Google Cloud Fundamentals: Core Infrastructure

Cloud Computing Overview

- Cloud computing provides on-demand, self-service computing resources accessible over the internet. Key traits include resource pooling, elasticity, and pay-per-use models.

Cloud Service Models

- IaaS (Infrastructure as a Service): Delivers on-demand infrastructure resources such as storage, raw compute, and network capabilities. Customers pay for what they allocate.
- PaaS (Platform as a Service): Manages all hardware and software through the cloud. Customers pay for what they use.
- SaaS (Software as a Service): Provides a complete application stack accessible to customers.

Google Cloud Network

- Google has a large global network with infrastructure in North America, South America, Europe, Asia, and Australia.
- Geographic locations contain regions, which in turn contain multiple zones where cloud resources are deployed.
- Resources can be run in different regions for redundancy and some services can run in multiple geographic locations.

- The network is designed for high throughput.

Environmental Impact

- Data centers consume a significant amount of the world's electricity.
- Google aims to improve efficiency and reduce waste, with data centers being ISO 14001 certified.
- Google has committed to carbon neutrality and using 100% renewable energy.

Security

- Hardware level: Custom-designed server boards, chips, and networking equipment with secure boot processes. Physical security is multi-layered.
- Service deployment level: Inter-service communication is encrypted.
- User Identity Level: Intelligent challenges and secondary factors like U2F are employed for user verification.
- Storage services level: Encryption is applied at the storage service layer using centrally managed keys. Hardware encryption is also enabled.
- Internet communication level: Google Front End (GFE) ensures secure TLS connections and has DoS protections.
- Operational security level: Intrusion detection systems and employee access monitoring are implemented. Software development practices also include central source control and two-party review.

Open APIs and Open Source

- Google Cloud is open-source friendly, allowing customers to run applications elsewhere.
- TensorFlow is a key open-source library for machine learning that works with Google Cloud.
- Interoperability is supported at multiple layers of the stack with options to mix and match microservices.
- Google Cloud Observability allows for monitoring workloads across multiple cloud providers.

Pricing and Billing

- Google Compute products are billed per second.
- Compute Engine offers billing flexibility with sustained-use discounts and custom virtual machine types.
- Google Cloud Pricing Calculator can be used for cost estimation.
- Billing tools help to budget and monitor usage via budgets, alerts, reports, and quotas.
- Quotas are allocated at the project level and govern the number of resources.

Google Cloud Resource Hierarchy

- Resources are organized hierarchically. The hierarchy includes an Organisation Node at the top, followed by Folders, and then Projects.
- The Organisation Node is the topmost resource, and everything attached to the account falls under it.
- Folders group projects and can contain subfolders. They allow for grouping resources on a per-department basis and facilitate policy inheritance.
- Projects are the basis for using Cloud services. They hold resources, can have different owners and users, and are billed separately.
- Project Attributes:
    - Project ID, Project Name, and Project Number are globally unique.
    - The Project ID is assigned by Google Cloud but is mutable during creation, while the Project Number is immutable after creation.
    - The Project Name is chosen by the user and need not be unique.
- Resource Manager: This tool manages projects. It can create new projects, gather a list of projects, update existing projects, delete projects, and recover previously deleted projects. Access is through RPC API and REST API.
- Policy Inheritance: Resources inherit policies and permissions assigned to folders, and projects inherit policies assigned to a folder. Identity and Access Management (IAM) manages and applies policies.

Identity and Access Management (IAM)

- IAM applies policies defining who can do what on which resources.
- Deny policies prevent specific IAM permissions and override any existing allow policy.
- There are three kinds of IAM roles:
  - Basic IAM roles are broad in scope.
  - Predefined IAM roles match specific job needs. For example, the Compute Engine instanceAdmin role includes specific actions like deleting, getting, listing, and setting machine types, as well as starting and stopping instances.
  - Custom IAM roles are more specific and flexible and can be applied at the project or organization level.
- Permissions need to be managed.

Service Accounts

- Service accounts are assigned roles and identified by email addresses.
- They are also managed by IAM.
- For example, a service account can be created to authenticate a VM to Cloud Storage.

Cloud Identity

- Cloud Identity manages team and organization access.
- It allows organizations to define policies and manage users and groups using the Google Admin console.
- It can integrate with existing Active Directory or LDAP systems.
- It is available in free and premium editions and is already available to Google Workspace customers.
- Interacting with Google Cloud: There are four main ways to interact with Google Cloud:

- Google Cloud Console: Provides a web-based graphical user interface to find resources, check their health, manage them, and set budgets. It also offers a search facility and allows connections to instances via SSH in the browser.

Cloud SDK and Cloud Shell

- The Cloud SDK is a collection of command-line tools to manage resources and applications. It includes the Google Cloud CLI and tools such as bq for BigQuery.
- Cloud Shell provides command-line access to cloud resources directly from a browser. It is a Debian-based virtual machine with the Cloud SDK tools pre-installed and authenticated.
- APIs: Google Cloud services offer APIs that allow code to control them. The Google APIs Explorer shows available APIs and versions. Google provides Cloud Client and Google API Client libraries in multiple languages.
- Google Cloud App: The app allows users to manage resources, including starting and stopping Compute Engine and Cloud SQL instances, administering App Engine applications, and viewing up-to-date billing information and alerts. It also provides customizable graphs and incident management.

Virtual Private Cloud (VPC)

- A VPC is a secure, individual, private cloud-computing model hosted within a public cloud.
- VPC networks connect Google Cloud resources to each other and to the internet.
- VPC networks can be segmented, with firewall rules to restrict access and static routes to forward traffic to specific destinations.
- Google VPC networks are global and can have subnets in any Google Cloud region worldwide.
- VPC subnets connect resources in different zones
- VPCs do not require a router to be provisioned as routing tables are built-in.
- VPCs also do not require a firewall to be provisioned.
- Firewalls restrict access to instances through both incoming and outgoing traffic and rules can be defined through network tags.

- VPC peering and sharing allow projects to communicate.

Compute Engine

- Compute Engine allows users to create and run virtual machines (VMs) on Google infrastructure.
- Each VM contains the power and functionality of a full-fledged operating system.
- VMs can run Linux and Windows Server images, or customized versions of these images, and can also be created using the Google Cloud console, the Google Cloud CLI, or the Compute Engine API.
- Compute Engine pricing is pay-for-what-you-need and offers discounts such as sustained-use, committed-use, and preemptible/spot VMs.
- Compute Engine also offers custom machine types and various storage options.

Scaling Virtual Machines

- VMs can be auto-scaled to meet demand.
- Autoscaling is useful for resilient, scalable applications.
- Big VMs can be used for memory- and compute-intensive applications.

Cloud Load Balancing

- Load balancing distributes traffic across instances.
- Cloud Load Balancing is a fully distributed, software-defined, managed service.
- It can be used for HTTP(S), TCP, SSL, and UDP traffic.
- It provides single and cross-region load balancing, including automatic multi-region failover, and does not require pre-warming for anticipated spikes in traffic.
- There are several load balancing options including Application Load Balancers and Network Load Balancers, which come in various types, such as global external, regional external, regional internal, and cross-region internal, as well as proxy and passthrough configurations.

Domain Name Service (DNS)

- Google provides public DNS services.
- Cloud DNS is a managed DNS service that runs on the same infrastructure as Google.
- It provides low latency, high availability, and a cost-effective way to make applications and services available to users.
- The DNS information is served from redundant locations around the world.
- Cloud DNS is programmable and allows users to publish and manage millions of DNS zones and records using the Google Cloud console, the command-line interface, or the API.

Content Delivery Network (CDN)

- Cloud CDN reduces network latency.
- It reduces the load on content origins and saves money.
- Cloud CDN is enabled with a single checkbox.

Connecting Networks to Google VPC

- Cloud VPN creates dynamic connections using a VPN tunnel and Cloud Router to exchange route information over the VPN using the Border Gateway Protocol, but this may not always be the best option due to security or bandwidth reliability concerns.
- Direct Peering routes traffic through a Google Point of Presence (PoP) using a router to exchange traffic between networks, connecting to more than 100 Google PoPs around the world.
- Carrier Peering gives direct access from an on-premises network through a service provider's network but is not covered by a Google Service Level Agreement.
- Dedicated Interconnect provides one or more direct, private connections to Google with the highest uptimes, potentially up to a 99.99% SLA with appropriate configurations. Connections can be backed up by a VPN for greater reliability.

- Partner Interconnect provides connectivity between an on-premises network and a VPC network through a supported service provider and can be configured to support mission-critical services, potentially with up to a 99.99% SLA.
- Cross-Cloud Interconnect establishes dedicated connectivity between Google Cloud and another cloud service provider, supporting the adoption of an integrated multi-cloud strategy, with connection sizes of 10 Gbps or 100 Gbps.

Data Storage Options

- Google Cloud offers many options for storing data, including structured, unstructured, transactional, and relational data.
- Google Cloud's core storage options include:
  - Cloud Storage: For scalable object storage.
  - Cloud SQL: For relational databases.
  - Spanner: A fully managed relational database.
  - Firestore: A NoSQL cloud database.
  - Bigtable: Google's NoSQL big data database service.

Types of File Storage

- Google Cloud provides different types of file storage:
  - Object storage.
  - File storage.
  - Block storage.

Cloud Storage

- Cloud Storage is a fully managed, scalable service suitable for website content, archival and disaster recovery, direct downloads, and storing binary large objects (BLOBs).
- Files are organized into buckets with unique names and geographic locations.
- Objects within Cloud Storage are immutable, and modifications create new versions.
- Object versioning keeps a record of modifications and allows reversion to older states.

- Access to data objects can be controlled using IAM roles and Access Control Lists (ACLs). For most purposes, IAM is sufficient.
- Cloud Storage can be used like a file system through third-party tools that "mount" the bucket.
- Lifecycle policies help save money by deleting older objects or keeping only the most recent versions.
- There are four basic Cloud Storage classes: Standard, Nearline, Coldline, and Archive.
- Each storage class offers unlimited storage, worldwide accessibility, low latency, high durability, geo-redundancy, and a uniform experience.
- Autoclass automatically transitions objects to appropriate storage classes to reduce costs.
- Additional features include encryption and a pay-for-what-you-use model with no prior capacity provisioning.
- Data can be brought into Cloud Storage using online transfers or a Transfer Appliance.

Cloud SQL

- Cloud SQL is a relational database service that lets you focus on building applications, supporting MySQL, PostgreSQL, and SQL Server.
- It doesn't require software installation or maintenance, supports automatic replication, and includes managed backups.
- Cloud SQL can scale up to 128 processor cores, 864 GB of RAM, and 64 TB of storage, and encrypts customer data.
- It also includes a network firewall.
- Cloud SQL instances are accessible by other Google Cloud and external services using standard drivers.

Spanner

- Spanner is a fully managed relational database that scales horizontally and provides strong consistency with SQL support.

- It is designed for applications requiring high availability, strong global consistency, and high input/output operations per second.

Firestore

- Firestore is a NoSQL cloud database ideal for mobile and web development.
- It scales horizontally and stores data as "documents" within "collections".
- Firestore uses online and offline data synchronization.
- It provides atomic batch operations, real transaction support, strong consistency guarantees, and automatic multi-region data replication.

Bigtable

- Bigtable is Google's NoSQL big data database service for handling massive workloads with consistently low latency and high throughput.
- It is suitable for semi-structured or structured data, time-series data, and big data with heavy read-and-write events.
- Bigtable is used for both operational and analytical applications.
- Bigtable uses APIs for reading and writing data and integrates with stream processing and batch processing frameworks.
- BigQuery and Bigtable are different products with different uses.

Comparing Storage Options

- Cloud Storage is best for storing immutable blobs larger than 10 MB, with a maximum unit size of 5 TB per object.
- Cloud SQL is best for full SQL support for an online transaction processing system, as well as for web frameworks and existing applications, and can scale up to 64 TB.
- Spanner is best for full SQL support for an online transaction processing system and horizontal scalability, reaching petabytes of storage.
- Firestore is best for massive scaling and predictability together with real-time query results and offline query support, with a maximum unit size of 1 MB per entity.

- Bigtable is best for storing large amounts of structured objects, and analytical data with heavy read and write events, but does not support SQL queries and multi-row transactions, with a maximum unit size of 10 MB per cell, 100 MB per row.

Introduction to Containers

- Containers group code and its dependencies into an invisible box with limited access to the file system and hardware.
- Containers only require a few system calls to create and start quickly.
- They need an OS kernel that supports containers and a container runtime on each host.
- Containers scale like PaaS but offer nearly the same flexibility as IaaS.
- Scaling can be achieved by duplicating single containers, which can be done in seconds.
- Applications can scale with multiple containers, which can be modular, easily deployable, and scaled independently across a group of hosts.

Kubernetes

- Kubernetes is an open-source platform for managing containerized workloads and services.
- It orchestrates many containers on many hosts, scales them as microservices, and deploys rollouts and rollbacks.
- Kubernetes provides a set of APIs to deploy containers on a cluster of nodes.
- It includes a control plane and nodes that run containers.
- Users can describe applications and how they should interact, and Kubernetes manages the process.

Kubernetes Components

- Containers run in groups called "pods".
- Pods have a unique IP address and can connect to services.
- Pods can have volumes attached.
- Deployments are replicas of a specific pod.

- Services have fixed IPs to connect to pods.
- Load balancers route traffic to pods.
- Deployments can be scaled on command.
- Configuration files describe how to create and scale deployments.
- Deployments can be modified by changing configuration files.
- Endpoints can be discovered using a kubectl command.
- Rolling updates can be triggered on the command line or by updating the configuration file.

Google Kubernetes Engine (GKE)

- GKE is Google's managed Kubernetes service.
- GKE has two modes of operation: Autopilot and Standard.
- In Autopilot mode, GKE manages the underlying infrastructure.
- In Standard mode, the user manages the infrastructure.
- Autopilot is optimized for production with strong security and operational efficiency.
- GKE can create customized Kubernetes clusters.
- GKE clusters benefit from Google Cloud's load balancing for Compute Engine instances.
- Node pools allow the designation of subsets of nodes within a cluster.
- GKE provides automatic scaling of node instance counts and automatic upgrades for node software.
- Node auto-repair maintains node health and availability.
- GKE integrates with Google Cloud Observability for logging and monitoring.

Cloud Run

- Cloud Run is a managed serverless compute platform that runs stateless containers.
- It removes the need for infrastructure management by being serverless.
- Cloud Run is built on Knative, an open API and runtime environment based on Kubernetes.
- It can automatically scale up and down from zero almost instantaneously.
- Users are only charged for the resources used.

- The Google Cloud Run workflow involves writing code, building, and packaging it into a container image, and deploying it.
- Cloud Run also supports a source-based workflow, using Buildpacks to deploy code directly to Cloud Run.
- Users are only charged when their container handles requests.
- Cloud Run can run any binary compiled for Linux 64-bit.

Cloud Run Functions

- Cloud Run Functions are integrated cloud functions that handle application events.
- They provide a lightweight, event-based, asynchronous compute solution.
- They allow for the creation of small, single-purpose functions that respond to cloud events without the need to manage a server or runtime environment.
- These functions can be used to construct applications from small pieces of business logic and extend cloud services.
- Billing is to the nearest 100 milliseconds, and only while the code is running.
- Cloud Run Functions support a variety of programming languages, including Node.js, Python, Go, Java, .Net Core, Ruby, and PHP.
- Events from Cloud Storage and Pub/Sub can trigger Cloud Run functions asynchronously, and HTTP invocation is available for synchronous execution.

Key Concepts

- Cloud Run is best suited for hosting dynamic web applications, whereas Cloud Run Functions are better for event-driven code.
- Cloud Run Functions are a scalable functions-as-a-service platform that can be used to extend Cloud services and are integrated with Cloud Logging. They do not require servers or VMs to be provisioned.
- A Google Cloud customer might choose to use Cloud Run functions when their application contains event-driven code and they do not want to provision compute resources.
- Regions contain multiple zones.
- Resources within zones can provide improved fault tolerance.

- Using resources in multiple regions improves fault tolerance and allows for localized application versions.
- A Kubernetes pod is a group of containers.
- The resources used to build Google Kubernetes Engine clusters come from Compute Engine.
- Services and APIs are enabled on a per-project basis.
- IAM roles, ordered from broadest to finest-grained, are: Basic roles, Predefined roles, and Custom roles.
- The project number is globally unique, permanent, and unchangeable but can be modified by the customer during creation.
- Cloud Storage is well-suited to providing durable and highly available object storage.
- Spanner is a relational database service that can scale to higher database sizes.
- Customers consider the Coldline storage class to save money on storing infrequently accessed data.
- VPC subnets have a regional scope.
- Customers choose Preemptible or Spot VMs primarily to reduce costs.
- A Service Level Agreement (SLA) is available for Dedicated Interconnect.
- Cloud Load Balancing allows you to balance HTTP-based traffic across multiple Compute Engine regions or across multiple virtual machine instances within a single Compute Engine region.

Further Reading

- Compute Engine: cloud.google.com/compute/docs
- Virtual Private Cloud: cloud.google.com/compute/docs/vpc
- Compute Engine Machine Types: cloud.google.com/compute/docs/machine-types
- Cloud Storage: cloud.google.com/storage
- Cloud SQL: cloud.google.com/sql/docs
- Bigtable: cloud.google.com/bigtable/docs
- Spanner: cloud.google.com/spanner/docs
- Firestore: firebase.google.com/docs/firestore
- Google Cloud Console: cloud.google.com/console
- Cloud SDK: cloud.google.com/sdk/gcloud

- Google APIs Explorer: console.cloud.google.com/apis/explorer
- Google Cloud App: cloud.google.com/app
- Google Cloud Locations: cloud.google.com/about/locations
- Google Cloud Security: cloud.google.com/security/security-design
- Google Cloud Pricing Calculator: cloud.google.com/products/calculator