flashgrid

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

White Paper

rev. 2024-08-02

Google Cloud Partner

Abstract

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 Google Cloud.
- Migration of existing Oracle databases from on-premises to Google Cloud without reducing uptime SLA.
- Design of new mission critical applications for the cloud using the proven database engine.

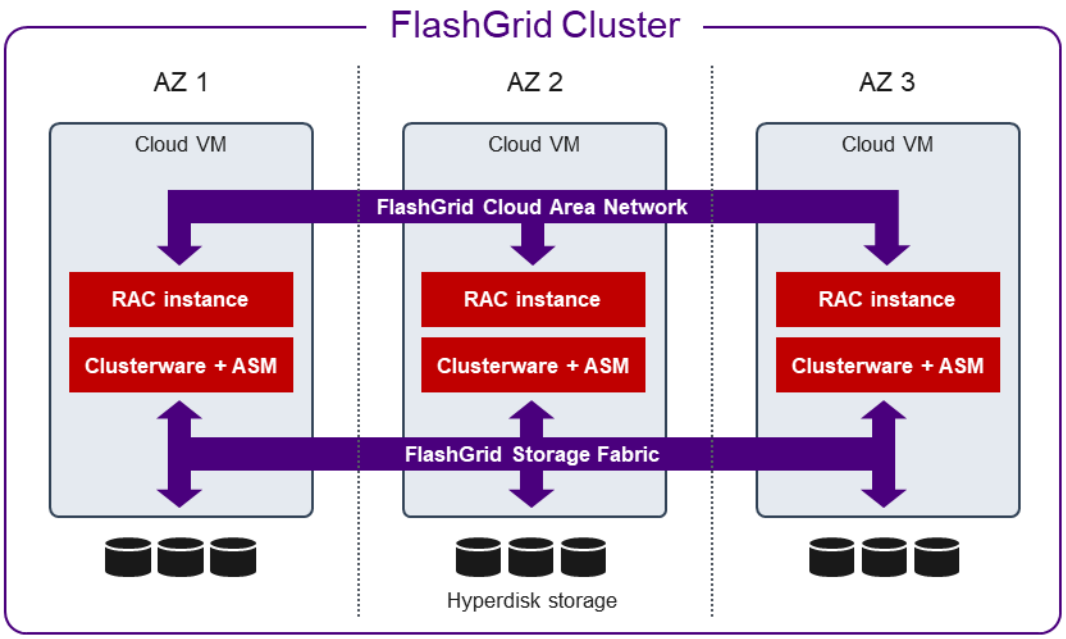
This paper provides architectural overview of FlashGrid Cluster for Oracle RAC on Google Cloud. It can be used for planning and designing high availability database deployments on Google Cloud.

Architecture Overview

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

Key components of FlashGrid Cluster for Oracle RAC on Google Cloud include:

- Google Cloud Engine (GCE) Virtual Machines
- Google Cloud block storage (Hyperdisk or Persistent Disks)
- FlashGrid Storage Fabric software
- FlashGrid Cloud Area Network software
- Oracle Grid Infrastructure software (includes Oracle Clusterware and Oracle ASM)
- Oracle RAC database engine



FlashGrid Cluster for Oracle RAC on Google Cloud: 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 block storage disks (Hyperdisk or Persistent Disk) attached to individual VMs into shared disks accessible from all nodes in the cluster.
- FlashGrid Read-Local™ Technology minimizes storage network overhead by serving reads from locally attached block storage disks (Hyperdisk or Persistent Disk).
- 2-way or 3-way mirroring of data across separate nodes and Availability Zones.
- Oracle ASM and Clusterware provide data protection and availability.

Why Oracle RAC Database Engine

Oracle RAC provides 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 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

Google Cloud consists of multiple independent *Regions*. Each Region is partitioned into 3+ *Availability Zones*. Each Availability Zone consists 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 1 ms. This makes the inter-AZ deployments compliant with the extended distance RAC guidelines.

Spreading cluster nodes across multiple Availability Zone helps 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

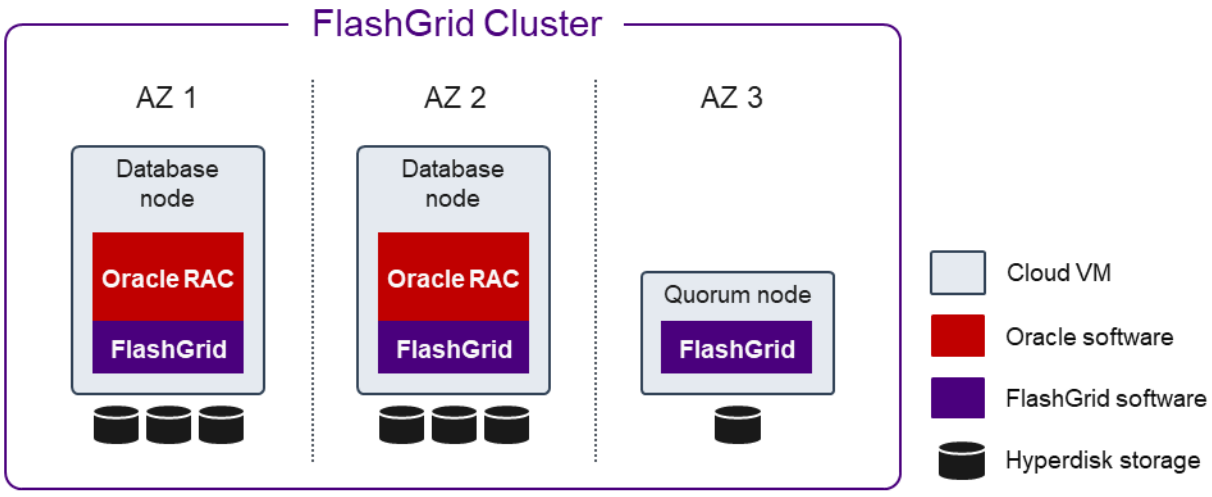
FlashGrid Cluster enables a variety of RAC cluster configurations on Google Cloud. Two or three node clusters are recommended in most cases. Clusters with four or more nodes can be used for extra-large (500+ TB) databases.

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 for running single-instance databases with automatic fail-over, 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 small VM (quorum node) is required to host quorum disks. Such cluster can tolerate loss of any one node without database downtime.

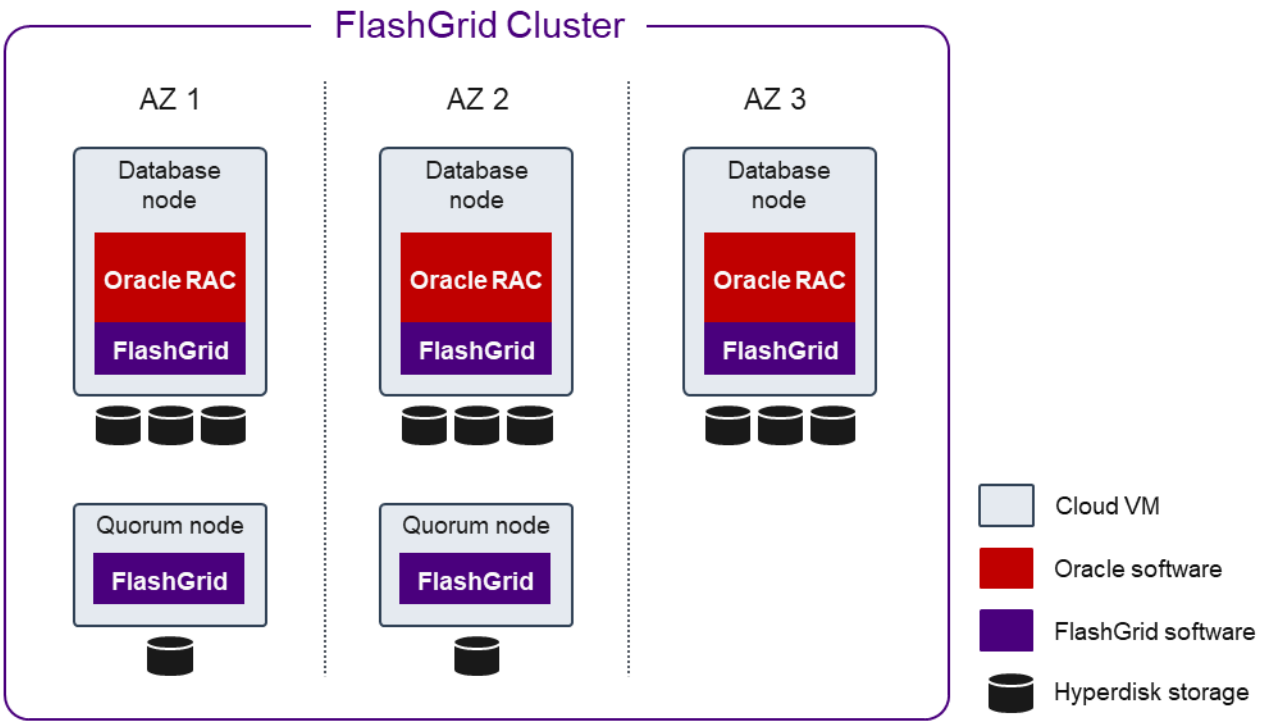


FlashGrid Cluster on Google Cloud with two RAC database nodes

Three RAC database nodes

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

The majority of Google Cloud regions have three availability zones. Because of this, placing the quorum nodes in separate availability zones is usually not possible. However, with three RAC nodes spanning three availability zones, placing the quorum nodes in the same availability zones as the RAC nodes still allows achieving the expected HA capabilities. Such a cluster can tolerate loss of any two nodes or loss of any one availability zone without database downtime.



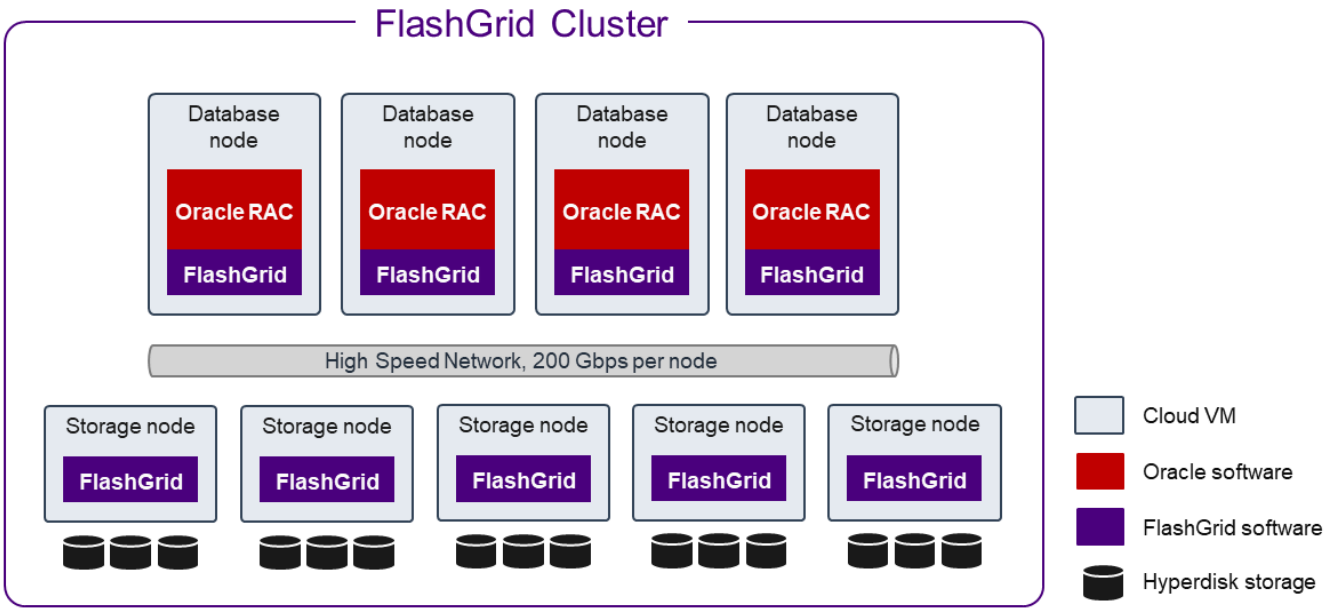
FlashGrid Cluster on Google Cloud with three RAC database nodes

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 block storage disks are attached to the storage nodes only. The disks are shared with the RAC database nodes over the high-speed network.

Each RAC database node can get up to 32,000 MBPS (VM sizes with 200 Gbps network) of storage throughput. Each storage node can provide up to 10,000 MBPS of throughput (e.g. *c3-highcpu-176 with Hyperdisk Balanced*).

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

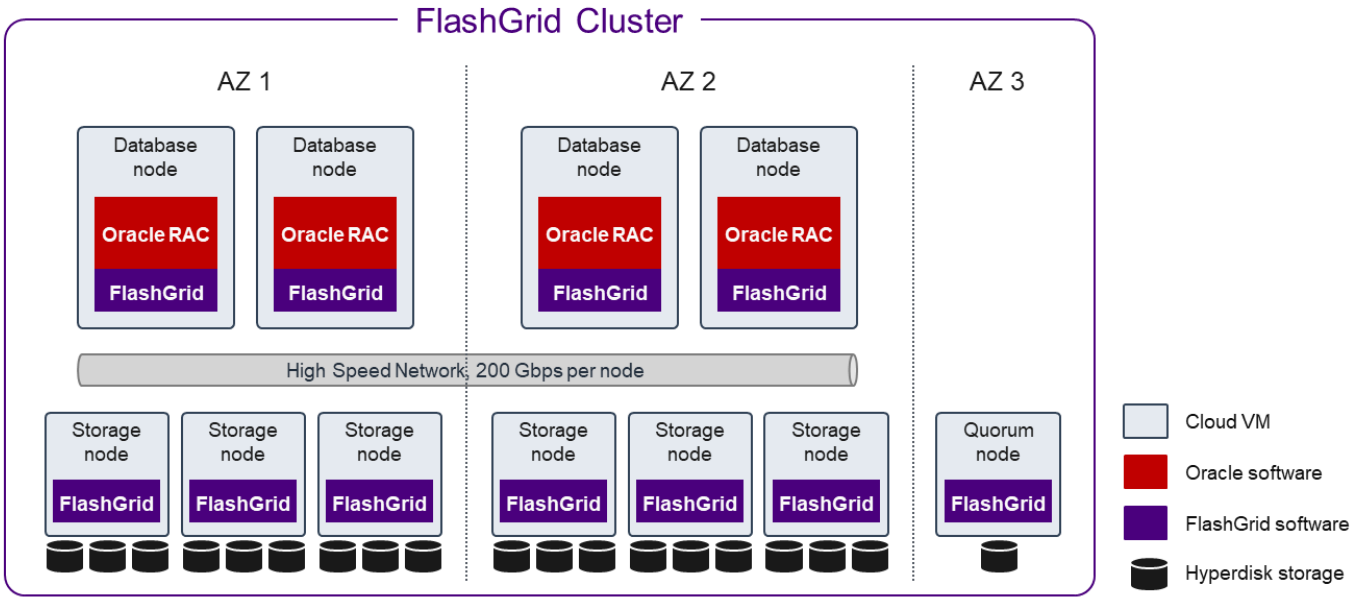


Extra-large database cluster on Google Cloud with 4+ RAC nodes and separate storage nodes

4+ RAC database nodes, multi-AZ

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

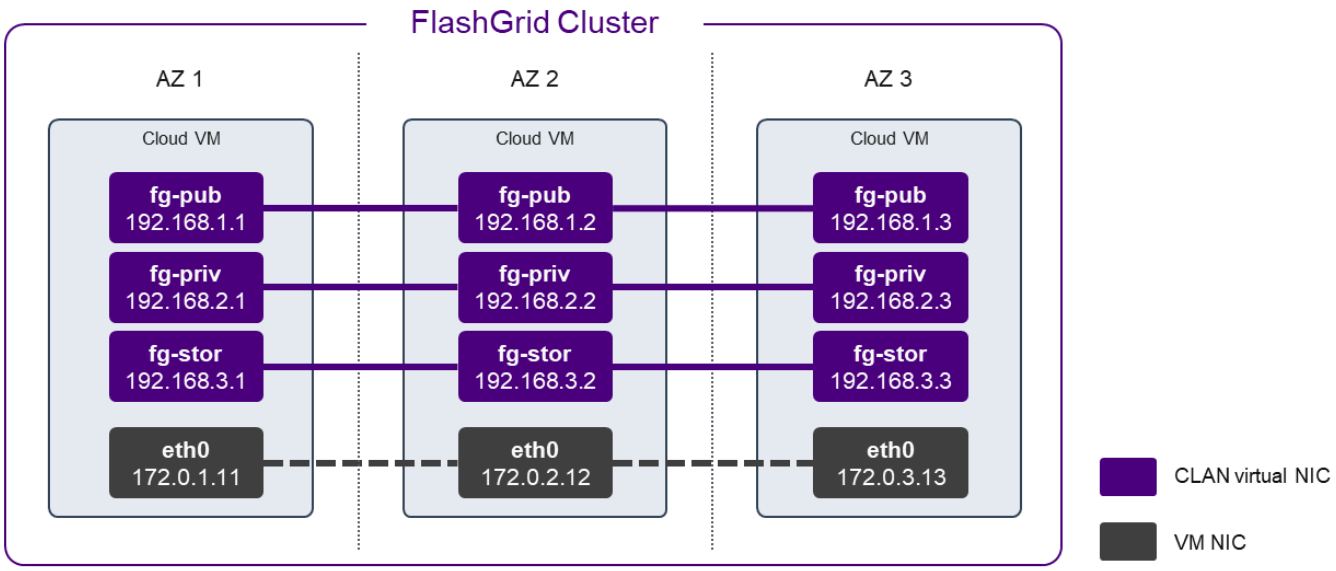


Extra-large database cluster with multi-AZ

Network Architecture

The standard network connecting GCE VMs is effectively a Layer 3 (Internet Protocol) 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 (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 Google Cloud.

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



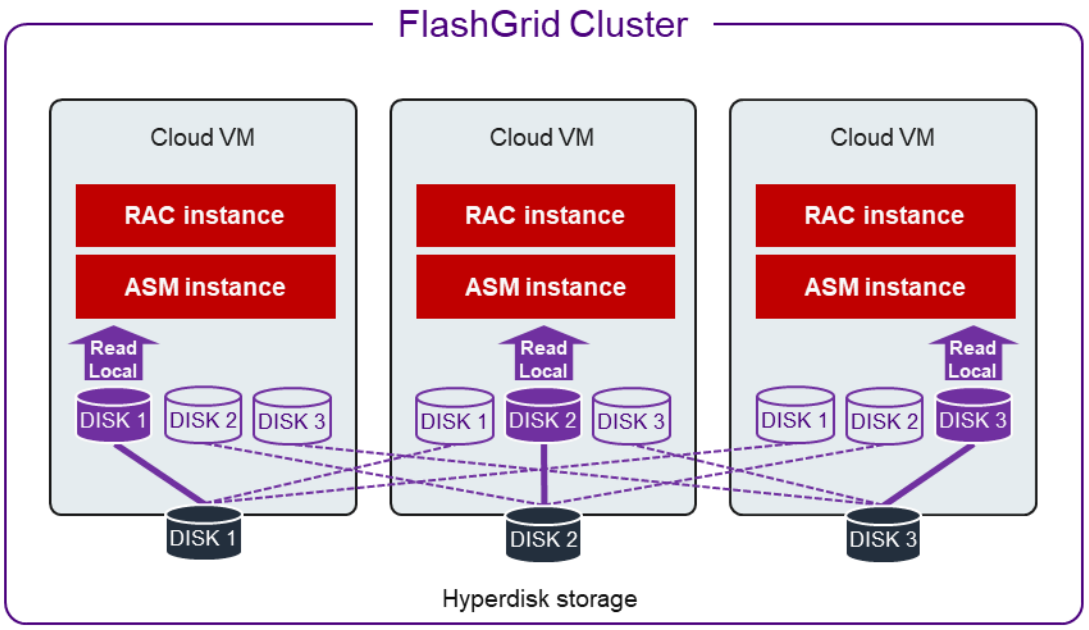
FlashGrid Cloud Area Network architecture on Google Cloud

Network capabilities enabled by FlashGrid CLAN for Oracle RAC on Google Cloud:

- 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 200 Gb/s total bandwidth per node (depends on the VM type and size)

Shared Storage Architecture

FlashGrid Storage Fabric software 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 Persistent disks. The sharing is done at the block level with concurrent access from all nodes.



FlashGrid Storage Fabric software architecture on Google Cloud

FlashGrid Read-Local Technology

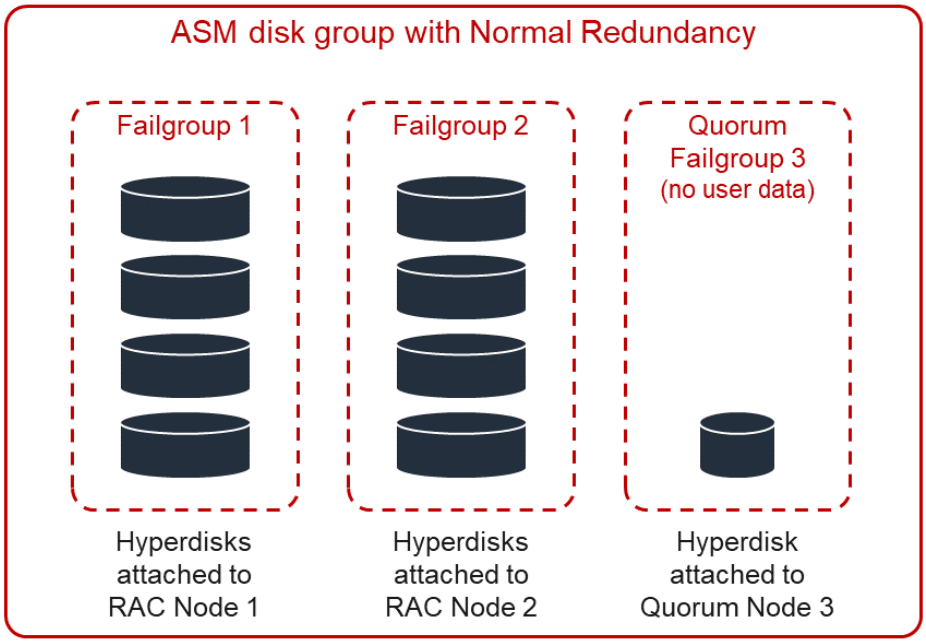
In 2-node or 3-node clusters each database node has a full copy of user data stored on Persistent disks attached to that database node. The FlashGrid Read-Local™ Technology allows serving all read I/O from the locally attached disks, which significantly improves both read and write I/O performance. The read requests avoid the extra network hop, thus reducing the 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.

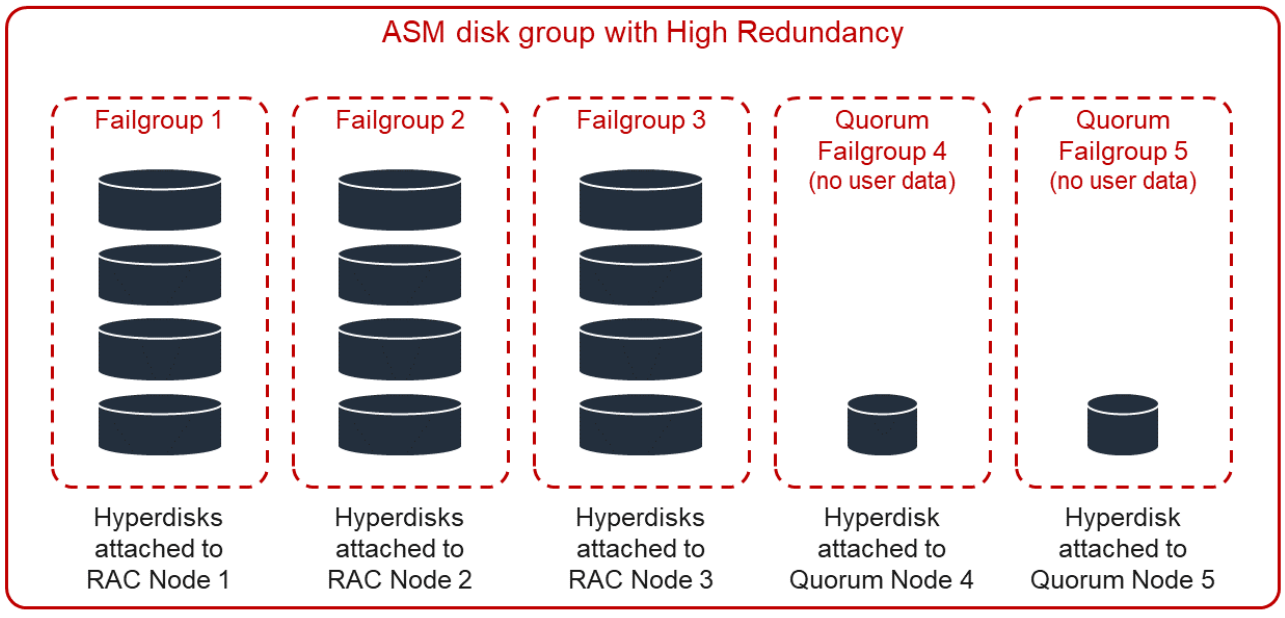
A typical Oracle RAC setup on Google Cloud will have three Oracle ASM disk groups: GRID, DATA, FRA.

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.



A Normal Redundancy disk group in a 2-node RAC cluster on Google Cloud

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.



A High Redundancy disk group in a 3-node RAC cluster on Google Cloud

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 GCE VMs 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 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.

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 virtually no interruption of data access except for a short period (seconds) required to detect the failure.

Data Availability

A Hyperdisk in Google Cloud 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.

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

Zero RPO

Data is mirrored across 2+ nodes in a synchronous manner. In case of a node failure, there is no loss of any committed data.

Performance Considerations

Multiple Availability Zones

Using multiple Availability Zones 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-east-1* region we measured latencies between all Availability Zones under 0.2 ms.

Read-heavy workloads will experience no or little impact because all read traffic is served locally and does not use the network between AZs.

Note that the different latency between different pairs of AZs provides an 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 lowest latency between them. See our knowledge base article for more details.

Storage Performance

Storage throughput per node can achieve up to 10,000 MBPS and 160,000 IOPS with C3 instance type and Hypedisk Balanced disks.

Read throughput is further multiplied with multiple nodes in a cluster. In a 2-node cluster read throughput can reach 320,000 IOPS and 20,000 MBPS. In a 3-node cluster read throughput can reach 480,000 IOPS and 30,000 MBPS.

For databases that require even higher storage throughput, multiple database nodes combined with multiple separate storage nodes may be used for achieving higher aggregate storage throughput.

Performance vs. on-premises solutions

The Hyperdisk storage is flash based and provides order of magnitude improvement in IOPS and latency compared to traditional spinning hard drive based storage arrays. With up to 160,000 IOPS and 10,000 MBPS per node, the performance is 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 Persistent disks, which ensures stable and predictable performance with no interference from noisy neighbors.

The extra-large database architecture using VMs with 200 Gbps network for RAC database nodes and multiple separate storage nodes provides up to 32,000 MBPS of storage throughput per RAC database node, thus enabling deployment of extra-large (500+ TB) databases, including migrations from large Exadata systems.

Disaster Recovery Strategy

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

FlashGrid Cluster in a Multi-AZ configuration 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 typical replication tool is (Active) Data Guard, but other replication tools may be used too.

The replica(s) can 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 GCE VMs in the customer's Google Cloud account and managed by the customer. The deployment model is similar to running your own GCE VMs and installing FlashGrid software on them. FlashGrid staff has no access to the systems or data.

System Control

Customer's assigned administrators have full (root) access to the GCE VMs and to 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 can choose to use their own hardening scripts or use FlashGrid provided scripts that are available for CIS Server Level 1 aligned hardening.

Data Encryption

All data on Persistent disks is encrypted at rest using Persistent disk Storage Server-Side Encryption (SSE).

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, 12.2, 12.1
- Oracle Grid Infrastructure: ver. 19c
- Operating System: Red Hat Enterprise Linux 8 or 9

Supported VM Types and Sizes

Database node VMs must have 4+ physical CPU cores (8+ vCPUs) and 32+ GB of memory. The following VM types are recommended for database nodes: C4, C3, C3D, M3.

Quorum nodes require fewer resources than database nodes, a single CPU core is sufficient. The c3-standard-2 (1 physical core) VM 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

Hyperdisk Balanced disks are recommended for the majority of deployments. With M3 VMs, Hyperdisk Extreme may be used to achieve the maximum MPBS throughput.

With older N2 or N2D VM types that do not support Hyperdisk Balanced, Persistent Disks may be used instead. *pd-balanced* is typically recommended in such cases. *pd-ssd*, *pd-standard* or *pd-extreme* may be considered in some special cases.

Only *zonal* disks are used. Data mirroring across availability zones is done at the ASM disk group level.

Database Features

FlashGrid Cluster does not restrict use of any database features. Customer's DBA can enable or disable database features based on the 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.

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.
- NFS can be mounted on database nodes for sharing files with other systems, e.g. application servers.
- File based access to object storage.

Automated Infrastructure-as-Code Deployment

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

The deployment process can be fully automated without the need to manually use the FlashGrid Launcher's web GUI. Instead, FlashGrid Launcher provides REST API for generating the Deployment Manager templates.

Conclusion

FlashGrid Cluster engineered cloud systems offer a wide range of highly available database cluster configurations on Google Cloud ranging from cost-efficient 2-node clusters to large high-performance clusters. Combination of the proven Oracle RAC database engine, Google Cloud 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