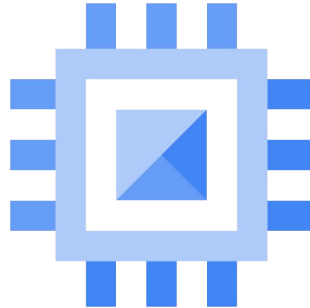

Compute Engine offers managed virtual machines

- High CPU, high memory, standard and shared-core machine types.
- Persistent disks .
- Standard, SSD, and local SSD.
- Snapshots
- Resize disks with no downtime.
- Instance metadata and startup scripts.



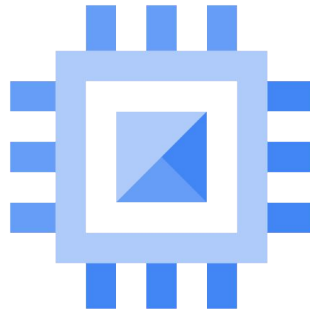
Virtual machines have the power and generality of a full-fledged operating system in each. You configure a virtual machine much like you build out a physical server: by specifying its amounts of CPU power and memory, its amounts and types of storage, and its operating system. Compute Engine lets you create and run virtual machines on Google infrastructure. There are no upfront investments, and you can run thousands of virtual CPUs on a system that is designed to be fast and to offer consistent performance.

You can flexibly reconfigure Compute Engine virtual machines. And a VM running on Google's cloud has unmatched worldwide network connectivity.

You can create a virtual machine instance by using the Cloud Console or the `gcloud` command-line tool. A Compute Engine instance can run Linux and Windows Server images provided by Google or any customized versions of these images. You can also build and run images of other operating systems.

Compute Engine offers customer friendly pricing

- Per-second billing, sustained use discounts, committed use discounts.
- Preemptible instances.
- High throughput to storage at no extra cost.
- Custom machine types: Only pay for the hardware you need.



Compute Engine bills by the second for use of virtual machines, with a one-minute minimum. And discounts apply automatically to virtual machines that run for substantial fractions of a month. For each VM that you run for more than 25% of a month, Compute Engine automatically gives you a discount for every incremental minute. You can get up to a 30% net discount for VMs that run the entire month.

Compute Engine offers the ability to purchase committed use contracts in return for deeply discounted prices for VM usage. These discounts are known as committed use discounts. If your workload is stable and predictable, you can purchase a specific amount of vCPUs and memory for up to a 57% discount off of normal prices in return for committing to a usage term of 1 year or 3 years.

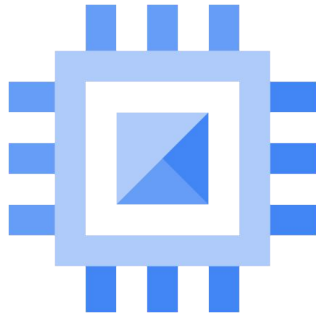
Suppose you have a workload that no human being is sitting around waiting to finish. Say, a batch job analyzing a large dataset. You can save money by choosing Preemptible VMs to run the job. A Preemptible VM is different from an ordinary Compute Engine VM in only one respect: you've given Compute Engine permission to terminate it if its resources are needed elsewhere. You can save a lot of money with preemptible VMs, although be sure to make your job able to be stopped and restarted.

You don't have to select a particular option or machine type to get high throughput

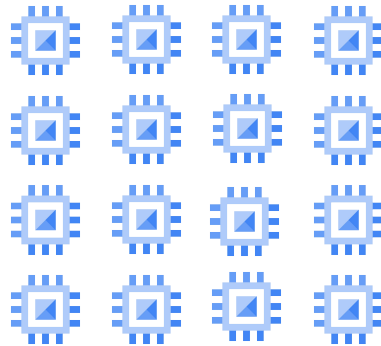
between your processing and your persistent disks. That's the default.

You can choose the machine properties of your instances, such as the number of virtual CPUs and the amount of memory, by using a set of predefined machine types or by creating your own custom machine types.

Scale up or scale out with Compute Engine



Use big VMs for memory- and compute-intensive applications



Use Autoscaling for resilient, scalable applications



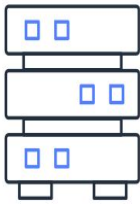
You can make very large VMs in Compute Engine. At the time this deck was produced, the maximum number of virtual CPUs in a VM was zone-dependent and at 96, and the maximum memory size was at 624 GB (6.5 GB per CPU).

You can also use a mega-memory machine type that scales to 1.4 TB memory.

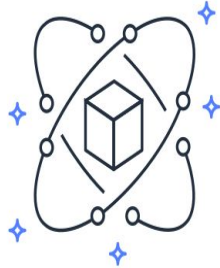
Check the Google Cloud website to see where these maximums are today.

These huge VMs are great for workloads like in-memory databases and CPU-intensive analytics. But most Google Cloud customers start off with scaling out, not up. Compute Engine has a feature called Autoscaling that lets you add and take away VMs from your application based on load metrics. The other part of making that work is balancing the incoming traffic among the VMs. And Google VPC supports several different kinds of load balancing! We'll consider those in the next section.

Compute Services on AWS



Instances



Containers



Serverless

Similarities between Compute Engine and Amazon EC2

- RAM, CPU, and GPU
- Boot disk and operating system
- Additional disks
- IP addresses
- Startup scripts with metadata



AWS and Google virtual machines have a lot in common.

- Both Compute Engine and Amazon Elastic Compute Cloud (Amazon EC2) allow you to choose and configure RAM, CPU, and GPU.
- Both offer a wide range of operating systems.
- Additional virtual disks can be added to both virtual machine types.
- Ephemeral and static public and private IP addresses can be used for both types of instances.
- Both instance types can use metadata and scripts for bootstrapping.

Differences between Compute Engine and Amazon EC2

- Faster spin-ups
- Regional persistent disks
- Preemptible VMs
- Discount pricing
- Custom machine types

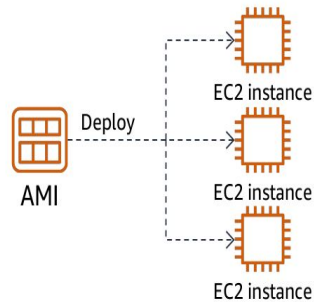


Let's look at some of the key differences between Compute Engine and Amazon EC2.

- It typically takes a Compute Engine instance about 30 seconds to start. An Amazon EC2 instance can take minutes.
- Google Cloud and AWS both offer block storage options as part of their compute services. Compute Engine provides persistent disks, and Amazon EC2 provides Elastic Block Store (EBS). Each service has several block storage types that cover a range of price and performance characteristics. Regional persistent disks provide durable storage and replication of data between two zones in the same region. If you are [designing robust systems](#) on Compute Engine, consider using regional persistent disks to maintain high availability for resources across multiple zones. AWS does not offer a regional persistent disk option.
- Amazon EC2 offers temporary instances called *spot instances*, and Compute Engine offers similar instances called *preemptible VMs*. Both provide similar functionality, but they have different pricing models. With Compute Engine, preemptible VMs pricing is fixed, although depending on the machine type, preemptible VM prices can be discounted to nearly 80% of the on-demand rate. If not reclaimed by Compute Engine, preemptible VMs run for a maximum of 24 hours and then are automatically terminated. Also, if you use a premium operating system with a license fee, you will be charged the full cost of the license while using that preemptible VM.
- Compute Engine and Amazon EC2 have similar on-demand pricing models for running instances. Both charge by the second with a minimum charge of one minute. In Amazon EC2, discounted pricing can be obtained by committing to

- provision reserved instances for one or three years. The more you pay up front, the greater the discount, subject to conditions. For Compute Engine instances, discounted pricing is obtained through sustained use, and the discount is automatically applied. The longer you use an instance in a given month, the greater the discount, with potential savings of as much as 30% of the standard on-demand rate.
- Although AWS has an extensive list of Amazon Machine Image, or AMI, options, you have to select a predefined instance to use. Google Cloud allows you to build a custom machine to your exact specification.

Amazon Machine Image (AMI)



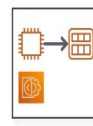
Where to get AMIs?



Use prebuilt AMIs offered by AWS.



Search the AWS Marketplace for a catalog with thousands of solutions.




Create your own AMIs manually, or use Amazon EC2 Image Builder

Amazon Machine Image (AMI)

- Provides the information required to launch an instance
- Launch multiple VMs from a single AMI with same configuration
- An AMI includes one or more Amazon Elastic Block Store (**Amazon EBS**) snapshots

Amazon EC2 Purchase Options

On-Demand	Reserved Instances	Savings Plans	Spot Instances
Pay for compute capacity by the second with no long-term commitments	Make a 1 or 3 year commitment and receive a significant discount off On-Demand prices	Same great discounts as Amazon EC2 RIs with more flexibility	Spare Amazon EC2 capacity at savings of up to 90% off On-Demand prices
			
Spiky workloads, to define needs	Committed and steady-state usage	Committed flexible access to compute	Fault-tolerant, flexible, stateless workloads

Savings Plans

- offer significant savings over On-Demand Instances, just like EC2 Reserved Instances, in exchange for a commitment to use a specific amount of compute power (measured in \$/hour) for a one or three-year period.

AWS offers two types of Savings Plans:

1. **Compute Savings Plans** provide the most flexibility and help to reduce your costs by up to 66%.
 - These plans automatically apply to EC2 instance usage regardless of instance family, size, AZ, Region, or OS.
 - For example, with Compute Savings Plans, you can change from C4 to M5 instances or

- shift a workload from EU (Ireland) to Europe (London), at any time and automatically continue to pay the Savings Plans price.

2. **EC2 Instance Savings Plans** provide the lowest prices, offering savings up to 72% (just like Standard RIs) in exchange for commitment to usage of individual instance families in a Region (for example, M5 usage in N. Virginia).

- This automatically reduces your cost on the selected instance family in that region regardless of AZ, size, or OS.
- For example, you can move from c5.xlarge running Windows to c5.2xlarge running Linux and automatically benefit from the Savings Plans prices.

Summary of Compute Engine and Amazon EC2 differences

	Compute Engine	Amazon EC2
<i>Machine RAM and CPU</i>	Machine types	Instance types
<i>Machine images</i>	Images	Amazon Machine Images
<i>Block storage</i>	Persistent disks	Elastic Block Store
<i>Local attached disk</i>	Local SSD	Ephemeral drives
<i>Discounts</i>	Preemptible VMs, Sustained-use discounts Committed-use discounts	Spot Instances, Reserved Instances Savings Plan



Let's summarize the key differences between Compute Engine and Amazon EC2.

- Compute Engine and Amazon EC2 both offer a variety of predefined instance configurations with specific amounts of virtual CPU, RAM, and network. Compute Engine refers to them as machine types; Amazon EC2 refers to these configurations as instance types.
- Compute Engine and Amazon EC2 both use machine images to create new instances. Amazon calls these images Amazon Machine Images (AMIs), and Compute Engine simply calls them images.
- Both offer block storage options as part of their compute services. Compute Engine provides persistent disks, and Amazon EC2 provides Elastic Block Store (EBS).
- On Compute Engine, local disks are referred to as local SSD and can be attached to almost any machine type. A maximum of 24 can be attached to a single instance. On Amazon EC2, local disks are called instance store or ephemeral store. These disks can be either HDD or SSD, depending on the instance type family. The number and size of these disks depends on the specific instance type and is not adjustable.
- Compute Engine and Amazon EC2 approach discount pricing in very different ways. Compute Engine offers discounts for preemptible VMs, sustained-use instances, and committed-use contracts. Amazon EC2 offers discounts for their temporary spot instances and by provisioning reserved instances. Savings Plans offer substantial savings in exchange for a commitment to consistent amount of usage for a 1- or 3-year term but is more flexible than

- Reserved Instances.

Similarities between GCE and Azure VM

- RAM, CPU, and GPU
- Boot disk and operating system
- Additional disks
- IP addresses
- Startup scripts with metadata

Azure and GCP virtual machines have a lot in common.

- Both Google Compute Engine (GCE) and Azure VM allow you to choose and configure RAM, CPU, and GPU.
- Both offer a wide range of operating systems.
- Additional virtual disks can be added to both virtual machine types.
- Ephemeral and static public and private IP addresses can be used for both types of instances.
- Both instance types can use metadata and scripts for bootstrapping.

Differences between GCE and Azure VM

- Faster spin-ups
- Regional persistent disks
- Preemptible VMs
- Discount pricing
- Custom machine types

Let's look at some of the key differences between Compute Engine and Azure VM.

- It typically takes a Compute Engine instance about 30 seconds to start. An Azure VM instance can take minutes.
- GCP and Azure both offer block storage options as part of their compute services.
 - Compute Engine provides Persistent Disks (PDs), and Azure VM provides **Azure Disk Storage**.
 - Each service has several storage types that cover a range of price and performance characteristics.
- **Regional PDs** provide durable storage and replication of data between two zones in the same region.

- If you are designing robust systems on Compute Engine, consider using regional persistent disks to maintain High Availability (HA) for resources across multiple zones.
- Azure does not offer a regional persistent disk option.
- Azure VM offers temporary instances called **low-priority VMs** and Compute Engine offers similar instances called **preemptible VMs**.
 - Both provide similar functionality, but they have different pricing models.
 - With Compute Engine preemptible VMs pricing is fixed, although depending on the machine type, preemptible VM prices can be discounted to nearly 90% of the on-demand rate.
 - If not reclaimed by Compute Engine, preemptible VMs run for a maximum of 24 hours and then are automatically terminated.
 - Also, if you use a premium operating system with a license fee, you will be charged the full cost of the license while using that preemptible VM.
- Compute Engine and Azure VM have similar on-demand pricing models for running instances.
 - Both charge by the second with a minimum charge of one minute.

- In Azure VM, discounted pricing can be obtained by committing to provision reserved instances for one or three years. The more you pay up front, the greater the discount.
- For Compute Engine instances, discounted pricing is obtained through sustained use and the discount is automatically applied.
 - The longer you use an instance in a given month, the greater the discount, with potential savings of as much as 30% of the standard on-demand rate.
- Azure has an extensive list of pre-configured Azure VMs, you have to select a predefined instance to use.
 - GCP allows you to build a custom machine to your exact specification.

Differences between GCE and Azure VM

	Google Compute Engine	Azure VM
Machine RAM/CPU	Machine types	Instance types
Machine images	Images	Azure VM Images
Block storage	Persistent disks	Azure Disk Storage
Local attached disk	Local SSD	Local SSD
Discounts	Preemptible VMs, Committed-use discounts Sustained-use discounts	Spot VMs (Unused compute capacity) Azure Reserved VM

Let's summarize the key differences between Compute Engine and Azure VM.

- Compute Engine and Azure VM both offer a variety of predefined instance configurations with specific amounts of virtual CPU, RAM, and network.
 - Compute Engine refers to them as **machine types**, Azure VM refers to these configurations as **instance types**.
- Compute Engine and Azure VM both use machine images to create new instances.
 - Azure calls these images **Azure VM Images**, and Compute Engine simply calls them **images**.
- Both offer block storage options as part of their compute services.

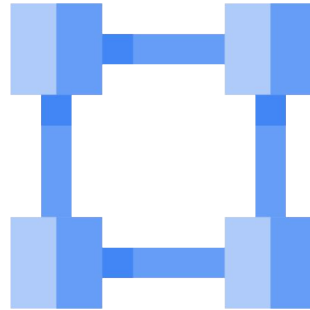
- Google Compute Engine provides persistent disks, and Azure VM provides **Azure Disk Storage**.
- On Compute Engine, local disks are referred to as **local SSD** and can be attached to almost any machine type, and a maximum of **24** can be attached to a single instance (each local SSD is **375 GiB**).
 - On Azure, local disks are also called **local SSD**. However, the number and size of these disks depend on the specific instance type.
- Compute Engine and Azure VM approach discount pricing in different ways. Azure VM has the following Pricing Models
 - Pay As You Go
 - Reserved VMs
 - Spot VMs

Some Google Cloud Updates

- Google Cloud now (2023) provides VMs up to 400+ vCPU and 12 TB
- There are now Spot VMs, which can reduce costs by up to 90%.
 - Unlike Preemptible VMs that Google can terminate for up to 24 hours,
 - Spot VMs do not have this restriction.
- **Committed use discounts** are now available as:
 - spend-based discounts
 - resource-based discounts

You control the topology of your VPC network

- Use its route table to forward traffic within the network, even across subnets.
- Use its firewall to control what network traffic is allowed.
- Use Shared VPC to share a network, or individual subnets, with other Google Cloud projects.
- Use VPC Peering to interconnect networks in Google Cloud projects.

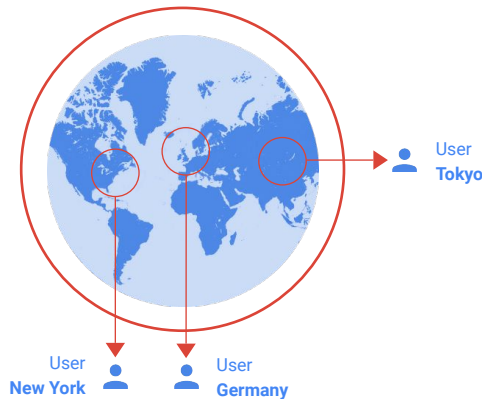


Much like physical networks, VPCs have routing tables. These are used to forward traffic from one instance to another instance within the same network, even across subnetworks and even between Google Cloud zones, without requiring an external IP address. VPCs' routing tables are built in; you don't have to provision or manage a router.

Another thing you don't have to provision or manage for Google Cloud: a firewall. VPCs give you a global distributed firewall you can control to restrict access to instances, both incoming and outgoing traffic. You can define firewall rules in terms of metadata tags on Compute Engine instances, which is really convenient. For example, you can tag all your web servers with, say, "WEB," and write a firewall rule saying that traffic on ports 80 or 443 is allowed into all VMs with the "WEB" tag, no matter what their IP address happens to be.

Recall that VPCs belong to Google Cloud projects. But what if your company has several Google Cloud projects, and the VPCs need to talk to each other? If you simply want to establish a peering relationship between two VPCs, so that they can exchange traffic, configure VPC Peering does. On the other hand, if you want to use the full power of IAM to control who and what in one project can interact with a VPC in another, configure Shared VPC.

With global Cloud Load Balancing, your application presents a single front-end to the world



- Users get a single, global anycast IP address.
- Traffic goes over the Google backbone from the closest point-of-presence to the user.
- Backends are selected based on load.
- Only healthy backends receive traffic.
- No pre-warming is required.



A few slides back, we talked about how virtual machines can autoscale to respond to changing load. But how do your customers get to your application when it might be provided by four VMs one moment and forty VMs at another? Cloud Load Balancing is the answer.

Cloud Load Balancing is a fully distributed, software-defined, managed service for all your traffic. And because the load balancers don't run in VMs you have to manage, you don't have to worry about scaling or managing them. You can put Cloud Load Balancing in front of all of your traffic: HTTP(S), other TCP and SSL traffic, and UDP traffic too.

With Cloud Load Balancing, a single anycast IP front-ends all your backend instances in regions around the world. It provides cross-region load balancing, including automatic multi-region failover, which gently moves traffic in fractions if backends become unhealthy. Cloud Load Balancing reacts quickly to changes in users, traffic, network, backend health, and other related conditions.

And what if you anticipate a huge spike in demand? Say, your online game is already a hit; do you need to file a support ticket to warn Google of the incoming load? No. No so-called "pre-warming" is required.