Essential GCP Infra: Foundation 2023 Ivan Vlad S.

- 1 GCP Overview
  - Cloud Computing Characteristics
    - Broad network access
      - > access resources over internet from anywhere
    - Rapid elasticity
      - > resources are elastic/flexible, can scale up or down based on load
    - Resource pooling
      - > CSP allocates resources from shared pool, economies of scale = cheaper service
    - On-demand self service
      - > provision and terminate resources using UI/CLI, without human interaction
    - Measured service
      - > pay for what you use, reserve as you go
  - Service Models
    - SaaS
      - > fully managed
      - > responsible for data
    - PaaS
      - > bind code to libraries that provide access to the infra the app needs
      - > responsible for runtime, data, and app
      - > pay for what you use
    - IaaS
      - > provides compute, storage, networking
      - > responsible for OS, container, runtime, data, and app
      - > pay for what you allocate
    - Serverless = managed infra

- > Cloud Functions: manages event-driven code as pay-as-you-go service
- > Cloud Run: deploy containerized microservices based app in a fully-managed environment

- GCP Network
  - 5 Geolocations
    > NA, SA, EU, AS, AU
  - Divided in to Regions >> Zones
    - Currently 37 regions and 112 zones
    - Can do multi-region hosting
- Environmental Impact
  - Google data centers were first to achieve ISO 14001 certification
  - carbon neutral, renewable energy
    by 2030 carbon free
- Security
  - Hardware infra layer
    - > custom hardware design and provenance
    - > secure boot stack
    - > physical security
  - Service deployment layer
    - > encryption of inter-service comms
  - User identity layer
    - > Google central Identify Service
  - Storage services layer
    - > encryption at rest
  - Internet comm layer
    - > Google Front End (GFE)
    - > DoS protection
  - Ops security layer
    - > intrusion detection
    - > reducing insider risk
    - > employee universal second factor (U2F) use

- > software dev practices

- Open source ecosystems
  - Google publishes key elements of tech using open source licenses to create ecosystems that provide customers with options other than GCP
  - great interoperability with different CSPs
- Pricing and Billing
  - per-second billing
  - Compute Engine offeres Sustained-use discounts
  - Online Pricing Calculator
  - GCP Tools
    - Budgets, Alerts, Reports, Quotas

2 Resources & Access

- Resource hierarchy
  - Order from bottom up:
    Resources > Project > Folder > Organization
  - directly relates to how policies are managed and applied in GCP
    - > policies can be defined at the project, folder, and org levels
      > some policies can be assigned at the resource level
    - > policies are applied downward
  - Projects are the basis for enabling GCP services
    - each resource belongs to exactly one project
    - can have different owners and users because they're billed and managed separately
    - 3 identifying attributes: ID, name, number
      > ID = globally unique identifier that CANT be changed after creation, immutable
    - Resource Manager Tool - manage projects
  - Folders let you assign policies to resources at a level of granularity you choose
    - Use folders to group projects under an organization in a hierarchy
  - At org node: can assign 2 special roles:
    > org policy admin and project creator

- > goog workspace customer = projects will automatically belong to your org node
  > non-goog workspace customer = use Cloud Identity to create an org node

- IAM
  - Admins can apply policies that define WHO can do WHAT on WHICH resources
  - IAM role = collection of permissions
    - 3 types: basic, predefined, custom
    - Basic = broad in scope
      > owner, editor, viewer, billing admind
    - Predefine = google defined roles
      > some for each service, e.g., Instance Admin Role
    - Custom = user defined
      > need to manage the permissions that define the custom role you create
      > can only be applied to either project or org level, not folder level
- Service Accounts
  - Accounts for machines/services, NOT humands
    - need to be managed, can have IAM roles attached to it
  - > named with an email address
    > use cryptographic keys instead of password
- Cloud Identity
  - can define policies and manage users and groups using the Google Admin console
    - log in and manage resources using the same creds used in existing AD or LDAP systems
    - Goog admin console can be used to disable user accounts and remove them from groups when they leave
    - available in free and premium editions
    - already available to goog workspace customers in goog admin console
- Interacting with GCP
  - Cloud console
    - web-based GUI, can connect to resources via SSH in the browser
  - SDK: set of tools to manage resources and apps hosted in GCP

- gcloud - main CLI for products and services
- gsutil - storage access from CLI
- bq - BigQuery CLI
- Cloud shell = access SDK CLI from browser

- API
  - Google APIs Explorer - shows what APIs are available
  - Google provides Cloud Client and Google API Client libraries
- Cloud Mobile App
  - Start, stop, and use SSH to connect into Compute Engine instances, and see logs
  - Stop and start Cloud SQL instances
  - Administer apps deployed on App Engine
  - up-to-date billing info and budget alerts

3️⃣ VPC

- A secure, individual, private cloud-computing model hosted within a public cloud. A VPC is hosted remotely by public cloud provider
  - Run code, store data, host websites, and anything else that can be done in an ordinary private cloud
- VPC networks connect GCP resources to each other and to the internet
  - segmenting networks
  - using firewall rules to restrict access to instances
  - creating static routes to forward traffic to specific destinations
- GCP VPC networks are global and can have subnets in any GCP region worldwide
- VPC objects
  - Projects
    - associates objects and services with billing
    - contains networks (15 max) that can be shared/peered
  - Networks: Default, auto mode, custom mode
    - has no IP address range
    - global and spans all available regions

- Default = every project, one subnet per region, default firewalls
- auto mode = default network, one subnet per region, regional IP allocation, fixed /20 subnet per region, expandable to /16
- custom = no default subnets created, full control of IP ranges, regional IP allocation, expandable to IP ranges you specify

- Subnetworks
- Regions
- Zones
- IP Addresses: internal, external, range
  - each VM can have 2 IPs assigned: internal + external
  - internal: allocated from subnet range to VMs by DHCP, lease renewed every 24hrs, VM name + IP is registered with network-scoped DNS
  - external: assigned from pool (ephemeral), reserved (static), BYOIP, VM doesn't know external IP; it is mapped to the internal IP
- VMs
- Routes
  - Apply traffic egressing a VM, fwd traffic to most specific route, created when subnet is created, enable VMs on same network to communicate, destination is in CIDR, traffic is delivered only if it matches a firewall fule
- Firewalls
  - VPC network functions as a distributed firewall, firewall rules are applied to the network as a whole, connections are allowed or denied at the instance level, firewall rules are stateful, implied deny all ingress and allow all egress

- Common Network Designs
  - increased availability with multiple zones
  - globalization with multiple regions
  - Best practice - only assign internal IP to VM instances whenever possible
  - Cloud NAT = GCP managed network address translation services
    - provides internet access to private instances
  - private google access to Google APIs and services

4 VMs

- Compute Engine

- IaaS solution, server autoscaling, use case is general workloads
- predefined or custom machines types
- Compute, Storage, Networking
- common compute engine actions
  - metadata and scripts for boot, run, maintenance, shudown
  - move an instance to a new zone: between regions is manual process, within = automatic
  - Snapshots: backup critical data, migrate data between zones, transfer to SSD to improve performance
  - resize persistent disk: you can grow size but never shrink
- VM access & lifecycle
  - > Linux: SSH , requires firewall rule to allow tcp: 22
    > Win: RDP , requires firewall rule to allow tcp: 3389
  - Lifecycle: provisioning > staging > running > stopping
    - Availability policy
      > automatic restart: auto VM restart due to crash or maintenance event
      > on host maintenance: determines whether host is live-migrated or terminated due to a maintenance event
      > live migration: during maintenance event, VM is migrated to different hardware without interruption
    - OS patch management:
      > Patch compliance reporting
      > Patch deployment
    - Charges for stopped/terminated VMs:
      > Attached Disks
      > Reserved IP addresses
      >>can't change image of stopped VM
- Compute options
  - Machine type structure: family >> series >> type
    - families:
      > General-purpose: E2, N2, N2D, N1, Tau T2D
      > Compute-optimized: C2, C2D
      > Memory-optimized: M1, M2
      > Accelerator-optimized: A2
    - Custom machine types: cost slightly more than predefined

- Compute pricing
  - per-second billing with 1min minimum
  - resource based pricing
  - discounts: sustained use, committed use, premptible VM instances
  - Recommendation Engine
    - notifies of underutilized instances
- Special compute configs
  - preemptible VMs
    - > lower price for interruptible service
      > VM might be terminated at any time
      > no live migrate or auto restart
  - spot VMas
    - > latest version of preemptible VMs
      > share same pricing model as preemp
      > no min or max runtime
      > Spot VMs are finite Compute Engine resources, so they might not always be available
      > no live migrate or auto restart
  - sole-tenant nodes physically isolate workloads
  - Shielded VMs offer verifiable integrity
    - > secure boot
      > vTPM - virtual trusted platform module
      > integrity monitoring
  - Confidential VMs allow you to encrypt data in use
    - encrypts data while being processed
- Images
  - Boot loader, OS, File system structure, Software, Customizations
  - > public base images
    > custom images
    - Public: Goog, 3P vendors, community, premium
      > Linux and Win
    - Custom: create new image from VM, import from on-prem, workstation, or another cloud
  - machine image

- most ideal for disk backups as well as instance cloning replication

- Disk options
  - every VM comes with a single root persistent disk; image is loaded onto root disk during first boot
  - Persistent disks
    - attached to vm through network interface
    - network storage appearing as block device
    - disk resizing, even running and attached
    - zonal or regional: pd-standard, pd-ssd, pd-balanced, pd-extreme (zonal only)
    - durable storage - can survive VM terminate
    - encryption keys: google managed, customer managed, customer supplied
  - local SSD: physically attached to a VM
    - more IOPS, lower latency, higher throughput than persistent
  - RAM disk
    - faster than local disk, slower than memory
    - very volatile, erase on start or stop
    - consider using a persistent to back up RAM disk data