Differences between clustering and high availability (HA)

In this article

Learn about the differences between deployment topologies for the virtual machines (VMs) that comprise a GitHub Enterprise Server instance.

GitHub determines eligibility for clustering, and must enable the configuration for your instance's license. Clustering requires careful planning and additional administrative overhead. For more information, see "About clustering."

About deployment topologies for GitHub Enterprise Server

You can deploy the virtual machines for a GitHub Enterprise Server instance in different topologies depending on your environment and user needs.

- To support a plan for disaster recovery and supplement backups, or to improve network and write performance for geographically distributed users, you can configure high availability. In a high-availability configuration, one node acts as a primary, while others act as replicas. For more information, see "About high availability configuration."
- To provide horizontal scaling for environments with tens of thousands of developers, a cluster topology is available. Clustering addresses situations where a single primary node would routinely experience resource exhaustion. This configuration requires careful planning and additional administrative overhead. GitHub will work with you to determine your eligibility for clustering. For more information, see "About clustering."

Failure scenarios

High availability (HA) and clustering both provide redundancy by eliminating the single node as a point of failure. They are able to provide availability in these scenarios:

- **Software crashes**, either due to operating system failure or unrecoverable applications.
- **Hardware failures**, including storage hardware, CPU, RAM, network interfaces, etc.

- **Virtualization host system failures**, including unplanned and scheduled maintenance events for AWS, Azure, or GCP.
- **Logically or physically severed network**, if the failover appliance is on a separate network not impacted by the failure.

Scalability

Clustering provides better scalability by distributing load across multiple nodes. This horizontal scaling may be preferable for some organizations with tens of thousands of developers. In HA, the scale of the appliance is dependent exclusively on the primary node and the load is not distributed to the replica server.

Differences in failover method and configuration

Feature	Failover configuration	Failover method
High availability configuration	DNS record with a low TTL pointed to the primary appliance, or load balancer.	You must manually promote the replica appliance in both DNS failover and load balancer configurations.
Clustering	DNS record must point to a load balancer.	If a node behind the load balancer fails, traffic is automatically sent to the other functioning nodes.

Backups and disaster recovery

Neither HA nor clustering should be considered a replacement for regular backups. For more information, see "Configuring backups on your instance."

Monitoring

Availability features, especially ones with automatic failover such as clustering, can mask a failure since service is usually not disrupted when something fails. Whether you are using HA or clustering, monitoring the health of each instance is important so that you are aware when a failure occurs. For more information about monitoring, see "Recommended alert thresholds" and "Monitoring the health of your cluster."