Mission-Critical Databases in the Cloud.
Oracle RAC on Amazon EC2
Enabled by FlashGrid® Software.

White Paper

rev. 2017-04-02

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. Running the entire IT infrastructure in the public cloud is important in order to fully realize the competitive and cost advantages that the public cloud provides.

Ensuring high availability of backend relational databases is a critical part of the cloud migration strategy. Oracle RAC is the trusted high-availability solution for running mission-critical databases and many IT organizations prefer to run it in the cloud without switching to cloud-native solutions.

Oracle RAC has the following infrastructure requirements that are not directly available in AWS:

- 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 on Amazon EC2. This paper provides architectural overview of the solution and can be used for planning and designing Oracle RAC deployments on Amazon EC2.

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 Amazon EC2.

FlashGrid Cloud Area Network™ software enables migration of mission-critical applications to the AWS by bridging the gap between the standard network capabilities of the Amazon Virtual Private Cloud (VPC) and the networking requirements of Oracle RAC.

Why Oracle RAC on AWS

AWS provides on-demand computing resources and services in the cloud, with pay-as-you-go pricing. AWS has been disrupting the traditional IT infrastructure world by removing the need to manage dedicated hardware and effectively shifting capital expenses (CapEx) to operational expenses (OpEx). After migrating to AWS, IT organizations may expect up to 70% reductions in the TCO and massive increases in elasticity and agility of development (IDC report). By using AWS, customers can free valuable resources used to manage complex and costly datacenters, and repurpose them to focus on other highly strategic areas of the business.

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 AWS 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 AWS. It also lowers entry barriers for new customers starting with small scale database deployments.

Supported Cluster Configurations

FlashGrid supports four RAC cluster configurations on Amazon EC2:

- Two RAC nodes in the same Availability Zone
- Three RAC nodes in the same Availability Zone
- Two RAC nodes across different Availability Zones
- Three RAC nodes across different Availability Zones

Support for configurations with 4+ RAC nodes for compute performance scaling is planned in the future.

Configurations with two RAC nodes

Configurations with two RAC 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.

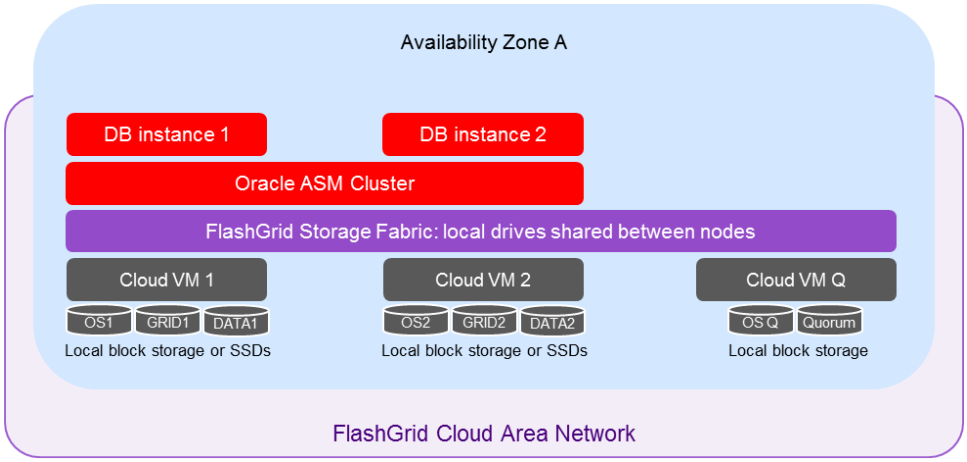

Figure 1. Two RAC nodes in the same Availability Zone

Configurations with three RAC nodes

Configurations with three RAC nodes have 3-way data mirroring using high redundancy ASM disk groups. However, a normal redundancy ASM disk group is used for clusterware files (the GRID disk group). Such a cluster can tolerate the loss of any one node without database downtime. However, loss of a second node will result in downtime. The main reason for using three (vs. two) RAC nodes are the additional CPU and memory resources and the additional storage read bandwidth.

It is possible to have a 3-node RAC cluster that can tolerate loss of two RAC nodes without database downtime, but such configurations are beyond the scope of this document. To learn more, contact FlashGrid. Our contact information is at the end of this document.

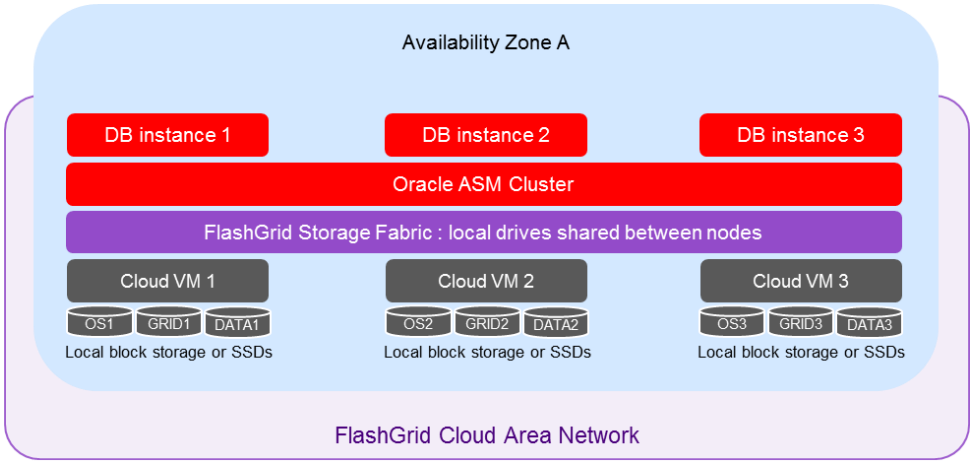

Figure 2. Three RAC nodes in the same Availability Zone

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 and in some cases lower network bandwidth, which may reduce write performance. Note that the read performance is not affected because all reads are served locally.

If you are placing nodes in separate Availability Zones then using a Region with at least three Availability Zones is highly recommended. The current number of Availability Zones for each Region can be found at https://aws.amazon.com/about-aws/global-infrastructure/ .

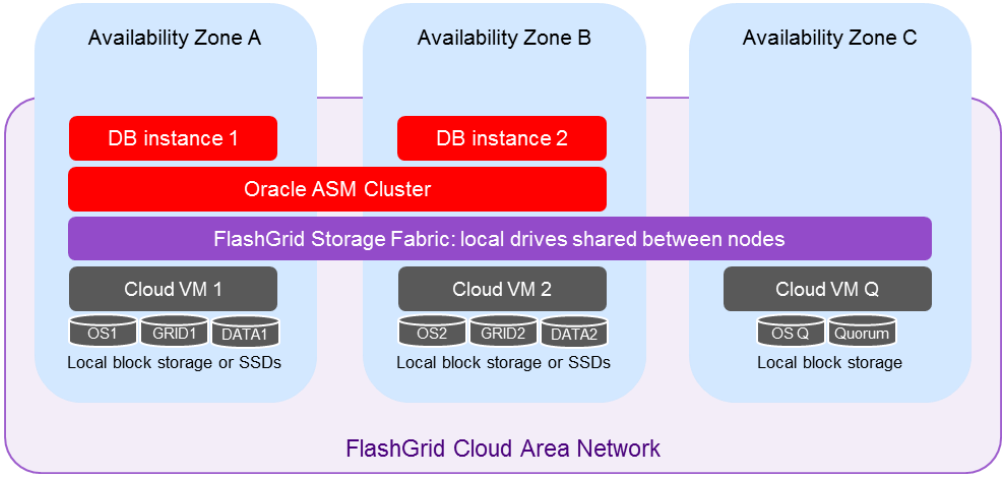

Figure 3. Two RAC nodes in separate Availability Zones

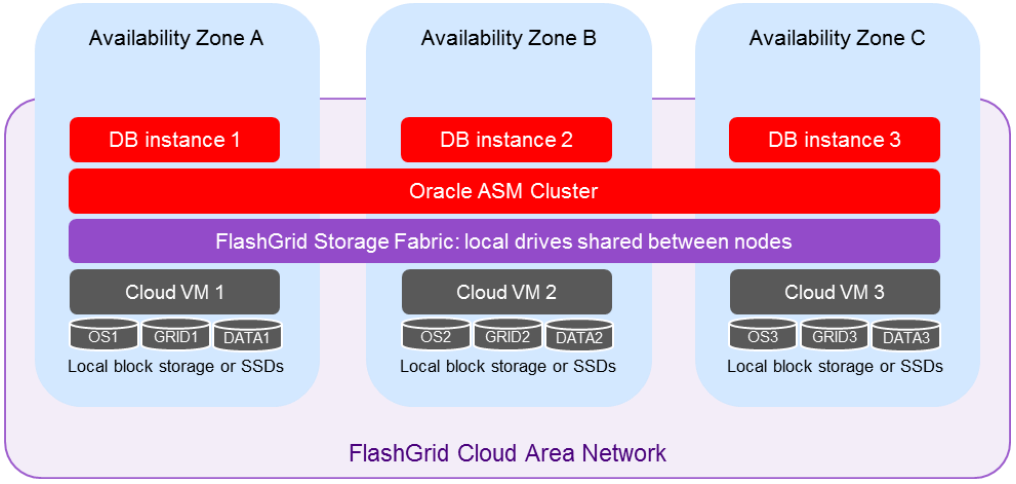

Figure 4. Three RAC nodes in separate Availability Zones

It is possible to deploy in a Region with only two Availability Zones for a 2-node RAC cluster. However, in such a case the quorum server must be located in a different Region to prevent network partitioning scenarios. This configuration is beyond the scope of this document. To learn more, contact FlashGrid. Our contact information is at the end of this document.

How It Works

Architecture Highlights

- FlashGrid Cloud Area Network™ enables high-speed overlay networks with multicast and bandwidth allocations, leveraging AWS networking features such as cluster placement groups, and Elastic Network Adapters rated up to 20Gbps
- FlashGrid Storage Fabric turns local drives (elastic block storage or local instance-store 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 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.

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

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 Amazon EBS volumes or LVM volumes. The sharing is done at the block level with concurrent access from all nodes.

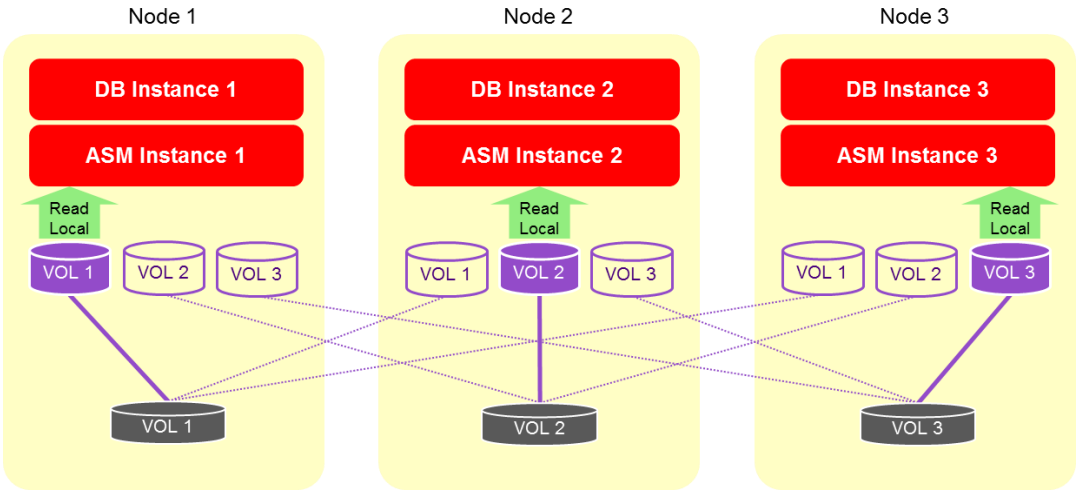

Figure 5. FlashGrid Storage Fabric with FlashGrid Read-Local Technology

Each database node has a full copy of user data stored on Amazon EBS volume(s) attached to that database node. The FlashGrid Read-Local™ Technology allows serving all read I/O from the locally attached volumes 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.

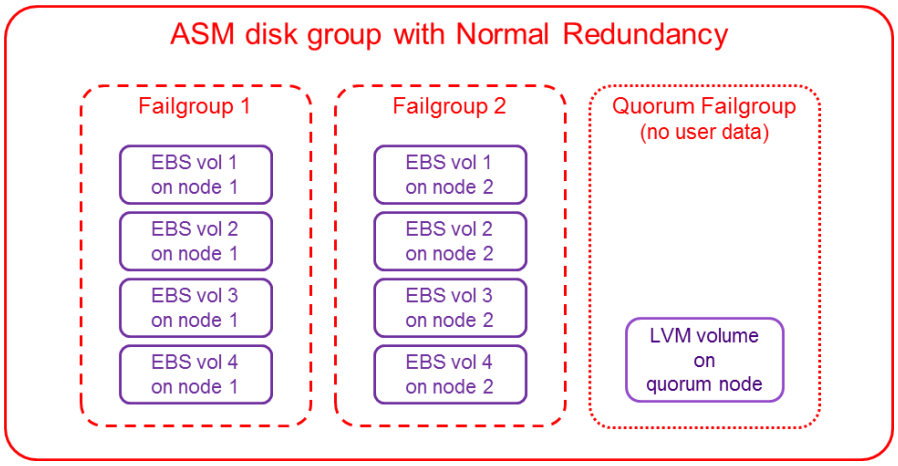

Figure 6. Example of a Normal Redundancy disk group in a 2-node RAC cluster

A typical Oracle RAC setup in Amazon EC2 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.

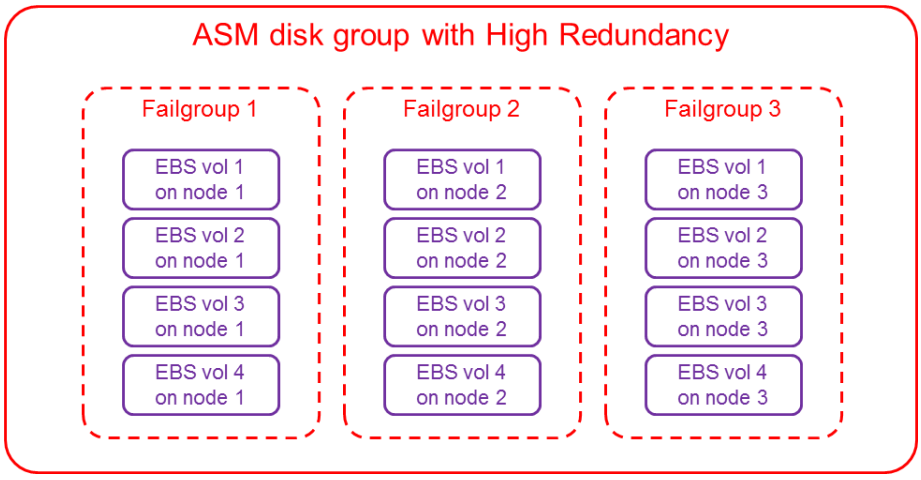

Figure 7. Example of a High Redundancy disk group in a 3-node RAC cluster

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.

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. Additionally, when using instance types that do not occupy an entire physical host, there is a chance that two instances may be running on the same physical host. A failure of that physical host will cause database downtime. There is no such risk for large instance types occupying an entire physical host, e.g. m4.10xlarge, m4.16xlarge, or r4.16xlarge.

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 latencies and, in certain cases, lower network bandwidth between the nodes.

Data Availability

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.

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: optimal memory to CPU ratio
- R4: high memory to CPU ratio, high peak network bandwidth
- I3: high memory to CPU ratio, high peak network bandwidth, local NVMe SSDs
- X1: large memory size, large number of CPU cores

Oracle Database Standard Edition 2 customers can use 2-node RAC clusters with up to 4 vCPUs per node, e.g. m4.xlarge or r4.xlarge instance types. More details about licensing Oracle software in the cloud computing environment: http://www.oracle.com/us/corporate/pricing/cloud-licensing-070579.pdf

Quorum servers require fewer resources than RAC. However, the above criteria are still important to ensure stable cluster operation. c4.large or m4.large instances can be used as quorum servers. Using T2 family for quorum servers is not recommended. 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 performance in the following ways:

- Network bandwidth between instances in different Availability Zones is limited to 5 Gb/s. Network speeds above 5 Gb/s are available only within a single Placement Group. A Placement Group is limited to a single Availability Zone. This constraint must be taken into account when using instances that normally have network bandwidth higher than 5 Gb/s.
- Latency between instances is increased. In the US-West-2 region for 8KB transfers we measured 0.6 ms between Availability Zones compared to 0.1 ms within a single Availability Zone.

The impact of inter-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 Volume Types

Use of Provisioned IOPS SSD (io1) volumes is recommended for data in order to achieve best performance. Each io1 volume can be configured with up to 20,000 IOPS.

General Purpose SSD (gp2) volumes are recommended for OS, GRID disk group, and quorum volumes. Normally, use of gp2 volumes for data is not recommended because of variability in their performance. However, for a volume of 1TB or larger the performance will be 3000 IOPS guaranteed. If this level of performance is sufficient then 1TB gp2 volumes can be used. Note that using N x 1TB volumes has N times performance advantage over using one N TB volume while the cost is the same.

Local SSDs

Use of local SSDs as the primary storage offers higher bandwidth and lower cost compared to Amazon EBS volumes. For example, the newly announced i3 instance type includes NVMe SSDs with up to 16GB/s and up to 3.3 mln IOPS. The use of local SSDs is not covered in this revision of this document, it is planned for future revisions.

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 EC2. 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 configuration:

- Two database nodes, M4.16xlarge
- Four io1 20000 IOPS 400GB volumes per node

Test script:

```
SET SERVEROUTPUT ON;
DECLARE
  lat INTEGER;
  iops INTEGER;
  mbps INTEGER;
BEGIN DBMS_RESOURCE_MANAGER.CALIBRATE_IO (8, 10, iops, mbps, lat);
DBMS_OUTPUT.PUT_LINE ('max_iops = ' || iops);
DBMS_OUTPUT.PUT_LINE ('latency = ' || lat);
DBMS_OUTPUT.PUT_LINE ('max_mbps = ' || mbps);
end;
/
```

Our results:

```
max_iops = 154864
latency = 0
max_mbps = 2219
```

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 and the private interconnect latency.

Test configuration:

- 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

The table below shows our results for tests performed in the same configuration (provided above) in one Availability Zone and different Availability Zones.

	Same AZ	**Different AZs**
Read+Write Database Requests (both nodes combined)	121,237 IOPS	111,546 IOPS
Read Database Requests (both nodes combined)	100,539 IOPS	92,081 IOPS
Write Database Requests (both nodes combined)	20,697 IOPS	19,465 IOPS
Interconnect Ping Avg Latency 500B msg	0.23 ms	0.69 ms
Interconnect Ping Latency Stddev 500B msg	0.05 ms	0.06 ms
Interconnect Ping Avg Latency 8K msg	0.27 ms	0.77 ms
Interconnect Ping Latency Stddev 8K msg	0.07 ms	0.15 ms

While there is a roughly 3X increase on the interconnect latency between the Availability Zones (and similar increase on the write latency), the difference in throughput is less than 10%. With over 100K IOPS in both cases, the performance results are comparable to having a dedicated all-flash storage array.

Compatibility

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

- Oracle Database: ver. 12.1.0.2 or 11.2.0.4 with the latest PSU
- Oracle Grid Infrastructure: ver. 12.1.0.2 with the latest PSU
- Operating System: Oracle Linux 7.3, Red Hat Enterprise Linux 7.3, CentOS 7.3
- FlashGrid software: ver. 17.01

The solution can be deployed on any EBS-Optimized Amazon EC2 instance type with Enhanced Networking.

Deployment Process

Below is a brief overview of the steps for automatic provisioning of an Oracle RAC cluster on Amazon EC2.

1) Optionally customize an AMI provided by FlashGrid or create your own AMI
2) Define cluster configuration: number of nodes, instance types, storage volumes, etc.
3) Generate CloudFormation template file
4) Create CloudFormation stack using the template file
5) Create a database

Conclusion

Running Oracle RAC clusters on Amazon EC2 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 wide 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