SKYCLUSTER

Mission-Critical Databases in the Cloud. Oracle RAC on Amazon EC2

Enabled by SkyCluster®.

White Paper

rev. 2020-08-16

Abstract

The use of Amazon Elastic Compute Cloud (Amazon EC2) in the Amazon Web Services (AWS) Cloud provides IT organizations with the flexibility and elasticity that are not available in the traditional data center. With AWS it is possible to bring new enterprise applications online in hours instead of months.

Ensuring high availability of backend relational databases is a critical part of the cloud strategy - whether it is a lift-and-shift migration or a green-field deployment of mission critical applications. SkyCluster is an engineered cloud system designed for database high availability. SkyCluster is delivered as a fully integrated Infrastructure-as-Code template that can be customized and deployed to AWS EC2 account with a few mouse clicks. Key components of SkyCluster for AWS include:

- Amazon EC2 VM instances
- Amazon EBS and/or local SSD storage
- FlashGrid Storage Fabric software
- FlashGrid Cloud Area Network software
- Oracle Grid Infrastructure software
- Oracle RAC database engine

By leveraging the proven Oracle RAC database engine SkyCluster enables the following use-cases:

- Lift-and-shift migration of existing Oracle RAC databases to AWS.
- Migration of existing Oracle databases from high-end on-premises servers to AWS without reducing availability SLAs.
- Design of new mission critical applications for the cloud based on the industry proven and widely supported database engine.

This paper provides architectural overview of SkyCluster and can be used for planning and designing high availability database deployments on Amazon EC2.

Why Oracle RAC Database Engine

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.

Supported Cluster Configurations

SkyCluster enables variety of RAC cluster configurations on Amazon EC2. 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 clusters with 4+ nodes containing several 2- or 3-node database sub-clusters. It is also possible

to use SkyCluster for running single-instance databases with automatic fail-over. Nodes of a cluster can be in one availability zone or can be spread across availability zones.

Configurations with two RAC database nodes

Configurations with two RAC database nodes have 2-way data mirroring using Normal Redundancy ASM disk groups. An additional EC2 instance is required to host quorum disks. Such cluster can tolerate loss of any one node without database downtime.

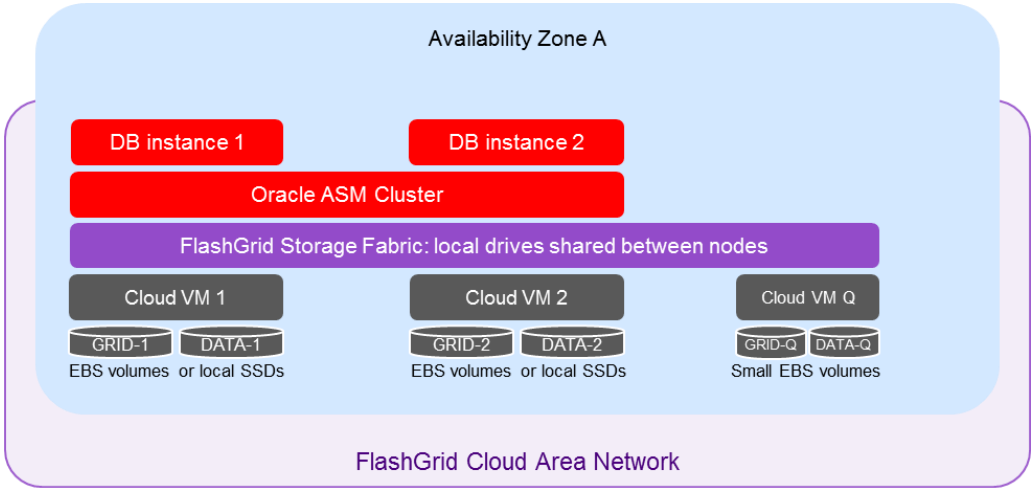

Figure 1. Two RAC database nodes in the same Availability Zone

In configurations where local NVMe SSDs are used instead of EBS volumes, High Redundancy ASM disk groups may be used to provide extra layer of data protection. In such cases the third node is configured as a *storage* node with NVMe SSDs or EBS volumes instead of the quorum node.

Configurations with three RAC database nodes

Configurations with three RAC database 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.

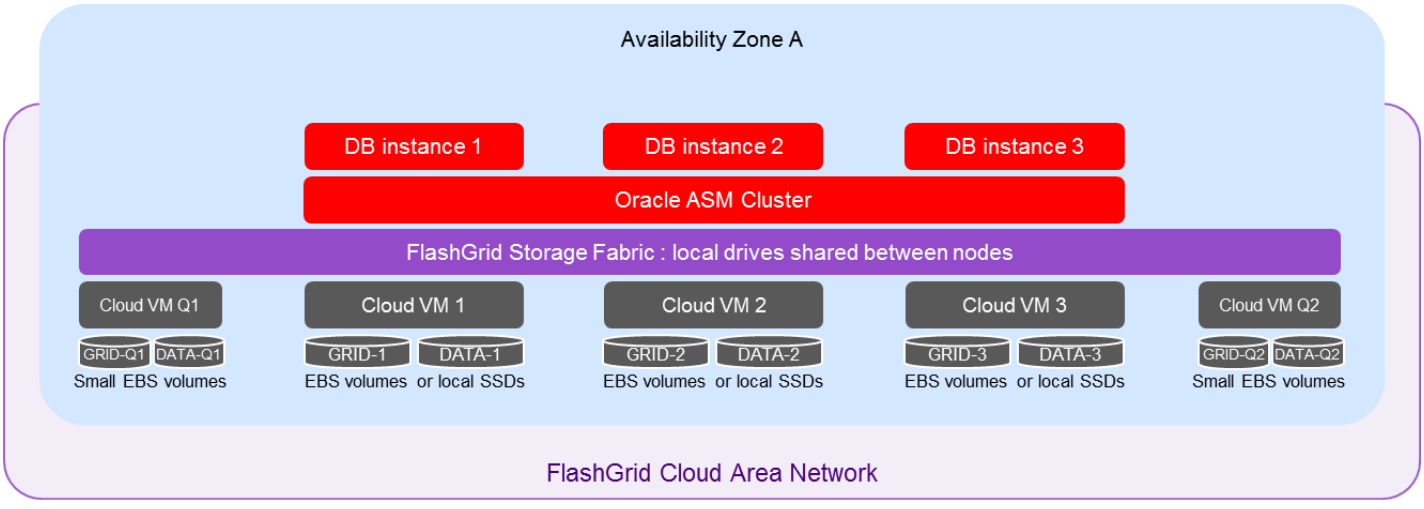

Figure 2. Three RAC database nodes in the same Availability Zone

Same Availability Zone vs. separate Availability Zones

Amazon Web Services consists of multiple independent Regions. Each Region is partitioned into several Availability Zones. Availability Zones consist of one or more discrete data centers, each with redundant power, networking and connectivity, housed in separate facilities. Availability Zones are physically separate, such that even extremely uncommon disasters such as fires, tornados or flooding would only affect a single Availability Zone.

Although Availability Zones within a Region are geographically isolated from each other, they have direct low-latency network connectivity between them. The network latency between Availability Zones is generally lower than 1ms. This makes the inter-AZ deployments compliant with the extended distance RAC guidelines.

Placing all nodes in one Availability Zone provides the best performance for write intensive applications by ensuring network proximity between the nodes. However, in the unlikely event of an entire Availability Zone failure, the cluster will experience downtime.

Placing each node in a separate Availability Zone helps avoid downtime, even when an entire Availability Zone experiences a failure. The trade-off is a somewhat higher network latency, which may reduce write performance. Note that the read performance is not affected because all reads are served locally.

If placing nodes in separate Availability Zones then using a Region with at least three Availability Zones is generally required. The current number of Availability Zones for each Region can be found at https://aws.amazon.com/about-aws/global-infrastructure/ . It is possible to deploy a 2-node RAC cluster in a Region with only two Availability Zones. However, in such a case the quorum server must be located in a different Region or in a data center with VPN connection to AWS to prevent network partitioning scenarios. This configuration is beyond the scope of this document. To learn more, contact FlashGrid.

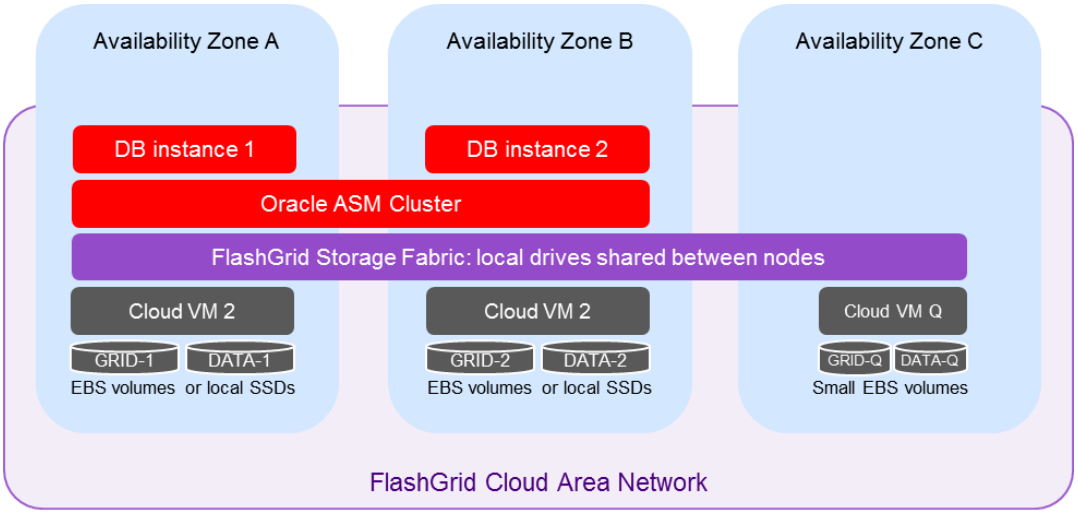

Figure 3. Two RAC database nodes in separate Availability Zones

Three RAC database nodes across availability zones

Most of the AWS regions are limited to 3 availability zones. Because of this, placing the additional quorum nodes in separate availability zones may not be possible. However, with three RAC nodes placing the quorum nodes in the same availability zones as the RAC nodes still allows achieving most of the expected HA capabilities. Such a cluster can tolerate loss of any two nodes or loss of any one availability zone without database downtime. Note, however, that simultaneous loss of two availability zones will cause database downtime.

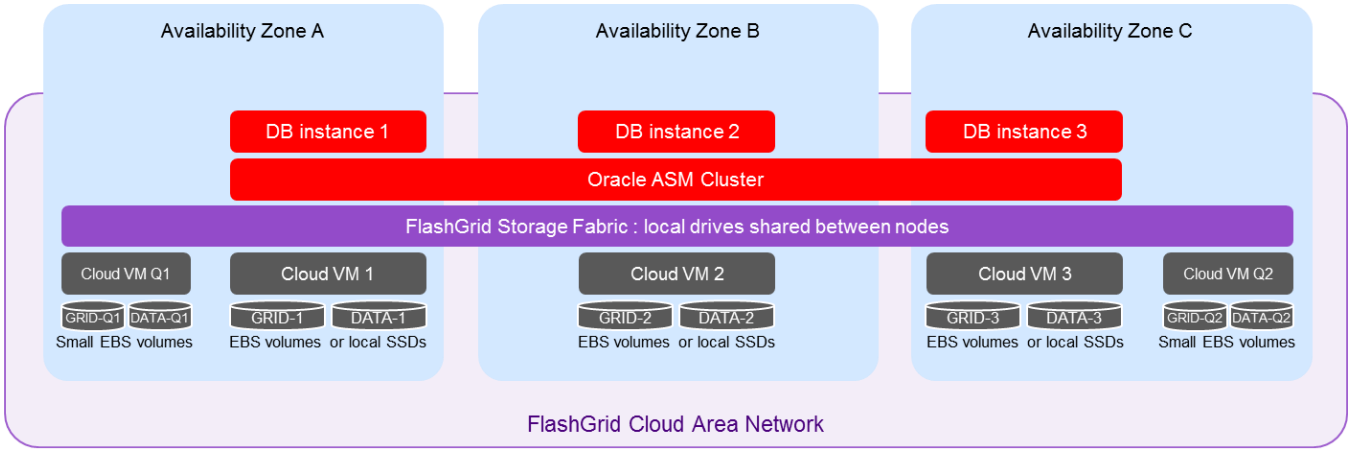

Figure 4. Three RAC database nodes in separate Availability Zones

Four or more RAC database 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.

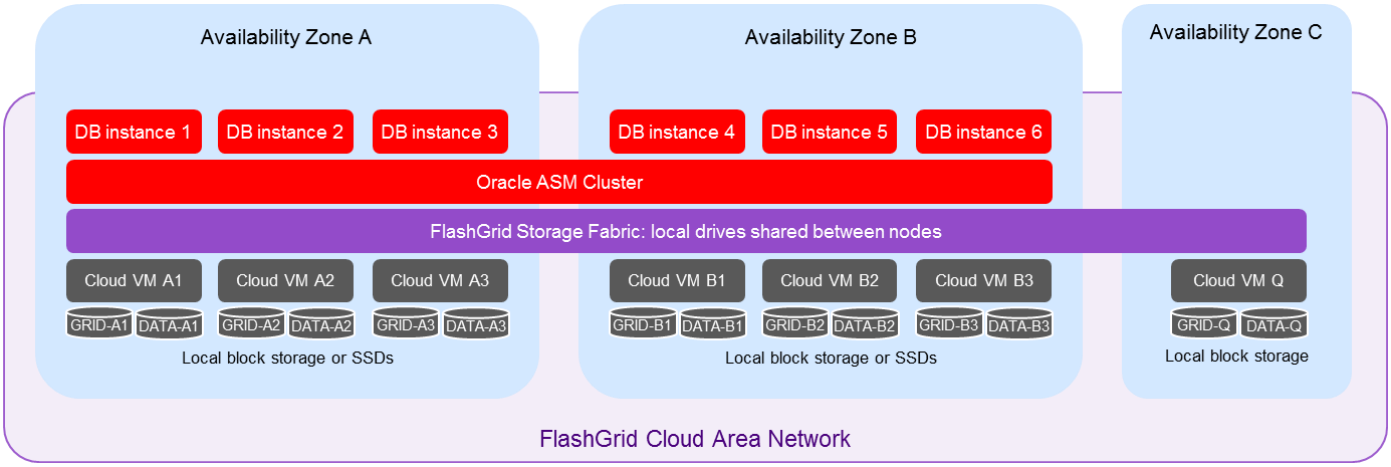

Figure 5. Example of a six-node RAC database cluster across availability zones

How It Works

SkyCluster Architecture Highlights

- Database clusters delivered as Infrastructure-as-Code templates for automated and repeatable deployments
- FlashGrid Cloud Area Network™ software enables high-speed overlay networks with advanced capabilities for HA and performance management
- FlashGrid Storage Fabric software turns locally attached disks (elastic block storage or local instance-store SSDs) into shared disks accessible from all nodes in the cluster
- FlashGrid Read-Local™ Technology minimizes network overhead by serving reads from locally attached disks
- 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 Amazon EC2 instances is effectively a single IP network with a fixed amount of network bandwidth allocated per instance for all types of network traffic (except for Amazon Elastic Block Storage (EBS) storage traffic on EBS-optimized instances). 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 Amazon EC2.

FlashGrid CLAN addresses the limitations described above by creating a set of high-speed virtual LAN networks and ensuring QoS between them.

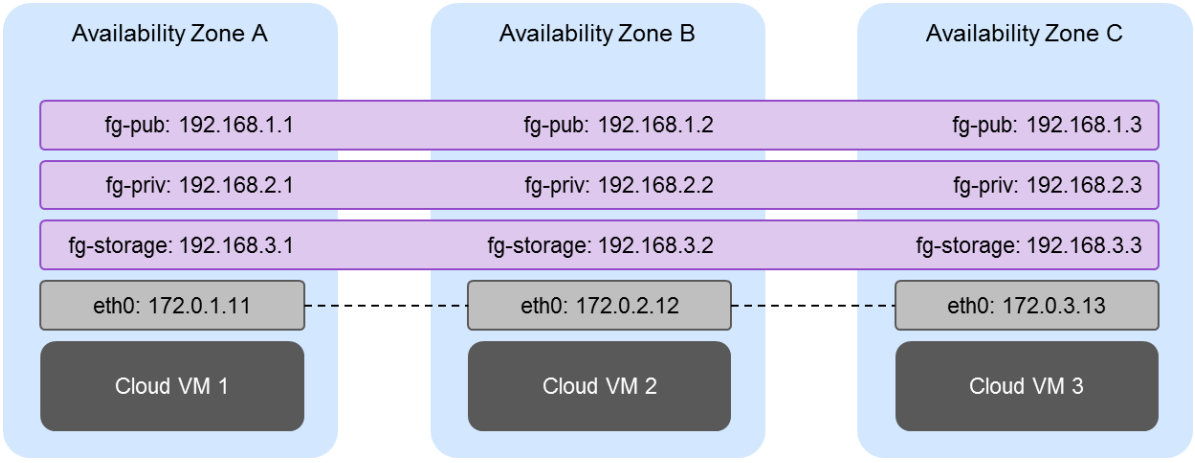

Figure 6. FlashGrid Cloud Area Network

Network capabilities enabled by FlashGrid CLAN for Oracle RAC in Amazon EC2:

- 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 connectivity across Availability Zones
- Multicast support
- Up to 100 Gb/s bandwidth per node

Shared Storage

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

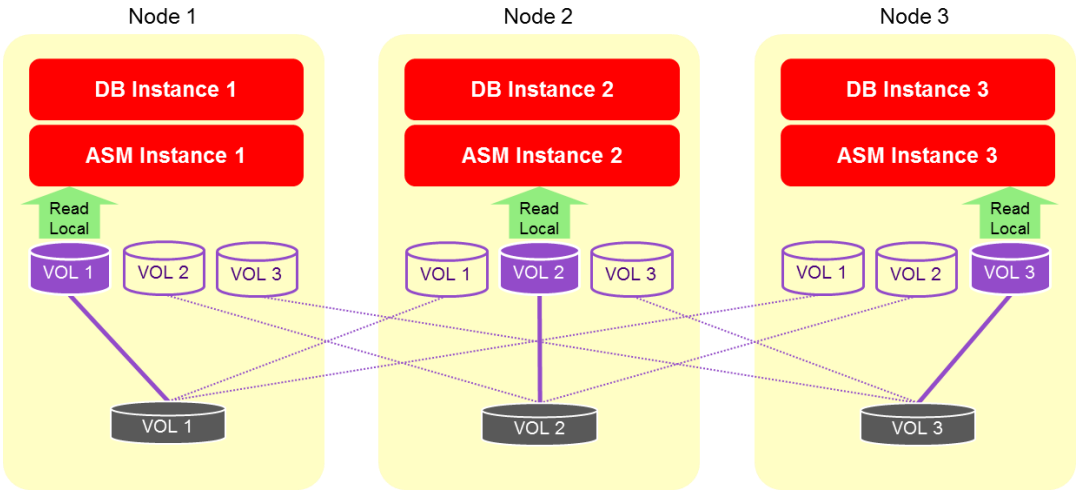

Figure 7. FlashGrid Storage Fabric with FlashGrid Read-Local Technology

In 2-node or 3-node clusters each database node has a full copy of user data stored on Amazon EBS volume(s) 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.

ASM Disk Group Structure and Data Mirroring

FlashGrid Storage Fabric 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.

A typical Oracle RAC setup in Amazon EC2 will have three Oracle ASM disk groups: GRID, DATA, FRA.

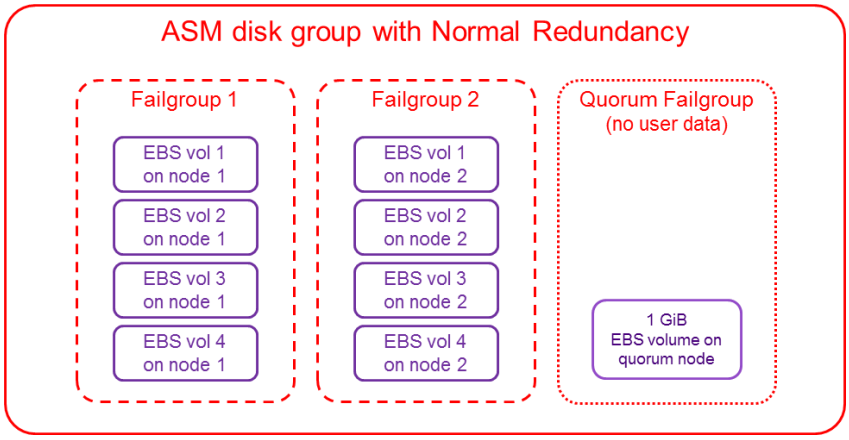

Figure 8. Example of a Normal Redundancy disk group in a 2-node RAC cluster

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 the quorum disks as they store a third copy of ASM metadata for better failure handling.

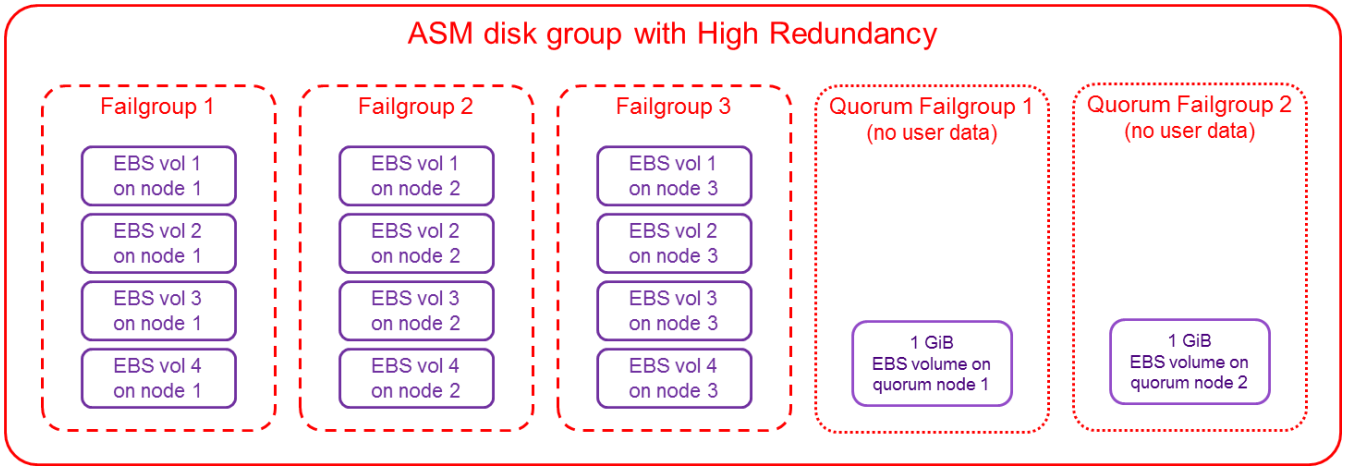

Figure 9. Example of a High Redundancy disk group in a 3-node RAC cluster

In a 3-node cluster all disk groups must have High Redundancy in order to enable full Read-Local capability. The GRID disk group containing voting files is required to have two additional quorum disks, so it can have five copies of the voting files. Other disk groups also benefit from having the quorum disks as they store additional copies of ASM metadata for better failure handling.

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.

Node Availability

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

A single Availability Zone configuration provides protection against loss of a database node. It is an efficient way to accommodate planned maintenance (e.g. patching database or OS) without causing database downtime. However, a potential failure of a resource shared by multiple instances in the same Availability Zone, such as network, power, or cooling, may cause database downtime.

Placing instances in different Availability Zones virtually eliminates the risk of simultaneous node failures, except for the unlikely event of a disaster affecting multiple data center facilities in a region. The trade-off is higher network latency. However, the network latency between AZs is less than 1ms in most cases and will not have critical impact on performance of many workloads.

Data Availability with EBS Storage

An Amazon EBS volume provides persistent storage that survives a failure of node instance where the volume is attached to. After the failed instance restarts on a new physical node all its volumes are attached with no data loss.

Amazon EBS volumes 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 Amazon EBS. Together Amazon EBS plus ASM

mirroring provide durable storage with two layers of data protection, which exceeds the typical level of data protection in on-premises deployments.

Data Availability with Local NVMe SSDs

Local NVMe SSDs are ephemeral (non-persistent), which means that in case an instance is stopped or fails-over to a different physical host, the data on the SSDs attached to that instance is not retained. FlashGrid Storage Fabric provides mechanisms for ensuring persistency of the data stored on local NVMe SSDs. Mirroring data across two or three instances ensures that there is a copy of data still available in the event of one instance losing its data. Placing the instances in different availability zones prevents the possibility of a simultaneous failures of more than one instance. Placing one or two copies of data on NVMe SSDs and one copy on EBS provides high read bandwidth of NVMe and an additional layer of persistency of EBS.

In the event of a loss of data on one of the instances with NVMe SSDs, FlashGrid Storage Fabric automatically reconstructs the affected disk groups and starts data re-synchronization process after the failed instance is back online. No manual intervention is required.

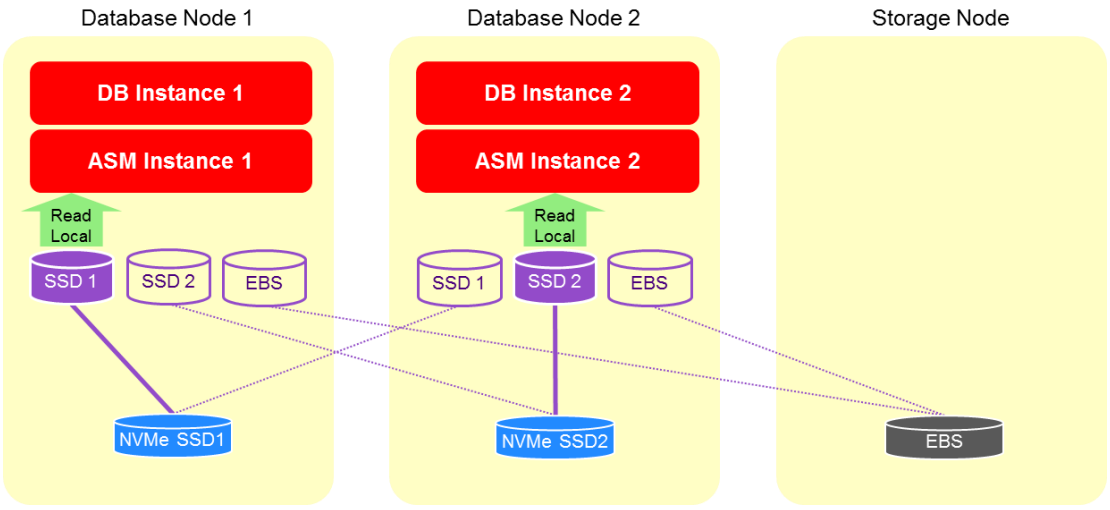

Figure 10. Two mirrors on NVMe SSDs plus one mirror on EBS

Performance Considerations

Recommended Instance Types

An instance type must meet the following criteria:

- At least two vCPUs
- Enhanced Networking – direct access to the physical network adapter
- EBS Optimized – dedicated I/O path for Amazon EBS, not shared with the main network

The following instance type families satisfy the above criteria and are optimal for database workloads:

- M4, M5: optimal memory to CPU ratio
- R4, R5: high memory to CPU ratio
- C5, Z1d: small memory to CPU ratio, for CPU intensive workloads
- i3: high memory to CPU ratio, up to 15 TB of local NVMe SSDs
- i3en: high memory to CPU ratio, up to 100 Gb/s network, up to 60 TB of local NVMe SSDs
- X1, X1E: large memory size, large number of CPU cores

Quorum servers require fewer resources than database nodes. However, the above criteria are still important to ensure stable cluster operation. For example, m5.large or c5.large instances can be used as quorum servers.

Using T2 instance family for quorum servers is not supported. Note that there is no Oracle Database software installed on the quorum servers, hence the quorum servers do not increase the number of licensed CPUs.

Single vs. Multiple Availability Zones

Using multiple Availability Zones provides substantial availability advantages. However, it does affect network latency. In the US-West-2 region for 8KB transfers we measured 0.3ms, 0.6 ms, and 1.0 ms between different pairs of Availability Zones compared to 0.1 ms within a single Availability Zone.

Note that the different latency between different pairs of AZs provides opportunity for optimizing selection of which AZs to use for database nodes. In a 2-node RAC cluster, it is optimal to place database nodes in the pair of AZs that has the minimal latency between them.

The latency impact in multi-AZ configurations may be significant for the applications that have high ratios of data updates. However, read-heavy applications will experience little impact because all read traffic is served locally and does not use the network.

EBS Volumes

Use of General Purpose SSD (gp2) volumes is recommended in most cases. However, use of gp2 volumes smaller than 1000 GB is not recommended due to expected variability in performance. Volumes of 1000 GB size and larger provide guaranteed level of performance of 3 IOPS/GB up to 10,000 IOPS per volume. In most cases the following number of volumes and volume sizes are recommended:

- Usable disk group capacities below 8 TB: up to 8 gp2 volumes, 1 TB each
- Usable disk group capacities over 8 TB: 8 gp2 volumes, 1 TB to 16 TB each

All volumes in the same disk group must be of equal size.

Use of Provisioned IOPS SSD (io1) volumes may be cost-efficient for configurations with very small capacities and small, but guaranteed, performance requirements below 1,000 IOPS.

Local NVMe SSDs

Use of local NVMe SSDs as the primary storage offers higher bandwidth and lower cost compared to Amazon EBS volumes. i3 instance family includes NVMe SSDs up to 8 x 1900 GB with up to 16GB/s of bandwidth and up to 3.3 mln IOPS. The new i3en instance family increases the local SSD capacity up to 8 x 7500 GB.

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 VMs. Therefore, in this paper we focus on the storage I/O and RAC interconnect I/O.

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 script:

```
SET SERVEROUTPUT ON;
DECLARE
  lat INTEGER;
  iops INTEGER;
  mbps INTEGER;
BEGIN DBMS_RESOURCE_MANAGER.CALIBRATE_IO (16, 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 with two database nodes:

Cluster configuration	Max_IOPS	Latency	Max_MB/s
EBS storage	154,864	0	2,219
NVMe storage	1,375,694	0	27,338

Note that the Calibrate_IO results do not depend on whether the database nodes are in the same or different Availability Zones.

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.

Our results with two database nodes:

Cluster configuration	Physical Write Database Requests	Physical Read Database Requests	Physical Read+Write Database Requests
EBS storage, same AZ*	20,697 IOPS	100,539 IOPS	121,237 IOPS
EBS storage, different AZs*	19,465 IOPS	92,081 IOPS	111,546 IOPS
NVMe storage, different AZs**	23,913 IOPS	429,660 IOPS	453,573 IOPS

Test configuration details

EBS storage, same AZ and different AZs

- Two database nodes, M4.16xlarge
- Four io1 20000 IOPS 400GB volumes per node
- SGA size: 2.6 GB (small size selected to minimize caching effects and maximize physical I/O)
- 8KB database block size
- Schemas: 30 x 240MB
- UPDATE_PCT= 20

NVMe storage, different AZs

- Two db nodes + storage node
- Instance type: i3.16xlarge
- (8) 1900GB NVMe SSDs per node
- SGA size: 4 GB (small size selected to minimize caching effects and maximize physical I/O)
- 8KB database block size

- Schemas: 200 x 240MB
- UPDATE_PCT= 5

Performance vs. on-premise solutions

Both EBS and NVMe SSD storage options are flash based and provide order of magnitude improvement in IOPS and latency compared to traditional spinning hard drive based storage arrays. With over 100K IOPS in both cases, the performance is comparable to having a dedicated all-flash storage array. It is important to note that the storage performance is not shared with other clusters or databases. Every cluster has its own dedicated set of EBS volumes or NVMe SSDs, which ensures stable and predictable performance with no interference from noisy neighbors.

NVMe SSDs enable speeds that are difficult or impossible to achieve even with dedicated all-flash arrays. Each NVMe SSD provides read bandwidth comparable to an entry-level flash array. The 27 GB/s bandwidth measured with 16 NVMe SSDs in a 2-node cluster is equivalent to a large flash array connected with 16 Fibre Channel links. Read-heavy analytics and data warehouse workloads can benefit the most from using the NVMe SSDs.

Compatibility

The following versions of software are supported with SkyCluster:

- Oracle Database: ver. 19c, 18c, 12.2, 12.1, or 11.2
- Oracle Grid Infrastructure: ver. 19c
- Operating System: Oracle Linux 7, Red Hat Enterprise Linux 7

Automated Infrastructure-as-Code Deployment

SkyCluster Launcher tool automates the process of deploying a cluster. The tool provides a flexible web-interface for defining cluster configuration and generating an Amazon CloudFormation template for it. The following tasks are performed automatically using the CloudFormation 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. Use of automatically generated standardized IaC templates prevents human errors that could lead to costly reliability problems and compromised availability.

Conclusion

SkyCluster offers a wide range of highly available database cluster configurations in AWS ranging from cost-efficient 2-node clusters to large high-performance clusters. Combination of the proven Oracle RAC database engine, AWS availability zones, and the fully automated Infrastructure-as-Code deployment provides high availability characteristics exceeding those of the traditional on-premises deployments.

Contact Information

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