# Elastic GCP Infra: Scaling & Automation
# 2023 Ivan Vlad S.

- 1 Interconnecting Networks
  - Cloud VPN
    - securely connects your on-prem network to your GCP VPC network
      - IPsec VPN tunnel
    - HA VPN: high-availability option
    - dynamic routing with Cloud Router
    - on-prem VPN gateway
  - Cloud Interconnect and Peering
    - Dedicated Interconnect
      - L3 - Direct Peering
      - L2 - Dedicated Interconnect
      - provides direct physical connections
        - connection in a google colo facility
        - if not near a google colo, partner interconnect provides connectivity through a supported service provider
      - SLAs in place
    - Shared
      - L3 - Carrier Peering
      - L2 - Partner Peering
      - provides direct connection between your business network and google's
        - done through service provider
      - no SLA
    - Cloud VPN is useful addition to direct and carrier peering (L3)
  - Choosing a connection
    - Interconnect

- Direct access to RFC1918 IPs in your VPC - with SLA
  - Dedicated interconnect
  - Partner interconnect
  - Cloud VPN
- Peering
  - Access to Google public IPs only -without SLA
    - Direct Perring
    - Carrier Peering
- Sharing VPC networks for multi-project networking
  - Shared VPC
    - centralized network administration
    - share a network across several projects in your GCP org
  - VPC Peering
    - decentralized
    - allows you to configure private communication

## 2 Load Balancing & Autoscaling

- Global vs Regional
  - Global
    - HTTP(S), SSL proxy, TCP proxy
  - Regional
    - Internal TCP/UDP, Network TCP/UDP, Internal HTTP(S)
- Managed instance groups
  - collection of identical VM instances that you control as a single entity using an instance template. Can easily update all instances in a group
  - work with load balancers to distribute network traffic to instances
    - used for autoscaling - based on increase and decrease in load
- Autoscaling and health checks
  - dynamically add/remove instances
    - increase or decrease in load
  - autoscaling policy

- cpu utilization
- load balancing capacity
- monitoring metrics
- queue-based workload

- health check
  - GCP computes health state for each instance based on config

- HTTP(S) load balancing
  - Global, anycast IP address, HTTP port 80 or 8080, HTTPs port 443, IPv4 or v6, autoscaling, URL maps
    - backend services contain: health check, session affinity, time out setting (30 sec default), one or more backends
  - cross region load balancing, content-based load balancing
  - Target HTTP(S) Proxy
  - requires one signed SSL certificate installed (minimum)
    - up to 15 certs per target proxy
    - client SSL session terminates at the load balancer
  - support the QUIC transport layer protocol
  - can use backend buckets
  - network endpoint group (NEG)
    - configuration object that specifies a group of backend endpoints or services

- Cloud CDN
  - Content delivery network, uses google globally distributed edge points of prescence to cache HTTP(s) load-balanced content to your users.
    - over 90 CDN nodes / cache sites
  - Cloud CDN cache modes
    - control the factors that determine whether or not Cloud CDN caches your content

- SSL Proxy / TCP Proxy Load Balancing
  - SSL proxy load balancing
    - glocal load balancing for encrypted, non-http traffic

- terminates SSL session at load balancing layer
- IPv4 or IPv6
- benefits: intelligent routing, certificate mgmt, security patching, SSL policies

- TCP proxy load balancing
  - global load balancing for unencrypted, non-http traffic
  - terminates TCP sessions at load balancing layer
  - IP v4 or v6
  - benefits: intelligent routing, security patching

- Network load balancing
  - regional, non-proxied load balancer
  - FWD'ing rules (IP protocol data)
  - Traffic
    - UDP, TCP/SSL ports
  - Architecture
    - backend service-based, target pool-based
- Internal load balancing
  - Internal TCP/UDP loadbalancing
    - regional, private load balancing
      - VM instances in same region, RFC1918 addresses
    - TCP/UDP traffic
    - reduce latency, simpler config
    - software-defined, fully distributed load balancing
  - Internal HTTP(S) load balancing
    - regional, private load balancing
      - VM instances in same region, RFC1918 addresses
    - http, https, or http/2 protocols
    - based on open source Envoy proxy
- Choosing a load balancer

- Summary of load balancers

| Load balancer | Traffic type | Global/ Regional | External/ Internal | External ports for load balancing |
|---|---|---|---|---|
| HTTP(S) | HTTP or HTTPS | Global IPv4 IPv6 | External | HTTP on 80 or 8080; HTTPS on 443 |
| SSL Proxy | TCP with SSL offload | | | 25, 43, 110, 143, 195, 443, 465, 587, 700, 993, 995, 1883, 5222 |
| TCP Proxy | • TCP without SSL offload<br>• Does not preserve client IP addresses | | | 25, 43, 110, 143, 195, 443, 465, 587, 700, 993, 995, 1883, 5222 |
| Network TCP/UDP | • TCP/UDP without SSL offload<br>• Preserves client IP addresses | Regional IPv4 | | Any |
| Internal TCP/UDP | TCP or UDP | | Internal | Any |
| Internal HTTP(S) | HTTP or HTTPS | | | HTTP on 80 or 8080; HTTPS on 443 |

## 3 Infrastructure Automation

- GCP supports many IaC tools
  - Terraform (main), Ansible, Chef, Puppet, Packer
- ⭐ Terraform
  - Used for IaC: allows quick provisioning and removing of instrastructures
    - build an infra when needed
    - destroy infra when not in use
    - create identical infras for dev, test , and prod
    - can be part of ci/cd pipeline
    - templates are the building blocks for disaster recovery procedures
    - manage resource dependencies and complexity
  - infra automation tool
    - repeatable deployment process
    - declarative language
    - focus on the app
    - parallel deployment
    - template-driven
  - terraform language is the interface to declare resources
    - resources are infra objects
    - the config file guides the management of the resource
  - can be used on multiple public and private cloud

- considered a first-class tool in GCP
- already installed in Cloud Shell

- GCP Marketplace
  - deploy production-grade solutions from 3P vendors
  - single bill for GCP and 3P services
  - manage solutions using terraform
  - notifications when a security update is available
  - direct access to partner support

## 4 Managed Services

- Managed service = outsource a lot of the admin and maintenance overhead to Google if your app reqs fit within the service offering, e.g., Serverless

- BigQuery
  - GCP serverless, highly scalable, and cost-effective cloud data warehouse
    - fully managed
    - petabyte scale
    - SQL interface
    - very fast

- Dataflow
  - use Dataflow to execute a wide variety of data processing patterns
    - serverless, fully managed data processing
    - batch and stream processing with autoscale
    - open source programming using Beam
    - intelligently scale to millions of QPS
  - manual or automatic provisioning of clusters

- Dataprep
  - use Dataprep to visually explore, clean, and prepare data for analysis and ML
    - serverless, works at any scale
    - suggests ideal data transformation
    - focus on data analysis
    - integrated partner service operated by Trifacta

- Dataproc
  - service for running Apache Spark and Apache Hadoop clusters
    - low cost (per-second, preemptible)
    - super fast to start, scale, and shut down
    - integrated with GCP
    - managed service
    - simple and familiar