Google Kubernetes Engine

The most automated and scalable managed Kubernetes platform.

New customers get $300 in free credits to spend on GKE. All customers get one Autopilot or zonal cluster free per month, not charged against your credits.

- Run your apps on a fully managed Kubernetes cluster with GKE Autopilot
- Start quickly with single-click clusters and scale up to 15000 nodes
- Leverage a high-availability control plane including multi-zonal and regional clusters
- Eliminate operational overhead with industry-first four-way auto scaling
- Secure by default, including vulnerability scanning of container images and data encryption

BENEFITS

Speed up app development without sacrificing security

Develop a wide variety of apps with support for stateful, serverless, and application accelerators. Use Kubernetes based CI/CD tooling to secure and speed up each stage of the build-and-deploy life cycle.

Streamline operations with release channels

Choose the channel that fits your business needs. Rapid, regular, and stable release channels have different cadences of node upgrades and offer support levels aligned with the channel nature.

Reduce Day 2 ops with help from Google SREs

Get back time to focus on your applications with help from Google Site Reliability Engineers (SREs). Our SREs constantly monitor your cluster and its computing, networking, and storage resources.

Key features

Two modes of operation, one GKE

GKE now offers two modes of operations: Autopilot and Standard. Autopilot mode is a hands-off, fully managed solution that manages your entire cluster's infrastructure without worrying about configuring and monitoring, while still delivering a complete Kubernetes experience. And with per-pod billing, Autopilot ensures you pay only for your running pods, not system components, operating system overhead, or unallocated capacity. Standard mode is the experience we've been building since the launch of GKE, enabling additional customization options over the nodes with the ability to fine tune and run custom administrative workloads for when you need low level controls.

Pod and cluster autoscaling

GKE is the industry's first fully managed Kubernetes service that implements full Kubernetes API, 4-way autoscaling, release channels and multi-cluster support. Horizontal pod autoscaling can be based on CPU utilization or custom metrics. Autopilot automatically scales your cluster capacity based on the resource requirements in your Pod specs. In the Standard mode of operation, cluster autoscaling works on a per-node-pool basis to scale up your nodes on demand. Vertical pod autoscaling continuously analyzes the CPU and memory usage of pods, automatically adjusting CPU and memory requests.

Prebuilt Kubernetes applications and templates

Get access to enterprise-ready containerized solutions with prebuilt deployment templates, featuring portability, simplified licensing, and consolidated billing. These are not just container images, but open source, Google-built, and commercial applications that increase developer productivity. Click to deploy on-premises or in third-party clouds from Google Cloud Marketplace.

Container focused networking and security

Privately networked clusters in GKE can be restricted to a private endpoint or a public endpoint that only certain address ranges can access. GKE Sandbox for the Standard mode of operation provides a second layer of defense between containerized workloads on GKE for enhanced workload security. GKE clusters inherently support Kubernetes Network Policy to restrict traffic with pod-level firewall rules.

Migrate traditional workloads to GKE containers with ease

Migrate to Containers makes it fast and easy to modernize traditional applications away from virtual machines and into containers. Our unique automated approach extracts the critical application elements from the VM so you can easily insert those elements into containers in Google Kubernetes Engine or Anthos clusters

without the VM layers (like Guest OS) that become unnecessary with containers. This product also works with GKE Autopilot.

All features

Backup for GKE	Backup for GKE is an easy way for customers running stateful workloads on GKE to protect, manage, and restore their containerized applications and data.
Identity and access management	Control access in the cluster with your Google accounts and role permissions.
Hybrid networking	Reserve an IP address range for your cluster, allowing your cluster IPs to coexist with private network IPs via Google Cloud VPN.
Security and compliance	GKE is backed by a Google security team of over 750 experts and is both HIPAA and PCI DSS compliant.
Integrated logging and monitoring	Enable Cloud Logging and Cloud Monitoring with simple checkbox configurations, making it easy to gain insight into how your application is running.
Cluster options	Choose clusters tailored to the availability, version stability, isolation, and pod traffic requirements of your workloads.
Auto scale	Automatically scale your application deployment up and down based on resource utilization (CPU, memory).
Auto upgrade	Automatically keep your cluster up to date with the latest release version of Kubernetes. Kubernetes release updates are quickly made available within GKE.

Auto repair	When auto repair is enabled, if a node fails a health check, GKE initiates a repair process for that node.
Resource limits	Kubernetes allows you to specify how much CPU and memory (RAM) each container needs, which is used to better organize workloads within your cluster.
Container isolation	Use GKE Sandbox for a second layer of defense between containerized workloads on GKE for enhanced workload security.
Stateful application support	GKE isn't just for 12-factor apps. You can attach persistent storage to containers, and even host complete databases.
Docker image support	GKE supports the common Docker container format.
Fully managed	GKE clusters are fully managed by Google Site Reliability Engineers (SREs), ensuring your cluster is available and up-to-date.
OS built for containers	GKE runs on Container-Optimized OS, a hardened OS built and managed by Google.
Private container registry	Integrating with Google Container Registry makes it easy to store and access your private Docker images.
Fast consistent builds	Use Cloud Build to reliably deploy your containers on GKE without needing to setup authentication.
Workload portability, on-premises and cloud	GKE runs Certified Kubernetes, enabling workload portability to other Kubernetes platforms across clouds and on-premises.
GPU and TPU support	GKE supports GPUs and TPUs and makes it easy to run ML, GPGPU, HPC, and other workloads that benefit from specialized hardware accelerators.

Built-in dashboard	Google Cloud console offers useful dashboards for your project's clusters and their resources. You can use these dashboards to view, inspect, manage, and delete resources in your clusters.
Spot VMs	Affordable compute instances suitable for batch jobs and fault-tolerant workloads. Spot VMs provide significant savings of up to 91% while still getting the same performance and capabilities as regular VMs.
Persistent disks support	Durable, high-performance block storage for container instances. Data is stored redundantly for integrity, flexibility to resize storage without interruption, and automatic encryption. You can create persistent disks in HDD or SSD formats. You can also take snapshots of your persistent disk and create new persistent disks from that snapshot.
Local SSD support	GKE offers always-encrypted local solid-state drive (SSD) block storage. Local SSDs are physically attached to the server that hosts the virtual machine instance for very high input/output operations per second (IOPS) and very low latency compared to persistent disks.
Global load balancing	Global load-balancing technology helps you distribute incoming requests across pools of instances across multiple regions, so you can achieve maximum performance, throughput, and availability at low cost.
Linux and Windows support	Fully supported for both Linux and Windows workloads, GKE can run both Windows Server and Linux nodes.

Hybrid and multi-cloud support	Take advantage of Kubernetes and cloud technology in your own data center. Get the GKE experience with quick, managed, and simple installs as well as upgrades validated by Google through Anthos.
Serverless containers	Run stateless serverless containers abstracting away all infrastructure management and automatically scale them with Cloud Run.
Usage metering	Fine-grained visibility to your Kubernetes clusters. See your GKE clusters' resource usage broken down by namespaces and labels, and attribute it to meaningful entities.
Release channels	Release channels provide more control over which automatic updates a given cluster receives, based on the stability requirements of the cluster and its workloads. You can choose rapid, regular, or stable. Each has a different release cadence and targets different types of workloads.
Software supply chain security	Verify, enforce, and improve security of infrastructure components and packages used for container images with Container Analysis.
Per-second billing	Google bills in second-level increments. You pay only for the compute time that you use.