performance Results
The main performance related concern when moving database workloads to the cloud tends to be around storage and network I/O performance. There is a very small to zero overhead related to the CPU performance between bare-metal and Azure cloud. Therefore, in this paper we focus on the I/O performance.

Calibrate_IO
The CALIBRATE_IO procedure provides an easy way for measuring storage performance including maximum bandwidth, random IOPS, and latency. The CALIBRATE_IO procedure generates I/O through the database stack on actual database files. The test is read-only and it is safe to run it on any existing database. It is also a good tool for directly comparing performance of two storage systems because the CALIBRATE_IO results do not depend on any non-storage factors, such as memory size or the number of CPU cores.

Test configuration:
- Two database nodes, E64s_v3
- Sixteen 1024 GB Premium SSD disks per node
Test script:

```sql
SET SERVEROUTPUT ON;
DECLARE
  lat INTEGER;
  iops INTEGER;
  mbps INTEGER;
BEGIN
  DBMSRESOURCE_MANAGER.CALIBRATE_IO (32, 10, iops, mbps, lat);
  DBMS_OUTPUT.PUT_LINE ('Max_IOPS = ' || iops);
  DBMS_OUTPUT.PUT_LINE ('Latency = ' || lat);
  DBMS_OUTPUT.PUT_LINE ('Max_MB/s = ' || mbps);
end;
/
```

Our results:

```
Max_IOPS = 140605
Latency = 1
Max_MB/s = 3250
```

**SLOB**

SLOB is a popular tool for generating I/O intensive Oracle workloads. SLOB generates database SELECTs and UPDATEs with minimal computational overhead. It complements Calibrate_IO by generating mixed (read+write) I/O load. AWR reports generated during the SLOB test runs provide various performance metrics. For the purposes of this paper we focus on the I/O performance numbers.

Test configuration:

- Database node VM type: E64s_v3
- Sixteen 1024 GB Premium SSD disks per node
- SGA size: 3 GB (small size selected to minimize caching effects and maximize physical I/O)
- 8KB database block size
- Schemas: 150 x 240MB per node
- UPDATE_PCT=15

The table below shows our results for tests performed in the same configuration (provided above) with 2- and 3-node RAC clusters compared to a single-instance database (on a single VM) as a baseline.

<table>
<thead>
<tr>
<th></th>
<th>Single-instance (as a baseline)</th>
<th>2-node RAC (both nodes combined)</th>
<th>3-node RAC (all nodes combined)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Read+Write Database Requests (IOPS)</td>
<td>62,887</td>
<td>99,020</td>
<td>143,897</td>
</tr>
<tr>
<td>Read Database Requests (IOPS)</td>
<td>53,923</td>
<td>84,349</td>
<td>122,646</td>
</tr>
<tr>
<td>Write Database Requests (IOPS)</td>
<td>8,964</td>
<td>14,671</td>
<td>21,251</td>
</tr>
</tbody>
</table>

At 100K or more IOPS per cluster the performance is comparable to using a dedicated flash storage array.
Software Compatibility

The following versions of software are supported by FlashGrid as part of the solution:

- Oracle Database: ver. 12.2.0.1, 12.1.0.2, or 11.2.0.4 with the latest PSU
- Oracle Grid Infrastructure: ver. 12.2.0.1 or 12.1.0.2 with the latest PSU
- Operating System: Oracle Linux 7, Red Hat Enterprise Linux 7

Deployment Process

FlashGrid Cloud Provisioning tool automates the process of deploying a cluster. The tool provides a flexible web-interface for defining cluster configuration and generating an Azure Resource Manager template for it. The following tasks are performed automatically using the Resource Manager template:

- Creating cloud infrastructure: VMs, storage, and optionally network
- Installing and configuring FlashGrid Cloud Area Network
- Installing and configuring FlashGrid Storage Fabric
- Installing, configuring, and patching Oracle Grid Infrastructure
- Installing and patching Oracle Database software
- Creating ASM disk groups

The entire deployment process takes approximately 90 minutes. After the process is complete the cluster is ready for creating a database.

Conclusion

Running Oracle RAC clusters in Microsoft Azure cloud out of the box has historically been challenging due to storage and network constraints. FlashGrid Cloud Area Network™ and FlashGrid Storage Fabric remove those constraints and enable a range of highly available database cluster solutions ranging from small cost-efficient and easy to deploy Oracle RAC Standard Edition 2 clusters to high-performance mission-critical Oracle RAC Enterprise Edition clusters with high availability characteristics exceeding those of the traditional on-premises deployments.

Contact Information

For more information please contact FlashGrid at info@flashgrid.io