Chapter 1

Overview of Google Cloud

THIS CHAPTER COVERS THE FOLLOWING OBJECTIVE OF THE GOOGLE ASSOCIATE CLOUD ENGINEER CERTIFICATION EXAM:

✓ **1.0 Setting up cloud projects and accounts**

Google Cloud is a public cloud service that offers some of the same technologies used by Google to deliver its own products. This chapter describes the most important components of Google Cloud and discusses how it differs from on-premises data center–based computing.

Types of Cloud Services

Public cloud providers such as Google, Amazon, and Microsoft offer a range of services for deploying computing, storage, networking, and other infrastructures to run a wide array of business services and applications. Some cloud users are new companies that start in the cloud. They have never owned their own hardware infrastructure. Other cloud customers are enterprises with multiple data centers that use public clouds to supplement their data centers. These different kinds of users have different requirements.

A company that starts on the cloud can choose services that best fit its application and architectural needs without having to consider existing infrastructure. For example, a startup could use Google Cloud's Cloud Identity, including Identity Access Management services, for all authentication and authorization needs. A company that has already invested in a Microsoft Active Directory solution for identity management may want to leverage that system instead of working solely with the cloud's identity management system. This can lead to additional work to integrate the two systems and keep them synchronized.

Another area of concern for enterprises with their own infrastructure is establishing and maintaining a secure network between their on-premises resources and their public cloud resources. If there will be high-volume network traffic between the on-premises systems and the public cloud, the enterprise may need to invest in dedicated networking between its data center and a facility of the public cloud provider. If the volume of traffic does not justify the cost of a dedicated connection between facilities, then the company may use a virtual private network that runs over the public Internet. This requires additional network design and management that a company that is solely in the cloud would not have to address.

Public cloud providers offer services that fall into four broad categories:

- Compute resources
- Storage
- Networking
- Specialized services such as machine learning services

Cloud customers typically make use of services in more than one of these categories.

Compute Resources

Computing resources come in a variety of forms in public clouds.

Virtual Machines

Virtual machines (VMs) are a basic unit of computing resources and a good starting point for experimenting with the cloud. After you create an account with a cloud provider and provide billing information, you can create a project, which is a logical grouping of Google Cloud resources. After you have a project, you can use a portal or command-line tools to create VMs in a project. Google Cloud offers a variety of preconfigured VMs with varying numbers of vCPUs and amounts of memory. You can also create a custom configuration if the preconfigured offerings do not meet your needs.

Once you create a VM, you can log into it and administer it as you like. You have full access to the VM, so you can configure filesystems, add persistent storage, patch the operating system, or install additional packages. You decide what to run on the VM, who else will have access to it, and when to shut down the VM. A VM that you manage is like having a server in your office that you have full administrator rights to.

You can, of course, create multiple VMs running different operating systems and applications. Google Cloud also provides services, such as load balancers, that provide a single access point to a distributed back end. This is especially useful when you need to have high availability for your application. If one of the VMs in a cluster fails, the workload can be directed to the other VMs in the cluster. Autoscalers can add or remove VMs from the cluster based on the workload. This is called *autoscaling*. This helps both control cost by not running more VMs than needed and ensure that sufficient computing capacity is available when workloads increase.

Managed Kubernetes Clusters

Google Cloud gives you all the tools you need to create and manage clusters of servers. Many cloud users would rather focus on their applications and not the tasks needed to keep a cluster of servers up and running. For those users, managed clusters are a good option.

Managed clusters make use of containers. A *container* is like a lightweight VM that isolates processes running in one container from processes running in another container on the same server. In a managed cluster, you can specify the number of servers you would like to run and the containers that should run on them. You can also specify autoscaling parameters to optimize the number of containers running.

In a managed cluster, the health of containers is monitored for you. If a container fails, the cluster management software will detect it and start another container.

Containers are good options when you need to run applications that depend on multiple microservices running in your environment. The services are deployed through containers, and the cluster management service takes care of monitoring, networking, and some security management tasks.

Serverless Computing

Both VMs and managed Kubernetes clusters require some level of effort to configure and administer computing resources. Serverless computing is an approach that allows developers and application administrators to run their code in a computing environment that does not require setting up VMs or Kubernetes clusters.

Google Cloud has three serverless computing options: App Engine, Cloud Run, and Cloud Functions. App Engine is used for applications and containers that run for extended periods of time, such as a website back end, point-of-sale system, or custom business application. Cloud Run is also used to run containers when the full features of Kubernetes Engine are not needed. Cloud Run is used for containers when you want a fully managed service and rapid autoscaling for your stateless applications. Cloud Functions is a platform for running code in response to an event, such as uploading a file or adding a message to a message queue. This serverless option works well when you need to respond to an event by running a short process coded in a function or by calling a longer-running application that might be running on a VM, managed cluster, or App Engine.

Storage

Public clouds offer a few types of storage services that are useful for a wide range of application requirements. These types include the following:

- Object storage
- File storage
- Block storage
- Caches

Enterprise users of cloud services will often use a combination of these services.

Object Storage

Object storage is a system that manages the use of storage in terms of objects or blobs. Usually these objects are files, but it is important to note that the files are not stored in a conventional filesystem. Objects are grouped into *buckets*. Each object is individually addressable, usually by a URL.

Object storage is not limited by the size of disks or solid-state drives (SSDs) attached to a server. Objects can be uploaded without concern for the amount of space available on a disk. Multiple copies of objects are stored to improve availability and durability. In some cases, copies of objects may be stored in different regions to ensure availability even if a region becomes inaccessible.

Another advantage of object storage is that it is serverless. There is no need to create VMs and attach storage to them. Google Cloud's object storage, called Cloud Storage, is accessible from servers running in Google Cloud as well as from other devices with Internet access.

File Storage

File storage services provide a hierarchical storage system for files. Filesystem storage provides network-shared filesystems. Google Cloud has a file storage service called Cloud Filestore, which is based on the Network File System (NFS) storage system.

File storage is suitable for applications that require operating system–like file access to files. The file storage system decouples the filesystem from specific VMs. The filesystem, its directories, and its files exist independent of VMs or applications that may access those files.

Block Storage

Block storage uses a fixed-size data structure called a *block* to organize data. Block storage is commonly used in ephemeral and persistent disks attached to VMs. With a block storage system, you can install filesystems on top of the block storage, or you can run applications that access blocks directly. Some relational databases can be designed to access blocks directly rather than working through filesystems.

In Linux filesystems, 4 KB is a common block size. Relational databases often write directly to blocks, but they often use larger sizes, such as 8 KB or more.

Block storage is available on disks that are attached to VMs in Google Cloud. Block storage can be either persistent or ephemeral. A persistent disk continues to exist and store data, even if it is detached from a virtual server or the virtual server to which it is attached shuts down. Ephemeral disks exist and store data only as long as a VM is running. Ephemeral disks are deleted when the VM is shut down. Persistent disks are used when you want data to exist on a block storage device independent of a VM. These disks are good options when you have data that you want available independent of the life cycle of a VM, and support fast operating system– and filesystem-level access.

Object storage also keeps data independent of the life cycle of a VM, but it does not support operating system– or filesystem-level access; you have to use higher-level protocols like HTTP to access objects. It takes longer to retrieve data from object storage than to retrieve it from block storage. You may need a combination of object storage and block storage to meet your application needs. Object storage can store large volumes of data that are copied to persistent disk when needed. This combination gives the advantage of large volumes of storage along with operating system– and filesystem-based access when needed.

Caches

Caches are in-memory data stores that maintain fast access to data. The time it takes to retrieve data is called *latency*. The latency of in-memory stores is designed to be submillisecond. To give you a comparison, here are some other latencies:

- Making a main memory reference takes 100 nanoseconds, or 0.1 microsecond.
- Reading 4 KB randomly from an SSD takes 150 microseconds.
- Reading 1 MB sequentially from memory takes 250 microseconds.

- Reading 1 MB sequentially from an SSD takes 1,000 microseconds, or 1 millisecond.
- Reading 1 MB sequentially from disk takes 20,000 microseconds, or 20 milliseconds.

Here are some conversions for reference:

- 1,000 nanoseconds equal 1 microsecond.
- 1,000 microseconds equal 1 millisecond.
- 1,000 milliseconds equal 1 second.

These and other useful timing data are available at Jonas Bonér's "Latency Numbers Every Programmer Should Know" at `https://gist.github.com/jboner/2841832`.

Let's work through an example of reading 1 MB of data. If you have the data stored in an in-memory cache, you can retrieve the data in 250 microseconds, or 0.25 millisecond. If that same data is stored on an SSD, it will take four times as long to retrieve at 1 millisecond. If you retrieve the same data from a hard disk drive (HDD), you can expect to wait 20 milliseconds, or 80 times as long as reading from an in-memory cache.

Caches are quite helpful when you need to keep read latency to a minimum in your application. Of course, who doesn't love fast retrieval times? Why don't we always store our data in caches? There are three reasons:

- Memory is more expensive than SSD or HDD storage. It's not practical in many cases to have as much in-memory storage as persistent block storage on SSDs or HDDs.
- Caches are volatile; you lose the data stored in the cache when power is lost or the operating system is rebooted. You can store data in a cache for fast access, but it should never be used as the only data store keeping the data. Some form of persistent storage should be used to maintain a "system of truth," or a data store that always has the latest and most accurate version of the data.
- Caches can get out of synchronization with the system of truth. This can happen if the system of truth is updated but the new data is not written to the cache. When this happens, it can be difficult for an application that depends on the cache to detect the fact that data in the cache is invalid. If you decide to use a cache, be sure to design a cache update strategy that meets your requirements for consistency between the cache and the system of truth. This is such a challenging design problem that it has become memorialized in Phil Karlton's well-known quip, "There are only two hard things in computer science: cache invalidation and naming things." (See `https://martinfowler.com/bliki/TwoHardThings.html` for riffs on this rare example of computer science humor.)

Real World Scenario

Improving Database Query Response Time

Users expect web applications to be highly responsive. If a page takes more than 2 to 3 seconds to load, the user experience can suffer. It is common to generate the content of a page using the results of a database query, such as looking up account information by

customer ID. When a query is made to the database, the database engine will look up the data, which is usually on disk. The more users query the database, the more queries it has to serve. Databases keep a queue for queries that need to be answered but can't be processed yet because the database is busy with other queries. This can cause longer latency response time, since the web application will have to wait for the database to return the query results.

One way to reduce latency is to reduce the time needed to read the data. In some cases, it helps to replace hard disk drives with faster SSD drives. However, if the volume of queries is high enough that the queue of queries is long even with SSDs, another option is to use a cache.

When query results are fetched, they are stored in the cache. The next time that information is needed, it is fetched from the cache instead of the database. This can reduce latency because data is fetched from memory, which is faster than disk. It also reduces the number of queries to the database, so queries that can't be answered by looking up data in the cache won't have to wait as long in the query queue before being processed.

This approach would require changes to the application code to store query results in the cache and to check the cache for data before querying the database.

Networking

When working in the cloud, you'll need to work with networking between your cloud resources and possibly with your on-premises systems.

When you have multiple VMs running in your cloud environment, you will likely need to manage IP addresses at some point. Each network-accessible device or service in your environment will need an IP address. In fact, devices within Google Cloud can have both internal and external addresses. Internal addresses are accessible only to services in your internal network. Your internal Google Cloud network is defined as a virtual private cloud (VPC). External addresses are accessible from the Internet.

External IP addresses can be either static or ephemeral. Static addresses are assigned to a device for extended periods of time. Ephemeral external IP addresses are attached to VMs and released when the VM is stopped.

In addition to specifying IP addresses, you will often need to define firewall rules to control access to subnetworks and VMs in your VPC. For example, you may have a database server that you want to restrict access to so that only an application server can query the database. A firewall rule can be configured to limit inbound and outbound traffic to the IP address of the application server or load balancer in front of the application cluster.

You may need to share data and network access between an on-premises data center and your VPC. You can do this using one of several types of *peering*, which is the general term for linking distinct networks. Google Cloud offers several types of peering, including VPNs, Interconnects, Shared VPC, VPC networking peering, and Direct or Carrier peering.

Specialized Services

Most public cloud providers offer specialized services that can be used as building blocks of applications or as part of a workflow for processing data. Common characteristics of specialized services are as follows:

- They are serverless; you do not need to configure servers or clusters.
- They provide a specific function, such as translating text or analyzing images.
- They provide an application programming interface (API) to access the functionality of the service.
- As with other cloud services, you are charged based on your use of the service.

These are some of the specialized services in Google Cloud:

- AutoML, a machine learning service
- Cloud Natural Language, a service for analyzing text
- Speech-to-Text for converting spoken language to text
- Recommendations AI for personalized product recommendations

Specialized services encapsulate advanced computing capabilities and make them accessible to developers who are not experts in domains, such as natural language processing and machine learning. Expect to see more specialized services added to Google Cloud.

Cloud Computing vs. Data Center Computing

Although it may seem that running VMs in the cloud is not much different from running them in your data center, there are significant differences between operating IT environments in the cloud and an on-premises or colocated data center.

Rent Instead of Own Resources

Corporate data centers are filled with servers, disk arrays, and networking equipment. This equipment is often owned or leased for extended periods by the company, a model that requires companies to either spend a significant amount of money up front to purchase equipment or commit to a long-term lease for the equipment. This approach works well when an organization can accurately predict the number of servers and other equipment it will need for an extended period and it can utilize that equipment consistently.

The model does not work as well when companies have to plan for peak capacity that is significantly higher than the average workload. For example, a retailer may have an

average load that requires a cluster of 20 servers but during the holiday season the workload increases to the point where 80 servers are needed. The company could purchase 80 servers and let 60 idle for most of the year to have resources to accommodate peak capacity. Alternatively, it could purchase or lease fewer servers and tolerate the loss in business that would occur when its compute resources can't keep up with demand. Neither is an appealing option.

Public clouds offer an alternative of short-term rental of compute capacity. The retailer, for example, could run VMs in the cloud during peak periods in addition to its on-premises servers. This gives the retailer access to the servers it needs when it needs them without having to pay for them when they are not needed.

The unit cost of running servers in the cloud may be higher than that of running the equivalent server in the data center, but the total cost of on-premises and short-term in the cloud mix of servers may still be significantly less than the cost of purchasing or leasing for peak capacity and leaving resources idle.

Pay-as-You-Go-for-What-You-Use Model

Related to the short-term rental model of cloud computing is the pay-as-you-go model. When you run a virtual server in the cloud, you will typically pay for a minimum period, such as 1 minute, and pay per second thereafter. The unit cost per second will vary depending on the characteristics of the server. Servers with more CPUs and memory will cost more than servers with fewer CPUs and less memory.

It is important for cloud engineers to understand the pricing model of their cloud provider. It is easy to run up a large bill for servers and storage if you are not monitoring your usage. In fact, some cloud customers find that running applications in the cloud can be more expensive than running them on-premises.

Elastic Resource Allocation

Another key differentiator between on-premises and public cloud computing is the ability to add and remove compute and storage resources on short notice. In the cloud, you could start 20 servers in a matter of minutes. In an on-premises data center, it could take days or weeks to do the same thing if additional hardware must be provisioned.

Cloud providers design their data centers with extensive compute, storage, and network resources. They optimize their investment by efficiently renting these resources to customers. With sufficient data about customer use patterns, they can predict the capacity they need to meet customer demand. Since they have many customers, the variation in demand of any one customer has little effect on the overall use of their resources.

Extensive resources and the ability to quickly shift resources between customers enables public cloud providers to offer elastic resource allocation more efficiently than can be done in smaller data centers.

Specialized Services

Specialized services are, by their nature, not widely understood. Many developers understand how to develop user interfaces or query a database, but fewer have been exposed to the details of natural language processing or machine learning. Large enterprises may have the financial resources to develop in-house expertise in areas such as data science and machine vision, but many others don't.

By offering specialized services, cloud providers are bringing advanced capabilities to a wider audience of developers. Like investing in large amounts of hardware, public cloud vendors can invest in specialized services and recover their costs and make a profit because the specialized services are used by a large number of customers.

Summary

Google Cloud offers a variety of services for compute, storage, networking, and specialized services. Compute services include virtual machines and Kubernetes clusters, while storage services support object and file storage along with caching. Networking services provide virtual private clouds and other services including VPNs, Interconnects, Shared VPC, VPC networking peering, and Direct or Carrier peering. Specialized services include machine learning, Speech-to-Text, and recommendation services.

Cloud computing has several advantages over on-premises computing including: renting rather than owning infrastructure, a pay-as-you-go model, elastic resource allocation, and specialized services.

Exam Essentials

Understand different ways of delivering cloud computing resources. Computing resources can be allocated as individual VMs or clusters of VMs that you manage. You can also use managed Kubernetes clusters that relieve you of some of the operational overhead of managing a Kubernetes cluster. Serverless computing options relieve users of any server management. Instead, developers run their code in a containerized environment managed by the cloud provider or in a compute platform designed for short-running code. Developers and DevOps professionals have the most control over resources when they manage their own servers and clusters. Managed services and serverless options are good choices when you do not need control over the computing environment and will get more value from not having to manage compute resources.

Understand the different forms of cloud storage and when to use them. There are four main categories of storage: object, file, block, and in-memory caches. Object storage is designed for highly reliable and durable storage of objects, such as images or data sets.

Object storage has more limited functionality than filesystem-based storage systems. Filesystem-based storage provides hierarchical directory storage for files and supports common operating system and filesystem functions. Filesystem services provide network-accessible filesystems that can be accessed by multiple servers. Block storage is used for storing data on disks. Filesystems and databases make use of block storage systems. Block storage is used with persistent storage devices, such as SSDs and HDDs. Caches are in-memory data stores used to minimize the latency of retrieving data. They do not provide persistent storage and should never be considered a "system of truth."

Understand the differences between running an IT environment on-premises or in the cloud. Running an IT environment in the cloud has several advantages, including short-term rental of resources, a pay-as-you-go model, elastic resource allocation, and the ability to use specialized services. The unit cost of cloud resources, such as the cost per minute of a mid-tier server, may be higher in the cloud than on-premises. It is important to understand the cost model of your cloud provider so that you can make decisions about the most efficient distribution of workload between cloud and on-premises resources.

Review Questions

You can find the answers in the Appendix.

1. Which of the following is an option for choosing a computing resource in Google Cloud?
   - **A.** Cache
   - **B.** Virtual machine (VM)
   - **C.** Block
   - **D.** Subnet

2. If you use a cluster that is managed by a cloud provider, which of these will be managed for you by the cloud provider?
   - **A.** Monitoring
   - **B.** Networking
   - **C.** Some security management tasks
   - **D.** All of the above

3. You need serverless computing for file processing and running the back end of a website; which two products can you choose from Google Cloud?
   - **A.** Kubernetes Engine and Compute Engine
   - **B.** Cloud Run and Cloud Functions
   - **C.** Cloud Functions and Compute Engine
   - **D.** Cloud Functions and Kubernetes Engine

4. You have been asked to design a storage system for a web application that allows users to upload large data files to be analyzed by a data analytics workflow. The files should be stored in a high-availability storage system. Filesystem functionality is not required. Which storage system in Google Cloud should be used?
   - **A.** Block storage
   - **B.** Object storage
   - **C.** Cache
   - **D.** Network File System

5. All block storage systems use what block size?
   - **A.** 4 KB.
   - **B.** 8 KB.
   - **C.** 16 KB.
   - **D.** Block size can vary.

6. You have been asked to set up network security in a virtual private cloud. Your company wants to have multiple subnetworks and limit traffic between the subnetworks. Which network security control would you use to control the flow of traffic between subnets?
   - **A.** Identity access management
   - **B.** Router
   - **C.** Firewall
   - **D.** IP address table

7. When you create a machine learning service to learn how to classify objects using tabular data, what type of servers should you choose to manage compute resources?
   - **A.** Virtual machines (VMs).
   - **B.** Clusters of VMs.
   - **C.** No servers; you should use specialized services, which are serverless.
   - **D.** VMs running Linux only.

8. When does investing in servers for extended periods of time, such as committing to use servers for three to five years, work well?
   - **A.** When a company is just starting up
   - **B.** When a company can accurately predict server need for an extended period of time
   - **C.** When a company has a fixed IT budget
   - **D.** When a company has a variable IT budget

9. Your company is based in North America and will be running a virtual server for batch processing invoices. What factor determines the unit per minute cost?
   - **A.** The time of day the virtual machine (VM) is run
   - **B.** The characteristics of the server
   - **C.** The application you run
   - **D.** None of the above

10. You plan to use AutoML to analyze sales data and predict product demand in the near future. You plan to analyze between 1,000 and 2,500 products per hour. How many VMs should you allocate to meet peak demand?
    - **A.** 1.
    - **B.** 10.
    - **C.** 25.
    - **D.** None; AutoML is a serverless service.

11. You have to run a number of services to support an application. Which of the following is a good deployment model?

A. Run on a large, single VM.

B. Use containers in a managed cluster.

C. Use two large VMs, making one of them read only.

D. Use a small VM for all services and increase the size of the VM when CPU utilization exceeds 90 percent.

12. You have created a VM. Which of the following system administration operations are you allowed to perform on it?

A. Configure the filesystem.

B. Patch operating system software.

C. Change file and directory permissions.

D. All of the above.

13. Cloud Filestore is based on what filesystem technology?

A. Network File System (NFS)

B. XFS

C. EXT4

D. ReiserFS

14. When creating resources in Google Cloud, those resources are always part of what?

A. Virtual private cloud

B. Subdomain

C. Cluster

D. None of the above

15. You need to store data for an application and are using a cache. How will the cache affect data retrieval?

A. A cache improves the execution of client-side JavaScript.

B. A cache will continue to store data even if power is lost, improving availability.

C. Caches can get out of sync with the system of truth.

D. Using a cache will reduce latency, since retrieving from a cache is faster than retrieving from SSDs or HDDs.

16. Why can cloud providers offer elastic resource allocation?

A. Cloud providers can take resources from lower-priority customers and give them to higher-priority customers.

B. Extensive resources and the ability to quickly shift resources between customers enables public cloud providers to offer elastic resource allocation more efficiently than can be done in smaller data centers.

C. They charge more the more resources you use.

D. They don't.

17. What is not a characteristic of specialized services in Google Cloud?
    A. They are serverless; you do not need to configure servers or clusters.
    B. They provide a specific function, such as translating text or analyzing images.
    C. They require monitoring by the user.
    D. They provide an API to access the functionality of the service.

18. Your client's transactions must access a drive attached to a VM that allows for random access to parts of files. What kind of storage does the attached drive provide?
    A. Object storage
    B. Block storage
    C. NoSQL storage
    D. SQL storage

19. You are deploying a new relational database to support a web application. Which type of storage system would you use to store data files of the database?
    A. Object storage
    B. Data storage
    C. Block storage
    D. Cache

20. A user prefers services that require minimal setup; why would you recommend Cloud Storage, Cloud Run, and Cloud Functions?
    A. They are charged only by time.
    B. They are serverless.
    C. They require a user to configure VMs.
    D. They can only run applications written in Go.