flashgrid

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

White Paper

rev. 2023-10-18

Microsoft Azure Certified

Microsoft Partner

Abstract

Microsoft Azure cloud provides IT organizations with the flexibility and elasticity that are not available in the traditional data center. With Azure 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 Azure.
- Migration of existing Oracle databases from on-premises to Azure without reducing uptime SLA.
- Design of new mission critical applications for the cloud using the proven database engine.

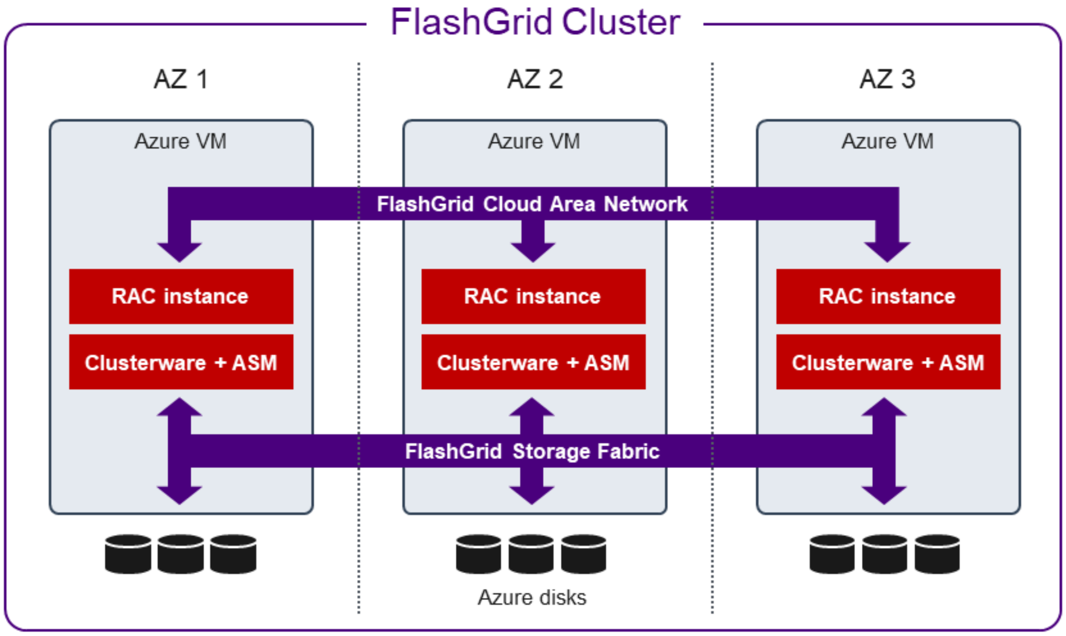
This paper provides architectural overview of FlashGrid Cluster for Oracle RAC on Azure. It can be used for planning and designing high availability database deployments on Azure.

Architecture overview

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

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

- Azure Virtual Machines
- Azure Managed Premium SSD block storage
- 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 Azure: 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 Azure disks 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 Azure disks.
- 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

Azure cloud consists of multiple independent *Regions*. Each Region (with some exceptions) 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 2ms. This makes the inter-AZ deployments compliant with the extended distance RAC guidelines.

Spreading cluster nodes across multiple Availability Zone 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.

Availability sets with fault domains can be used in those regions that do not currently support availability zones.

Typical cluster configurations

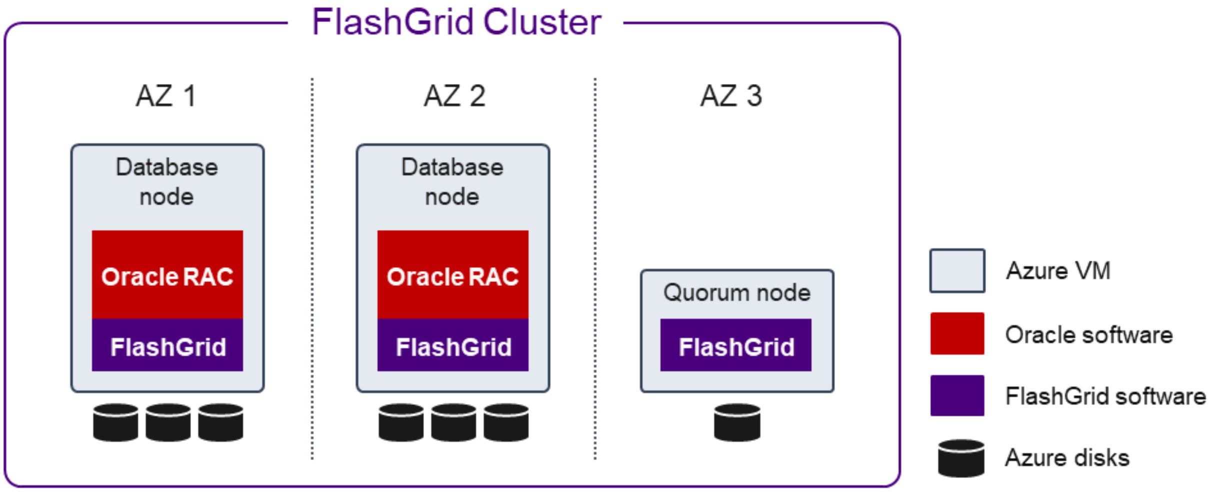
FlashGrid Cluster enables variety of RAC cluster configurations on Azure. 2 or 3-node clusters are recommended in most cases. Clusters with four or more nodes can be used for extra-large (200+ TB) databases.

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 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 the loss of any one node without incurring database downtime.

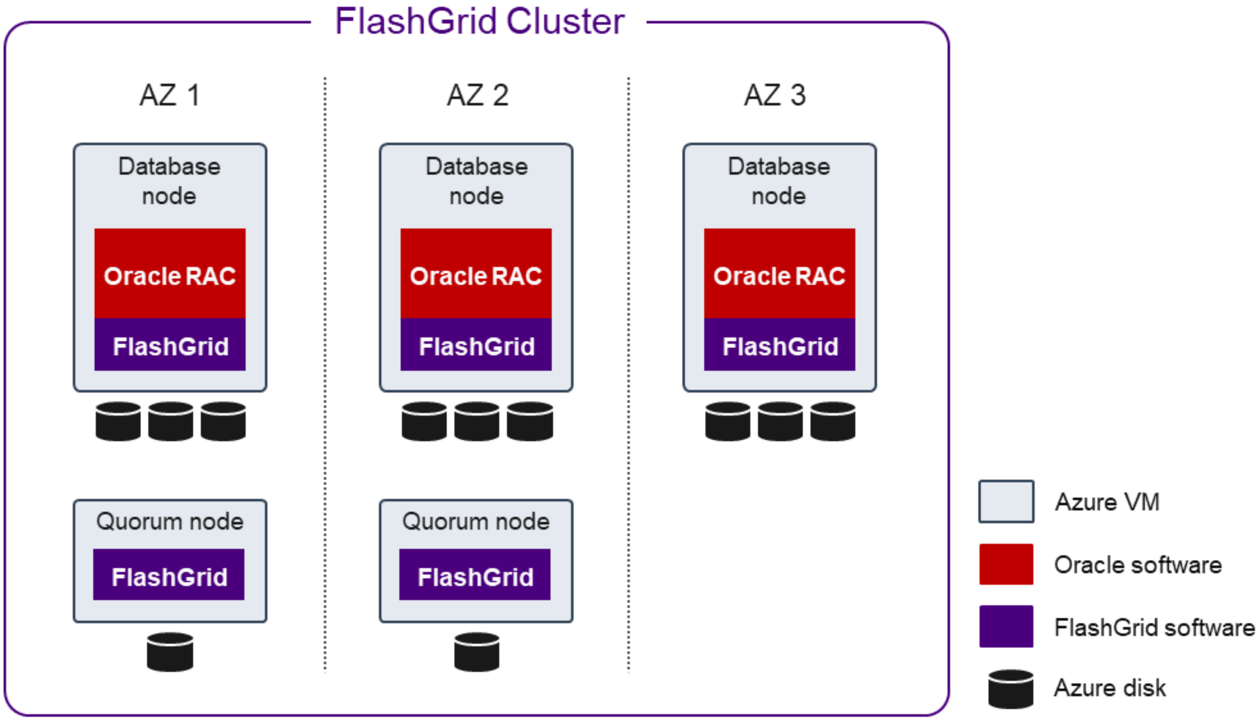


FlashGrid Cluster on Azure with two RAC database nodes

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.

Each Azure region has three availability zones. Because of this, placing the quorum nodes in separate availability zones is 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 the loss of any two nodes or loss of any one availability zone without incurring database downtime.



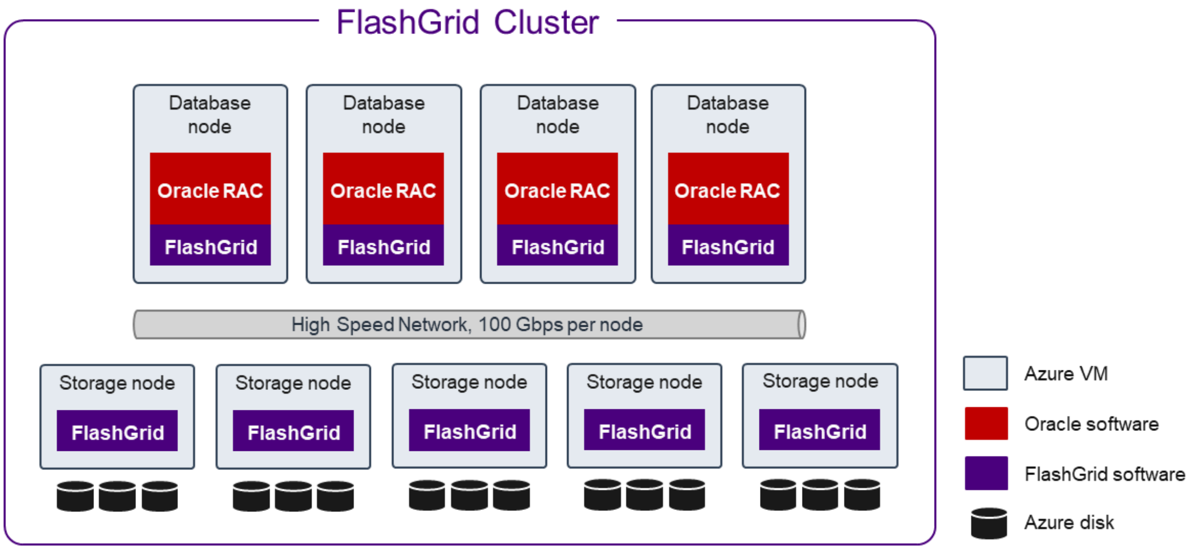
FlashGrid Cluster on Azure with three RAC database nodes

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 Azure 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 16,000 MBPS (E104is_v5 VM) of storage throughput. Each storage node can provide up to 4,000 MBPS (D96s_v5) of throughput.

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 respectively.

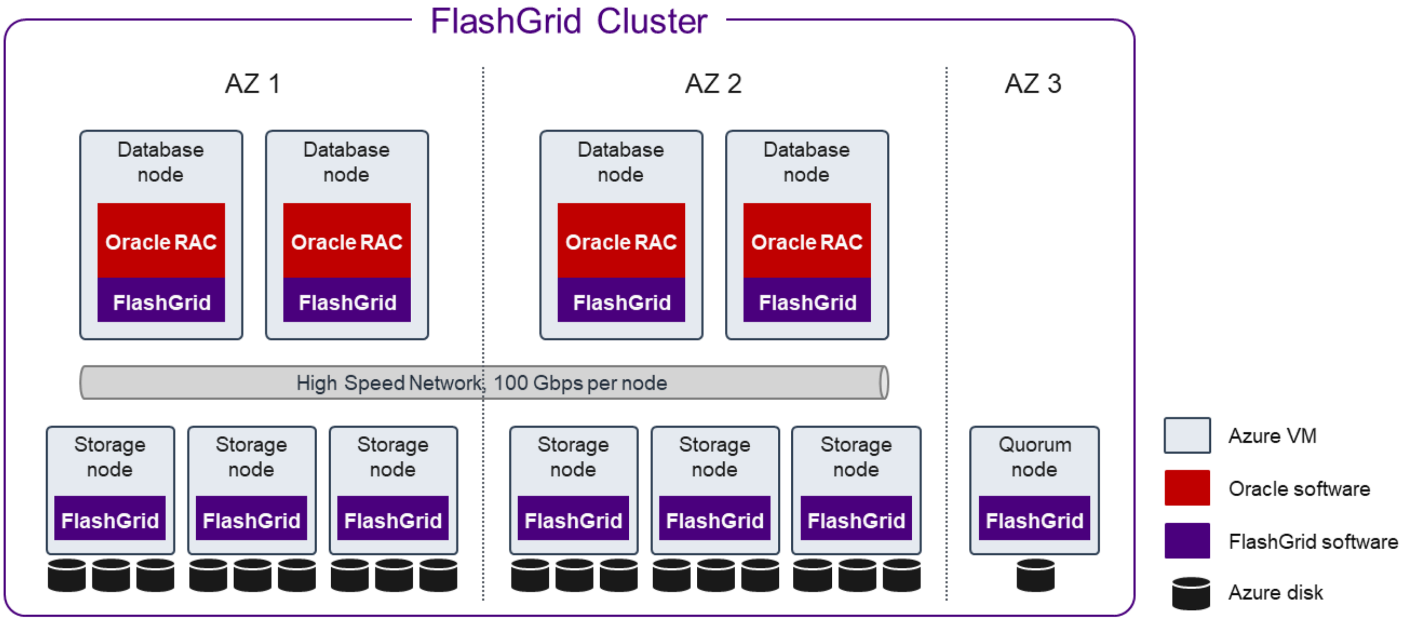


Extra-large database cluster on Azure with 4+ RAC nodes and separate storage nodes

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 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.

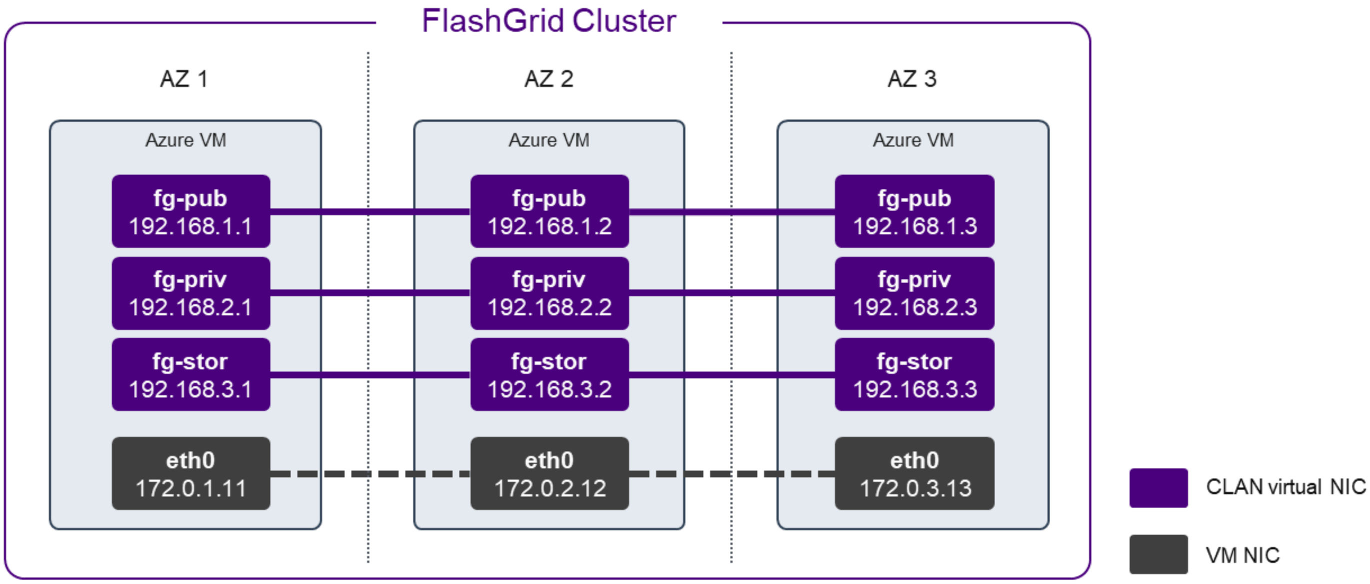


Extra-large Oracle RAC database cluster on Azure with multi-AZ

Network architecture

The standard network connecting Azure 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 Azure.

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



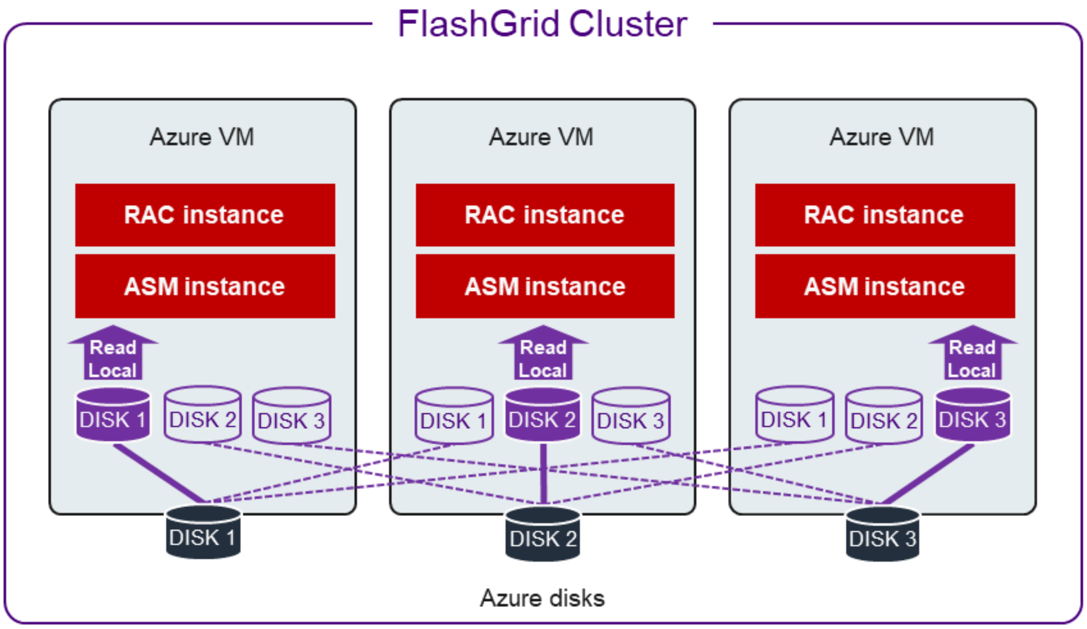
FlashGrid Cloud Area Network architecture on Azure

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

- 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 100 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 Azure disks. The sharing is done at the block level with concurrent access from all nodes.



FlashGrid Storage Fabric software architecture on Azure

FlashGrid Read-Local Technology

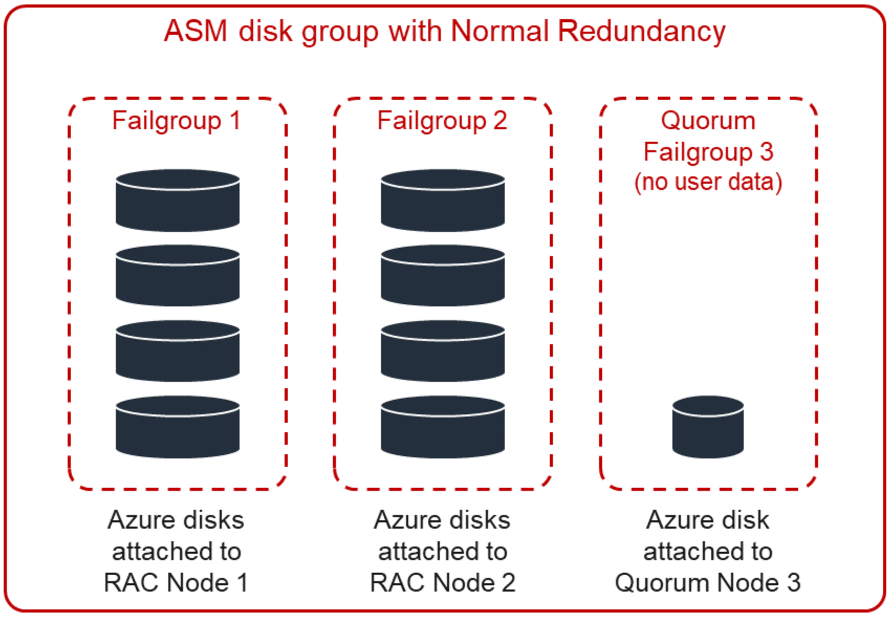
In 2-node or 3-node clusters each database node has a full copy of user data stored on Azure disks attached to that database node. The FlashGrid Read-Local™ Technology allows serving all read I/O from the locally attached disks, which significantly improves both read and write I/O performance. Read requests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

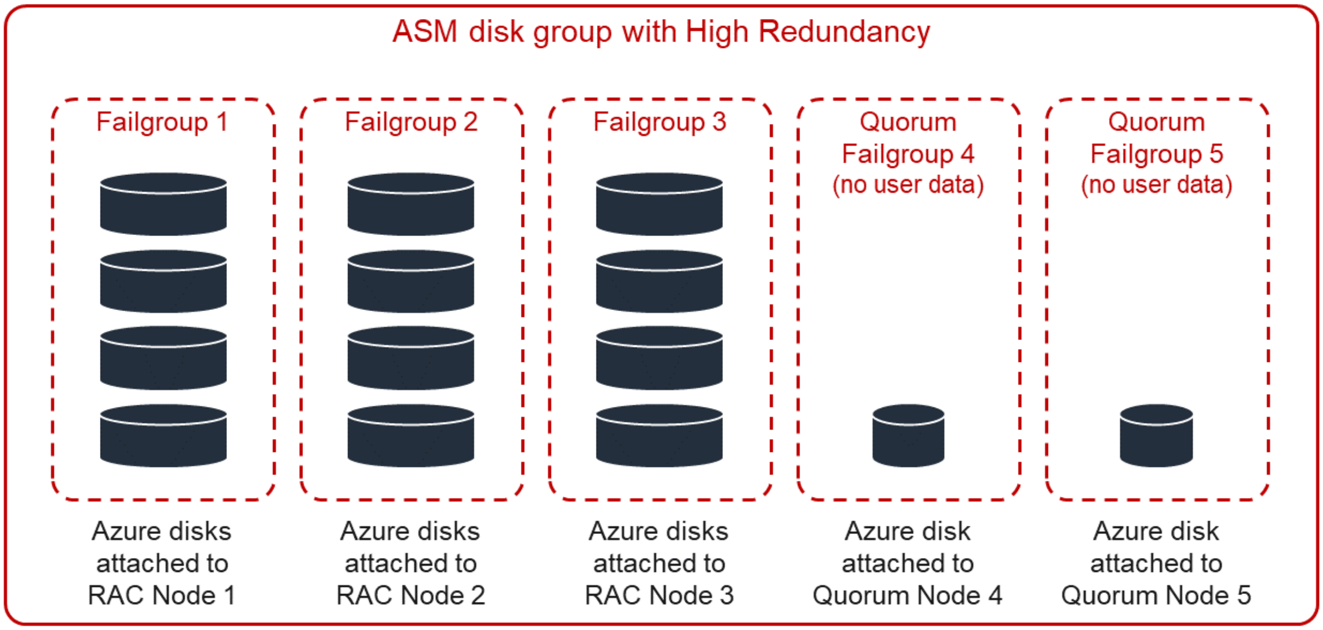
A typical Oracle RAC setup on Azure will have three Oracle ASM disk groups: GRID, DATA, FRA.

In a 2-node RAC cluster all disk groups must have Normal Redundancy. The GRID disk group containing voting files is required to have a quorum disk for storing a third copy of the voting files. Other disk groups also benefit from having 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 Azure

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 Azure

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

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

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

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

Because Azure VMs can move between physical hosts, the failure of a physical host causes only a short outage to the affected node. The node VM will automatically restart on another physical host. This significantly reduces the risk of double 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 virtually no interruption of data access except for a short period (seconds) required to detect the failure.

Data availability

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

Premium SSD disks 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 Premium SSD disks. Together Premium SSD disks and ASM 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 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 2ms in most cases and will not have critical impact on performance of many workloads. For example, in the *West US 3* region for 8KB transfers we measured 0.7 ms, 1.0 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 the 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

Each Premium SSD (v1) disk provides up to 20,000 IOPS and 900 MB/s depending on its size. The maximum performance of 20,000 IOPS is available for 32 TB disks. For databases that require high performance with smaller capacity, use of multiple 1024 GB or 2048 GB disks may be optimal to maximize the total IOPS and MB/s.

The new Premium SSD v2 provide up to 80,000 IOPS and 1200 MBPS. The IOPS and MBPS can be configured independent of the disk size.

By using multiple disks per disk group attached to each database node, the per node throughput can reach the maximum of 120,000 IOPS and 4,000 MBPS per VM with Esv5 and Ebsv5 VM types. (Additionally, the new NVME-enabled Ebsv5 VMs, which are currently in preview, will provide up to 260,000 IOPS and 8,000 MBPS).

Read throughput is further multiplied with multiple nodes in a cluster. In a 2-node cluster read throughput can reach 240,000 IOPS and 8,000 MBPS. In a 3-node cluster read throughput can reach 360,000 IOPS and 12,000 MBPS.

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

Performance vs. on-premises solutions

The Premium SSD storage is flash based and provides an order of magnitude improvement in IOPS and latency over traditional spinning HDD based storage arrays. With up to 120,000 IOPS and 4,000 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 Azure disks, which ensures stable and predictable performance with no interference from noisy neighbors.

The extra-large database architecture using the E104is_v5 VMs and separate storage nodes provides up to 16,000 MBPS of storage throughput per RAC database node, thus enabling 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 configuration:

- Two database nodes, E104ids_v5
- Sixteen 2048 GB Premium SSD disks per node

Test script:

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

Calibrate_IO results measured by FlashGrid:

Cluster configuration	Max IOPS	Latency, ms	Max MBPS
2 RAC nodes (E104ids_v5 with 16 x 2048 GB Premium SSD each)	377,000	0.532	7683
3 RAC nodes (E104ids_v5 with 16 x 2048 GB Premium SSD each)	587,000	0.545	11494

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. AWR reports generated during a SLOB test provide various performance metrics but here we will focus on the I/O performance numbers.

SLOB results measured by FlashGrid:

Cluster configuration	Physical Read Database Requests	Physical Write Database Requests	Physical Read+Write Database Requests
2 RAC nodes, single AZ	288,023 IOPS	32,525 IOPS	320,548 IOPS
2 RAC nodes, multi-AZ	260,257 IOPS	29,027 IOPS	289,284 IOPS
3 RAC nodes, multi-AZ	376,064 IOPS	42,828 IOPS	418,892 IOPS

Test configuration details

- SGA size: 3 GB (small size selected 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: E104ids_v5
- Disks: 16 x 2048 GB Premium SSD per database node

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) 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 Azure VMs in the customer's Azure account and managed by the customer. The deployment model is similar to running your own Azure VMs 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 Azure 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's scripts that are available for CIS Server Level 1 aligned hardening.

Data Encryption

All data on Azure disks is encrypted at rest using Azure 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, 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

Supported VM types and sizes

Database node VMs must have 4+ physical CPU cores (8+ vCPUs), 32+ GB of memory, and Premium storage support. The following VM types are recommended for database nodes: Ebsv5, Esv5, Dsv5, Msv2, Mdsv2, FX.

Quorum nodes require fewer resources than database nodes, a single CPU core is sufficient. However, the smallest VM sizes with a single CPU core are limited to 4 disks, which is not sufficient in many cases. Therefore, the D4s_v5 (2 physical cores) 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

Premium SSD v2 disks are recommended for use in the regions where they are available. Premium SSD v2 offer more flexibility and potentially significant cost savings compared to v1. If Premium SSD v2 is not yet available in your target region then Premium SSD (v1) can be used. Migration from Premium SSD v1 to v2 is possible when v2 becomes available in your region.

Database features

FlashGrid Cluster does not restrict the use of any database features. DBAs 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, and 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

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

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

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

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 ARM templates.

Conclusion

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

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. Oracle and Java are registered trademarks of Oracle and/or its affiliates. Red Hat is a registered trademark of Red Hat Inc. Microsoft and Azure are registered trademarks of Microsoft Corporation. Other names may be trademarks of their respective owners.