Google Cloud

GKE Autopilot

Autopilot Overview

GKE Autopilot is a mode of operation in Google Kubernetes Enging where Google manages your cluster configuration, nodes, scaling, security, and other preconfigured settings.

Autopilot clusters are optimized to run most production workloads, and provision compute resources based on your Kubernetes manifests. The streamlined configuration follows GKE best practices and recommendations for cluster and workload setup, scalability, and security.

Feature comparison

The table includes some of the standard and autopilot features of GKE. For the features below:

Default values are automatically configured if not otherwise specified. They can be changed.
Pre-configured values are always enabled and set by Google. They cannot be changed.
Optional settings are not enabled by default, and can be configured for use.

Feature	Autopilot	Standard
Versions and releases	Regular release channel (default) GKE version in channels (optional)	Release channel, Static GKE version (Optional)
Location availability	Regional (pre-configured)	Regional or zonal (optional)
Node provisioning and scaling	Node resource management and	Manual new node provisioning and specifying resources (default)

	scaling (pre-configured) Horizontal and vertical pod autoscaling (optional)	Node auto-provisioning, cluster autoscaler, Horizontal and vertical pod autoscaling (optional)
Node compute configuration	General-purpose platform optimized for most workloads (default) Compute classes for workloads that have specific needs, such as Arm (optional) GPUs (optional)	General-purpose compute engine machine types (default) Choose specific compute engine machine types, hardware (optional) GPUs (optional)

For a complete list of features, read the documentation.

Benefits of GKE Autopilot

There are many benefits of using Autopilot for your Kubernetes workloads. Here are some that are listed below:

- **Focus on your apps**: Google manages the infrastructure, so you can focus on building and deploying your applications.
- **Security**: Clusters have a default hardened configuration, with many security settings enabled by default. GKE automatically applies security patches to your nodes when available, adhering to any maintenance schedules you configured.
- **Node management**: Google manages worker nodes, so you don't need to create new nodes to accommodate your workloads or configure automatic upgrades and repairs.
- **Scaling**: When your workloads experience high load and you add more Pods to accommodate the traffic, such as with Kubernetes Horizontal Pod Autoscaling, GKE automatically provisions new nodes for those Pods, and automatically expands the resources in your existing nodes based on need.
- **Resource management**: If you deploy workloads without setting resource values such as CPU and memory, Autopilot automatically sets pre-configured default values and modifies your resource requests at the workload level.
- **Networking**: Autopilot enables some networking security features by default, such as ensuring that all Pod network traffic passes through your Virtual Private Cloud firewall rules, even if the traffic is going to other Pods in the cluster.
- **Release management**: All Autopilot clusters are enrolled in a GKE release channel, which ensures that your control plane and nodes run on the latest qualified versions in that channel.
- **Managed flexibility**: If your workloads have specific hardware or resource requirements, such as high CPU or memory, Autopilot offers pre-configured compute classes built for those workloads. You can also use GPUs to accelerate workloads like batch or AI/ML applications.
- **Reduced operational complexity**: Autopilot reduces platform administration overhead by removing the need to continuously monitor nodes, scaling, and scheduling operations.

Autopilot comes with a SLA that covers both the control plane and the compute capacity used by your Pods.

Plan your Autopilot clusters

Before you create a cluster, plan and design your Google Cloud architecture. In Autopilot, you request hardware in your workload specifications. GKE provisions and manages the corresponding infrastructure to run those workloads. For example, if you run machine learning workloads, you request hardware accelerators. If you develop Android apps, you request Arm CPUs.

Plan and request quota for your Google Cloud project or organization based on the scale of your workloads. GKE can only provision infrastructure for your workloads if your project has enough quota for that hardware.

Consider the following factors during planning:

- Estimated cluster size and scale
- Workload type
- Cluster layout and usage
- Networking layout and configuration
- Security configuration
- Cluster management and maintenance
- Workload deployment and management
- Logging and monitoring

For more information, view the documentation.

Pricing

You only pay for the CPU, memory, and storage that your workloads request while running on GKE Autopilot.

You aren't billed for unused capacity on your nodes, because GKE manages the nodes. You also aren't charged for system Pods, operating system costs, or unscheduled workloads. For detailed pricing information, refer to Autopilot pricing.