Revision notes for GCP Exam

Plan

18 chapters → two a week and you'd finish the content in 2 months

Started December 3rd

Objectives

https://cloud.google.com/certification/guides/cloud-engineer/

Chapter 1

- Google Cloud Platform (GCP) is a public cloud service that offers some of the same technologies used by Google to deliver its own products.

Types of cloud services

- Different users have different requirements with regard to their cloud needs
- A startup may join GCP from the start and therefore rely solely on leveraging GCP's services to help build their product. They may use things like GCP's Cloud identity (for **authentication**) and Access Management services (for **authorization**)
- Another company may have already invested heavily in another platform and may want to integrate the two securely. This can be done using a VPN, this requires additional networking design and maintenance which a company that is solely on the cloud would not need to take into account.
- Cloud providers offer services that fall into 4 broad categories
    - Compute resources
    - Storage
    - Networking
    - Specialized services such as ML Services

Compute Resources

Compute resources come in many different forms, the ones we will be discussing are the following:

- Virtual machines
- Managed kubernetes clusters
- Serverless computing

Virtual machines

Virtual machines are a basic unit of computing resource that can be used to mess around with the cloud at the beginning. You can SSH into a virtual machine and use your CLI to work with it as you please. You can do things like install packages, patch the OS, configure the file system. A VM is like a server that you have in your office that you have full admin rights to.

(Add drawing here of user SSHing into VM and using it)

GCP also provides services like load balancer. This is like a single access point to a distributed backend. The load balancer can strategically be used to spread out the traffic coming in equally amongst the computing resources available to ensure the load on any particular resource is not too high.

(Add diagram showing what a load balancer looks like)

Autoscalers can add and remove VM's from the cluster based on workload. This is known as autoscaling. This helps control costs by not running more VMs than is needed and ensures that sufficient computing capacity is available when workload increases.

(Add diagram of how autoscaling could work)

Managed Kubernetes clusters

GCP provides you with all the tools needed to create and manage clusters of servers. This means that the users that just want to develop great applications and don't really want to bother with maintenance and upkeep of servers, Google makes this possible by providing managed cluster solutions.

Managed clusters make use of containers. These are essentially very similar to VMs except they are much more lightweight and processes are isolated between each one. In a managed cluster, you can specify the number of servers you want running as well as the different containers you want on each server. You can also specify things like autoscaling paramters such that the number of containers and servers you have running always stays optimal.

In a managed cluster, the health of containers is also managed for you. So if a container fails, the system will automatically detect this and start a new one.

Containers are a good option for when you have an application that relies on many microservices. Each microservice can be assigned its own container and the cluster management system will take care of everything from monitoring, networking and some security tasks aswell.

Serverless computing

Serverless computing is the approach where developers can run their applications without the need for setting up a VM or clusters.

GCP offers two options for serverless computing:

1) App Engine
2) Cloud functions

App Engine - run something for a long time

App engine is used for applications or containers that run for extended periods of time. This includes things like a website backend, point of sales system etc.

Cloud functions - run a function in response to an event

Cloud functions is a platform used to run code in response to an event such as uploading a file or adding a message to a message queue. This works well when you need to respond to an event by running a short process coded in a function or by calling a longer application that might be running on a VM, managed cluster or app engine.

Storage

Public clouds offer a few types of storage options. These include the following:

- Object storage
- File storage
- Block storage
- Caches

Users will usually make use of a few of these at any given time.

Object storage

Object storage is a system that manages the use of storage in terms of objects or blobs. Objects are not stored in a conventional file system rather they are grouped into buckets. Each object is then addressable usually by a URL.

Object storage is not limited by size of disk or SSDs attached to a server. Objects can be uploaded without concern for the amount of available space on a disk. Multiple copies of an object can be stored to improve things like availability and durability. Copies can even be stored in different regions such that the object is available even when one region is down.

Another advantage to object storage is the fact that it is serverless. You dont need to configure a VM and add storage to it like that. GCP's object storage called ***Cloud Storage*** is accessible from servers running in GCP as well as other devices with internet access.

Access control can be applied at an object level so that the owner can control who access the object and what they can do to it.

File Storage

File storage provides a hierachical storage system for files. GCP provides users with a service called ***Cloud filestore*** which is based on the Network file system (NFS)

File storage is useful for applications that require an OS like file storage system. The file system, 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 known as a "block" to organise data. Block storage is available on disks that are attached to VMs in Google Cloud Platform. Block storage can either be ephemeral or persistent (short lived or long lasting). 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. Persistent disks are used when you want data to exist independent of the existence of a VM.

It takes longer to retrieve something from object storage than from block storage. Object storage also does not support OS level access or file based access and has to be retrieved using a high level protocol like HTTP (can only access object storage by using http)

Usually you would use a combination of Object and block storage to suit your storage needs. Object storage can store large volumes of data that are copied over to persistent disks when needed - allowing for them to be accessed via the OS or file based access.

Caches

Cache are in memory data stores that maintain fast access to data. The time it takes to retrieve data is known as the latency. The latency of the cache is meant to be orders of magnitude faster than other storage options.

Cache is volatile → you lose data when power is lost or the OS rebooted.

Problems occur when the cache is not storing the most up to date data. This is known as cache invalidation. What happens is that the cache stores a value from the “system of truth” (this is the persistent storage like a database) and keeps it cached. So whenever that value is needed, the cache gives it over. The problem occurs when the value in the “system of truth” gets updated however the value in the cache doesnt. This problem is known as cache invalidation.

If cache is going to be used, then you need to design a cache update stategy to avoid such problems with cache invalidation.

Caches have many real world examples from browsers to optimising database queries.

Networking

When working in the cloud, you will need to work with networking between the resources in the cloud and those on premise resources. When you have multiple VMs in the cloud, you will need to discuss IP addresses at some point. Each device or service that is accessible from a network will have an IP address. Resources within GCP can have both internal and external IP address. Internal ones can only be accessed from other devices within the internal GCP network known as the virtual private cloud (VPC). External addresses are accessible from the internet.

External IP’s can be static or ephemeral. Static means that the IP is assigned for long periods of time. Ephemeral means that the IP is assigned to the VM and released when the VM shuts down.

In addition to specifying IP address, you will also most likely need to define firewalls to control access into subnets and VMs in your virtual private cloud.

Specialised services

Most public cloud providers offer building blocks to improve their user’s applications. Common characteristics of speicalised services include:

- They are serverless, you dont need to set up the servers or clusters yourself

- They provide a specific function such as translating text or analysing images.
- They provide an API to access the service
- You are charged based on your usage of the service

Examples on GCP

- AutoML ; machine learning service
- Cloud natural language ; service to analyse text
- Cloud vision ; analyse images

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 Platform.

Cloud vs Data centre computing

Rent instead of own resources

Corporate data centres are filled with servers, disk arrays and networking equipment. Companies often need to spend alot of money upfront to purchase this equipement. This approach works well when you have a constand predicatble workload. However if you have peaks and troughs in your workload, you may not be using your resources to their full potential or may not be meeting demands of the workload.

Public clouds offer the desirable option to rent resources as you need them. For example if demand is high, you could run VMs in the cloud to accomodate for the extra demand.

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

When you run a virtual server in the cloud, you will typically pay for a minimum period, such as 10 minutes, and then pay per minute used thereafter. The unit cost per minute 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.

Elastic resource allocation

Another key differentiator between on-premise 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-premise 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.

Specialised services

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.