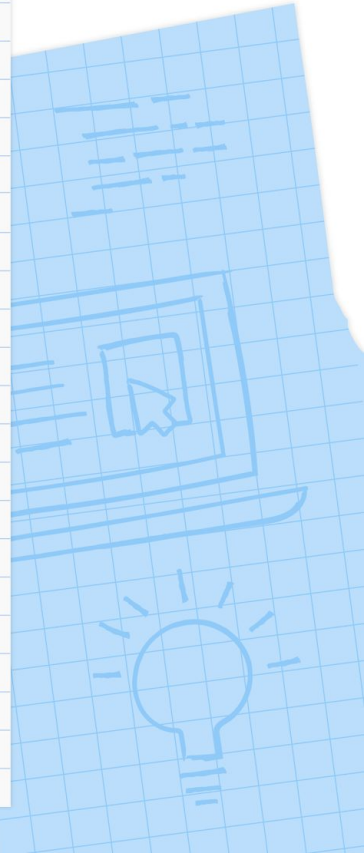
Google Developer Groups

GDG Cloud Team: Google Cloud Associate

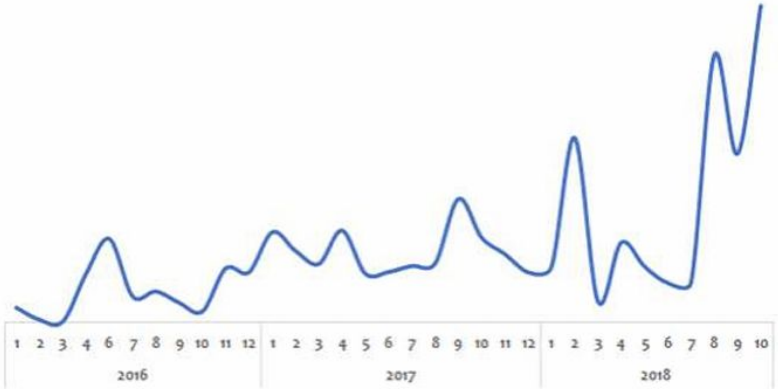
2024/11/11
Jinsuk Park

Index



Why do we need the cloud?

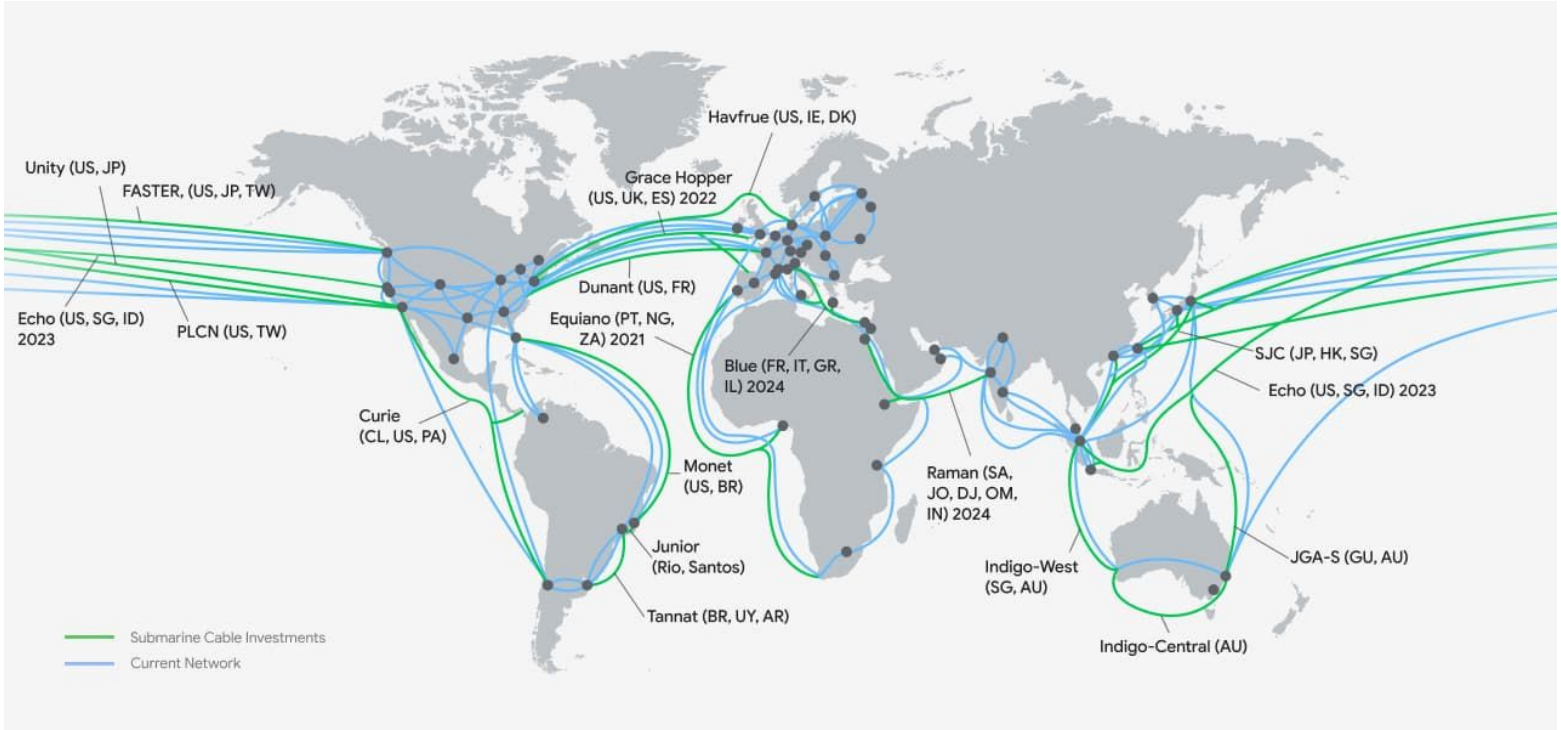
- Expensive maintaining infrastructure
- Needs to plan ahead
- Low infrastructure utilization
- Need dedicated infrastructure team



Why do we need the cloud?

- Trade "capital expense" for "variable expense"
- Benefit from massive economies of scale
- Stop guessing capacity
- No money in maintaining data centers
- "Go global" in minutes!

1. Regions & Zones



What is a Region?

Specific geographical locations to host our resources!

Pros:

1. High Availability
2. Low Latency
3. Global Footprint
4. Adhere to government regulations

40 Regions, 121 Zones, 187 Network Edge locations, Available in 200+ countries & territories

What is a Zone?

A deployment area within a region

Features:

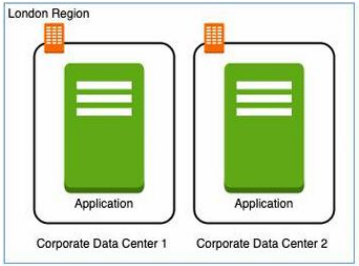
1. **3+ zones** within a region
2. **1+ cluster** for each zone
3. Zones within region are connected via **low-latency links**

Zones	**Region**	Machine Types	CPUs
asia-northeast3-a	**Seoul, South Korea, APAC**	E2, N2, N2D, N1, M1, M2, C2, A3, A2, G2	Intel Ivy Bridge, Sandy Bridge, Haswell, Broadwell, Skylake, Cascade Lake, Ice Lake, AMD EPYC Rome, AMD EPYC Milan
asia-northeast3-b	**Seoul, South Korea, APAC**	E2, N2, N2D, N1, M1, M2, C2, A2, G2	Intel Ivy Bridge, Sandy Bridge, Haswell, Broadwell, Skylake, Cascade Lake, Ice Lake, AMD EPYC Rome, AMD EPYC Milan
asia-northeast3-c	**Seoul, South Korea, APAC**	E2, N2, N2D, N1, C2, A3	Intel Ivy Bridge, Sandy Bridge, Broadwell, Skylake, Cascade Lake, Ice Lake, AMD EPYC Rome, AMD EPYC Milan

What are the challenges we can solve?

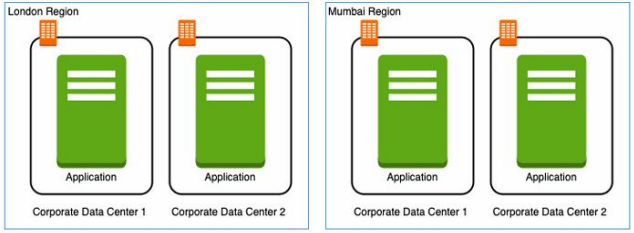
Problem 1: What if the data center crashes?

→ Solution? **Multiple data centers!** → Our application will still run from another datacenter within the same region!



Problem 2: What if the entire Seoul region is unavailable?

→ Solution? **Multiple Regions** → Our application is served in Los Angeles!



Problem 3: Slow access for users from other parts of the world

→ Solution? **Add deployments for our applications in other regions**

2. Google Compute Engine

What is Google Compute Engine (GCE)?

Provision & Manage Virtual Machines

Features:

1. Create and manage lifecycle of instances
2. Load balancing and auto scaling for multiple instances
3. Attach storage (& network storage) to your instances
4. Manage network connectivity and configuration for your VM instances

Compute Engine Machine Family

Machine configuration

General purpose | Compute optimized | ✓ Memory optimized | Storage optimized | GPUs

Machine types for workloads with higher memory-to-vCPU ratios, like in-memory databases

	Series	Description	vCPUs	Memory	Platform
○	X4	Extra Large, in-memory databases	960 - 1920	16,384 - 32,768 GB	Intel Sapphire Rap
○	M3	High memory, memory-intensive workloads	32 - 128	976 - 3,904 GB	Intel Ice Lake
◉	M2	Ultra-high memory, in-memory databases	208 - 416	5,888 - 11,776 GB	Intel Cascade Lak
○	M1	High memory, memory-intensive workloads	40 - 160	961 - 3,844 GB	Intel Skylake

Machine type

Choose a machine type with preset amounts of vCPUs and memory that suit most workloads.

m2-megamem-416 (416 vCPU, 208 core, 5,888 GB memory)

vCPU
416 (208 cores)

Memory
5,888 GB

Compute Engine Machine Family

Choose type of hardware we want to run

Types:

1. **General Purpose:**

Best price-performance ratio Web and application servers, Small-medium databases, Dev environments

2. **Memory Optimized:**

Ultra high memory workloads Large in-memory databases and In-memory analytics

3. **Compute Optimized:**

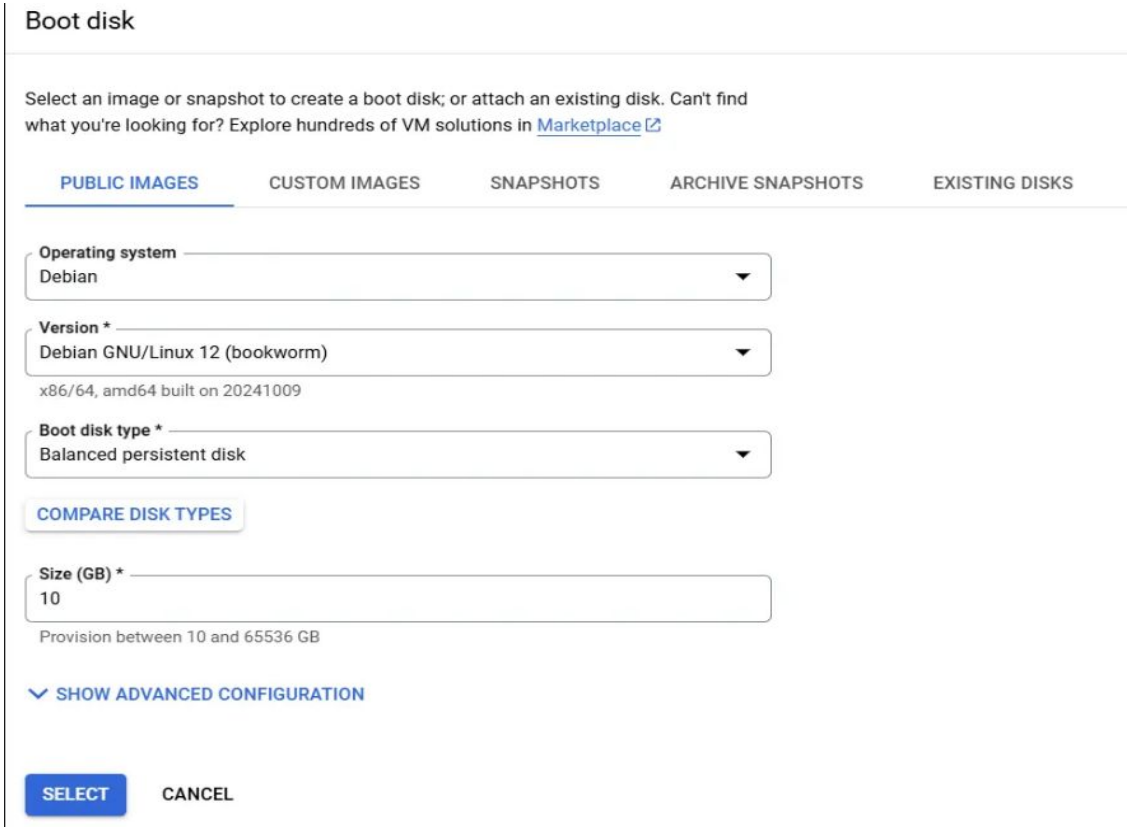
Compute intensive workloads Gaming applications

Compute Engine Machine Family

Machine name	vCPUs[1]	Memory (GB)	Max number of persistent disks (PDs)[2]	Max total PD size (TB)	Local SSD	Maximum egress bandwidth (Gbps)[3]
e2-standard-2	2	8	128	257	No	4
e2-standard-4	4	16	128	257	No	8
e2-standard-8	8	32	128	257	No	16
e2-standard-16	16	64	128	257	No	16
e2-standard-32	32	128	128	257	No	16

Let's take an example : **e2-standard-2**
e2 - Machine Type Family
standard - Type of workload
2 - Number of vCPUs

Images


Boot disk
Select an image or snapshot to create a boot disk; or attach an existing disk. Can't find what you're looking for? Explore hundreds of VM solutions in Marketplace
PUBLIC IMAGES
CUSTOM IMAGES
SNAPSHOTS
ARCHIVE SNAPSHOTS
EXISTING DISKS
Operating system
Debian
Version *
Debian GNU/Linux 12 (bookworm)
x86/64, amd64 built on 20241009
Boot disk type *
Balanced persistent disk
COMPARE DISK TYPES
Size (GB) *
10
Provision between 10 and 65536 GB
SHOW ADVANCED CONFIGURATION
SELECT
CANCEL

Images

What operating system and what software do you want on the instance?

Public Images:
Provided & maintained by Google or Open source communities or third party vendors

Custom Images:
Created by you for your projects

Internal & External IP Addresses

External (public) IP Addresses can be accessed via internet

- NO two external IP addresses are alike
- **Ephemeral**: Doesn't exist when instance is stopped. New address when restarted
- **Static**: Remains constant with resource

Internal (private) IP Addresses are internal to corporate network

- Assigned at least 1 internal address
- Not accessible from internet directly

Static IP Addresses

Constant External IP Address for a VM instance

- Can allocate to other instance **WITHIN** same project

Caution!

- **Remains attached even if we stop instance**! We have to manually detach it! (Money…)

How do we simplify creating a VM instance?

- Startup Script
- Instance Template
- Custom Image

How do we simplify creating a VM instance?

1. Startup Script

```
#!/bin/bash
apt update
apt -y install apache2
echo "Hello world from $(hostname) $(hostname -I)" > /var/www/ht
```

Bootstrapping: Install OS patches/software when the VM instance is launched!

How do we simplify creating a VM instance?

2. Instance Templates

Simplifies VM instance launch process -> convenient way to Create similar instances!

- Cannot be updated!

How do we change? Copy an existing template & modify!

How do we simplify creating a VM instance?

3. Custom Image

Creates custom image with OS patches & software pre-installed!

- Can be shared across projects!
- Hardening an Image: customize image to match company security standards
- Prefor to use Custom image than Startup Script

How do we simplify creating a VM instance?

3. Custom Image

Creates custom image with OS patches & software pre-installed!

- Can be shared across projects!
- Hardening an Image: customize image to match company security standards
- Prefor to use Custom image than Startup Script

3. Google Compute - Optimizing Costs & Performance

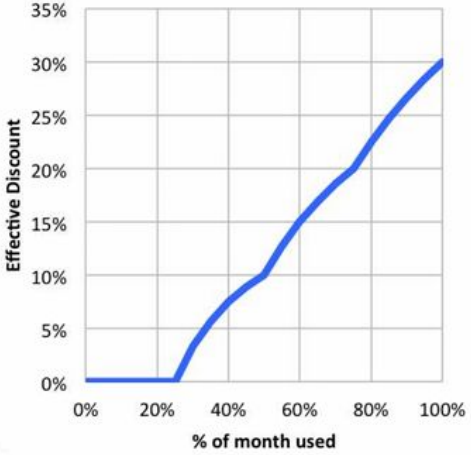
Sustained use discounts

Automatic savings provided by Google Cloud for running VM instances for significant portion of month

- ex) If we use n1 for more than 25% a month, we get 20~50% discount on every incremental minute
- Discount increases with usage
- Instances created by GKE, GCE

Not Applied on following

- Machine type E2, A2,
- App Engine Flexible and Dataflow created VM instances



Source: https://cloud.google.com

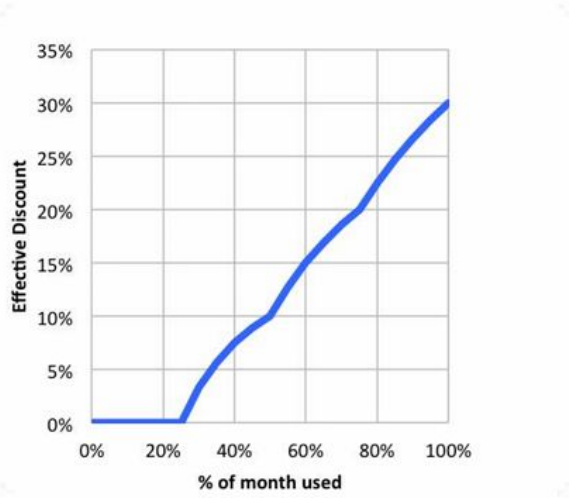
Committed use discounts

Workloads with predictable usage needs

- Commit for **1 or 3 years**
- Up to 70% discount based on machine type & GPUs
- Instances created by GKE, GCE

Not Applied on following

- App Engine Flexible and Dataflow created VM instances



Source: https://cloud.google.com

Spot Virtual Machines

Preemptible virtual machines (**NO max runtime**)

- Can be stopped by GCP anytime within 24 hours
- Instances get 30 second warning (to save anything they want to save)

When to use?

- Application is fault tolerant (i.e batch processing jobs)
- Save costs (60 ~ 91% cheaper than on-demand VMs)
- Workload is not immediate

Restrictions

- NOT always available,
- NO automatic restarts
- Cannot be migrated to regular VMs
- Free tier credits not applicable

Google Compute Engine - Billing policy

Billed by the second! (after min 1 minute)

- Not billed when Compute Engine is STOPPED (however, storage can be charged)

Always create budget alerts!

How can we save money?

1. Choose right machine type, image for our workload
2. Sustained user discounts
3. Committed user discounts
4. Discounts for preemptible VM instances

Google Compute Engine - Billing policy

Billed by the second! (after min 1 minute)

- Not billed when Compute Engine is STOPPED (however, storage can be charged)

Always create budget alerts!

How can we save money?

1. Choose right machine type, image for our workload
2. Sustained user discounts
3. Committed user discounts
4. Discounts for preemptible VM instances

Compute Engine : Live Migration & Availability Policy

How do we keep your VM instances running when a host system needs to be updated?

Live Migration

1. Your running instance is migrated to another host in the **same zone**
1. Does NOT change any attributes or properties of the VM
2. SUPPORTED for instances with local SSDs
3. NOT SUPPORTED for GPUs and preemptible instances

Availability Policy

1. What should happen during periodic infrastructure maintenance?

-> Migrate or terminate

2. How should we handle non-user initiated terminated instances?

-> Automatic restart!

Compute Engine Features: Custom Machine Types

What if predefined machine type options are not suitable for our workload???

Custom Machine Type!

- Adjust vCPUs, GPUs, and memory
- Choose between E2, N1, N2 machine types
- Supports wide variety of OS
- Billed per vCPUs, memory provisioned to each instance

  ex) Example Hourly Price: $0.033174 / vCPU + $0.004446 / GB

Compute Engine Features: GPUs

How do you accelerate math intensive and graphics-intensive workloads for AI tasks???

Add GPU to Virtual Machine!!

- Expensive…
- Use images with GPU libraries installed or GPU won't be utilized
- Not supported on memory-optimized, shared-core machine types
- On host maintenance can only have the value "Terminate VM instance"
- Availability policy: Automatic restart - on

Compute Engine: Key takeaways

- Associated with a **project**
- Machine type availability can vary from **region to regions**
- We can only change the machine type (adjust the number of vCPUs and memory) of a **stopped** instance **NOT a running instance**
- VM's can be filtered by various properties (Name, Zone, Machine Type, Internal/External IP, Network, Labels etc)
- Instances are Zonal (Run in a specific zone (in a specific region))
    - -Images are global (You can provide access to other projects - if needed)
    - -Instance templates are global (Unless you use zonal resources in your templates)
- **Automatic Basic Monitoring** is enabled
    - -Default Metrics: CPU utilization, Network Bytes (in/out), Disk Throughput/IOPS
    - -For Memory Utilization & Disk Space Utilization- **Cloud Monitoring agent** is needed

Compute Engine: Key takeaways

Scenario	Solution
What are the pre-requisites to be able to create a VM instance?	1. Project 2. Billing Account 3. Compute Engines APIs should be enabled
You want dedicated hardware for your compliance, licensing, and management needs	Sole-tenant nodes
I have 1000s of VM and I want to automate OS patch management, OS inventory management and OS configuration management (manage software installed)	Use "VM Manager"
You want to login to your VM instance to install software	You can SSH into it
You do not want to expose a VM to internet	Do NOT assign an external IP Address
You want to allow HTTP traffic to your VM	Configure Firewall Rules

Quiz!

You have an application that looks for its licensing server on the IP 10.0.3.21. You need to deploy the licensing server on Compute Engine. You do not want to change the configuration of the application and want the application to be able to reach the licensing server. What should you do?

A. Reserve the IP 10.0.3.21 as a static internal IP address using gcloud and assign it to the licensing server.

B. Reserve the IP 10.0.3.21 as a static public IP address using gcloud and assign it to the licensing server.

C. Use the IP 10.0.3.21 as a custom ephemeral IP address and assign it to the licensing server.

D. Start the licensing server with an automatic ephemeral IP address, and then promote it to a static internal IP address.

Quiz!

You want to configure 10 Compute Engine instances for availability when maintenance occurs. Your requirements state that these instances should attempt to automatically restart if they crash. Also, the instances should be highly available including during system maintenance. What should you do?

A. Create an instance template for the instances. Set the 'Automatic Restart' to on. Set the 'On-host maintenance' to Migrate VM instance. Add the instance template to an instance group.

B. Create an instance template for the instances. Set 'Automatic Restart' to off. Set 'On-host maintenance' to Terminate VM instances. Add the instance template to an instance group.

C. Create an instance group for the instances. Set the 'Autohealing' health check to healthy (HTTP).

D. Create an instance group for the instance. Verify that the 'Advanced creation options' setting for 'do not retry machine creation' is set to off.

Quiz!

You have a virtual machine that is currently configured with 2 vCPUs and 4 GB of memory. It is running out of memory. You want to upgrade the virtual machine to have 8 GB of memory. What should you do?

A. Rely on live migration to move the workload to a machine with more memory.

B. Use gcloud to add metadata to the VM. Set the key to required-memory-size and the value to 8 GB.

C. Stop the VM, change the machine type to n1-standard-8, and start the VM.

D. Stop the VM, increase the memory to 8 GB, and start the VM.

Q&A