PLURALSIGHT CLOUD GUIDE

GCP
Guidebook

PLURALSIGHT

GCP Guidebook

Last updated: May 2023

Whether you're just diving into the cloud, looking for a refresher, or expanding your knowledge of the different cloud platforms, this guidebook explains the most important terms to help you talk like a Google Cloud Platform (GCP) local.

We provide an overview explanation for each term to help you understand the lay of the land. Then we dive into the secrets only the GCP locals know—what to avoid and where to spend the most time. When you want to know more, check out the related courses and hands-on labs.

Ready to explore the world of GCP? Dive right in.

INDEX

Cloud Identity and Access Management (IAM)

Overview

The first step to securing your cloud environments is defining who gets access to which parts. Within Google Cloud Platform, you can manage permissions and identity and access management (IAM) with Cloud IAM and Cloud Identity. Cloud IAM is a unified resource access management system for users and services. Identities come in many forms, such as Google accounts, unmanaged accounts, service accounts, and collections of these (e.g., Google Groups and G-Suite domains).

There are three critical elements to IAM:

- **Policies**: A set of rules that governs who can perform tasks using specific resources
- **Roles**: Permissions assigned to identities or members
- **Resources**: Projects, folders, cloud services, or parts of those services, like instances or buckets

Google refers to Cloud Identity as a nice alphabet soup acronym dubbed “identity as a service” (IDaaS). It’s a stand-alone product that allows you to manage users and groups within your Google environment, unify identities between Google and other cloud providers, and manage access and compliance for Google accounts outside of your corporate environment. Once you’ve adopted Cloud Identity, you can use IAM to define access to resources for each Cloud Identity account.

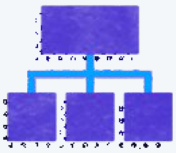
In addition to defining and managing access, Cloud Identity offers single-sign on (SSO) and 2-Step Verification (2SV) to further secure your cloud accounts.

Off the record

Cloud Identity creates what Google calls “folders” for one or more Google Cloud projects. You might have a folder for your developer team and live team. You can apply policies to your folders where all resources within the projects in those folders inherit permissions from those policies.

On top of this, IAM can quickly become one of your points of failure in managing security. The scope of your IAM policies represents a large part of your attack surface. You might misconfigure resource permissions for third parties or set policies too broad. Or you might lose the password altogether. Setting specific, robust IAM policies is one of the best ways to protect your cloud operations from breaches and bad actors.

Want to learn more? Our Google Cloud Identity and Access Management (IAM) Deep Dive course offers over seven hours of course material and hands-on labs with all the information you need to understand and implement IAM in a GCP environment.



Cloud Load Balancing

Overview

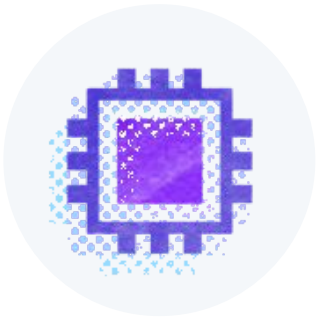
Cloud Load Balancing is a fully managed incoming traffic service responsible for distributing traffic across several virtual machine (VM) instances. It's autoscaling set by policy, CPU utilization, or serving capacity (or all of them). It supports heavy traffic and is instrumental in routing traffic to the closest instances. It can also detect and remove unhealthy instances (determined by percentage of failure).

Off the record

Part of the challenge is understanding the type of traffic you need to balance across your instances. There are three main types of load balancing supported by Google Cloud:

- **Global external:** HTTP/S, SSL, and TCP
- **Regional external:** TCP/UDP within a region
- **Regional internal:** Between groups of instances within a region

Get hands-on with this Global Load Balancing with Google Compute Engine lab. Learn to set up a managed instance group and use a Google Cloud load balancer to manage incoming requests from the outside world.



Compute Engine

Overview

Compute Engine is a type of infrastructure as a service (IaaS). It consists of VMs that run the operations for your applications. Combined with Google Cloud Storage, Compute Engine is your operation running in the cloud without you ever needing to touch a power button.

Compute Engine abstracts the complicated processes of managing server hardware. You select the amount of compute power you need for your operations and pay for what you need without buying or taking care of the actual hardware. Google Cloud offers several machine types that vary in compute power, memory, and cost.

These VMs can run public disk images pre-configured and offered by Google or private disk images accessible by you. And Google Cloud configures VMs to work with containers. They also come with virtual private network (VPN) connectivity.

Off the record

While you can't see it, it's still your computer. You still have to maintain parts of the operation without ever looking at the hardware. You have to provision the right number of resources, configure your machine types correctly, and make sure you're optimizing for performance and cost.

Fortunately, Google offers some price estimates in your dashboard when configuring your Compute Engine VM. Google Cloud, in particular, offers four machine types:

- **General-purpose machines** that work well for day-to-day needs. You'll see types like E2, N2, N2D, and N1—all of which may work well in different scenarios.
- A set of **memory-optimized machines** when you might have specific requirements around managing memory. These offer more memory per core.
- **Compute-optimized instances** for when you need to optimize for computing power, even if it may cost more or sacrifice memory performance
- **Shared-core machine types**, which timeshare a physical core and may be more cost-effective

There are other configuration options. You have to set your zone, determine whether to deploy a container image to your VM, select a type of disk image, pick sizes for those disks, etc.

Take a two-hour deep dive into our Advanced Google Cloud Compute Engine course for a look at designing, planning, and implementing virtual machines and related services in Google Cloud Platform.

Google App Engine

Overview

App Engine is one of the original four Google Cloud services. It requires less management than the other compute products, so you're abstracting more of the hardware from your work.

It is a platform as a service (PaaS), so it requires no manual server setup or provisioning. It provides automatic scaling and load balancing. It's great for websites, mobile apps, and line-of-business apps that need to get up and running quickly. Developers often use App Engine for HTTP applications.

It's exceptionally straightforward to get started with, code-centric, and provides automatic scaling regionally. It has easy access to the rest of the platform as a full-fledged product within Google Cloud. App Engine also has access to Cloud SQL for relational work and works with Cloud Storage.

Off the record

App Engine excels when working with scalable mobile apps—primarily gaming. Rapid scalability is a crucial factor when creating and handling online games, and App Engine can scale up when needed to handle surges. Unfortunately, you can't customize App Engine like Compute Engine.

App Engine has two environments: standard and flexible.

Standard is Google's original App Engine environment. It's more proprietary, and source code must be written in specific versions of the supported programming languages. It has the fastest spin-up time (milliseconds) and works well if you're expecting spikes in traffic. Because it's the standard package, it's also the least expensive option.

Flexible is Google's newer App Engine environment. It's less proprietary but more standardized than Standard. Flexible runs on Docker containers, so it works with more languages. It also allows access to your Google Cloud project resources residing in the Compute Engine network. But with great power comes slower spin-up times. And the inability to scale down to zero like Standard.

Get hands-on with our Deploying to Google Cloud App Engine lab. In just 30 minutes, learn to deploy an app to App Engine that allows users to upload photos and enter details into Cloud Datastore, a NoSQL database.

Cloud Functions and Cloud Run

Overview

App Engine isn't Google's only serverless product. Google also has three others: Cloud Functions, Cloud Run, and Cloud Run for Anthos.

- **Cloud Functions**: Event-driven code, like third-party apps and API integrations such as Twilio and Stripe. It also works well with IoT products for real-time processing, data collection, and processing. And it integrates with a large number of services in Google Cloud, making it work well for real-time processing.
- **Cloud Run**: Container-based operations, including mature technology stacks. It's useful for internal company-only applications in addition to regular robust websites. You could also use it, for example, to create monthly invoices with a cloud scheduler task.
- **Cloud Run for Anthos**: Good for enterprise-grade CI/CD pipelines so you can continuously build new content. It's suitable for integrating on-premise services and executing your applications closest to customers (or "at the edge").

Off the record

There are a lot of serverless services from Google Cloud, each designed for different scenarios. Your operation may work best with one, multiple, or none. You'll have to determine the best approach when choosing whether to go serverless

Venture into Google's serverless world with our Introduction to Serverless on Google Cloud course. Learn about the differences and use cases for Google Cloud's top four serverless contenders: App Engine, Cloud Functions, Cloud Run, and Cloud Run for Anthos.

Cloud SQL

Overview

Cloud SQL is a relational database service that works across various needs for data-rich organizations, ranging from flight services to ed-tech startups. Cloud SQL manages, maintains, and administers PostgreSQL, SQL Server, and MySQL (including two generations of MySQL). It removes the need to deal with hardware so you can focus on working with the data.

Data is automatically encrypted and has a default firewall. Cloud SQL also automates and replicates your backups so you get 99.95% uptime. Easily copy data to another zone or region for high availability so any time your source database crashes, you can failover to your backup. It's also accessible from Compute Engine and App Engine.

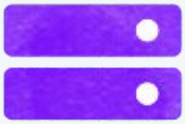
Once you have the database instances created, all you have to do is use them. Even the console is pretty sparse.

Off the record

Everything comes down to configuration for your Cloud SQL databases. You have to add high availability, specify a region, and select a zone. You also have to set up the times for auto-backups and a maintenance schedule—which, if set automatically, may end up causing service disruptions.

In theory, you could go with best-practice versions or defaults, but your service probably has unique needs that justify going deep into the config options.

Learn to create and manage Cloud SQL instances on Google Cloud Platform with our Google Cloud SQL Deep Dive course.



Cloud Storage

Overview

Cloud Storage is the object storage service on Google Cloud. All the data uploaded to your Google Cloud environment is stored in buckets to organize and control access to that data.

A project can have one or more buckets assigned to it, and it observes the inherited format for IAM. You can apply an IAM role to an individual bucket. Objects in those buckets inherit those buckets' permissions, and those buckets can inherit permissions from the project.

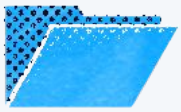
You can upload a near infinite number of objects to your Cloud Storage as long as you have a credit limit high enough. And you pay only for the amount of data currently in the bucket. You also don't have to allocate space in advance

Off the record

There are different types of storage classes. Regional and multiregional offer high-performance storage, whereas nearline and coldline are for longer term backup or archival storage. You can apply a storage class to an entire bucket or the objects in those buckets. They all have the same performance, but they differ in geolocation, service level agreements (SLAs), and operations restrictions.

- **Multiregional**: This storage class spans your data across multiple regions in a continental region (geo-redundant) and is suitable for hot data from different geographic regions. It has no retrieval cost and the highest SLA.
- **Regional**: Similar to a multiregional bucket, except buckets in this storage class reside in a single region. If you're using services in that region, you may have slightly better performance accessing the data in a regional bucket and lower costs.
- **Nearline**: Nearline storage is suitable for regular backups you plan to access periodically. The cost drops to a much lower per-gigabyte fee but comes with a retrieval fee.
- **Coldline**: This storage class is suitable for data you don't intend to touch more than three or four times a year. It's even cheaper than nearline at a per-gigabyte level but has an even higher cost per retrieval.
- **Archive**: This is the newest (and coldest) storage class designed for cold data storage and disaster recovery. Unlike traditional cold storage methods, your data is available in milliseconds should you need it. But it comes at a cost.

Finding the right storage class for your data is one thing, but there are also restrictions on how that data can move once you make a choice. You can change the storage class for your data except from multiregional to regional and vice versa. But if you change that class, it affects only new objects coming into that bucket—you'll have to change the storage class for objects inside the bucket.



Instance groups

Overview

Instance groups are buckets of instances where you can try to make changes to a bunch of instances at once. Google offers two kinds of instance groups:

- **Managed instance groups:** collections of identical instances you can manage as a single unit. If you need to make changes to all of them at once, it makes more sense to run it as an instance group. If there are health problems, the managed instance group will automatically recreate that instance.
- **Unmanaged instance groups:** collections of instances with different configurations you can use to apply load balancing to existing configurations

Off the record

There are advantages and disadvantages to both. If you want the benefits of managed groups, you have to ensure that all the instances are identical. And you'll probably run into plenty of scenarios where you're running multiple identical instances. That makes it friendly and convenient when you want to handle autoscaling and health checks for your instances.

But you might frequently run into scenarios where you'll need unmanaged instance groups. They're not uniform, so you won't get some of the features of managed groups, like autoscaling or up-to-date support. But you can still enable some batch processes to make the grouping worthwhile.

Get an overview of managed instance groups in this multi-course track that prepares you for the role of a GCP network engineer.

Snapshots

Overview

A snapshot is the primary backup method where a compute engine takes a read-only picture of an instance or disk as a backup. You can capture snapshots of boot disks, attached disks, or actively running instances. Once you have the snapshot, you can create a new instance or disk in a new zone or in response to a disaster recovery scenario.

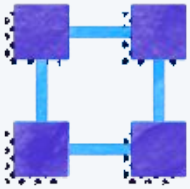
Off the record

Snapshots operate as incremental backups, which can save you a fair bit of money. It's an incremental backup where the second snapshot copies the changes from the first, the third from the second, and so forth.

Because of this, you generally want to reduce activity whenever you can. You can create a snapshot when an instance runs, but your disk shouldn't change too much while the snapshot is in process. This might require you to pause any options that write data, unmount the disk, or schedule the snapshot when not much is happening inside your instance.

Get hands-on with our Working with Snapshots on Google Compute Engine lab. Work with snapshots on Compute Engine in the web console and command line format.

AWS Amazon Virtual Private Cloud (Amazon VPC) | **Azure** Azure Virtual Network (VNet)



Virtual Private Cloud (VPC)

Overview

Virtual Private Cloud (VPC) is a virtualized network that provides IPv4 and IPv6 connectivity for GCP resources like Compute Engine VMs and Google Kubernetes Engine (GKE) clusters. They're the foundational component of all other networking functions. VPCs create a virtualized global network instance to connect resources across all GCP locations.

VPC is a software-defined private network implemented on Google Cloud that eliminates routers, switches, servers, or tripping over tangled cables. This allows you to customize and scale your services rapidly while ensuring no communication between any resources on your network has exposure to the public internet.

VPC networks are divided into subnets, each with their own specific IP range and assigned region (not just a zone). Unlike other cloud providers, subnets can span multiple availability zones within a region. Resources in different regions can speak to one another because of default routes that allow communication between all subnets, regardless of where they're located.

Off the record

Although projects group various resources so that different users can manage them, resources within the same project can't communicate with each other by default. For different resources to communicate, they need to use public communication via the internet or private communication via a VPC. Although it's common for a project to contain a single VPC, one project may contain multiple VPCs. You can also share VPCs across multiple projects, enabling resources from different projects to communicate with each other.

Let's say you have multiple VPCs in the same project, each with some Compute Engine instances. The instances (resources) in one VPC can't communicate with the resources in another VPC without VPC Network Peering. However, resources within one VPC can communicate over this private network across multiple regions.

There are two tiers of networking: premium and standard. Here are a few differences:

- **Premium-tier data** goes through Google's network as much as possible, with the internet-based user's traffic entering (and exiting) Google's network at the location nearest them. This is sometimes referred to as "cold potato" routing.
- **Standard-tier traffic** goes through Google's network only within the region where you deployed GCP resources. So an internet-based user's traffic will travel across the public internet to reach that region. This is sometimes referred to as "hot potato" routing.

Get hands-on with our novice-level Create Our First VPC in Google Cloud lab.

About Pluralsight

Pluralsight helps organizations around the globe advance their technology workforce. Because the hardest part of building a business isn't building software and technology. It's building up the people who grow your business. That's why everyone from CIOs to developers trust Pluralsight—the only partner that helps leaders build better teams and better products, all at the same time.

Our software and solutions are purpose-built to address your top challenges and outcomes:

- Onboard new engineers faster
- Build products faster and improve the developer experience
- Develop internal cloud talent and enable cloud transformation
- Improve retention and cut hiring costs
- Improve cycle times and reduce burnout for remote teams
- Develop teams that deliver on key tech initiatives
- Increase delivery speed and overcome Agile roadblocks
- Hire job-ready, diverse talent
- Build fluency and collaboration organization-wide

Our cloud transformation solutions help you create the cloud talent you need, when you need it, to deliver on your biggest, boldest vision. Pluralsight Skills delivers expert-authored courses in the latest cloud technologies, paired with unlimited access to hands-on labs, sandboxes, and certification prep. Upskilling your teams with Skills equips your teams to execute on strategic cloud investments that ultimately drive innovation, automation, and efficiency.

PLURALSIGHT