flashgrid

Mission-critical databases in the cloud. Oracle RAC on Amazon EC2 enabled by FlashGrid® Cluster engineered cloud system.

White Paper

rev. 2023-02-15

© 2017-2023 FlashGrid Inc.

Abstract

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. FlashGrid Cluster is an engineered cloud system designed for database high availability.

By leveraging the proven Oracle RAC database engine, FlashGrid Cluster enables the following use-cases:

- Lift-and-shift migration of existing Oracle RAC databases to AWS.
- Migration of existing Oracle databases from on-premises to AWS without reducing uptime SLA.
- Design of new mission critical applications for the cloud using the proven database engine.

This paper provides architectural overview of FlashGrid Cluster for Oracle RAC on AWS. It can be used for planning and designing high availability database deployments on Amazon Elastic Compute Cloud (Amazon EC2).

Architecture overview

FlashGrid Cluster is delivered as a fully integrated Infrastructure-as-Code template that can be customized and deployed to your AWS account in a few clicks.

Key components of FlashGrid Cluster for Oracle RAC on AWS include:

- Amazon EC2 instances
- Amazon EBS storage
- FlashGrid Storage Fabric™ software
- FlashGrid Cloud Area Network™ software
- Oracle Grid Infrastructure software (includes Oracle Clusterware and Oracle ASM)
- Oracle RAC database engine.

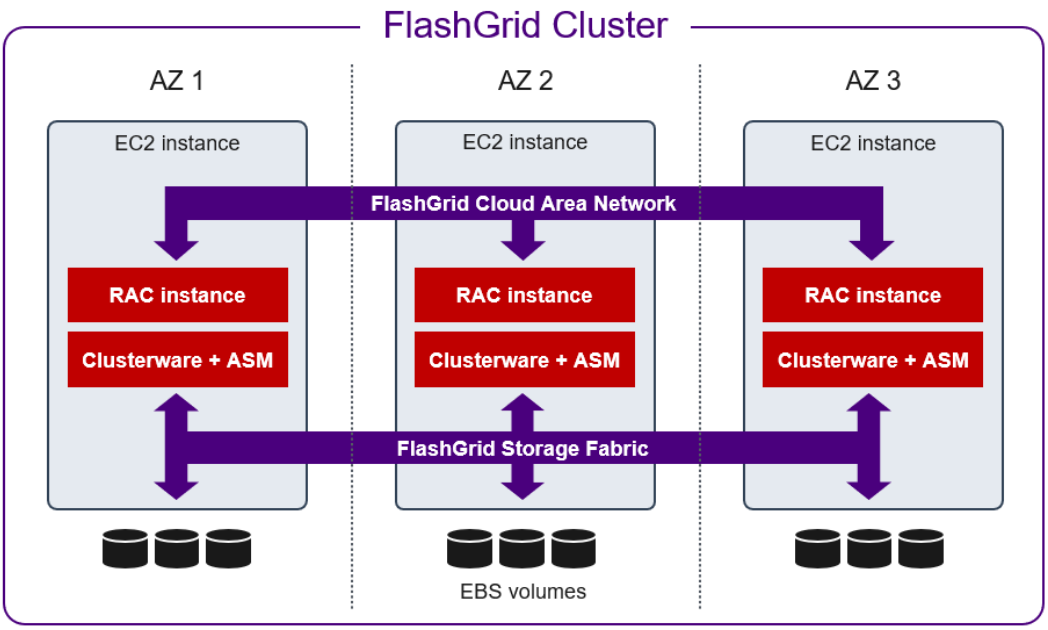

Figure 1. FlashGrid Cluster software architecture

FlashGrid Cluster architecture highlights:

- Active-active database HA with Oracle RAC and 2 or more database nodes.
- No single point of failure.
- Zero RPO and near-zero RTO for maximum uptime SLA.
- Spreading RAC database nodes across availability zones (multi-AZ) protects against failures affecting an entire data center.
- FlashGrid Cloud Area Network™ software enables high-speed overlay networks with advanced capabilities for HA and performance management.
- FlashGrid Storage Fabric™ software turns EBS volumes attached to individual EC2 instances into shared disks accessible from all nodes in the cluster.
- FlashGrid Read-Local™ Technology minimizes storage network overhead by serving reads from locally attached EBS volumes.
- 2-way or 3-way mirroring of data across separate nodes and availability zones.
- Oracle ASM and Clusterware provide data protection and availability.

Advantages of an Oracle RAC database engine

Oracle RAC provides an advanced technology for database high availability. Many organizations, such as financial institutions and telecom operators, use Oracle RAC to run their mission-critical applications that have the strictest requirements for uptime and data integrity.

Oracle RAC has an active-active distributed architecture with shared database storage. Shared storage plays a central role in enabling zero RPO, near-zero RTO, and maximum application uptime. These HA capabilities minimize outages due to unexpected failures, as well as during planned maintenance.

Multi-AZ architecture options

Amazon Web Services consists of multiple independent *regions*. Each region is partitioned into several availability zones. Each availability zone consists of one or more discrete data centers housed in separate facilities, each with redundant power, networking, and connectivity. Availability zones are physically separate, such that even extremely uncommon disasters such as fires 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.

Spreading cluster nodes across multiple availability zones helps to avoid downtime even when an entire availability zone experiences a failure. FlashGrid recommends using multi-AZ cluster configurations unless there is a specific need to use a single availability zone.

Typical cluster configurations

FlashGrid Cluster enables a variety of RAC cluster configurations on Amazon EC2. 2 or 3-node clusters are recommended in most cases. Clusters with four or more nodes can be used for extra-large (200+ TB) databases.

Multiple databases can share one FlashGrid Cluster as separate databases or as pluggable databases in a multitenant container database. For larger databases and for high-performance databases, dedicated clusters are typically recommended for minimizing interference.

It is also possible to use FlashGrid Cluster to run single-instance databases with automatic failover, including Standard Edition High Availability (SEHA).

Two RAC database nodes

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

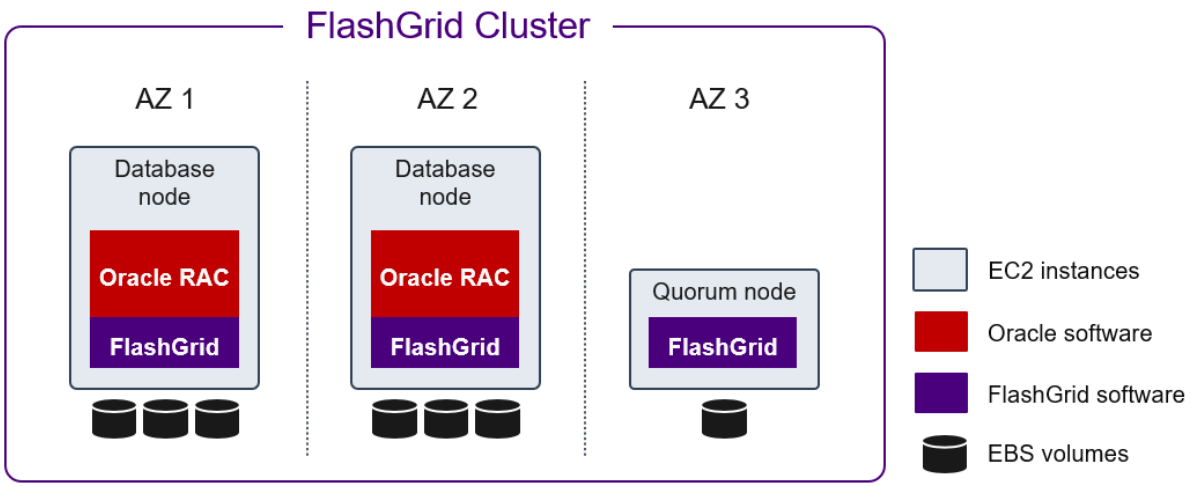

Figure 2. FlashGrid Cluster with two RAC database nodes

Three RAC database nodes

Clusters with three RAC database nodes have 3-way data mirroring using high redundancy ASM disk groups. Two additional EC2 instances (*quorum* nodes) are required to host quorum disks. Such a cluster can tolerate the loss of any two nodes without database downtime.

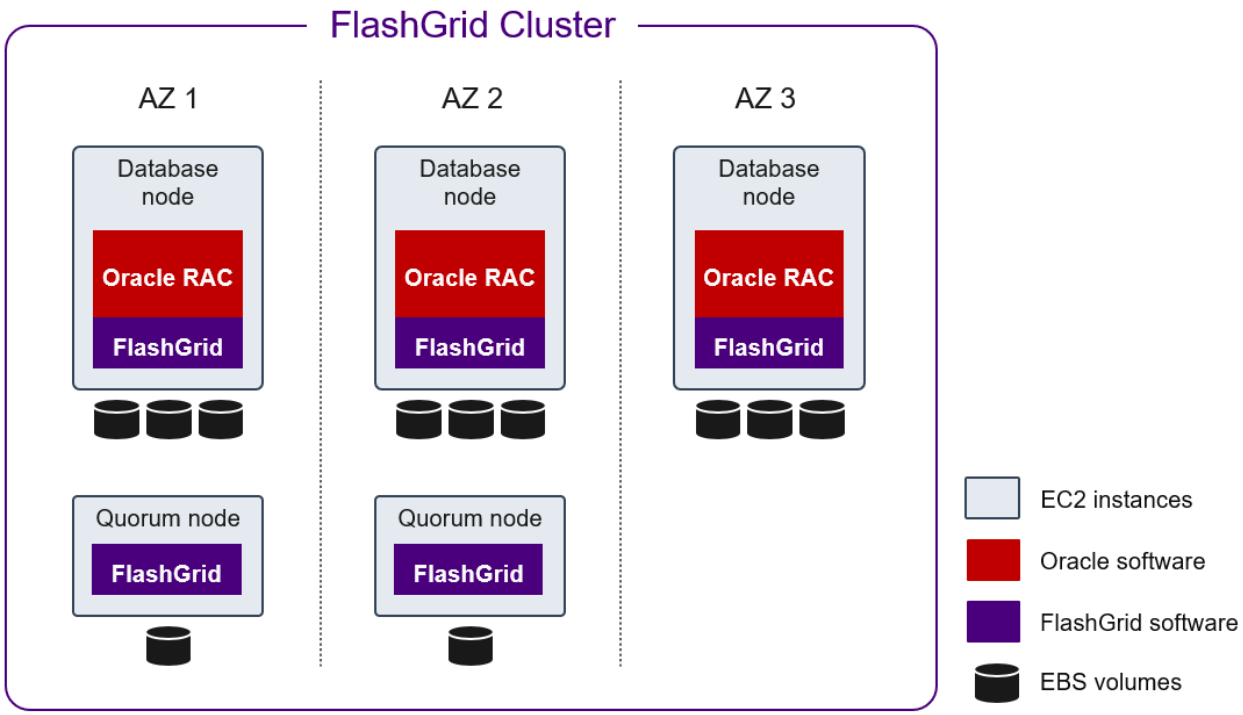

Figure 3. FlashGrid Cluster with three RAC database nodes

Many AWS regions have only three availability zones. Because of this, placing the quorum nodes in separate availability zones may not be possible. However, with three RAC nodes spanning three availability zones, placing the quorum nodes in the same availability zones as the RAC nodes still achieves the expected HA capabilities. Such a cluster can tolerate the loss of any two nodes or the loss of any one availability zone without incurring database downtime.

4+ RAC database nodes, single AZ

Extra-large (200+ TB) databases or databases requiring extreme performance may benefit from having four or more RAC database nodes and separate storage nodes. In this architecture the EBS storage volumes are attached to the storage nodes only. The storage volumes are shared with the RAC database nodes over the high-speed network.

Each RAC database node can achieve up to 30,000 MBPS of storage throughput. Each storage node can provide up to 10,000 MBPS of throughput.

ASM disk groups are configured with either Normal Redundancy (2-way mirroring), or High Redundancy (3-way mirroring). This provides protection against the loss of either one or two storage nodes respectively.

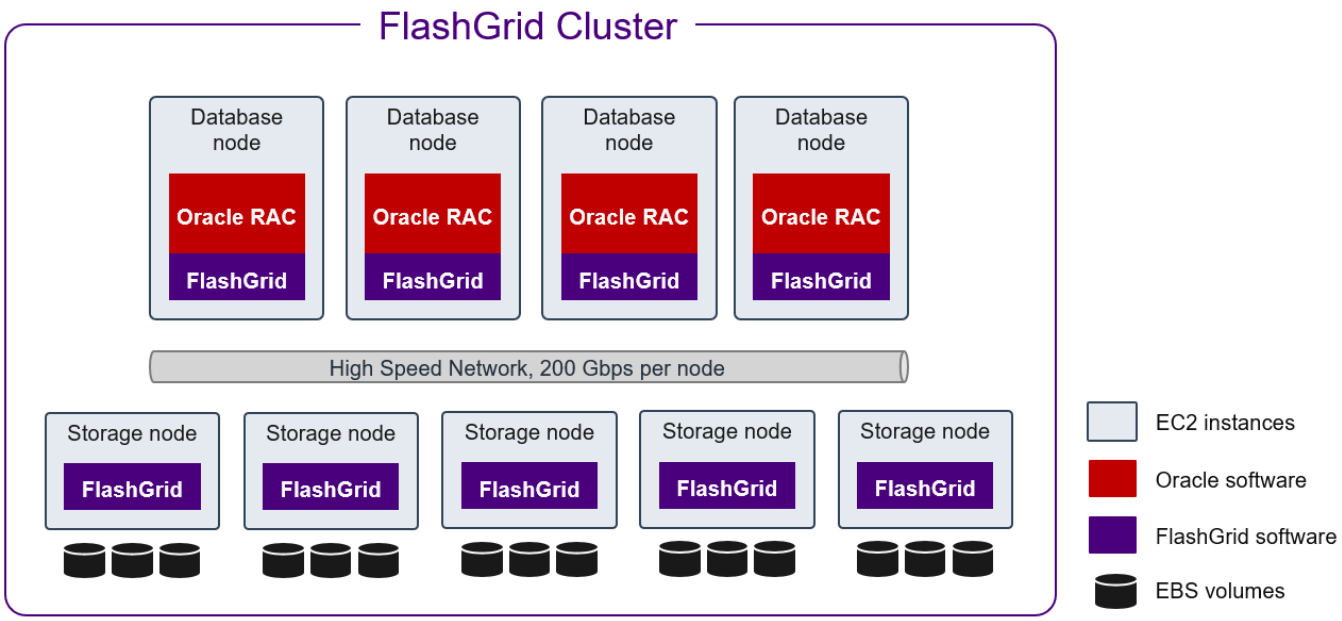

Figure 3. Extra-large database cluster with 4+ RAC nodes and separate storage nodes

4+ RAC database nodes, multi-AZ

It is possible to configure a cluster with four or more RAC database nodes across three availability zones. The database nodes are spread across two availability zones. The third availability zone is used for a *quorum* node. Such a cluster can tolerate the loss of an entire availability zone.

ASM disk groups are configured with either Normal Redundancy (2-way mirroring), or Extended Redundancy (4-way mirroring). This provides protection against the loss of either one, or three storage nodes respectively.

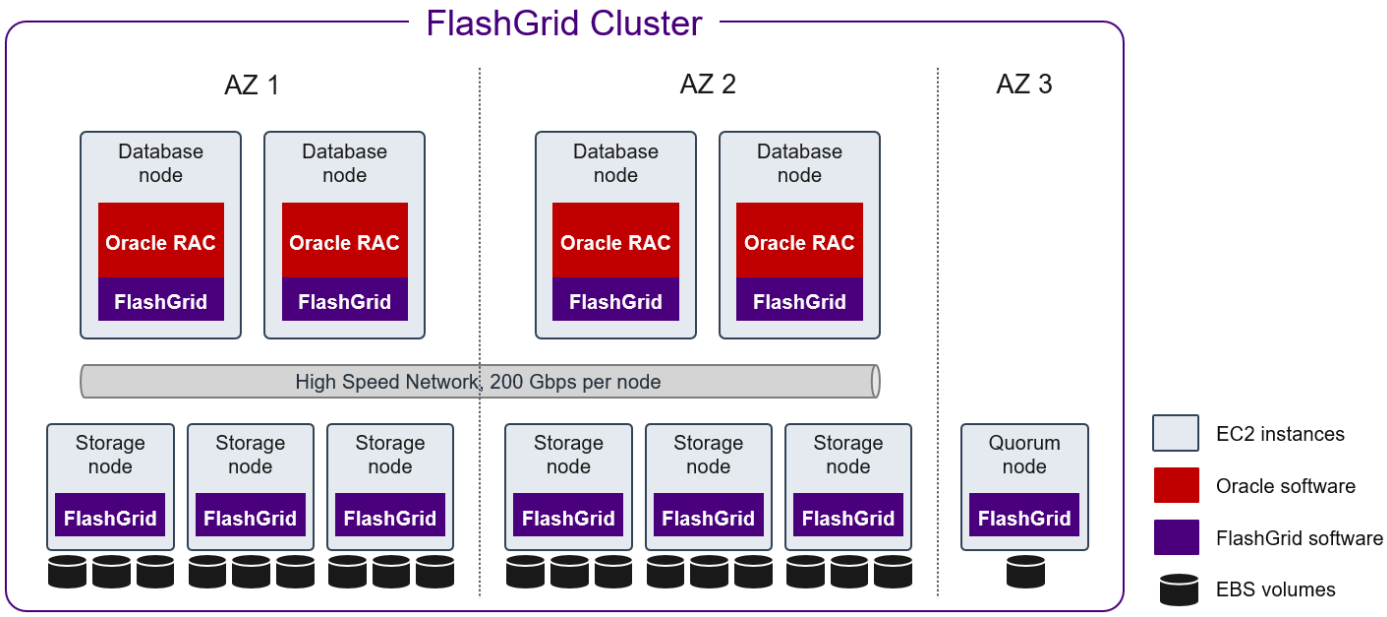

Figure 4. Extra-large database cluster with multi-AZ

Network architecture

The standard network connecting Amazon EC2 instances is essentially a Layer 3 (Internet Protocol) network with a fixed amount of bandwidth allocated per instance for all types of traffic. However, Oracle RAC architecture requires separate networks for client connectivity (a.k.a. *public network*) and for the private cluster interconnect (a.k.a. *private network*) between the cluster nodes. Additionally, Oracle RAC requires a network with multicast capability, which is not available in Amazon EC2.

FlashGrid Cloud Area Network™ (CLAN) software addresses the gaps in the EC2 networking capabilities by creating a set of high-speed virtual LAN networks and ensuring quality of service (QoS) between them.

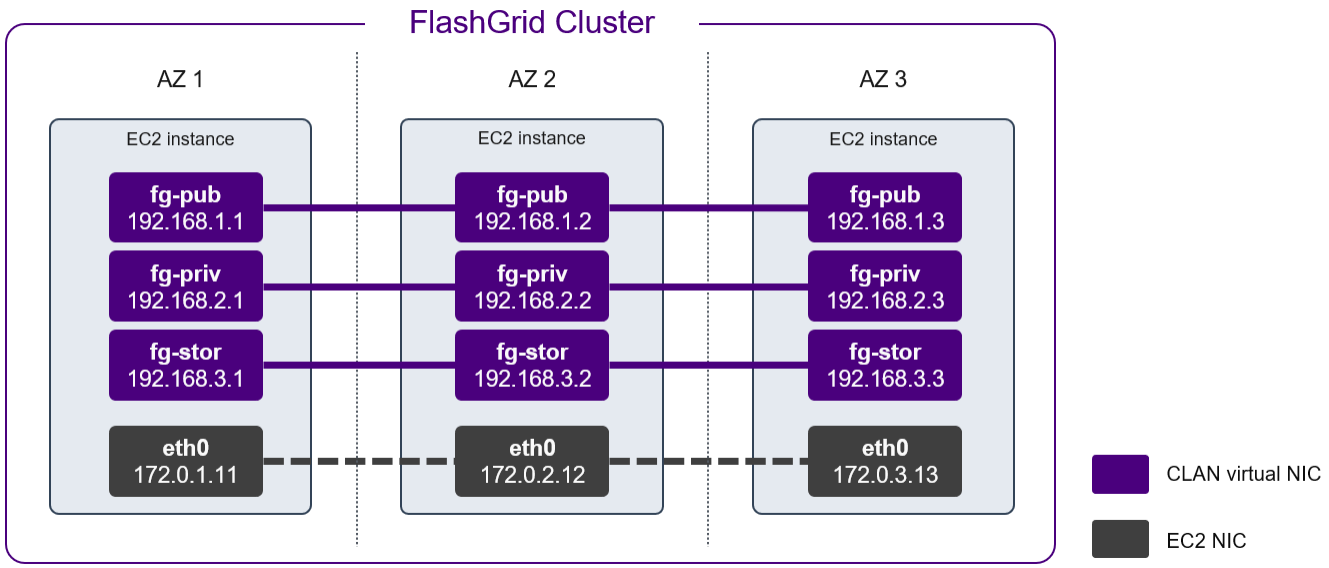

Figure 5. FlashGrid Cloud Area Network architecture

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

- Transparent layer 2 connectivity between cluster nodes and across availability zones
- Each type of traffic has its own virtual LAN with a separate virtual NIC, e.g. *fg-pub*, *fg-priv*, *fg-storage*
- Guaranteed bandwidth allocation for each traffic type
- Negligible latency overhead compared to the raw network
- Low latency of the cluster interconnect in the presence of large volumes of traffic of other types
- Multicast support
- Up to 170 Gbps total bandwidth per node (depends on the EC2 instance type and size)

Shared storage architecture

FlashGrid Storage Fabric software turns local disks into shared disks accessible from all nodes in the cluster. 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.

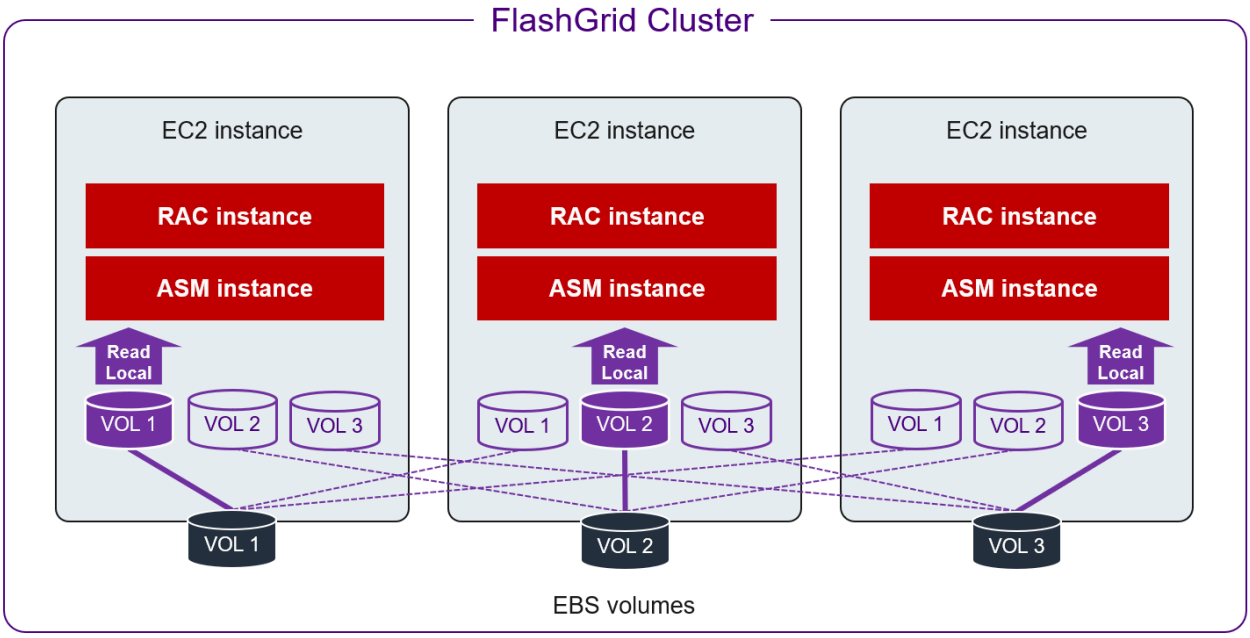

Figure 6. FlashGrid Storage Fabric with FlashGrid Read-Local Technology

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 locally attached disks. This significantly improves both read and write I/O performance. Read requests avoid the extra network hop, thus reducing latency and the amount of 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 – typically 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.

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 to store a third copy of ASM metadata for better failure handling.

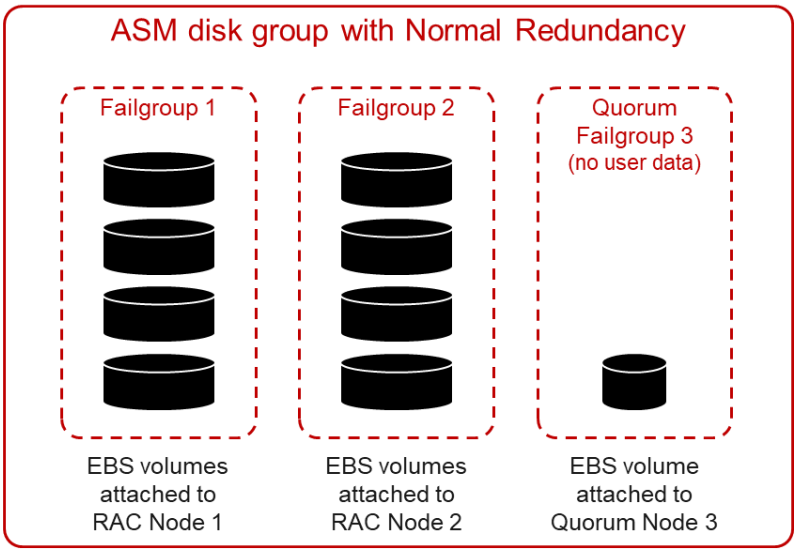

Figure 7. Example of a Normal Redundancy disk group in a 2-node RAC cluster

In a 3-node cluster all disk groups must have *High Redundancy* 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 to store additional copies of ASM metadata for better failure handling.

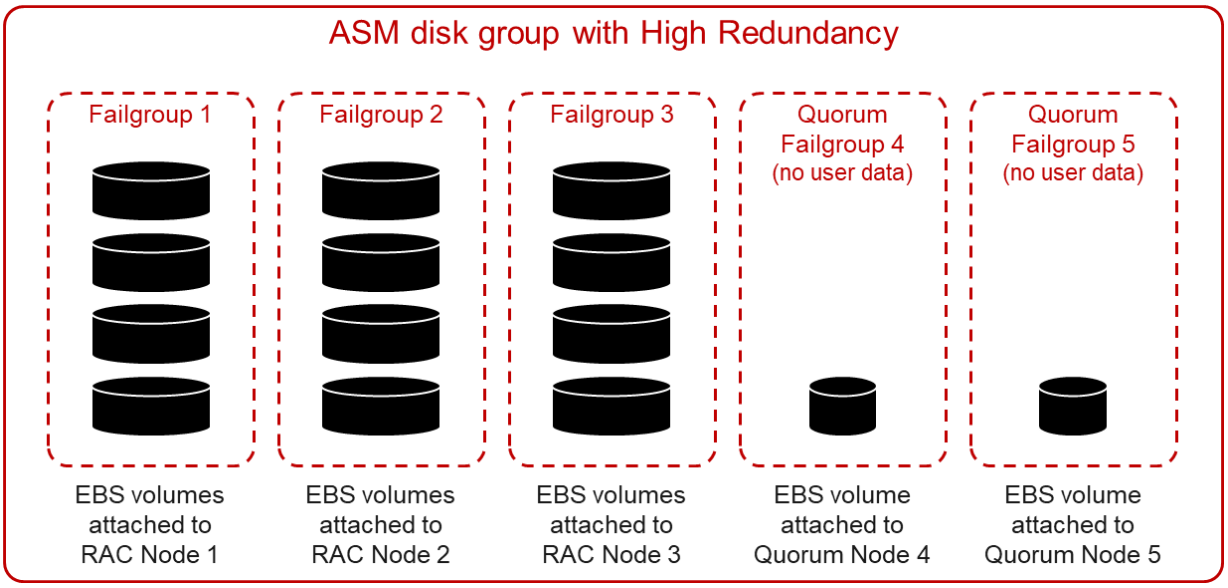

Figure 8. Example of a High Redundancy disk group in a 3-node RAC cluster

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 EC2 instances can move between physical hosts, a 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-AZ configuration provides protection against loss of a database node. It is an efficient way to accommodate planned maintenance (e.g. database or OS patching) without causing database downtime. However, the 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 minimizes the risk of simultaneous node failures.

Near-zero RTO

Thanks to the active-active HA, when a RAC node fails, the other RAC node(s) keep providing access to the database. The client sessions can fail over transparently for the application. There is essentially no interruption of data access, except for the short period (seconds) required to detect the failure.

Data availability with EBS Storage

An Amazon EBS volume provides persistent storage that survives a failure of the EC2 instance that the volume is attached to. After the failed instance restarts on a new physical node, all its volumes are re-attached with no data loss.

Amazon EBS volumes have built-in redundancy that protects data from failures in the underlying physical media. ASM performs data mirroring on top of the built-in protection of Amazon EBS. Together, Amazon EBS and ASM's mirroring provide durable storage with two layers of data protection, which exceed the typical levels of data protection in on-premises deployments.

Zero RPO

Data is mirrored across 2+ nodes in a synchronous manner. In case a node fails, no committed data is lost.

Performance considerations

Multiple availability zones

Using multiple availability zones (AZs) provides substantial availability advantages. However, it does increase network latency because of the distance between the AZs. The network latency between AZs is less than 1ms in most cases and will not have critical impact on performance of many workloads. For example, 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.

Read-heavy workloads will experience zero or little impact because all read traffic is served locally and does not use the network between AZs.

Note that differences in latency between different pairs of AZs provides an opportunity for optimization by choosing which AZs to place database nodes in. For example, in a 2-node RAC cluster, it is optimal to place database nodes in the two AZs with the lowest latency between them. See our knowledge base article for more details.

Storage performance

In most cases, the use of General Purpose SSD (gp3) volumes is recommended. The performance of each gp3 volume can be configured from 3,000 to 16,000 IOPS and 125 to 1,000 MBPS. By using multiple volumes per disk group attached to each database node, the per node throughput can reach the maximum of 350,000 IOPS and 10,000 MBPS (with r6in.24xlarge or r6in.metal instances).

With multiple nodes in a cluster, read throughput is further multiplied. In a 2-node cluster read throughput can reach 700,000 IOPS and 20,000 MBPS. In a 3-node cluster, read throughput can reach 1,050,000 IOPS and 30,000 MBPS.

Performance vs. on-premises solutions

EBS GP3 storage is SSD based and provides an order of magnitude of improvement in IOPS and latency over traditional spinning HDD based storage arrays. With up to 350,000 IOPS and 10,000 MBPS per node, the performance is even higher than a typical dedicated all-flash storage array. It is important to note that the storage performance is not shared between multiple clusters. Every cluster has its own dedicated set of EBS volumes, which ensures stable and predictable performance with no interference from noisy neighbors.

An extra-large database architecture, that uses R6in or M6in instances and separate storage nodes, provides up to 30,000 MBPS of storage throughput per RAC database node. Thus, enabling the deployment of extra-large (200+ TB) databases and migrations from large Exadata systems.

Reference performance results

When moving database workloads to the cloud, the main areas of concern regarding performance tend to be around storage and network I/O. Because the CPU performance overhead between bare-metal and VMs is close to zero, here we will focus instead on storage I/O and RAC interconnect I/O.

Calibrate_IO

The CALIBRATE_IO procedure provides a convenient way to measure storage performance, including maximum bandwidth, random IOPS, and latency. The CALIBRATE_IO procedure generates I/O through the database stack on actual database files. It is also useful for comparing the performance between two storage systems because CALIBRATE_IO's results are not influenced by non-storage factors such as memory size or number of CPU cores.

The test is read-only and safe to run on an existing database. However, do not run it on a production system because it will cause severe performance degradation of the applications using the database.

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;
/
```

Calibrate_IO results measured by FlashGrid:

Cluster configuration	Max IOPS	Latency, ms	Max MBPS
EBS storage, 2 RAC nodes (r5b.24xlarge)	456,886	0.424	14631
EBS storage, 3 RAC nodes (r5b.24xlarge)	672,931	0.424	21934
Local SSD storage, 2 RAC nodes (i3en.24xlarge)	1,565,734	0.103	24207

Note that Calibrate_IO's results are not influenced by whether the database nodes are in the same availability zone or not.

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 a mixed (read+write) I/O load. Database AWR reports generated during a SLOB test provide various performance metrics but we will focus on I/O performance.

SLOB results measured by FlashGrid:

Cluster configuration	Physical Write Database Requests	Physical Read Database Requests	Physical Read+Write Database Requests
EBS storage, 2 RAC nodes, single AZ	40,425 IOPS	372,481 IOPS	412,906 IOPS
EBS storage, 2 RAC nodes, multi-AZ	39,172 IOPS	364,264 IOPS	403,436 IOPS
EBS storage, 3 RAC nodes, multi-AZ	54,927 IOPS	507,230 IOPS	562,157 IOPS
Local SSD storage, 2 RAC nodes, multi-AZ	104,646 IOPS	988,426 IOPS	1,093,072 IOPS

Test configuration details

- SGA size: 3 GB (We chose a smaller SGA to minimize caching effects and maximize physical I/O)
- 8KB database block size
- 240 schemas, 240MB each
- SLOB UPDATE_PCT= 10 (10% updates, 90% selects)
- Database nodes:
  - EBS storage configurations: r5b.24xlarge
  - Local SSD storage configuration: i3en.24xlarge
- Disks:
  - EBS storage configurations: (20) gp3 volumes per node, 16000 IOPS, 1000 MBPS each
  - Local SSD storage configuration: (8) 7500 GB local SSDs

Optimizing Oracle Database licenses

For customers on a per-CPU Oracle licensing model, optimizing the number of Oracle licenses may be an important part of managing costs. With FlashGrid Cluster, the following options are available to optimize Oracle Database licensing:

- **The X2iezn instance type** has smaller number of CPU cores at higher frequency (up to 4.5 GHz) paired with a large memory size and high network bandwidth. Separate storage nodes can also be used for extra storage throughput.
- **Bare-metal instances** allow the use of server hardware without the virtualization layer.
- **Dedicated hosts** allow management of licenses at the physical host level instead of the VM level. Multiple VM instances (belonging to different clusters) can share the same physical host and associated licenses.
- **Consolidating** multiple smaller databases on a single cluster (and scheduling CPU intensive jobs at different times) has the potential for reducing the total number of CPU cores.
- **Optimizing** system and software configuration often allows reducing CPU consumption, thus reducing the required number of CPU cores.

Disaster Recovery strategy

An optimal Disaster Recovery (DR) strategy for Oracle databases will depend on the higher-level DR strategy for the entire application stack.

In a Multi-AZ configuration, FlashGrid Cluster provides protection against a catastrophic failure of an entire data center. However, it cannot protect against a region-wide outage or against an operator error causing destruction of the cluster resources. The most critical databases may benefit from having one or more replicas as part of the DR strategy. The most common replication tool is (Active) Data Guard but there are other tools that can be used.

The replica(s) may be placed in a different region and/or in the same region:

- **Remote standby** in a different region protects against a region-wide outage or disaster. Asynchronous replication should be used.
- **Local standby** in the same region protects against a logical destruction of a database cluster caused by an operator error, software bugs, or malware. Synchronous replication should be used for zero RPO.
- A combination of both remote and local standby may be used for most critical systems.

A single-instance (non-RAC) database may be used as a standby replica. However, using an identical clustered setup for the standby provides the following benefits:

- Consistent performance in case of a DR scenario.
- Ability to routinely switch between the two replicas.
- Ability to apply software updates and configuration changes on the standby first.

Security and control

System and data access

FlashGrid Cluster is deployed on EC2 instances in the customer's AWS account and managed by the customer. The deployment model is similar to running your own EC2 instances and installing FlashGrid software on them. FlashGrid staff has no access to the systems or data.

System control

Customer assigned administrators have full (root) access to the EC2 instances and the operating system. Additional 3rd party monitoring or security software can be installed on the cluster nodes for compliance with corporate or regulatory standards.

OS hardening

OS hardening can be applied to the database nodes (as well as to quorum/storage nodes) for security compliance. Customers may choose to use their own hardening scripts or FlashGrid's scripts that are available for CIS Server Level 1 aligned hardening.

Data Encryption

All data on EBS storage can be encrypted. By default, encryption is enabled and uses AWS managed keys. Optionally, customer managed symmetric KMS encryption key can be used.

Oracle Transparent Data Encryption (TDE) can be used as a second layer of data encryption if the corresponding Oracle license is available.

TCPS

Customers requiring encrypted connectivity between database clients and database servers can configure TCPS for client connectivity.

Compatibility

Software versions

The following versions of software are supported with FlashGrid Cluster:

- Oracle Database: ver. 19c, 18c, 12.2, 12.1, or 11.2
- Oracle Grid Infrastructure: ver. 19c
- Operating System: Oracle Linux 7/8, Red Hat Enterprise Linux 7/8

EC2 instance types

The following EC2 instance types are typically recommended for database nodes:

- R6i: high memory to CPU ratio
- M6i: higher CPU to memory ratio
- R6in, M6in: high speed network, best for extra-large database clusters
- X2idn, X2iedn, X2iezn: highest memory to CPU ratio, high speed network
- i3en: local SSDs for extreme storage performance

Database nodes must have at least four physical CPU cores (8 vCPUs with hyperthreading) to ensure reliable operation.

Quorum nodes require fewer resources than database nodes, a single CPU core is sufficient. The c6i.large instance type is recommended for quorum servers. 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.

Database features

FlashGrid Cluster does not restrict the use of any database features. DBAs can enable or disable database features based on their requirements and available licenses.

Database tools

Various database tools from Oracle or third parties can be used with Oracle RAC databases running on FlashGrid Cluster. This includes RMAN and RMAN-based backup tools, Data Guard, GoldenGate, Cloud Control (Enterprise Manager), Shareplex, DBvisit, and AWS DMS.

Shared file systems

The following shared file access options can be used with FlashGrid Cluster:

- ACFS or DBFS for shared file access between the database nodes.
- Amazon EFS or NFS can be mounted on database nodes for sharing files with other systems, e.g. application servers.
- File based access to S3.

Automated Infrastructure-as-Code deployment

The FlashGrid Launcher tool automates the process of deploying a cluster. It 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 databases. Human errors that could lead to costly reliability problems and compromised availability are avoided by the use of automatically generated and standardized Infrastructure-as-Code templates.

Terraform

If Terraform is preferred as a deployment tool then the CloudFormation template can be embedded into a Terraform template.

Native Terraform template support is expected by March 2023.

Generating templates via REST API

The entire deployment process can be fully automated without needing to manually use the FlashGrid Launcher's web GUI, by using its REST API instead to generate CloudFormation templates.

Conclusion

FlashGrid Cluster engineered cloud systems offer 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

Copyright © 2017-2023 FlashGrid Inc. All rights reserved.

This document is provided for information purposes only, and the contents hereof are subject to change without notice. This document is not warranted to be error-free, nor subject to any other warranties or conditions, whether expressed orally or implied in law, including implied warranties and conditions of merchantability or fitness for a particular purpose. We specifically disclaim any liability with respect to this document, and no contractual obligations are formed either directly or indirectly by this document.

Nothing in this document shall be interpreted as an advice pertaining to licensing of any third-party software products, including the Oracle Database family of products. It is the responsibility of a third-party software licensee to maintain compliance with all applicable licensing terms and conditions. FlashGrid Inc does not sell, distribute, or provide access to Oracle Database software licenses.

FlashGrid is a registered trademark of FlashGrid Inc. Amazon and Amazon Web Services are registered trademarks of Amazon.com Inc. and Amazon Web Services Inc. Oracle and Java are registered trademarks of Oracle and/or its affiliates. Red Hat is a registered trademark of Red Hat Inc. Other names may be trademarks of their respective owners.