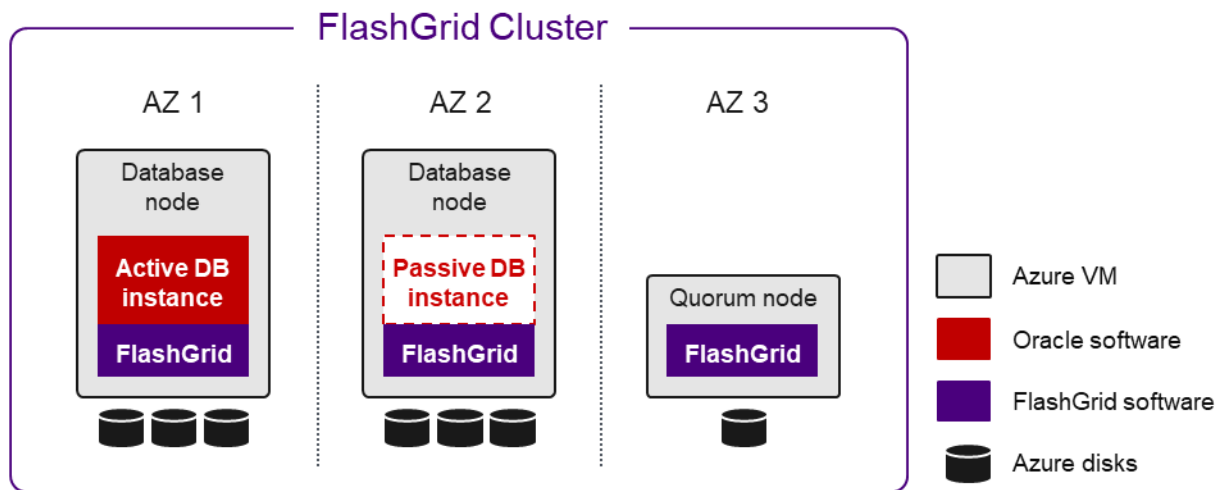


Architecture Overview

FlashGrid Cluster for Oracle Failover HA enables achieving uptime SLA of 99.95% for Oracle Databases running on Azure virtual machines through enhanced reliability of the database servers, fast failure isolation, and rapid failover between availability zones. FlashGrid Cluster is delivered as a fully integrated Infrastructure-as-Code template that can be customized and deployed to your Azure account in a few clicks. It is fully customer-managed and does not restrict any Oracle Database functionality.

Key components of FlashGrid Cluster for Oracle Database Failover HA on Azure include:

- Azure VMs
- Azure disks
- FlashGrid Storage Fabric™ software
- FlashGrid Cloud Area Network™ software
- FlashGrid Reliability Services
- Oracle Grid Infrastructure software (includes Oracle Clusterware and Oracle ASM)
- Oracle Database (Enterprise Edition or Standard Edition)



FlashGrid Cluster for Oracle Failover HA on Azure

FlashGrid Cluster architecture highlights:

- Active-Passive database HA
- 1+1 or 2+1 redundancy
- Zero RPO
- Spreading database nodes across availability zones (multi-AZ) protects against data center outages and local disasters.
- FlashGrid Cloud Area Network™ software enables high-speed overlay networks with advanced capabilities for HA.
- FlashGrid Storage Fabric™ software shares Azure disks between nodes and availability zones.
- FlashGrid Read-Local™ Technology minimizes storage network overhead by serving reads from locally attached Azure disks.
- Oracle ASM and Clusterware provide data protection and availability.
- Data is mirrored across separate nodes and availability zones.

Multi-AZ for Maximum Uptime and Fault Tolerance

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 it possible to use synchronous mirroring of data between AZs and achieving zero RPO in case of a failover.

Spreading cluster nodes across multiple availability zones helps to minimize downtime even when an entire data center experiences an outage or a local disaster. FlashGrid recommends using multi-AZ cluster configurations unless there is a specific need to use a single availability zone.

Possible Cluster Configurations

In most cases a multi-AZ cluster with two database nodes (1+1 redundancy) is recommended for Failover HA (see the diagram above). 2-way data mirroring is used with Normal Redundancy ASM disk groups. An additional small Azure VM (*quorum* node) is required to host quorum disks. Such a cluster can tolerate the loss of any one node without incurring database downtime.

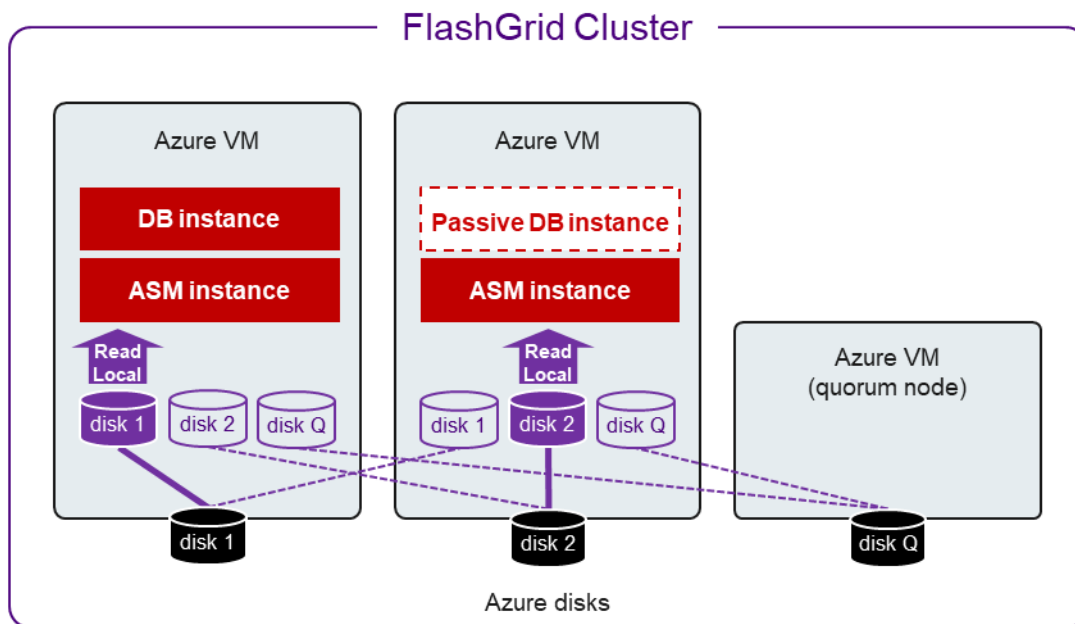
FlashGrid Cluster can also be deployed in other configurations for specific use-cases:

- Three database nodes (2+1 redundancy) may be used for consolidating multiple databases and reducing overprovisioning of the compute resources.
- Separate storage nodes may be configured for extra capacity and storage throughput for running extra-large (200+ TB) databases.
- Single AZ configurations may be used to meet specific requirements (for example, for test clusters to avoid inter-AZ traffic fees).

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.

Shared Storage Architecture

FlashGrid Storage Fabric software turns Azure disks attached to individual Azure VMs into shared disks accessible from all nodes in the cluster. The sharing is done at the block level with concurrent access from all nodes.



FlashGrid Storage Fabric with FlashGrid Read-Local Technology

FlashGrid Read-Local Technology

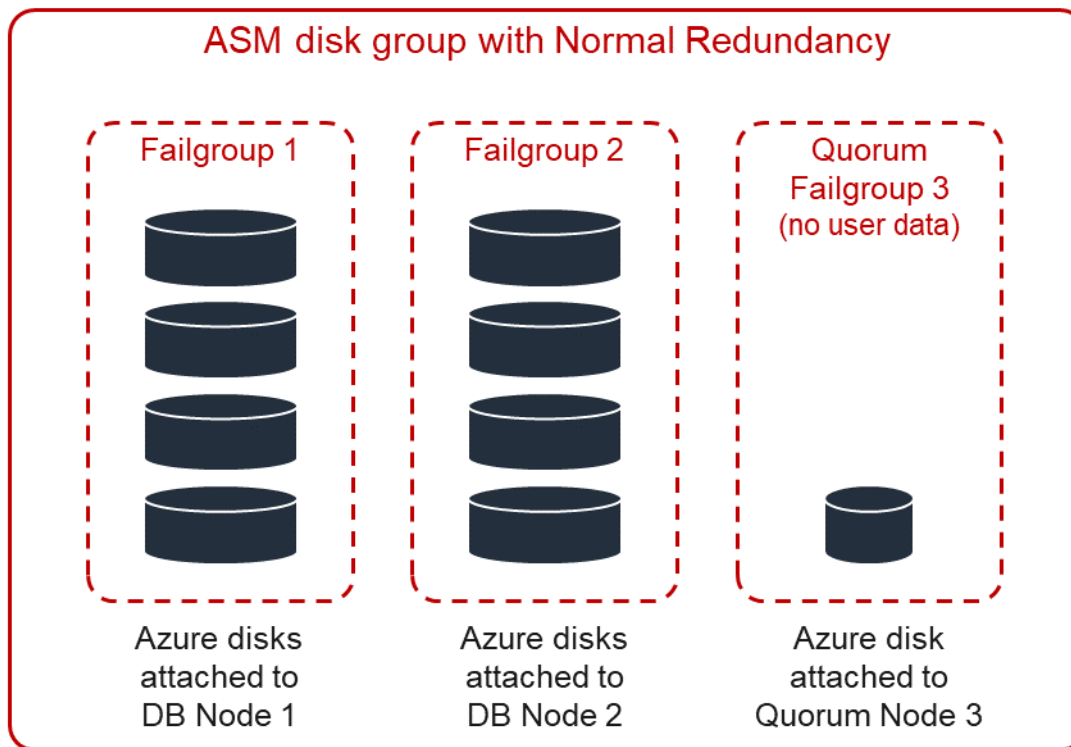
In a cluster with two or three database nodes, each database node has a full copy of user data stored on Azure disks attached to that database node. The FlashGrid Read-Local™ Technology allows serving all read I/O from the locally attached Azure disks. Read requests avoid the extra network hop, thus reducing read 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 a cluster with two database nodes all disk groups are configured with *Normal Redundancy*. In the *Normal Redundancy* mode each block of data has two mirrored copies. Each ASM disk group is divided into failure groups – one *regular* failure group per database node plus one *quorum* failure group. Each disk is configured to be a part of a failure group that corresponds to the node where the corresponding Azure disk is attached.

ASM stores mirrored copies of each block in different failure groups. Thus, the cluster can continue running and access data when any one of the disks fails or any one of the nodes goes down. Since the data mirroring is done synchronously, there is no data loss in the event of an abrupt failure.



A Normal Redundancy disk group in a Failover HA cluster on Azure

High Availability and Uptime Considerations

FlashGrid Storage Fabric and FlashGrid Cloud Area Network™ have a fully distributed architecture with no single point of failure. Additionally, the FlashGrid Cluster architecture leverages HA capabilities built in Oracle Clusterware and ASM.

Node availability

Because Azure VMs can move between physical hosts, a failure of a physical host causes only a short outage for the affected node. The node VM 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 VMs in the same availability zone, such as network, power, or cooling, may cause database downtime.

Placing VMs in different availability zones minimizes the risk of simultaneous node failures.

Data availability with Azure disks

An Azure disk provides persistent storage that survives a failure of the VM that the disk is attached to. After the failed VM restarts on a new physical node, all its disks are re-attached with no data loss.

Azure disks have built-in redundancy that protects data from failures in the underlying physical media. In the FlashGrid Cluster architecture, Oracle ASM performs data mirroring on top of the built-in protection of Azure disks. Together, Azure disks 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.

Fast failure detection and failover time

In the event of a failure of an active database node or database instance, the database instance can be immediately started (failed over) on the second node.

The time to restore the database service after an unexpected failure consists of the *time to detect the failure* plus the *time to restart the database instance on the second node*.

The time to detect the failure is between 30 to 60 seconds. FlashGrid Cluster uses several techniques to reduce the detection time in various scenarios including VM failure, disk failure, network failure, out-of-memory, or a different type of OS-level failure. The fast failure detection is an important benefit of the FlashGrid Cluster architecture. Other implementations of Oracle Database on Azure may suffer from long outages because of their limited abilities to detect certain failure types.

The time to start the database instance on the second node depends on the size of redo log that must be re-applied and the number of uncommitted transactions that must be undone. Essentially, this time depends on the transaction rate at the time of the failure and the sizing of the Azure VM and the Azure disks. The recovery time can be reduced by using higher performance Azure VMs and Azure disks.

With proper sizing of resources, FlashGrid Cluster enables Recovery Time Objective (RTO) as low as 120 seconds for essentially all types of failures, including the more complicated scenarios, such as system disk failure, running out-of-memory, etc.

Uptime SLA

Azure states an uptime SLA of 99.99% for two VMs running in different availability zones. By using the multi-AZ architecture, FlashGrid Cluster enables achieving the 99.95% or greater uptime SLA at the database service level.

The 99.95% uptime SLA assumes downtime of no more than 4 hours per year. In the FlashGrid Cluster for Failover HA architecture, database instance startup time during a failover is the biggest potential contributor to the database downtime. Therefore, measuring the startup time during crash-recovery under typical and peak application loads and proper sizing of the VM and disk resources is important for minimizing the downtime and meeting the uptime SLA.

FlashGrid Health Checker tool helps to minimize preventable failures by checking for possible mistakes in the database and cluster configuration.

Note: For an uptime SLA greater than 99.99% and for zero-downtime maintenance FlashGrid recommends using [FlashGrid Cluster for Oracle RAC on Azure](#) with 3+ RAC database nodes and active-active HA.

FlashGrid Reliability Services

Ensuring reliable operation of each database node is an integral part of maintaining highly available operation and achieving the target uptime SLA. Also, unexpected failures must be quickly contained to minimize disruption.

FlashGrid Reliability Services provide a layer of failure protection and prevention at the database node level:

- OS configuration optimized for HA clustering
- Disk heartbeat services for fast Azure disk failure detection
- Out-of-memory prevention services
- Monitoring of network connectivity, Azure disk responsiveness, memory utilization, CPU utilization, etc.

- Alerting on failure events
- “Call home” alerts sent directly to FlashGrid support
- Collecting diagnostics information and logs for troubleshooting
- Health Checker tool for analyzing cluster and database configuration for identifying possible misconfiguration and preventing failures.

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 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 differences in latency between different pairs of AZs provides an opportunity for optimization by choosing which AZs to place database nodes in. In a 2-node cluster it is optimal to place database nodes in the two AZs with the lowest latency between them.

Storage performance

In most cases, the use of Premium SSD v2 disks is recommended. The performance of each Premium SSD v2 disk can be configured from 3,000 to 80,000 IOPS and 125 to 1,200 MBPS. By using multiple disks per disk group attached to each database node, the storage throughput can reach the maximum of 260,000 IOPS and 8,000 MBPS per VM with Ebsv5 VM types.

Performance vs. on-premises solutions

The Premium SSD v2 storage provides an order of magnitude of improvement in IOPS and latency over traditional spinning HDD based storage arrays. With up to 260,000 IOPS and 8,000 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 Azure disks, which ensures stable and predictable performance with no interference from noisy neighbors.

Reference performance results

When moving database workloads to the cloud, the main areas of concern regarding performance tend to be around storage I/O because the CPU and memory can be sized like any on-premises system. Hence, for performance measurements we use Calibrate_IO and SLOB tools that exercise storage 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 NUMBER;
  iops INTEGER;
  mbps INTEGER;
BEGIN DBMS_RESOURCE_MANAGER.CALIBRATE_IO (20, 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:

Max IOPS	Max MBPS	Latency, ms
261,547	6,559	0.351

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:

Physical Read Database Requests	Physical Write Database Requests	Physical Read+Write Database Requests
219,713 IOPS	24,939 IOPS	244,652 IOPS

Test configuration details

- Two database nodes: E112ibs_v5
- Azure disks: (10) Premium SSD v2 per node, 50000 IOPS, 1200 MBPS each
- SGA size: 3 GB (We chose a smaller SGA to minimize caching effects and maximize physical I/O)
- 8KB database block size
- SLOB configuration: 240 schemas, 240MB each, UPDATE_PCT= 10 (10% updates, 90% selects)

Full Control of the Database and OS

FlashGrid Cluster is fully customer managed. This allows more interoperability, control, and flexibility than DBaaS offerings. Customer assigned administrators have full (root) access to the VMs and the operating system. Additional third-party tools, connectors, monitoring, or security software can be installed on the cluster nodes for compliance with corporate or regulatory standards.

Short Learning Curve for Oracle DBAs

Management of the Oracle databases running on FlashGrid Cluster is done using standard Oracle tools and services that include Oracle Clusterware, ASM, RMAN, Data Guard, GoldenGate, etc. All database features are available. The patching process is the same as in a typical on-premises environment. FlashGrid Cluster adds tools

and services for managing the cluster configuration and maintaining its reliable operation. But there is no need to learn any new proprietary procedures for performing standard DBA functions.

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:

- **Dedicated hosts** allow management of licenses at the physical host level instead of the VM level. Multiple VMs (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 the most critical systems.

A standalone (no clustering) database server 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

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.

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 Azure disks has server-side encryption (SSE). Platform-managed keys are used by default. Optionally, customer managed keys 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

Azure VM types

The following Azure VM types are typically recommended for database nodes:

- Ebsv5: high memory to CPU ratio and high storage throughput
- Esv5, Easv6 (preview): high memory to CPU ratio
- Dsv5, Dasv6 (preview): lower memory to CPU ratio
- Msv2, Mdsv2: highest memory to CPU ratio

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 D4s_v5 VM size 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, 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 (ARM) template for it. The following tasks are performed automatically using the ARM 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 using 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.