☐ ~~Dan Book is "master~~	
☐ Dan Book Summary	
☐ Prepare all the questions	
☐ Prepare notes	Associate Cloud Engineer - Study Note… 229 kB
☐ Qwicklabs	
☐ Whizlabs	
☑ ~~Udemy Test for all authors~~	
☑ ~~Google's exam~~	
☐ Repeat bookmarked mock exams discussions	
☑ ~~Notes / Github/ Cheatsheet~~	
☐ Kubernetes revision, roles for app engine, deployment session missing., app engine scaling	

Projects Billing

- **Organization**
  - Gsuite Domain - > if your company has gsuite domain the organization can be created
  - Identity as a service - > Cloud Identity
  - Every identity at identity has Organization Admin role
    - Defining the structure of the resource hierarchy
    - Defining identity access management policies over the resource hierarchy
    - Delegating other management roles to other users
  - All users in that organization are granted Project Creator and Billing Account Creator roles
  - If a parent policy is **less** restrictive, it overrides a more restrictive policy applied on the resource. If a parent policy is **more** restrictive, it does not override a less restrictive policy applied on the resource

- **Projects**
  - resourcemanager.projects.create
  - constraints/compute.disableSerialPortAccess to TRUE => This is example of constrains on resources.

- **Roles** [principle of least privilege]
  - Types
    - Primitive
      1. Viewer
      2. Editor
      3. Owner
    - Pre-defined
    - Custom
  - It is important to know that permissions cannot be assigned to users. They can be assigned only to roles. Roles are then assigned to users.
  - **roles/browser** Read access to browse the hierarchy for a project, including the folder, organization, and Cloud IAM policy
  - *roles/accessapproval.approver*

- **Service Accounts**
  - It's identity and resource too. When we assign role then it's identity and when we assign user permission to service accounts, it's resource.
  - **Types**
    - User managed
    - Google managed.
  - 100 accounts per project
  - Both the Compute Engine and App Engine service accounts are granted editor roles on the projects in which they are created
  - iam.serviceAccountUser => either on project level or user level.
  - iamserviceAccountAdmin => Creation plus management plus deletion.
  - Encrypted keys as authentication mechanism

- **Billing**
  - self-serve and invoiced

Billing Account Creator	Who can create self service billing account
Administrator	who manages billing account but cannot create th
User	Which enables a user to link projects to billing acc
Viewer	Can view txns

- If you would like to respond to alerts programmatically, you can have notifications sent to a Pub/Sub topic by checking the appropriate box in the Manage Notification sections
- Please note notifications are not sent to projects owners
- Billing data can be exported to either a BigQuery database or a Cloud Storage file.

- **Workspace**
  - Tool for monitoring resources contained in one or more Google Cloud projects or AWS accounts. A Workspace accesses metric data from its monitored projects, but the metric data remains in those projects.
  - To stop all Cloud Monitoring charges for metrics usage, do one of the following:
    1. Disable the Monitoring APIs
    2. Stop Cloud Monitoring agents, Cloud Logging agents, and other software modules from sending metrics or logs to your Google Cloud project, or to the AWS connector projects.

- **Roles**
  - ALPHA -> BETA ->GA
  - TESTING, SUPPORTED, NOT_SUPPORTED
  - Google Cloud Directory Sync (GCDS)

Compute Resources

- Compute Engine -> GKE -> App Engine / Functions

- **Compute Engine**
  - Comes with pre-defined sizes but custom configuration is possible
  - remember custom GPU is possible but memory size per GPU is not possible.
  - TAGS- firewall importance.
  - --labels option with labels in the format of KEYS=VALUE
  - Scopes - Default is read only storage and write to stackdriver
  - Sole tenancy
  - VMs are changed at 1 second increment but minimum one minute.
  - Attached Disk resize is possible without downtime .. but no type conversion
  - Preemptive
    - 80% savings
    - will be shut down by google at any time (24 Hours)
    - Image formatting, continuous integration, financial modelling, rendering, big data.
    - If shutdown in first 10 minutes then no charge
    - no automatic restart
    - 30 seconds shutdown notification - you can use this for shutdown notification.

- SSH 22 / RDP 3389 firewall rule should be allowed. Remember default network has already enabled these f/w rules.
- VM lifecycle attached below. remember that reset meaning machine is still running.
- image.png 293 kB
- Even if VM is stopped image can be changed once it is configured.
- Maximum 128 persistent disk and 64 TB of PD can be attached to any single VM. Shared core machine types 16 PD and 3 TB.
- Discounts
  - Preemptive
  - Committed use - 1 to 3 years
  - Sustained use - region collectively
  - Remember , these can not be combined together.
- startup-script-url / shutdown-script-url [metadata]
- startup-script/ shutdown-script[metadata]
- #movement within region gcloud compute instances move #different regions - then follow snapshot approach
- **Snapshot service**
  - not available for SSD
  - does not backup metadata , tags etc\
  - To work with snapshots, a user must be assigned the Compute Storage Admin role.
- Images
  - Images are similar to snapshots in that they are copies of disk contents. The difference is that snapshots are used to make data available on a disk, **while images are used to create VM**s.
  - You can't directly create a VM from a snapshot without the disk. You can use the snapshot to create a disk for the new instance, but you can't create the instance directly from a snapshot without the disk.
- GPU - Math or machine learning
  - Image must contain GPU supported libraries.
  - GPUs cannot be attached to shared memory machines
  - When you add a GPU to an instance, you must ensure that: The instance is set to terminate during maintenance.
- **Instance Groups**
  - Command applied to group is applied to all instances
  - **Managed**
    - identical configuration
  - **Un managed**
    - different config within group
  - Zonal and regional possible

        - If it fails, The likely causes are
            1. A persistent disk already exists with the same name as VM Instance
            2. disks.autoDelete option is set to false
            3. instance template might be invalid
    - **Load Balancer**
        - When autoscaling, ensure you leave enough time for VMs to boot up or shut down before triggering another change in the cluster configuration. If the time between checks is too small, you may find that a recently added VM is not fully started before another is added. This can lead to more VMs being added than are actually needed.
        - CPU utilization, monitoring metric, load-balancing capacity, or queue-based workloads.

    - To perform this task, you must have the following permissions:
        - compute.instances.setMetadata on the instance if setting metadata on the instance
        - compute.projects.setCommonInstanceMetadata on the project if setting project-wide metadata
        - iam.serviceAccounts.actAs on the project if setting project-wide metad
    - Instances need to be stopped before you can make changes to their network interfaces.
- **Kubernetes Engine**
    - Cluster is built from VM and also health checked. These VMs run specialized operating systems optimized to run containers
    - Default vm is n1s1
    - YAML format
    - The nodes run an agent called kubelet, which is the service that communicates with the cluster master
    - . Kubernetes reserves CPU capacity according to the following schedule: 1. 6 percent of the first core 2. 1 percent of the next core (up to two cores) 3. 0.5 percent of the next two cores (up to four cores) 4. 0.25 percent of any cores above four cores
    - Containers also help you for easy migration from on prem to cloud as container designed applications are anyways loose coupled.
    - Containers start much faster than virtual machines and use fewer resources, because each container does not have its own instance of the operating system.
    - Google Cloud container builder & google cloud container registry
    - Deployment - > expose the deployment which results in service -> scale the deployment ->
    - This cluster does not have persistent volumes but uses standard storage. Persistent volumes are durable disks that are managed by Kubernetes and implemented using Compute Engine persistent disks
    - Understand **resize vs update**
    - Sandboxes for untrusted code execution. *gvisor*
    - **Objects**
        - Pods
            1. Multiple containers are used when two or more containers must share resources

2. Pods are considered ephemeral.
3. Pods run containers but is not set of containers.
4. use shared networking and storage across containers.
5. Each pod gets a unique IP address and a set of ports. Containers connect to a port.
6. Multiple containers in a pod connect to different ports and can talk to each other on localhost. This structure is designed to support running one instance of an application within the cluster as a pod.
7. A pod allows its containers to behave as if they are running on an isolated VM, sharing common storage, one IP address, and a set of ports
8. Pending, which indicates the pod is downloading images;Succeeded, indicates the pod terminated successfully; Failed, indicates at least one container failed; and Unknown, means the master cannot reach the node and status cannot be determined.
9. Ideally you should not change pods but change configuration.

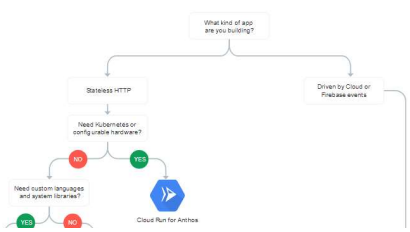
- Services
  1. A service, in Kubernetes terminology, is an object that provides API endpoints with a stable IP address that allow applications to discover pods running a particular application
- Volumes
- Deployment
  1. Deployments are sets of identical pods
- Namespaces
- ReplicaSet
  1. Where would you look for details on the number of pods that should be running
- StatefulSet
  1. StatefulSets are like deployments, but they assign unique identifiers to pods.
  2. This enables Kubernetes to track which pod is used by which client and keep them together.
  3. Used when an application needs a unique network identifier or stable persistent storage.
- Secrets are for passwords and config map for unencrypted and non sensitive data.

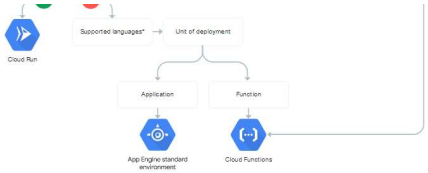
- **App Engine**
  - PaaS offering, well suited for mobile and web backed
  - All resources associated with an App Engine app are created in the region specified when the app is created
  - Standard & Flexible model
  - Each service can have multiple version running at the same time which helps to rollout
  - Permanent resident instances are possible along with this dynamic addition and removal of services.
  - Daily Spending limits are possible

- Types
  - Standard [python php node java go, ruby] but specific versions only
    1. Restrictions, you deploy code. network is not possible.
    2. No writing to local, 60s timeout, no SSH, no third party library
    3. Fast startup time
    4. Background processes - not possible
    5. VPN not possible as it's really serverless
  - Flexible
    1. Container and native support.
    2. remember you can change runtime through docker file.
    3. Differences with k8s is - in case of k8s you monitor the cluster and in flexible app google manages it.
- The App Engine standard environment scales down to no running instances if there is no load, but this is not the case with the flexible environment. There will always be at least one container running with your service, and you will be charged for that time even if there is no load on the system
- App Engine offers NoSQL databases, in-memory caching, load balancing, health checks, logging, and user authentication to applications running in it.
- If you configure auto-scaling or basic scaling, then instances will be dynamic. If you configure manual scaling, then your instances will be resident.
- Type, Bandwidth, execution time and invocations - these are considered as important points while cost estimations.
- Scaling
  - only parameters for basic scaling are **idle_timeout and max_instances.**
  - If you prefer to use manual scaling because you need to control scaling, then specify the manual_scaling parameter and the number of instances to run.
  - Autoscaling - rest other parameters like CPU, latency etc
- Splitting
  - IP - IP address hash based on users IP address - some stickiness.
  - Cookie - GOOGAPPUID a user will access the same version of the app even if the user's IP address changes.
  - Random is good for even traffic distribution.
  - --no-promote option just to make sure that you do not publish the features
- Migrate
  - 1. Set env: flex in app.yaml
  - 2. gcloud app deploy --no-promote --version=[NEW_VERSION]
  - 3. Validate [NEW_VERSION] in App Engine Flex
  - 4. gcloud app versions migrate [NEW_VERSION]
  - ***gcloud app versions migrate v3 --service="pt-createOrder"***

Limit	Free app	Paid app
Maximum services per app	**5**	**105**
Maximum versions per app	**15**	**210**

- **Cloud Functions**
  - Short Living execution, triggering or events , and not like app engine where you have multiple versions.
  - Parallel execution if there are multiple triggers. TRY not to have stateless.
  - Default timeout is one minute but you can customize it till 9 minutes
  - node python go java
  - 128 to 2 GB
    - **Triggers**
      - Cloud Storage
        - file created, deleted, archived or meta data changed
      - Cloud Pub/Sub
      - HTTP
      - Firebase
      - Stackdriver Logging
- **Cloud Run**
  - Cloud Run is a managed compute platform that enables you to run stateless containers that are invocable via web requests or Pub/Sub events. Cloud Run is serverless
  - Cloud Run for Anthos - when container as well as websockets
- **Cloud Endpoints and APIGEE Edge**
  - **Cloud Endpoints**
    - Backend must be in GCP
    - Authentication , logging, monitoring, proxy deployment based on NGINX
  - **APIGEE Edge**
    - Business problem
    - Not necessary to have back-end in GCP
    - rate limiting , quota, etc.





Storage

MemoryStore

- Redis protocol compatible
- 1gb to 300GB
- Basic (no HA) , standard (fail over) - but downgrade not possible
- The Redis instance will be accessible from the default network unless you specify a different network

Object Storage / Cloud Storage

1. It's server-less & Objects can be accessed using URL.
2. Works well in GCP as well as with other VMs using internet.
3. Object level access possible.
4. remember it's global
5. no concurrency so last written file is correct .. Also partial objects not possible .. It is not file system.
6. GCP Users and other users can access objects inside.
7. Life Cycle is available for objects
    1. Lifecycle management policies are applied to buckets and affect all objects in the bucket.
    2. Multiregional and regional storage objects can be changed **to nearline or coldline**.
        1. Nearline can be changed **only to coldline**.
        2. regional to mutli regional not possible
    3. Conditions are often based on **age**. Once an object reaches a certain age, it can be **deleted or moved to a lower-cost storage class**
    4. A. Nearline and coldline incur access charges.
8. Regional , Multi Regional, near-line, cold-line
9. Images, Videos etc. partial object retrieval is not possible as this not block storage.
10. Rewrite means to any new class - this is not possible in lifecycle
11. Versioninng
    1. last version is archieved
    2. whenever updated or deleted.

12. Location can not be changed once it is created
13. -m flag means parallel but split and parallel upload is also good.

Standard Storage (per GB per Month)	Nearline Storage (per GB per Month)	Coldline Storage (per GB per Month)	Archive Storage (per GB per Month)
$0.020	$0.010	$0.004	$0.0012
Frequent used	Stored for 30 d	90 d	365 d

File Storage

1. Cloud filestore
2. Hierarchical
3. NFS - Network file system

Block Storage / Persistent

1. Persistent SSD or HDD
2. Block Size can vary
3. Disks can also be resized as needed while in use without the need to restart your VMs.
4. Local SSD does not survive machine termination.
5. Each local SSD is 375 GB in size, but you can attach a maximum of 24 local SSD partitions for 9 TB per instance.
6. Persistent disks automatically encrypt data on the disk.
7. Persistent disks can be created blank or from an image or snapshot.
    1. Use the image option if you want to create a persistent boot disk.
    2. Use a snapshot if you want to create a replica of another disk.

Databases

Relational

- Cloud SQL
  1. Vertical scaling possible
  2. business intellligance
  3. ecommerce
  4. remember on demand vs scheduled backs .. command is totally different.. backups create vs patch
     1. Cloud SQL retains up to 7 backups for an instance.
     2. A Cloud SQL instance configured for HA is called a regional instance because it's primary and secondary instances are in the same region..
  5. Export
     1. First describe
     2. get service account and then assign write access on bucket
     3. then gcloud sql export **sql** [INSTANCE_NAME] gs://[BUCKET_NAME]/[FILE_NAME] --database=[DATABASE_NAME]
     4. gcloud sql export **csv** [INSTANCE_NAME] gs://[BUCKET_NAME]/[FILE_NAME] --database=[DATABASE_NAME]

- Cloud Spanner

  1. 99.999
  2. HA
  3. Globally Available
  4. Financial Data or supply chain
  5. Cloud spanner instance, Google recommends implementing this base on the Cloud Monitoring metrics on CPU or storage utilization in conjunction with Cloud Functions.
  6. https://cloud.google.com/spanner/docs/schema-design#primary-key-prevent-hotspots
  7. To import data, click the Import tab to display the Import form. You will need to specify a source bucket, a destination database, and a region to run a job.

- BigQuery

  1. Designed for a data warehouse and analytic applications. and to store petabytes of data.
  2. works with large numbers of rows and columns of data
  3. **not** suitable for transaction-oriented applications
  4. Though this is managed you need to care about cost and status of job.
  5. It can not be scaled out for performance as it is possible in the bigtable by adding the nodes
  6. Flat-rate customers purchase dedicated resources for query processing and are not charged for individual queries. BigQuery offers flat-rate pricing for customers who prefer a stable cost for queries rather than paying the on-demand price per TB of data processed
  7. Export to GCS or Data Studio

1. Avro is a compact binary format that supports complex data structures. When data is saved in the Avro format, a schema is written to the file along with data. **Schemas are defined in JSON**. Avro is a good option for large data sets, especially when importing data into other applications that read the Avro format, including Apache Spark, which is available as a managed service in Cloud Dataproc. Avro files can be compressed using either the deflate or snappy utilities. Deflate produces smaller compressed files, but snappy is faster.

8. bq extract --destination_format CSV --compression GZIP 'mydataset.mytable' gs://example-bucket/myfile.zip
9. bq load --autodetect --source_format=CSV mydataset.mytable gs://ace-exam-biquery/mydata.csv

Non - Relational

◦ Cloud Bigtable

1. NoSQL
2. Petabyte scale
3. Hbase API compatable
4. wide column concept
5. you need to manage this remember.
6. also , import export is done using dataflow and java jar .. beam jar

◦ Cloud Datastore

1. NoSQL document database.
2. Although it is a NoSQL database, Cloud Datastore supports transactions, indexes, and SQL-like queries but **not** joins**.**
3. Entity is the unit stored, not necessary that each entity is common as this is schema less.
4. GQL
5. Backup
    1. it requires storage
    2. backups - datastore.databases.export permission.
    3. importing data, - datastore.databases.import. The Cloud Datastore Import Export Admin has both the rights

Cloud Firestore

1. NoSQL database service designed as a backend
2. for highly scalable web and mobile applications. A distinguishing feature of Cloud Firestore is its client libraries that provide offline support, synchronization, and other features
3. for managing data across mobile devices, IoT devices, and backend data stores. example, applications on mobile devices can be updated in real time as data in the backend changes.
4. Cloud Firebase includes a Datastore mode, which enables applications written for Datastore to work with Cloud Firebase as well.
5. When running in Native mode, Cloud Firestore provides real-time data synchronization and offline support.

Cloud Memorystore

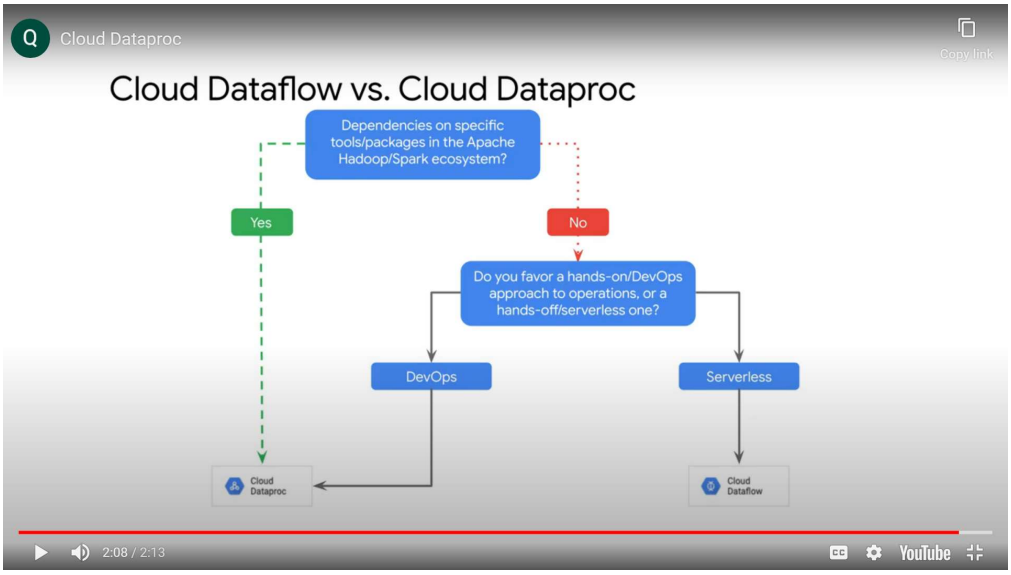
1. Managed Redis cache service

PubSub

- pub sub has push or pull model for subscription.
- The time to wait can range from 10 to 600 seconds.

Dataproc

- Standard has one master and good for development
- HA uses 3 masters
- The time to wait can range from 10 to 600 seconds.
- gcloud beta dataproc clusters export ace-exam-dataproc-cluster --destination=gs://ace-exam-bucket1/mydataproc.yaml
- gcloud beta dataproc clusters import gs://ace-exam-bucket1/mydataproc.yaml

- 

Networking

VPC

1. is Global even without internet connectivity
2. Subnets are regional
3. Even if machines are in same n/w and in different region they can communicate with each other using private IPs
4. Even if machines are in same region and in different n/w they can not communicate with each other using private IPs
5. **Subnets**
    1. Subnets can be in different zones..... as they are regional
    2. Every subnet has 4 reserved addresses
    3. Range can be expanded but no shrink
6. Internal IP DNS names are scoped to networks only...

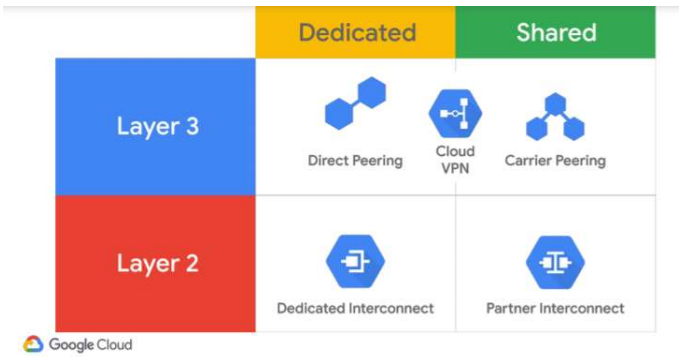
1. Network Types
    1. default
        1. subnet per region
        2. These firewall rules allow **ICMP**, **RDP**, and **SSH** ingress traffic from anywhere (0.0.0.0/0) and all **TCP**, **UDP**, and **ICMP** traffic within the network (10.128.0.0/9)
        3. The default network is an auto mode VPC network with pre-populated firewall rules.
    2. Auto
        1. Auto to custom possible but other way round is not possible
        2. subnet per region
        3. starts with /20 and can be expanded upto /16
        4. New zone available / then subnet is added automatically but this is not OK sometimes as overlapping might happen.
    3. Custom
        1.

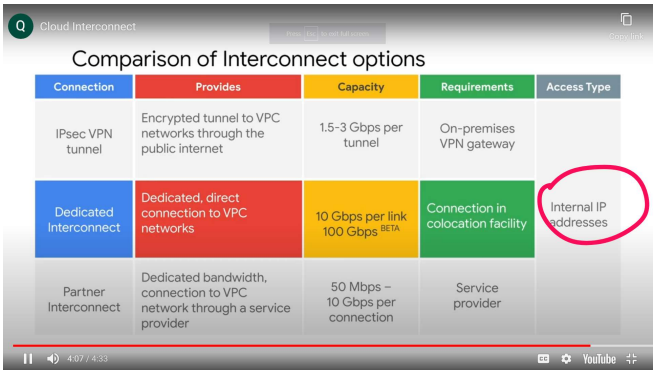
Armor

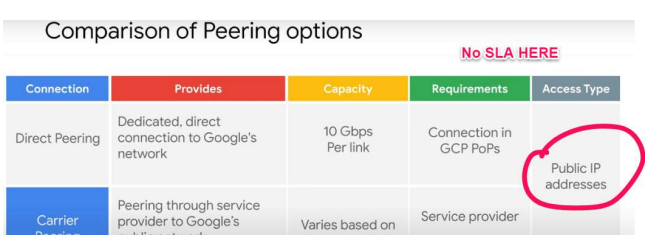
1. Services exposed to the Internet can become targets of distributed denial-of-service (DDoS) attacks. Cloud Armor is a Google network security service that builds on the Global
2. HTTP(s) Load Balancing service. Cloud Armor features include the following:
    - Ability to allow or restrict access based on IP address
    - Predefined rules to counter cross-site scripting attacks
    - Ability to counter SQL injection attacks
    - Ability to define rules at both level 3 (network) and level 7 (application)
    - Allows and restricts access based on the geolocation of incoming traffic

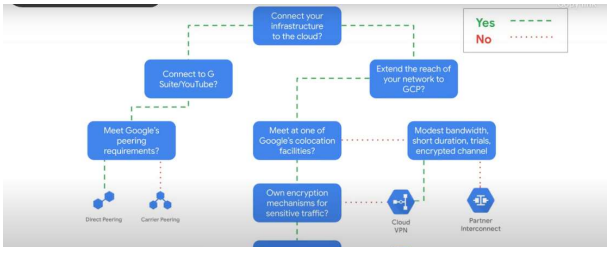
VPN

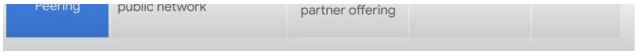
- securely connects on-premises network to your GCP VPC network through an IPsec VPN tunnel.
- Traffic traveling encrypted by one VPN gateway, then decrypted by the other VPN gateway.
- This protects your data as it travels over the public internet, and that's why Cloud VPN is useful for low-volume data connections.
- Post VPN you can communicate over**private IP address**.
- SLA of 99.9%
- Supports
    1. Site-to-site VPN
    2. ○ Static routes
    3. ○ Dynamic routes (Cloud Router)
    4. ○ IKEv1 and IKEv2 ciphers
- MTU = 1460 Bytes
- VPN gateways on both sides and then tunnels [1-2 and 2-1]
    - You are using one tunnel to pass traffic in each direction. And if both tunnels are not established, you won't be able to ping the remote server on its internal IP address. The ping might reach the remote server, but the response can't be returned.
- VPN gateway is regional resources and uses regional IP address.
- Cloud Router can manage routes for a Cloud VPN tunnel using Border Gateway Protocol, or BGP. This routing method allows for routes to be updated and exchanged without changing the tunnel configuration.
- **Interconnect** - Direct connection to GCP meaning private IP access though BGB tunneling ... Direct Interconnect (Google colocation) or partner interconnect (Partner connection)

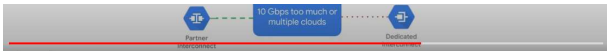












VPC Peering and VPC shared

- VPC Shared
  - Organization must be same but **across** projects (Not same project and multiple VPC)
  - Different projects
  - One host project and other worker projects
  - Centralized admin
  - Can be created at organization or folder level.
  -

```
- gcloud organizations add-iam-policy-binding [ORG_ID]--member='user:[EMAIL_ADDRESS]' --
role="roles/compute.xpnAdmin"
- gcloud resource-manager folders add-iam-policy-binding [FOLDER_ID] --member='user:[EMAIL_ADDRESS]' --
role="roles/compute.xpnAdmin"
- gcloud resource-manager folders list --organization=[ORG_ID]
- gcloud compute shared-vpc enable [HOST_PROJECT_ID]
- gcloud compute shared-vpc associated-projects add [SERVICE_PROJECT_ID] --host-project [HOST_PROJECT_ID]

                                  gcloud compute networks peerings create peer-ace-exam-1 \
                                --network ace-exam-network-A \
                                --peer-project ace-exam-project-B \
                                --peer-network ace-exam-network-B \
                                --auto-create-routes
                                And then create a peering on the other network using:
                                gcloud compute networks peerings create peer-ace-exam-1 \
                                --network ace-exam-network-B \
                                --peer-project ace-exam-project-A \
                                --peer-network ace-exam-network-A \
                                --auto-create-routes
```

- VPC Peering
  - Projects can be within same org or different org
  - Everyone managers their own VPC
  - if you want to configure private communication between VPC networks in the **same project**

Load Balancer

- **global load balancers**
  - are the HTTP(S), SSL proxy, and TCP proxy load balancers.
  - These load balancers leverage the Google frontends, which are software-defined, distributed

- These load balancers leverage the Google frontends, which are software-defined, distributed systems that sit in Google's points of presence and are distributed globally.
- Therefore, you want to use a global load balancer when your users and instances are globally distributed, your users need access to the same applications and content, and you want to provide access using a single anycast IP address.

- **regional load balancers**
  - are the internal and network load balancers,
  - they distribute traffic to instances that are in a single GCP region.
  - The internal load balancer uses Andromeda, which is GCP's software-defined network virtualization stack, and the network load balancer uses Maglev, which is a large, distributed software system.
- **HTTP(S) Load Balancer**
  - Understand URL Maps like deliver /*.video file from this instance and rest from another ..
  - Default timeout of 30 seconds ..
  - HTTPS - uses target https proxy.. requires at-least one SSL certificate on proxy.
  - SSL is terminated at load balancer.
- SSL
  - Non HTTP SSL traffic
  - Termination at Load balancer
- TCP
  - unEncrypted non http
- Network
  - No proxy but traffic is passed through
  - regional so resource must be in same region
  - forwarding rule to **instance groups or Target pools**
- **Internal**
  - Regional & private load balacer for TCP , UDP
  -