Quickstart: Deploy an app to a GKE cluster

About cluster configuration choices

AUTOPILOT (/KUBERNETES-ENGINE/DOCS/CONCEPTS/AUTOPILOT-OVERVIEW)

STANDARD (/KUBERNETES-ENGINE/DOCS/CONCEPTS/TYPES-OF-CLUSTERS)

This page explains the two modes of operation and main cluster configuration choices you can make when creating a cluster in Google Kubernetes Engine (GKE). As a rule, the choices discussed here cannot be changed after a cluster is created. These choices impact a cluster's availability (#availability), version stability (#version-choices), and network (#networking-choices).

Level of cluster management

Before we can talk about types of clusters, it's important to understand the level of flexibility, responsibility, and control that you require for your cluster. The level of control that you require determines the mode of operation (#modes) to use in GKE, and the cluster configurations (#config-choices) that you need to make.

Modes of operation

When you create a cluster in GKE, you do so by using one of the following modes of operation:

- **Autopilot**: Provides a fully-provisioned and managed cluster configuration. For clusters created using the Autopilot mode, the cluster configuration options are made for you. Autopilot clusters are pre-configured with an optimized cluster configuration that is ready for production workloads.
- **Standard**: Provides advanced configuration flexibility over the cluster's underlying infrastructure. For clusters created using the Standard mode, you determine the configurations needed for your production workloads.

For more information about these modes, and to learn more about Autopilot, see the Autopilot overview (/kubernetes-engine/docs/concepts/autopilot-overview).

Cluster configuration choices

Based on the operation mode you chose, you then select which configurations you require for your cluster. In Autopilot mode, most of the choices are made for you. In Standard mode, you need to select the configurations that work best for your production workloads.

Cluster choices	Mode	
	Autopilot	Standard
Availability type (#availability)	Regional (#regional_clusters)	Regional (#regional_clusters) or Zonal (#zonal_clusters)
Version (#version-choices)	Release channel (#release_channel)	Release channel (#release_channel) , Default (#default_version), or Specific (#specific_version)
Network routing (#vpc-clusters)	VPC-native (#vpc-clusters)	VPC-native (#vpc-clusters) or Routes-based (#vpc-clusters)
Network isolation (#isolation-choices)	Private (#isolation-choices) or Public (#isolation-choices)	Private (#isolation-choices) or Public (#isolation-choices)
Kubernetes features (#kubernetes_feature_choices)	Production (#kubernetes_feature_choices)	Production (#kubernetes_feature_choices) or Alpha (#alpha_cluster)

Cluster availability type

With GKE, you can create a cluster tailored to the availability requirements of your workload and your budget. The types of available clusters include: zonal (single-zone or multi-zonal) and regional.

Note: Clusters created in the Autopilot mode are regional.

To help you choose which available cluster to create in the Standard mode, see Choosing a regional or zonal control plane (/kubernetes-engine/docs/concepts/scalability#choosing_a_regional_or_zonal_control_plane).

After you create a cluster, you cannot change it from zonal to regional, or from regional to zonal. Instead, you must create a new cluster then migrate traffic to it.

Zonal clusters

Zonal clusters have a single control plane in a single zone (/compute/docs/regions-zones). Depending on your availability requirements, you can choose to distribute your nodes for your zonal cluster in a single zone or in multiple zones.

Important: Use regional clusters for production workloads clusters as they offer higher availability than zonal clusters.

To create a zonal cluster in the Standard mode, see Creating a zonal cluster (/kubernetes-engine/docs/how-to/creating-a-zonal-cluster).

Single-zone clusters

A single-zone cluster has a single control plane running in one zone (/compute/docs/regions-zones). This control plane manages workloads on nodes running in the same zone. If you run a workload in a single zone, this workload is unavailable in the event of a zonal outage.

Multi-zonal clusters

A multi-zonal cluster has a single replica of the control plane running in a single zone, and has nodes running in multiple zones. During an upgrade of the cluster or an outage of the zone where the control plane runs, workloads still run. However, the cluster, its nodes, and its workloads cannot be configured until the control plane is available. Multi-zonal clusters balance availability and cost for consistent workloads. If you want to maintain availability and the number of your nodes and node pools are changing frequently, consider using a regional cluster (#regional_clusters). If you run a workload in multiple zones and there is a zonal outage, the workload is disrupted in that zone but remains available in other zones.

Regional clusters

A regional cluster (/kubernetes-engine/docs/concepts/regional-clusters) has multiple replicas of the control plane, running in multiple zones within a given region. Nodes in a regional cluster can run in multiple zones or a single zone depending on the configured node locations. By default, GKE replicates each node pool across three zones of the control plane's region. When you create a cluster or when you add a new node pool, you can change the default configuration by specifying the zone(s) in which the cluster's nodes run. All zones must be within the same region as the control plane.

Use regional clusters to run your production workloads, as they offer higher availability than zonal clusters.

To create a regional cluster in the Standard mode, see Creating a regional cluster (/kubernetes-engine/docs/how-to/creating-a-regional-cluster).

To create a regional cluster in the Autopilot mode, see Creating an Autopilot cluster (/kubernetes-engine/docs/how-to/creating-an-autopilot-cluster).

Cluster version

When you create a cluster, you can choose the cluster's specific Kubernetes version or you can make choices about its overall mix of stability and features.

Regardless of how you manage your cluster's version, it is recommended that you enable auto-upgrade (/kubernetes-engine/docs/concepts/cluster-upgrades) for the cluster and its nodes. GKE automatically upgrades nodes for Autopilot clusters.

Note: Clusters created in the Autopilot mode are pre-configured to use release channels.

Release channel

If you know the level of stability you need for a given cluster, you can enroll the cluster in a release channel (/kubernetes-engine/docs/concepts/release-channels). By default, new clusters are enrolled in the Regular release channel. Google automatically upgrades the cluster and its nodes when an update is available in that release channel. The Rapid channel receives multiple updates a month, while the Stable channel only receives a few updates a year.

Specific version

If you know that you need to use a specific supported version of Kubernetes for a given workload, you can specify it when creating the cluster (/kubernetes-engine/versioning-and-upgrades#specifying_cluster_version).

If you do not need to control the specific patch version you use, consider enrolling your cluster in a release channel (#release_channel) instead of managing its version directly.

Default version

If you choose to use a static version instead of enrolling the cluster in a release channel (that is, no channel), and you do not set a specific version (#specific_version) for the cluster, the current default version is used. The default version is selected based on usage and real-world performance, and is changed regularly. The default version for static is typically aligned with the Regular release channel. While the default version is the most mature one, other versions being made available are GA versions that passed internal testing and qualification. Changes to the default version are announced in a release note (/kubernetes-engine/docs/release-notes).

Cluster networking

When creating a GKE cluster, you can specify the network routing mode, and how to isolate your cluster network.

VPC-native and routes-based clusters

In Google Kubernetes Engine, clusters can be distinguished according to the way they route traffic from one Pod to another Pod. A cluster that uses Alias IPs is called a *VPC-native cluster* (/kubernetes-engine/docs/how-to/alias-ips). A cluster that uses Google Cloud routes (/vpc/docs/routes) is called a *routes-based cluster* (/kubernetes-engine/docs/how-to/routes-based-cluster).

VPC-native is the recommended network mode for new clusters. This is the default for clusters created in the Autopilot mode.

For clusters created in the Standard mode, the default network mode depends on the GKE version and the method you use to create the cluster.

For details, see the Default cluster network mode (/kubernetes-engine/docs/concepts/alias-ips#default_cluster_network_mode) chart.

Network isolation choices

By default, you can configure access from public networks to your cluster's workloads. Routes are not created automatically. Private clusters assign internal addresses to Pods and nodes, and workloads are completely isolated from public networks.

To create a private cluster, see Creating a private cluster (/kubernetes-engine/docs/how-to/private-clusters).

Kubernetes features

New features in Kubernetes are listed as Alpha, Beta, or Stable (https://github.com/kubernetes/community/blob/master/contributors/devel/sig-architecture/api_changes.md#alpha-beta-and-stable-versions) , depending upon their status in development. In most cases, Kubernetes features that are listed as Beta or Stable are included with GKE clusters. Kubernetes Alpha features are available in special GKE alpha clusters.

Note: Alpha features are not available in clusters created in the Autopilot mode.

Alpha cluster

An alpha cluster (/kubernetes-engine/docs/concepts/alpha-clusters) has all Kubernetes alpha APIs (sometimes called feature gates) enabled. You can use alpha clusters for early testing and validation of Kubernetes features. Alpha clusters are not supported for production workloads, cannot be upgraded, and expire within 30 days.

To create an alpha cluster in the Standard mode, see Creating an alpha cluster (/kubernetes-engine/docs/how-to/creating-an-alpha-cluster).

Previous
← Choose a GKE mode of operation
(/kubernetes-engine/docs/concepts/choose-cluster-mode)

Last updated 2023-11-20 UTC.