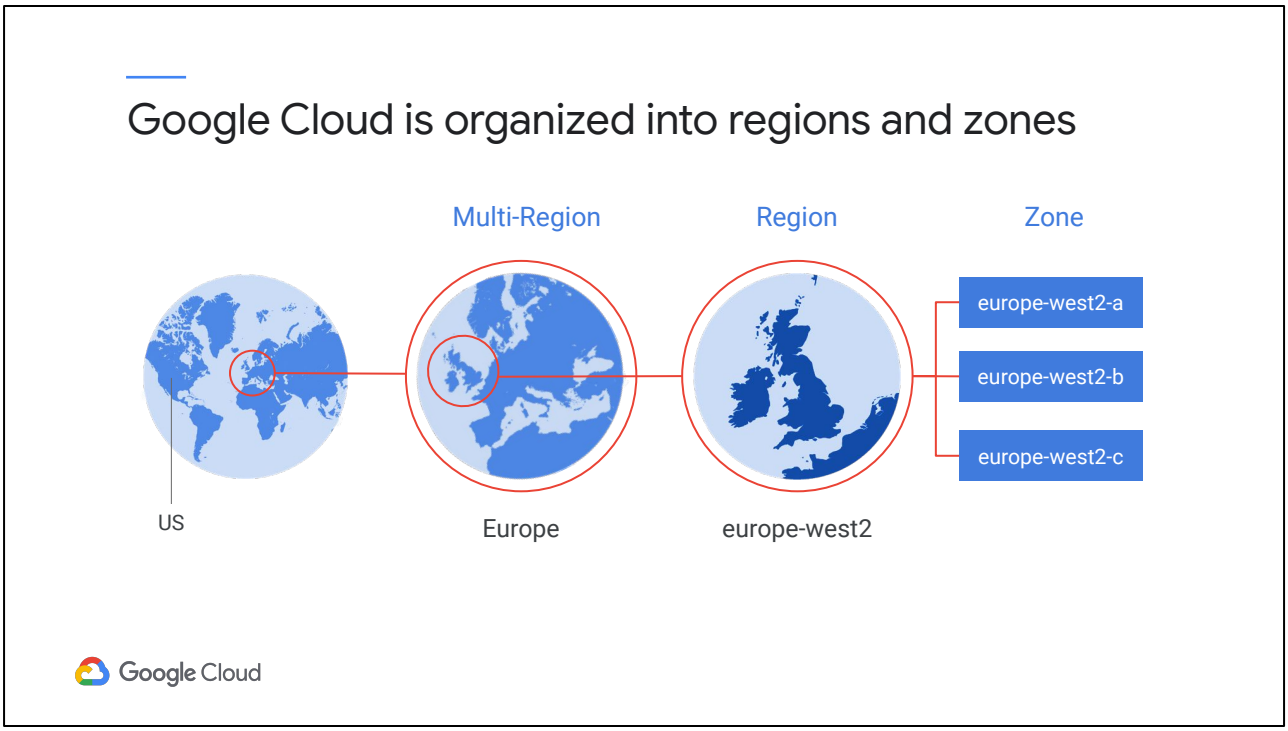


Regions and zones
Regions are independent geographic areas that consist of zones. Locations within regions tend to have round-trip network latencies of under 5 milliseconds on the 95th percentile.

A zone is a deployment area for Google Cloud resources within a region. Think of a zone as a single failure domain within a region. In order to deploy fault-tolerant applications with high availability, you should deploy your applications across multiple zones in a region to help protect against unexpected failures.

To protect against the loss of an entire region due to natural disaster, you should have a disaster recovery plan and know how to bring up your application in the unlikely event that your primary region is lost.
For more information on the specific resources available within each location option, see Google's Global Data Center Locations.

Google Cloud's services and resources can be zonal, regional, or managed by Google across multiple regions. For more information on what these options mean for your data, see geographic management of data.

Zonal resources
Zonal resources operate within a single zone. If a zone becomes unavailable, all of the zonal resources in that zone are unavailable until service is restored.

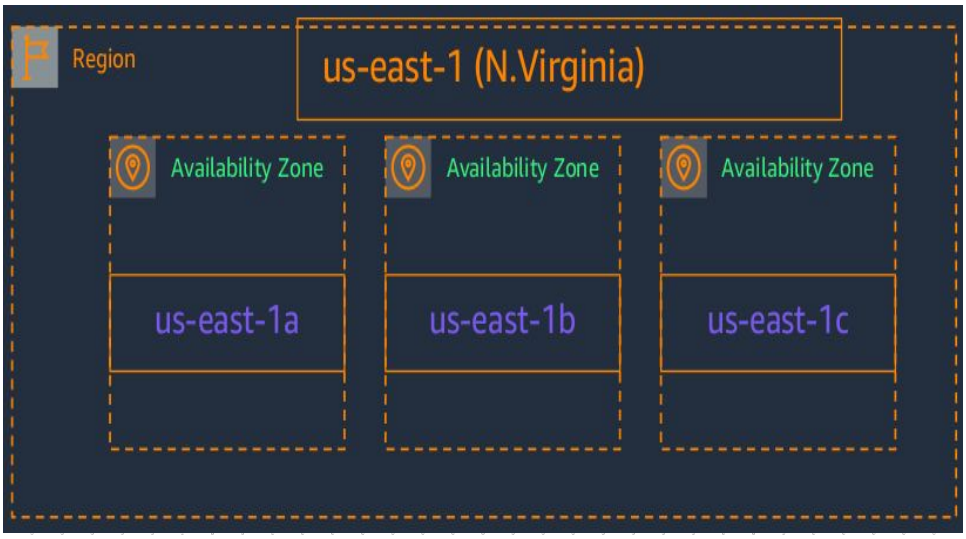
- Compute Engine VM instance resides within a specific zone.

Regional resources
Regional resources are deployed with redundancy within a region. This gives them higher availability relative to zonal resources.

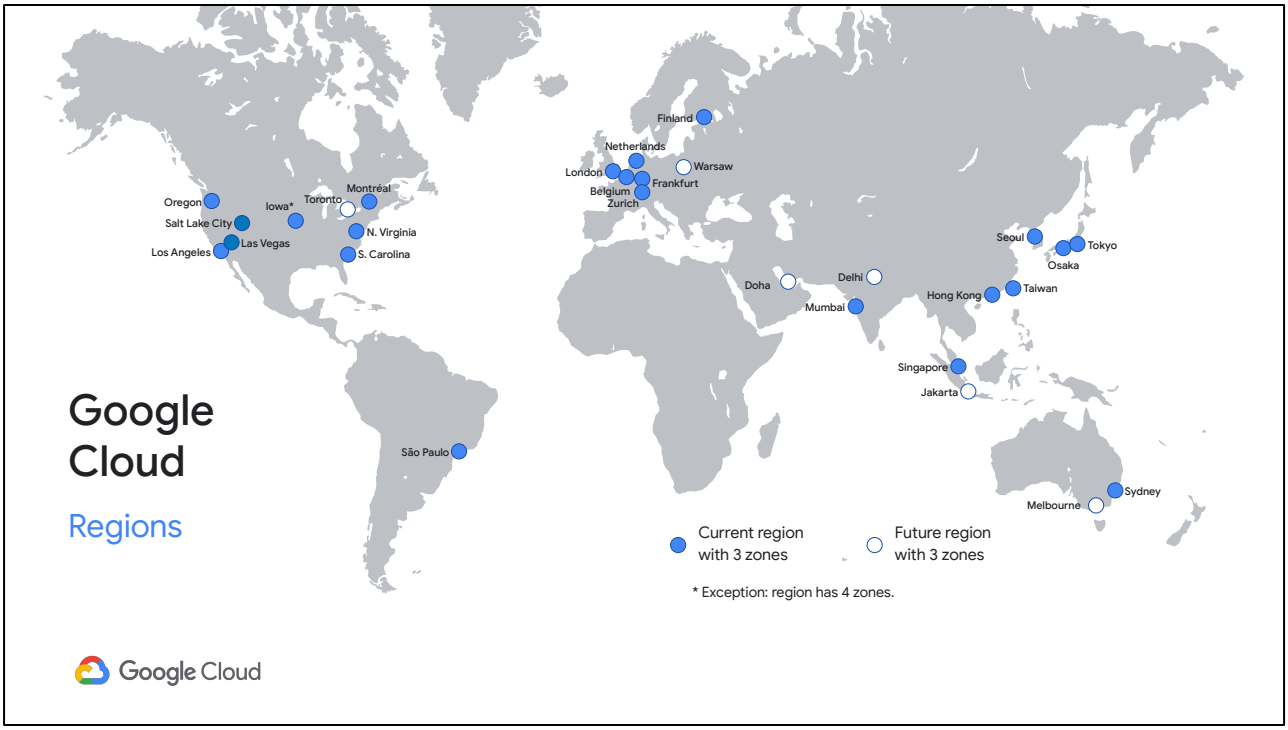
Multi-regional resources
A few Google Cloud services are managed by Google to be redundant and distributed within and across regions. These services optimize availability, performance, and resource efficiency. As a result, these services require a trade-off on either latency or the consistency model. These trade-offs are documented on a product-specific basis. The following services have one or more multi-regional deployments in addition to any regional deployments:

- App Engine and its features
- Firestore
- Cloud Storage
- BigQuery

AWS Availability Zones (AZs)



- Each AWS Region consists of multiple, isolated, and physically separate AZs within a geographic area
- AWS AZ is one or more discrete data centers with redundant power, networking, and connectivity in an AWS Region
- There is a high throughput and low latency (< 10 ms) network between AWS AZs
- There is a distance between AWS AZs of several kilometers, although all are within 100 km
- All traffic between AWS AZs is encrypted
- A minimum of 2 AZs and a maximum of 6 AZs in an AWS region, while in a Google Cloud region, the minimum is 3 zones and the maximum is 4 zones
- The number of regions is 39 in Google Cloud, 32 in AWS, and 42 in Azure.



As of mid-2020, Google Cloud has 24 regions and 73 zones with more to come.

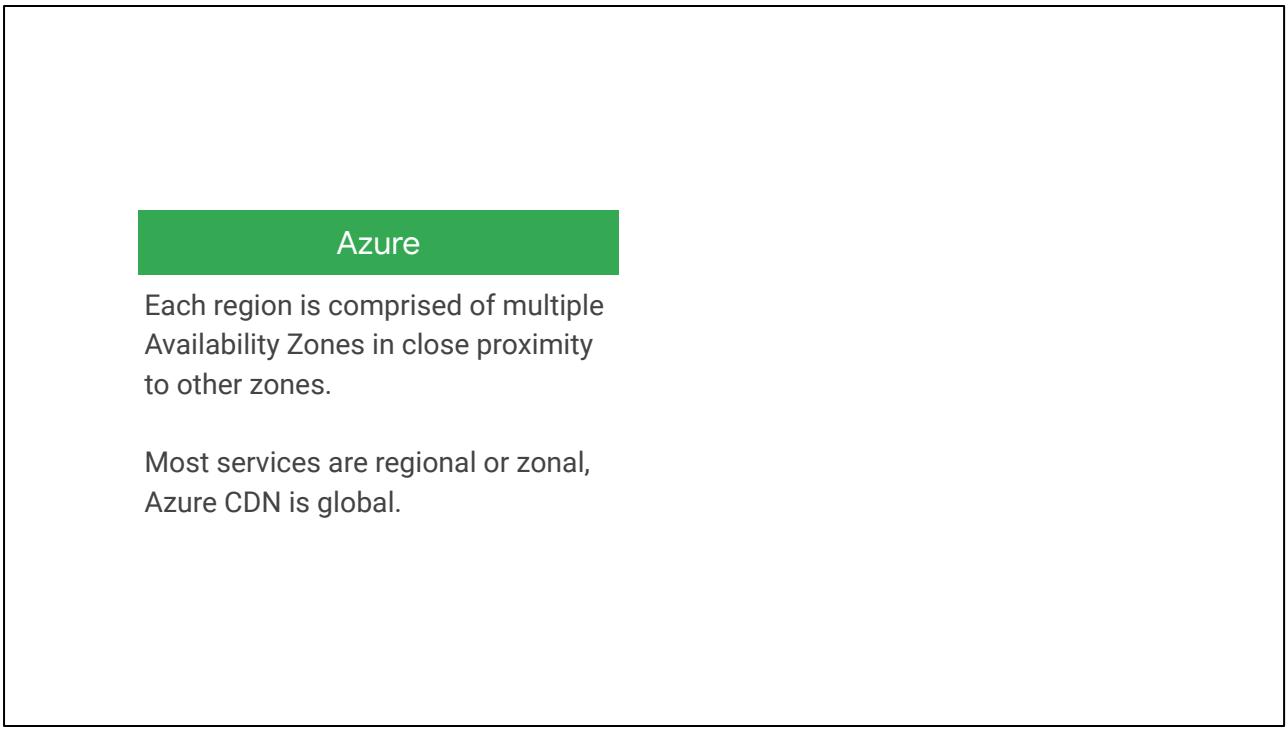
AWS Regions

Regions and zones are different in AWS

Google Cloud	AWS
Each region is composed of zones in close proximity to other zones. Google offers services that are multi-regional, global, or zonal.	Each region is composed of multiple Availability Zones in close proximity to other zones. Most services are regional or zonal; Cloudfront is global.

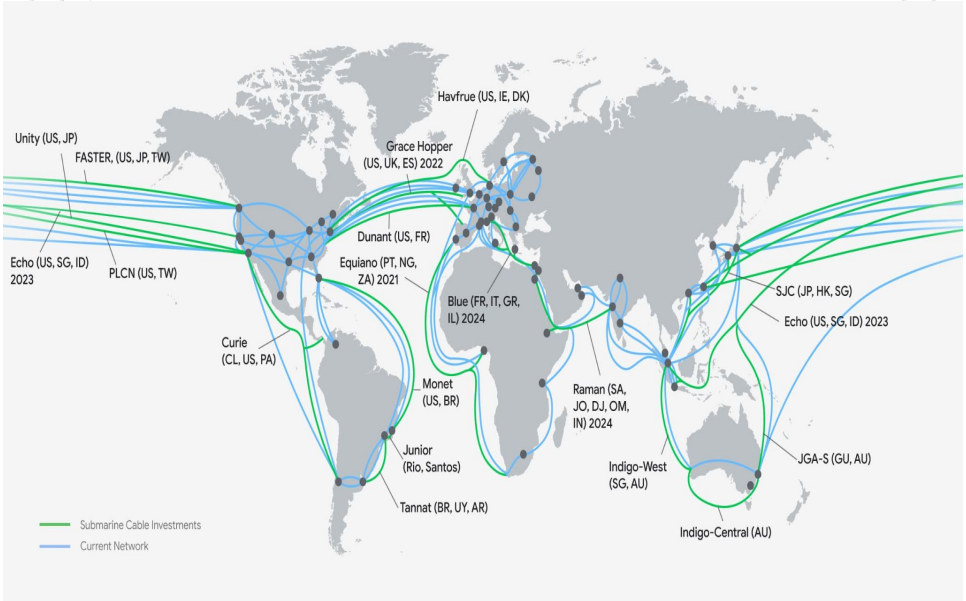
Google Cloud

Google and AWS both use regions as a way to provide cloud services to customers. One difference is that Google also uses zones to provide data center services, and every region will have at least 3 zones.
AWS uses clusters of data centers called *Availability Zones* as a way to provide high availability. Every region will have at least two availability zones.



Azure uses clusters of data centers called Availability Zones as a way to provide high availability. Every region will have at least three availability zones.

Google Cloud PoPs

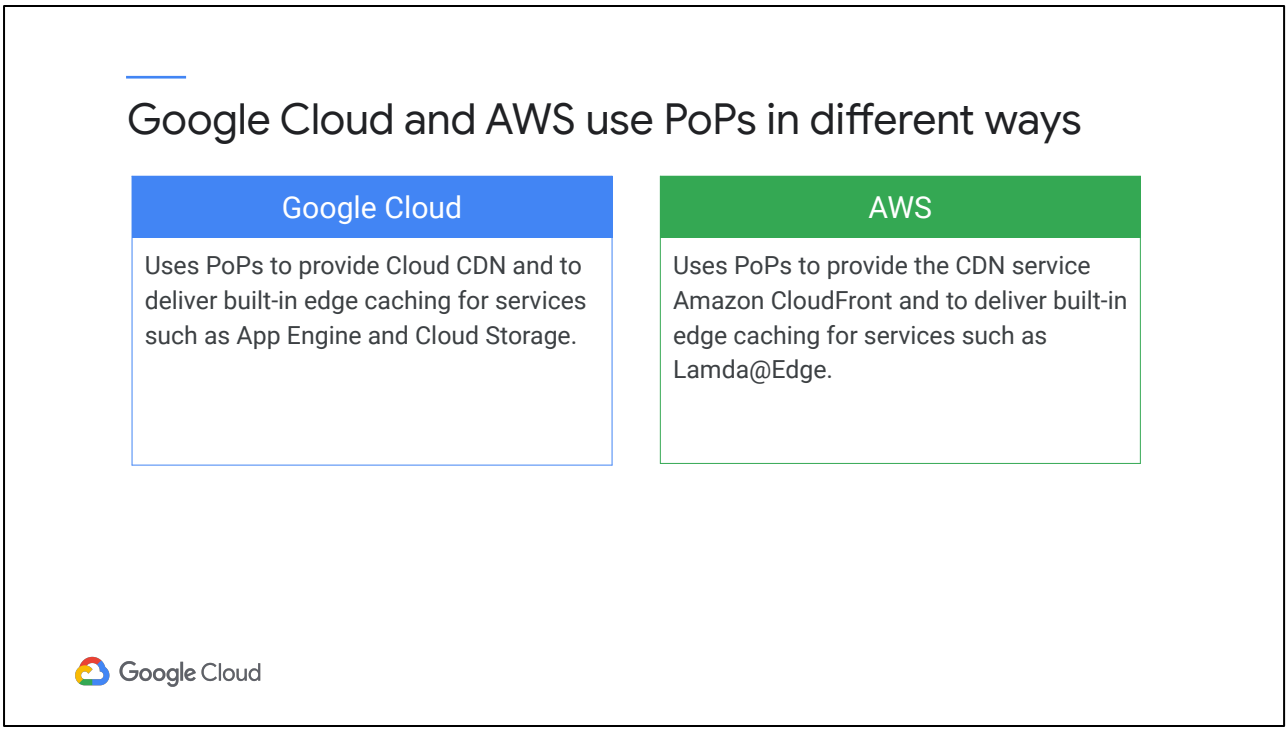


According to some publicly available estimates, Google's network carries as much as 40% of the world's internet traffic every day. Google's network is the largest network of its kind on Earth. Google has invested billions of dollars over the years to build it.

It is designed to give customers the highest possible throughput and lowest possible latencies for their applications.

The network interconnects at more than 90 Internet exchanges and more than 100 points of presence worldwide. When an Internet user sends traffic to a Google resource, Google's edge caching nodes respond to users requests from an Edge Network location that will provide the lowest latency.

AWS PoPs

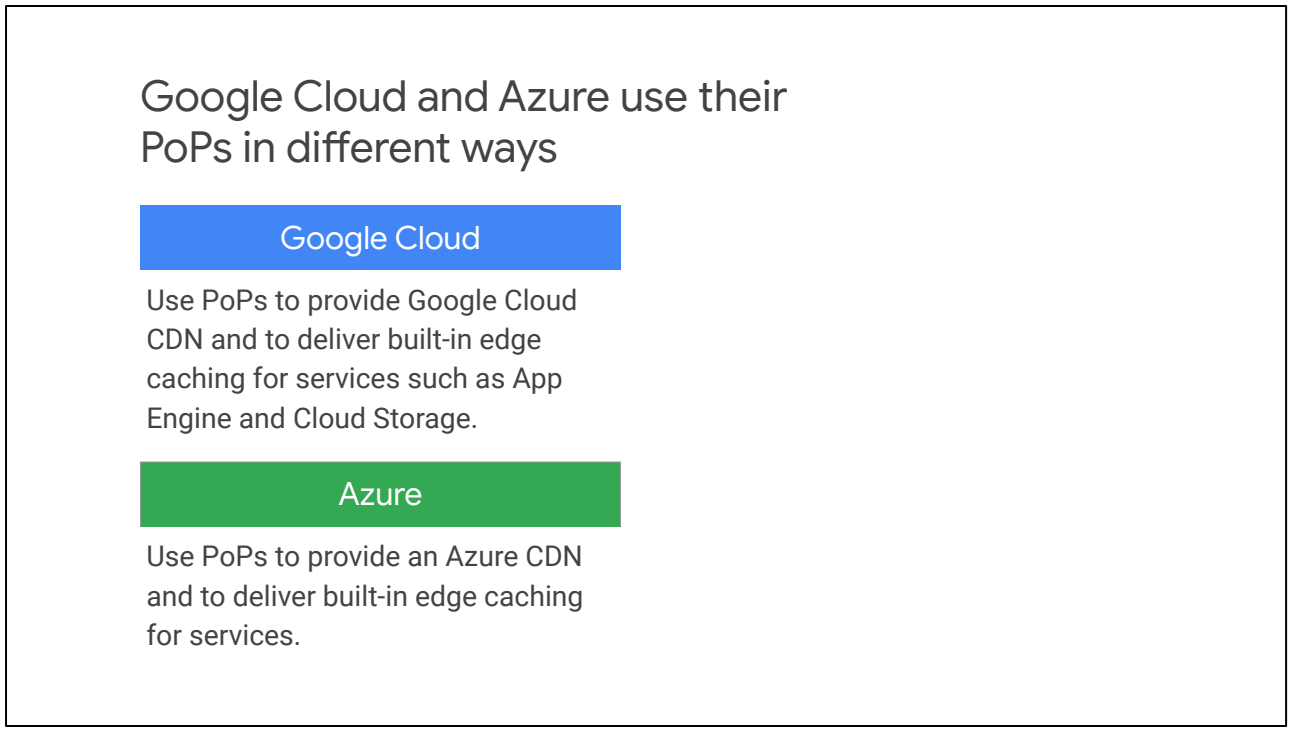


Google Cloud and AWS both have points of presence (PoPs) located in many more locations around the world. These points of presence locations help cache content closer to end users. However, each platform uses their respective points of presence locations in different ways.

Google Cloud uses points of presence to provide Cloud CDN and to deliver built-in edge caching for services such as App Engine and Cloud Storage.

AWS uses points of presence to provide the content delivery network service Amazon CloudFront and for edge caching services like Lamda at the edge.

Google Cloud's points of presence connect to data centers through Google-owned fiber. This unimpeded connection means that Google Cloud-based applications have fast, reliable access to all of the services on Google Cloud.



Azure uses points of presence to provide a content delivery network (CDN) service across 4 products (Azure CDN Standard from Microsoft, Azure CDN Standard from Akamai, Azure CDN Standard from Verizon, and Azure CDN Premium from Verizon).

GCP's points of presence connect to data centers through Google-owned fiber. This unimpeded connection means that GCP-based applications have fast, reliable access to all of the services on GCP.

- Google Cloud and Azure both use regions as a way to provide cloud services to customers.
- Google Cloud and Azure both have PoPs located in many more locations around the world.
    - These Points of Presence locations help cache content closer to end users.
- Azure announced about 60+ regions, including the Reserved Access Regions
    - The physical distance between the Azure AZs is undefined and varies
    - Azure regions provide AZs, which are separated groups of data centers within a region
    - AZs are close enough to have low-latency connections to other AZs in the same region
    - A high-performance network connects Azure AZs with a round-trip latency of less than 2ms in the same region

Summary of region and zones terminology

Concept	Google Cloud term	AWS term
Cluster of data centers and services	Region	Region
Abstracted data center	Zone	Availability Zone
Edge caching	PoP (multiple services)	PoP (multiple services)

Google Cloud

To summarize let's look at the terminology associated with region and zones. Both Google Cloud and AWS products use the term *region* to define a cluster of data centers and services that are relatively close to each other.

Data center services and availability can be abstracted into Zones in Google Cloud, which are the equivalent of Availability Zones in AWS.

Google Cloud uses POPs to deliver built-in edge caching for multiple services, such as App Engine and Cloud Storage. AWS delivers edge caching in a similar way.

Concept	Google Cloud term	Azure term
Cluster of data centers and services	Region	Region

To summarize the terminology associated with region and zones, both GCP and Azure products use the term region to define a cluster of data centers and services that are in relatively close proximity to each other.

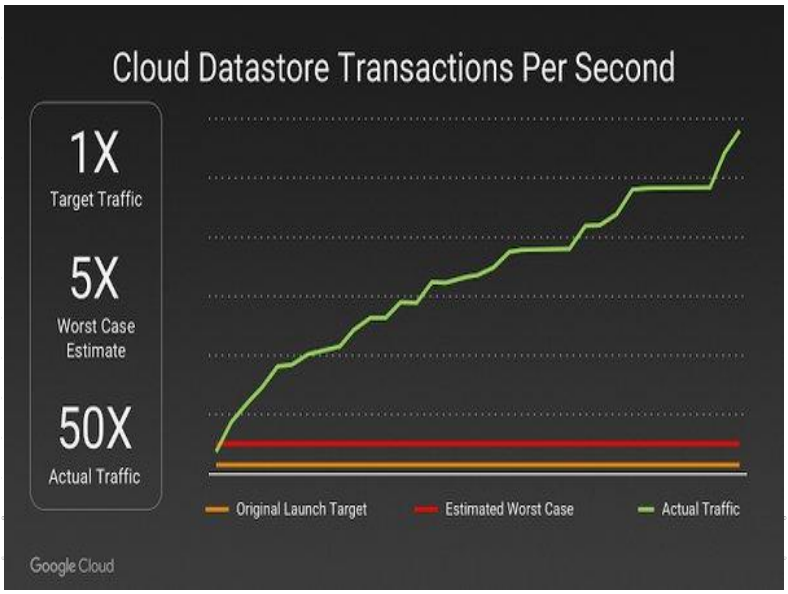
Concept	Google Cloud term	Azure term
Cluster of data centers and services	Region	Region
Abstracted data center	Zone	Availability Zone

Data Center services and availability can be abstracted into Zones in GCP which are the equivalent of Availability Zones in Azure.

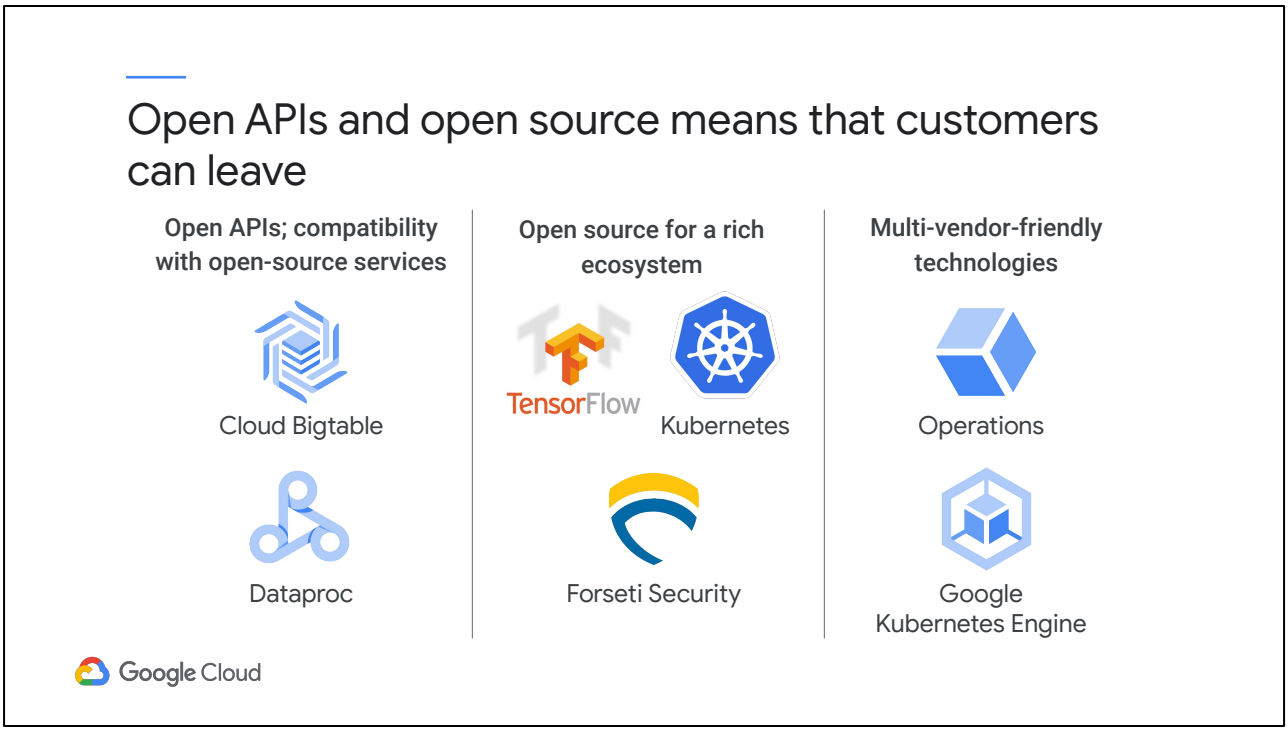
Concept	Google Cloud term	Azure term
Cluster of data centers and services	Region	Region
Abstracted data center	Zone	Availability Zone
Edge caching	PoP (multiple services)	PoP (multiple services)

Both Google Cloud and Azure use POPs to deliver built-in edge caching for multiple services.

Google App Engine (GAE) Use Case: Pokémon GO



In the first week since its release, Pokémon Go was downloaded over 10 million times in the United States.



Google gives customers the ability to run their applications elsewhere if Google becomes no longer the best provider for their needs.

This includes:

- Using Open APIs. Google services are compatible with open-source products. For example, Cloud Bigtable, a horizontally scalable managed database: Bigtable uses the Apache HBase interface, which gives customers the benefit of code portability. Another example: Dataproc offers the open-source big data environment Hadoop as a managed service.
- Google publishes key elements of its technology, using open-source licenses, to create ecosystems that provide customers with options other than Google. For example, TensorFlow, an open-source software library for machine learning developed inside Google, is at the heart of a strong open-source ecosystem.
- Google provides interoperability at multiple layers of the stack. Kubernetes and Google Kubernetes Engine give customers the ability to mix and match microservices running across different clouds. Google Cloud's operations suite lets customers monitor workloads across multiple cloud providers.

Security is designed into Google's technical infrastructure

Layer	Notable security measures (among others)
Operational security	Intrusion detection systems; techniques to reduce insider risk; employee U2F use; software development practices
Internet communication	Google Front End; designed-in Denial of Service protection
Storage services	Encryption at rest
User identity	Central identity service with support for U2F
Service deployment	Encryption of inter-service communication
Hardware infrastructure	Hardware design and provenance; secure boot stack; premises security

Google Cloud

Hardware design and provenance: Both the server boards and the networking equipment in Google data centers are custom-designed by Google. Google also designs custom chips, including a hardware security chip that is currently being deployed on both servers and peripherals.

Secure boot stack: Google server machines use a variety of technologies to ensure that they are booting the correct software stack, such as cryptographic signatures over the BIOS, bootloader, kernel, and base operating system image.

Premises security: Google designs and builds its own data centers, which incorporate multiple layers of physical security protections. Access to these data centers is limited to only a very small fraction of Google employees. Google additionally hosts some servers in third-party data centers, where we ensure that there are Google-controlled physical security measures on top of the security layers provided by the data center operator.

Encryption of inter-service communication: Google's infrastructure provides cryptographic privacy and integrity for remote procedure call ("RPC") data on the network. Google's services communicate with each other using RPC calls. The infrastructure automatically encrypts all infrastructure RPC traffic which goes between data centers. Google has started to deploy hardware cryptographic accelerators that will allow it to extend this default encryption to all infrastructure RPC traffic inside Google data centers.

User identity: Google's central identity service, which usually manifests to end users as the Google login page, goes beyond asking for a simple username and password. The service also intelligently challenges users for additional information based on risk factors such as whether they have logged in from the same device or a similar location in the past. Users also have the option of employing second factors when signing in, including devices based on the Universal 2nd Factor (U2F) open standard

Encryption at rest: Most applications at Google access physical storage indirectly via storage services, and encryption (using centrally managed keys) is applied at the layer of these storage services. Google also enables hardware encryption support in hard drives and SSDs.

Google Front End ("GFE"): Google services that want to make themselves available on the Internet register themselves with an infrastructure service called the Google Front End, which ensures that all TLS connections are ended using correct certificates and following best practices such as supporting perfect forward secrecy. The GFE additionally applies protections against Denial of Service attacks.

Denial of Service ("DoS") protection: The sheer scale of its infrastructure enables Google to simply absorb many DoS attacks. Google also has multi-tier, multi-layer DoS protections that further reduce the risk of any DoS impact on a service running behind a GFE.

Intrusion detection: Rules and machine intelligence give operational security engineers warnings of possible incidents. Google conducts Red Team exercises to measure and improve the effectiveness of its detection and response mechanisms.

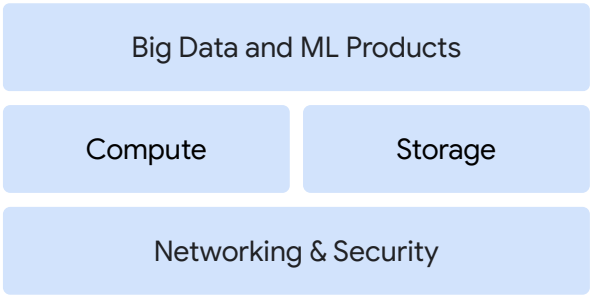
Reducing insider risk: Google aggressively limits and actively monitors the activities of employees who have been granted administrative access to the infrastructure.

Employee U2F use: To guard against phishing attacks against Google employees, employee accounts require use of U2F-compatible Security Keys.

Software development practices: Google employs central source control and requires two-party review of new code. Google also provides its developers libraries that prevent them from introducing certain classes of security bugs. Google also runs a Vulnerability Rewards Program where we pay anyone who is able to discover and inform us of bugs in our infrastructure or applications.

For more information about Google's technical-infrastructure security, see https://cloud.google.com/security/security-design/

The Google Cloud infrastructure



You can think of the Google Cloud infrastructure in terms of three layers.

- At the base layer is **networking and security**, which lays the foundation to support all of Google's infrastructure and applications.
- On the next layer sit **compute** and **storage**. Google Cloud separates, or decouples, as it's technically called, compute and storage so they can scale independently based on need.
- And on the top layer sit the **big data and machine learning products**, which enable you to perform tasks to ingest, store, process, and deliver business insights, data pipelines, and ML models.

And thanks to Google Cloud, these tasks can be accomplished without needing to manage and scale the underlying infrastructure.