

ĐIỆN TOÁN ĐÁM MÂY (Cloud Computing)

PRACTICES Practice 5 – Serverless Computing (Compute Engine)

Presenter: **Dr. Nguyen Dinh Long**

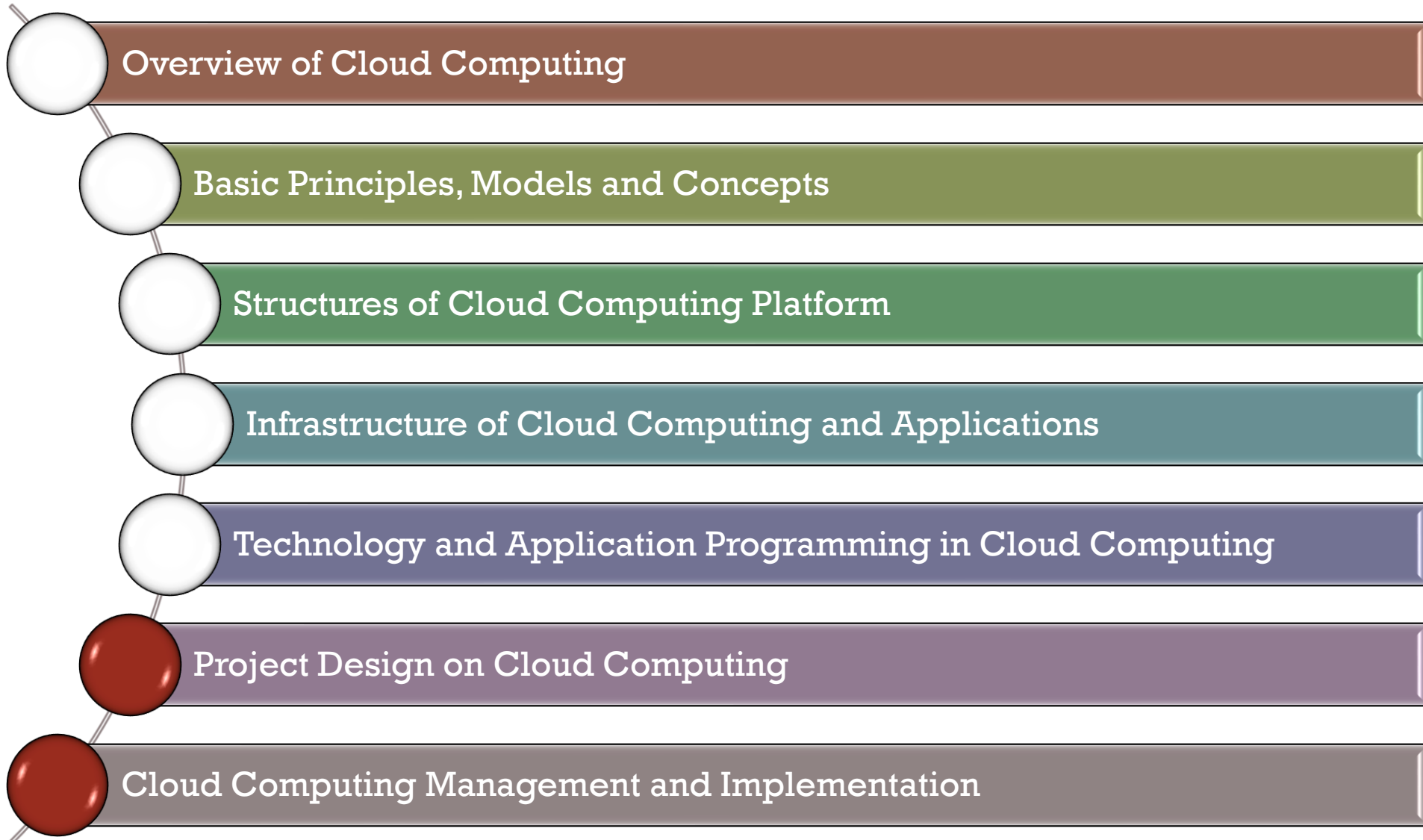
Email: dinhlonghcmut@gmail.com

Phone: +84 947 229 599

Google-site: <https://sites.google.com/view/long-dinh-nguyen>

Nov. 2022

Outline



References

Main:

- Thomas Erl, Zaigham Mahmood, and Ricardo Puttini. 2013. *Cloud Computing Concepts, Technology & Architecture*. Prentice Hall.
- Michael J. Kavis. 2014. *Architecting the Cloud: Design Decisions for Cloud Computing Service Models*. Wiley
- Arshdeep Bahga, and Vijay Madisetti. 2013. *Cloud Computing: A Hands-On Approach*. CreateSpace Independent Publishing Platform

More:

- Rajkuma Buyya, Jame Broberg and Andrzej Goscinski. 2011. *Cloud Computing –Principles and paradigms*, Wiley
- Nick Antonopoulos, and Lee Gillam. 2010. *Cloud Computing - Principles, Systems and Applications*, Springer-Verlag London Limited.
- Slides here are modified from several sources in Universities and Internet.

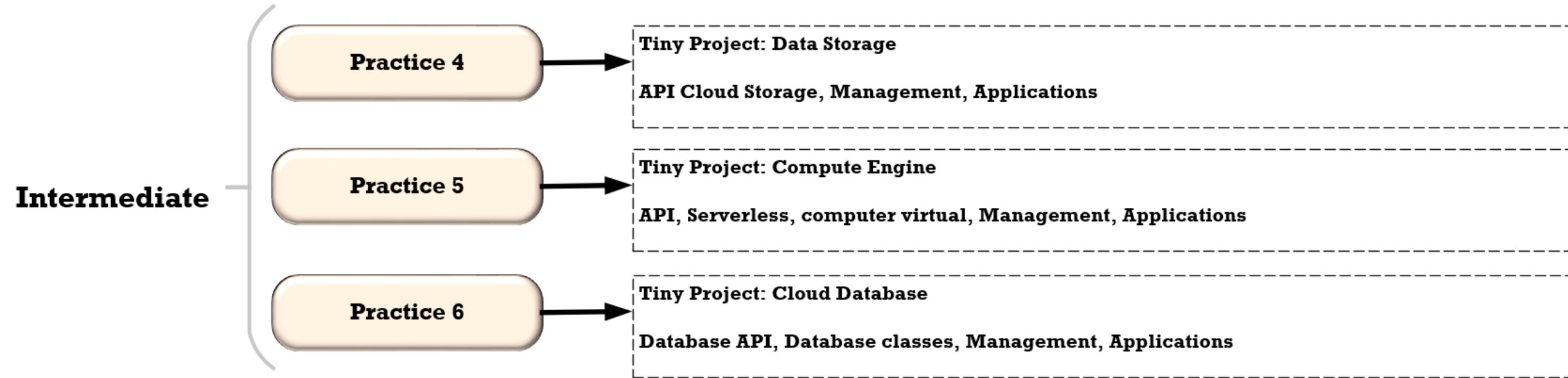
Cloud Computing: Practices

Levels: Beginning (3 weeks) – **Intermediate (3 weeks)** – Advanced (3 weeks)

Groups: 9 with 5 person/group

Practice: submit a report for each group, submit to our Google Classroom

Cloud Computing: Practices



Content of Practice 4

1. Resource settings
2. Resource API
3. Resource logging
4. Resource billing



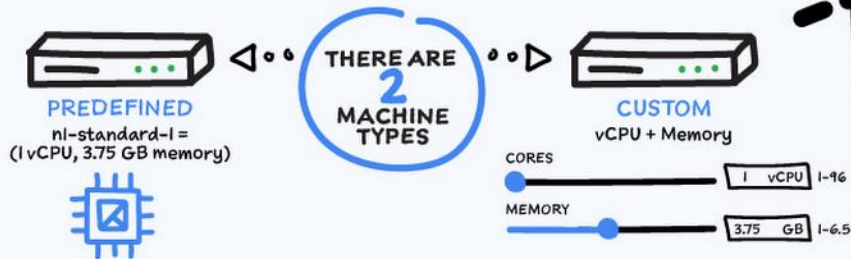
Compute Engine

#GCPsketchnotes

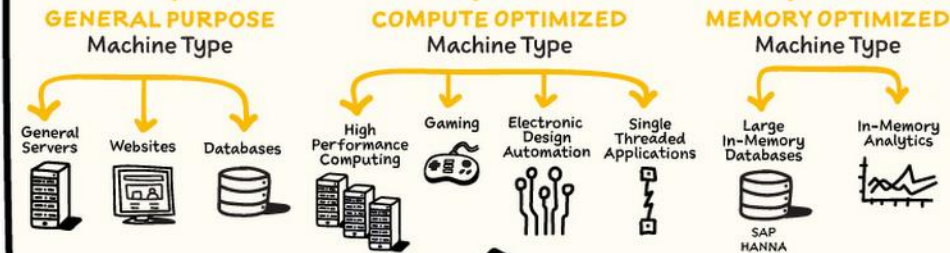
@PVERGADIA THECLOUDGIRL.DEV

What is Compute Engine?

CUSTOMIZABLE VIRTUAL MACHINES IN GOOGLE CLOUD



THERE ARE **3** MACHINE TYPE FAMILIES



Compute Engine Use case (example)



FIG. 1

✓ Websites



FIG. 2

✓ Databases



FIG. 3

✓ Legacy Monolithic Apps



FIG. 4

✓ Containers



FIG. 5

✓ Windows Apps

Compute Engine

PRICING

SUSTAINED USE SAVINGS

Automatic discounts for running VMs a significant portion of the month



PREEMPTIBLE VMs

Up to 80% savings and run batch jobs & fault-tolerant workloads



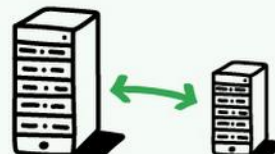
COMMITTED USE DISCOUNT

Up to 57% savings with no up-front cost



RIGHT SIZE RECOMMENDATIONS

Suggests resizing for efficiency and cost



How does it **WORK?** ???

CREATE

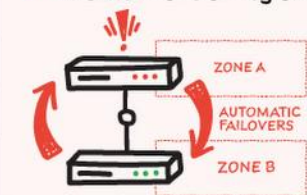
region
+ zone
+ machine type
(cpu & memory)
= Instance

BACKUPS

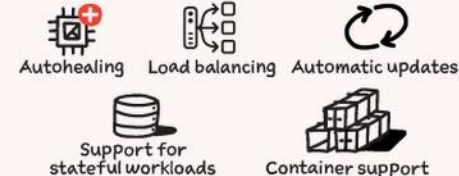


HIGH AVAILABILITY

Automatic failover to another zone or region



MANAGED INSTANCE GROUPS (MIGs)



AUTOSCALER - 3 types of policies:

- 1.** CPU utilization
= more than 60%
→ create new instance
- 2.** HTTP(S) load balancers service capacity
Requests per second or utilization
- 3.** Cloud monitoring metrics

Compute Engine

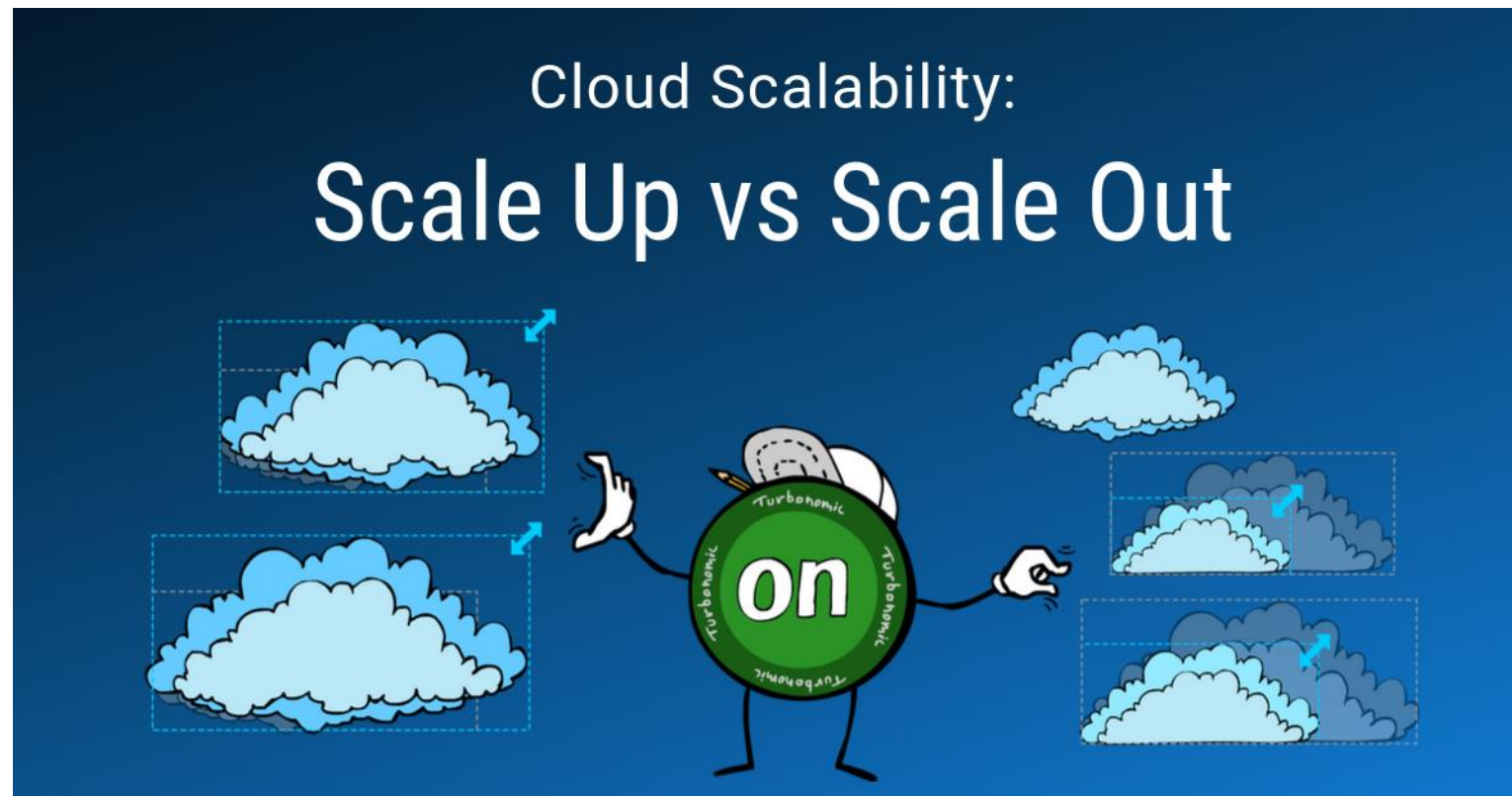
- ❑ **Secure and customizable compute service that lets you create and run virtual machines on Google's infrastructure.**
- **Predefined machine types:** Start running quickly with pre-built and ready-to-go configurations.
- **Custom machine types:** Create VMs with optimal amounts of vCPU and memory, while balancing cost.
- **Spot machines:** Reduce computing costs by up to 91%.
- **Confidential computing:** Encrypt your most sensitive data while it's being processed.
- **Rightsizing recommendations:** Optimize resource utilization with automatic recommendations.

Compute Engine

❑ Choosing the right virtual machine type:

- Scale out workloads (T2A, T2D):

Tau VMs are the lowest cost solution for scale-out workloads on Compute Engine, with up to 42% higher price-performance compared to general-purpose VMs of any of the leading public cloud vendors. Choose from x86 or Arm-based VMs to meet your workload and business requirements.

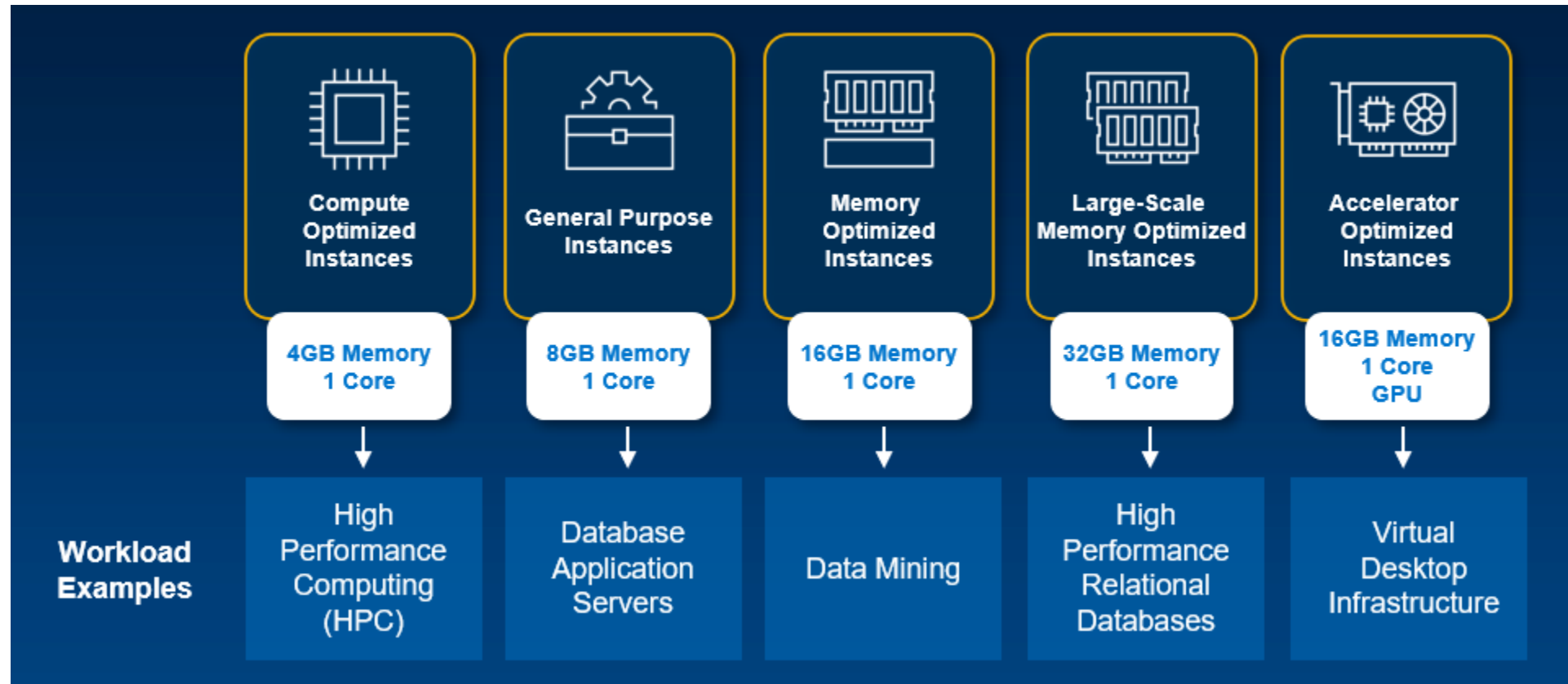


Compute Engine

❑ Choosing the right virtual machine type:

- General purpose workloads (E2, N2, N2D, N1)

E2, N2, N2D, and N1 are general-purpose machines offering a good balance of price and performance, and are suitable for a wide variety of common workloads including databases, development and testing environments, web applications, and mobile gaming. They support up to 224 vCPUs and 896 GB of memory.

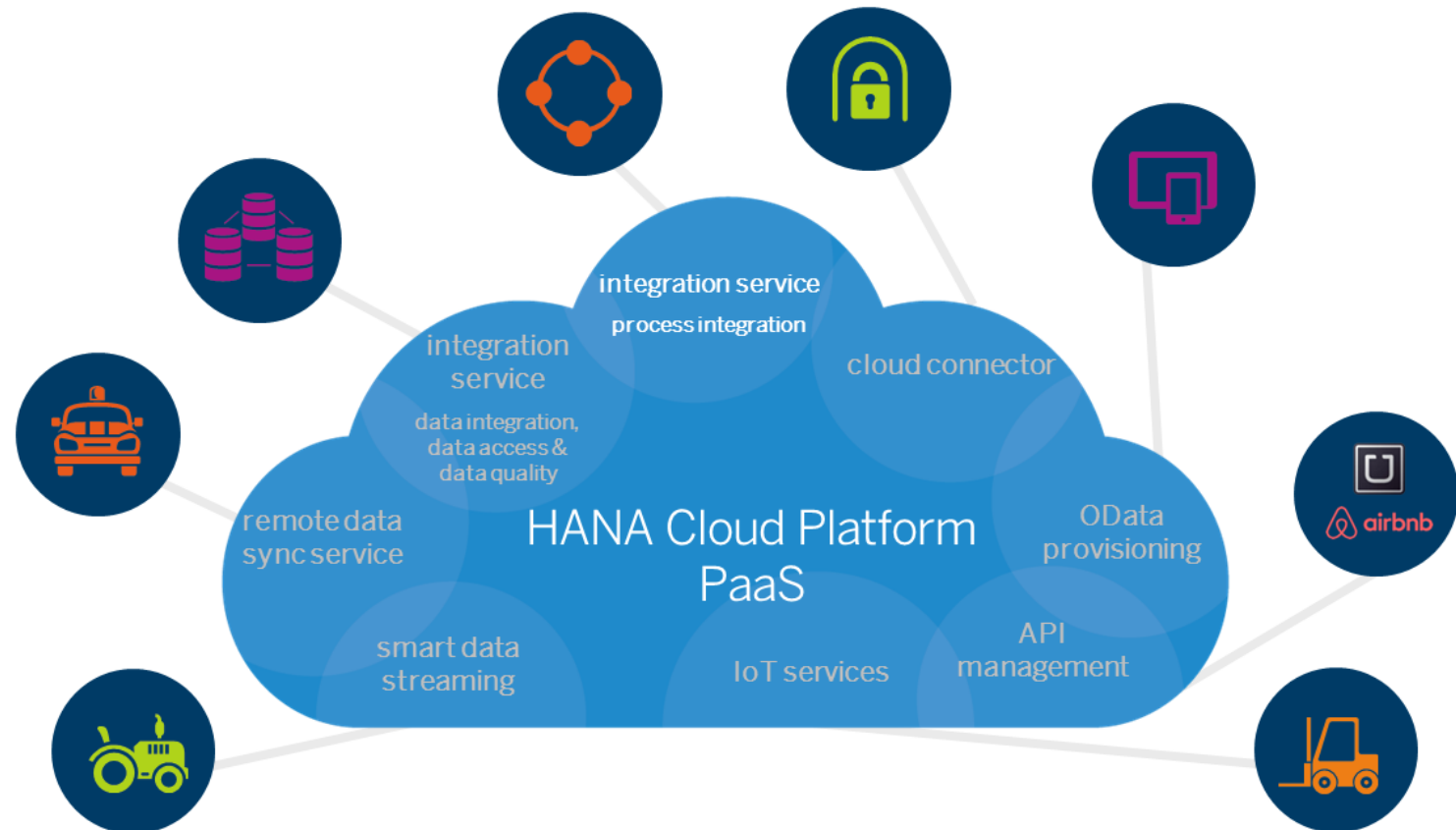


Compute Engine

❑ Choosing the right virtual machine type:

- Ultra-high memory (M2, M1)

Memory-optimized machines offer the highest memory configurations with up to 12 TB for a single instance. They are well suited to memory-intensive workloads such as large in-memory databases like SAP HANA, and in-memory data analytics workloads.

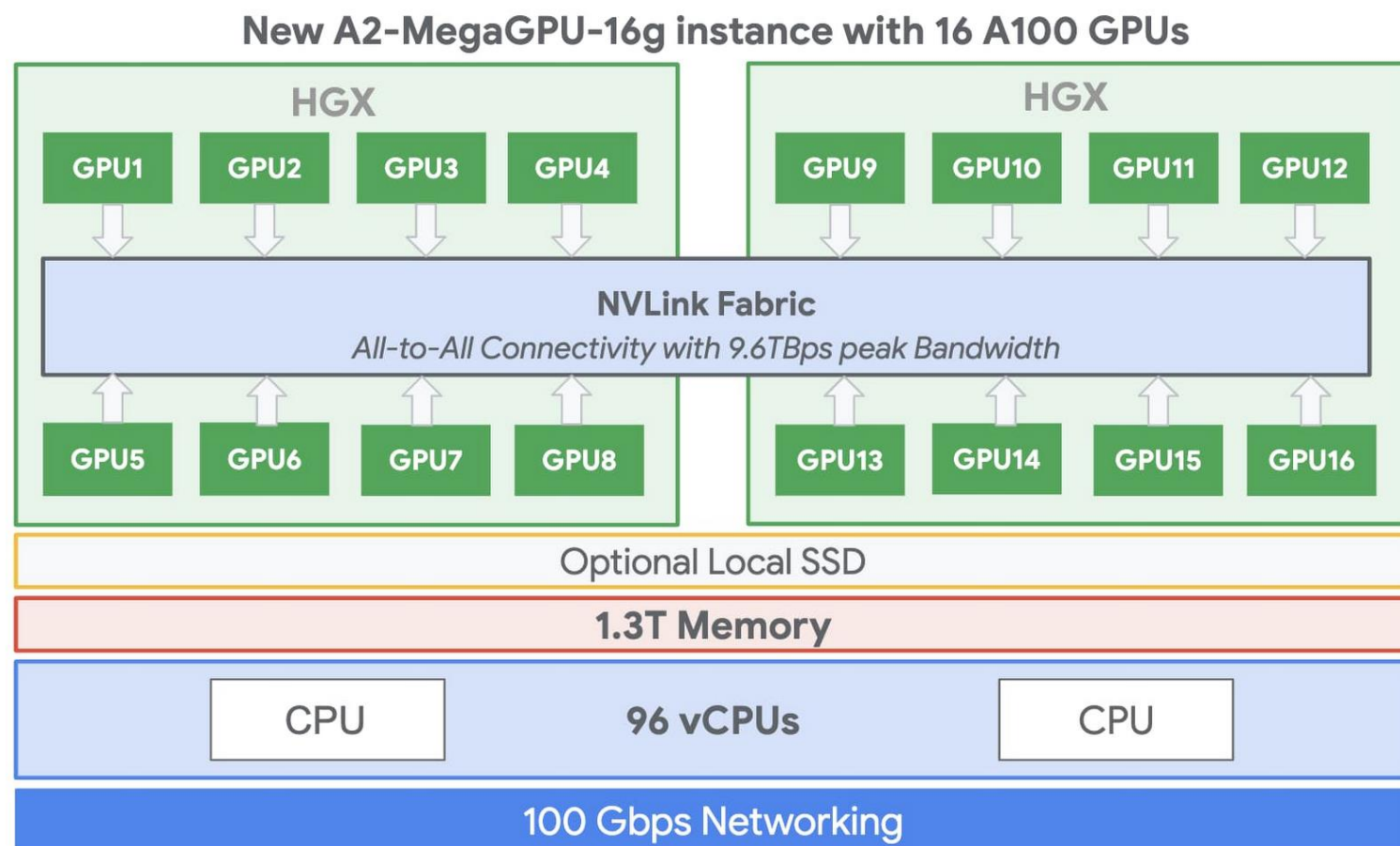


Compute Engine

❑ Choosing the right virtual machine type:

- [Compute-intensive workloads \(C3, C2, C2D\)](#)

Compute-optimized machines provide the highest performance per core on Compute Engine and are optimized for workloads such as high performance computing (HPC), game servers, and latency-sensitive API serving.



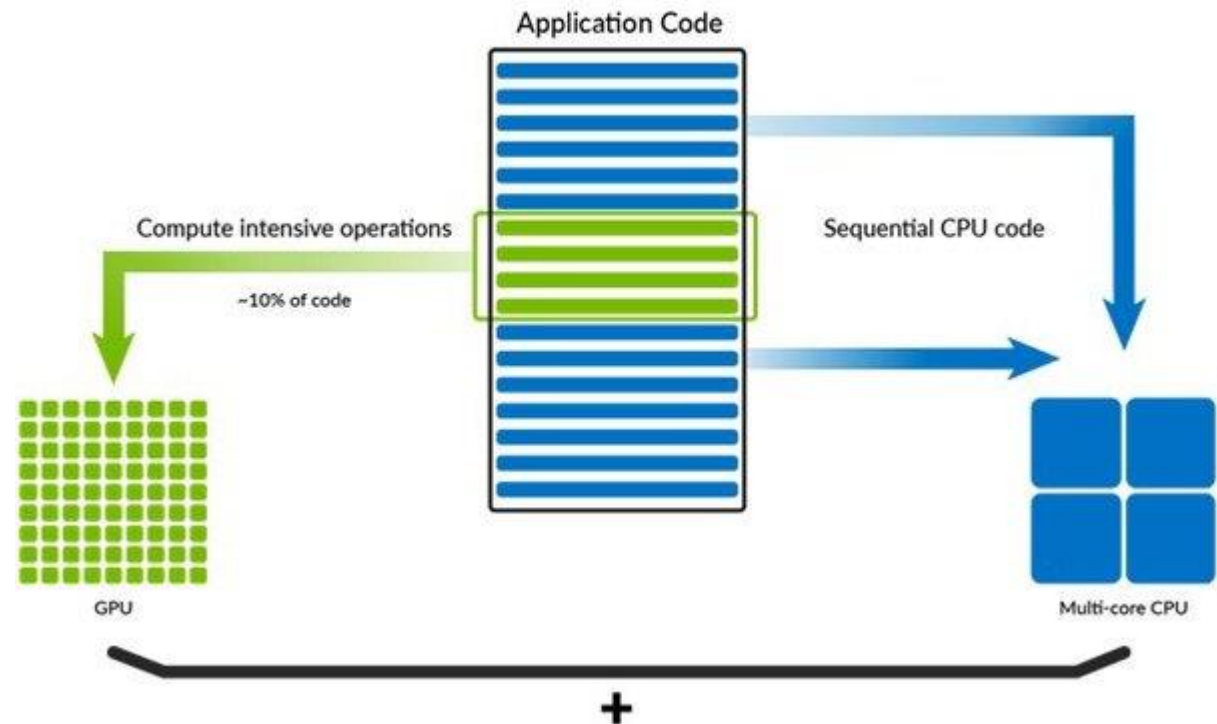
Compute Engine

❑ Choosing the right virtual machine type:

- [Most demanding applications and workloads \(A2\)](#)

Accelerator-optimized machines are based on the NVIDIA Ampere A100 Tensor Core GPU. Each A100 GPU offers up to 20x the compute performance compared to the previous generation GPU. These VMs are designed for your most demanding workloads such as machine learning and high performance computing.

How GPU Acceleration Works



Compute Engine

Machine family and series recommendations:

Workload type					
General-purpose workloads			Optimized workloads		
Cost-optimized	Balanced	Scale-out optimized	Memory-optimized	Compute-optimized	Accelerator-optimized
E2	N2, N2D, N1	Tau T2D, Tau T2A	M3, M2, M1	C2, C2D	A2
Day-to-day computing at a lower cost	Balanced price/performance across a wide range of VM shapes	Best performance/cost for scale-out workloads	Ultra high-memory workloads	Ultra high performance for compute-intensive workloads	Optimized for high performance computing workloads
<ul style="list-style-type: none">• Web serving• App serving• Back office apps• Small-medium databases• Microservices• Virtual desktops• Development environments	<ul style="list-style-type: none">• Web serving• App serving• Back office apps• Medium-large databases• Cache• Media/streaming	<ul style="list-style-type: none">• Scale-out workloads• Web serving• Containerized microservices• Media transcoding• Large-scale Java applications	<ul style="list-style-type: none">• Medium-large OLAP and in-memory databases such as SAP HANA• In-memory databases and in-memory analytics• Microsoft SQL Server and similar databases• Genomic modeling• Electronic design automation	<ul style="list-style-type: none">• Compute-bound workloads• High-performance web serving• Gaming (AAA game servers)• Ad serving• High-performance computing (HPC)• Media transcoding• AI/ML	<ul style="list-style-type: none">• CUDA-enabled ML training and inference• HPC• Massive parallelized computation

Compute Engine

❑ CPU platforms:

When you create a virtual machine (VM) on Compute Engine, you specify a machine series and a machine type for the VM. Each machine series is associated with one or more CPU platforms. If there are multiple CPU platforms available for a machine type, you can select a minimum CPU platform for the VM.

A CPU platform offers multiple physical processors, and each of these processors are referred to as a core. For all processors available on Compute Engine, a single CPU core can run as multiple hardware multithreads through Simultaneous multithreading (SMT), which is known on Intel processors as Intel Hyper-Threading Technology. On Compute Engine, each hardware multithread is called a virtual CPU (vCPU).

The machine type of your VM specifies its number of vCPUs, and you can infer its number of physical CPU cores using the default vCPU per core ratio for that machine series:

- For the Tau T2D and Tau T2A machine series, VMs always have one vCPU per core.
- For all other machine series, VMs have two vCPUs per core by default.

Compute Engine

GPU platforms:

Compute Engine provides graphics processing units (GPUs) that you can add to your virtual machine (VM) instances. You can use these GPUs to accelerate specific workloads on your VMs such as machine learning and data processing.

Compute Engine provides NVIDIA GPUs for your VMs in passthrough mode so that your VMs have direct control over the GPUs and their associated memory.

GPU regions and zones availability

Select a location ▼

Select a GPU model ▼

Clear all

Zones ▼	Location	GPU platforms	NVIDIA RTX virtual workstations
asia-east1-a	Changhua County, Taiwan, APAC	T4, P100, K80	T4, P100
asia-east1-b	Changhua County, Taiwan, APAC	K80	
asia-east1-c	Changhua County, Taiwan, APAC	T4, V100, P100	T4, P100
asia-east2-a asia-east2-b asia-east2-c	Hong Kong, APAC		
asia-northeast1-a	Tokyo, Japan, APAC	A100 40GB, T4	T4
asia-northeast1-b	Tokyo, Japan, APAC		
asia-northeast1-c	Tokyo, Japan, APAC	A100 40GB, T4	T4
asia-northeast2-a asia-northeast2-b asia-northeast2-c	Osaka, Japan, APAC		

Compute Engine

❑ Regions and zones:

Compute Engine resources are hosted in multiple locations worldwide. These locations are composed of regions and zones. A region is a specific geographical location where you can host your resources. Regions have three or more zones.

For example, the us-west1 region denotes a region on the west coast of the United States that has three zones: us-west1-a, us-west1-b, and us-west1-c.

Resources that live in a zone, such as [virtual machine instances](#) or [zonal persistent disks](#), are referred to as [zonal resources](#).

Compute Engine

❑ Choosing a region and zone:

You choose which region or zone hosts your resources, which controls where your data is stored and used. Choosing a region and zone is important for several reasons:

Handling failures

Distribute your resources across multiple zones and regions to tolerate outages. Google designs zones to minimize the risk of correlated failures caused by physical infrastructure outages like power, cooling, or networking. Thus, if a zone becomes unavailable, you can transfer traffic to another zone in the same region to keep your services running.

Decreased network latency

To decrease network latency, you might want to choose a region or zone that is close to your point of service. For example, if you mostly have customers on the East Coast of the US, then you might want to choose a primary region and zone that is close to that area and a backup region and zone that is also close by.

Compute Engine

❑ Virtual machine instances:

An instance is a virtual machine (VM) hosted on Google's infrastructure. You can create an instance or create a group of managed instances by using the [Google Cloud console](#), the [Google Cloud CLI](#), or the [Compute Engine API](#).

Compute Engine instances can run the public images for [Linux](#) and [Windows Server](#) that Google provides as well as private custom [images](#) that you can create or import from your existing systems. You can also deploy [Docker containers](#), which are automatically launched on instances running the Container-Optimized OS public image.

You can choose the machine properties of your instances, such as the number of virtual CPUs and the amount of memory, by using a set of predefined machine types or by creating your own custom machine types.

Compute Engine

❑ Virtual machine instances - Characteristics:

- [Instances and projects](#)

Each instance belongs to a Google Cloud console project, and a project can have one or more instances.

- [Instances and storage options](#)

By default, each Compute Engine instance has a small boot persistent disk that contains the operating system. you can add additional storage options to your instance.

- [Instances and networks](#)

Each network interface of a Compute Engine instance is associated with a subnet of a unique VPC network.

- [Instances and containers](#)

Compute Engine instances support a declarative method for launching your applications using containers. When creating a VM or an instance template, you can provide a Docker image name and launch configuration.

- Tools to manage instances

To create and manage instances, you can use a variety of tools, including the [Google Cloud console](#), the [gcloud command-line tool](#), and the [REST API](#). To configure applications on your instances, connect to the instance using Secure Shell ([SSH](#)) for Linux instances or Remote Desktop Protocol ([RDP](#)) for Windows Server instances.

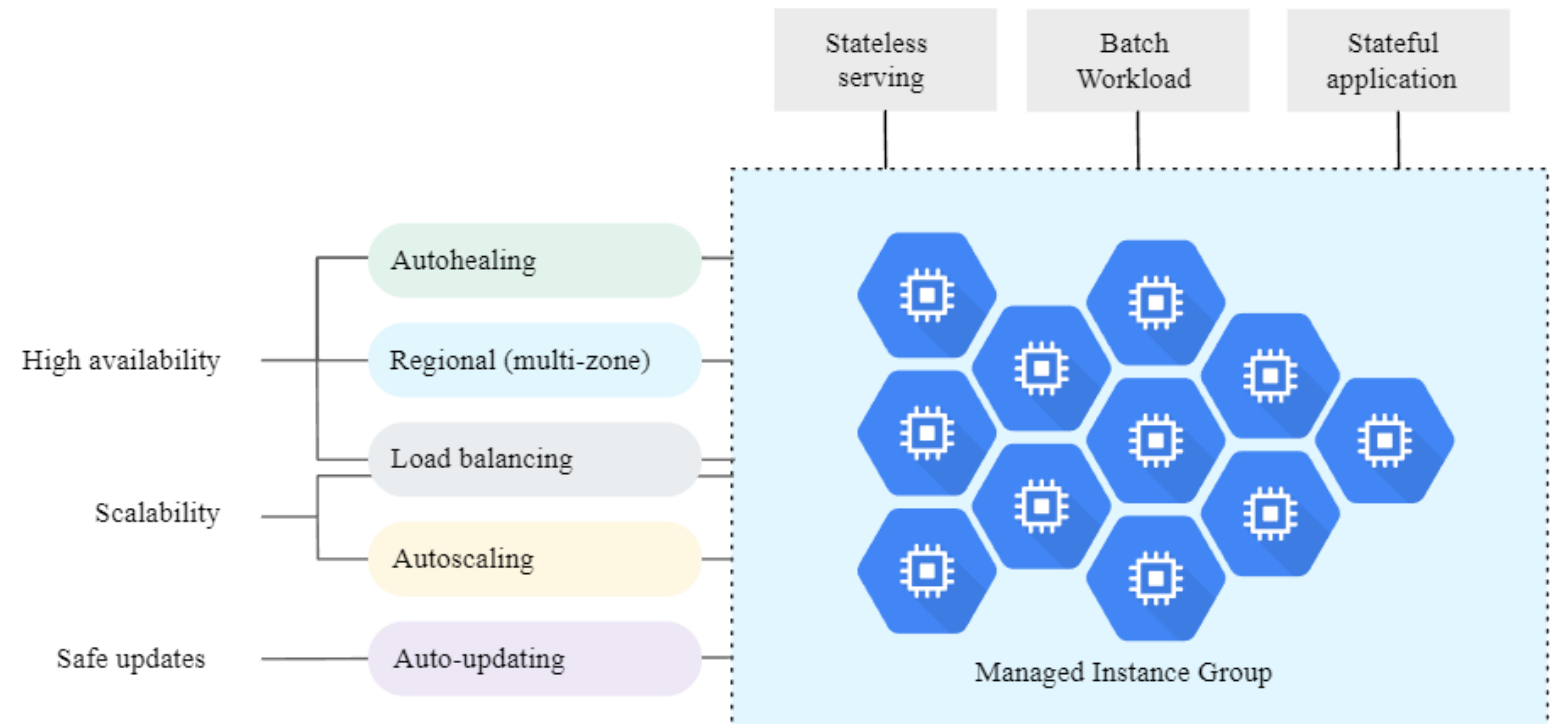
Compute Engine

Instance groups:

An instance group is a collection of virtual machine (VM) instances that you can manage as a single entity.

Compute Engine offers two kinds of VM instance groups, managed and unmanaged:

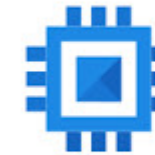
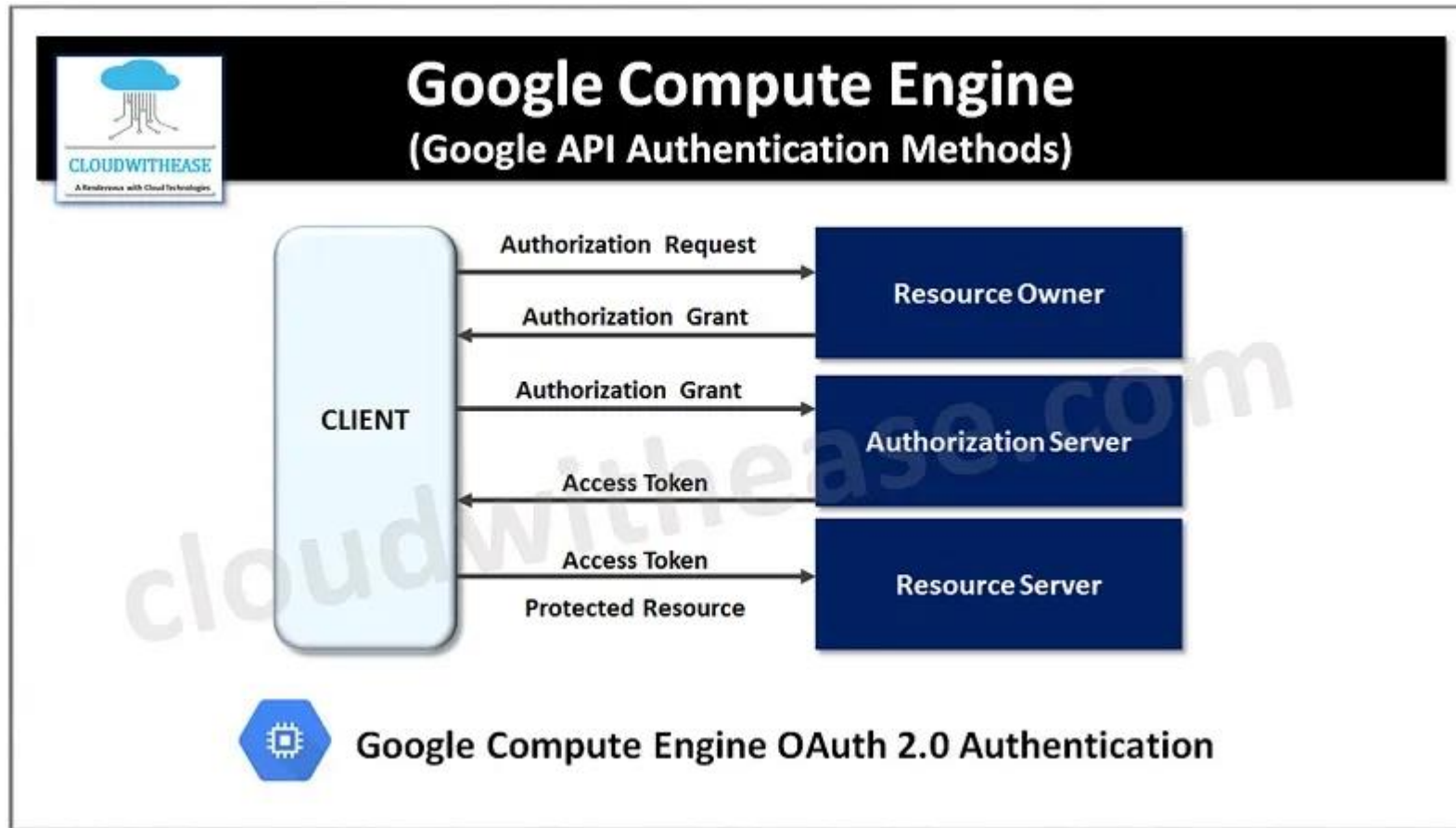
- Managed instance groups (MIGs) let you operate apps on multiple identical VMs. You can make your workloads scalable and highly available by taking advantage of automated MIG services, including: autoscaling, autohealing, regional (multiple zone) deployment, and automatic updating.
- Unmanaged instance groups let you load balance across a fleet of VMs that you manage yourself.



Compute Engine

❑ Compute Engine API:

Creates and runs virtual machines on Google Cloud Platform.



Compute Engine API

Google Enterprise API

Compute Engine API

ENABLE

TRY THIS API

Compute Engine

❑ Create and start a VM instance:

While creating your VM, you can create one or more disks for it. You can also add more disks to the VM after it's created. Compute Engine automatically starts the VM instance after you create it.

Before you begin

- If you want to use the command-line examples in this guide, do the following:
 1. Install or update to the latest version of the [Google Cloud CLI](#).
 2. [Set a default region and zone](#).
- If you want to use the API examples in this guide, [set up API access](#).
- When creating VMs from images or disks by using the Google Cloud CLI or the Compute Engine API, there's a limit of 20 VM instances per second. If you need to create a higher number of VMs per second, [request a higher quota limit](#) for the **Images** resource.

<https://cloud.google.com/compute/docs/instances/create-start-instance>

Compute Engine – Method 1

❑ Create a VM with a custom machine type:

Compute Engine offers predefined machine types that you can use when you create a VM instance. A predefined machine type has a preset number of vCPUs and amount of memory, and is charged at a set price.

Pricing

Google charges for custom VMs based on the number of vCPUs and memory hours that the VM uses. This is different from how predefined machine types are charged. See [VM pricing](#) for details.

Custom VMs are subject to the same 1-minute minimum charge as any other instance, but sustained use discounts for custom machine types are calculated differently. For more information, see [sustained use discounts](#) for custom VMs.

Expressing memory in GB or MB

For Google Cloud tools and documentation, machine type memory is calculated in gigabytes (GB), where 1 GB is 2^{30} bytes. This unit of measurement is also known as a [gibibyte \(GiB\)](#). When converting memory from GB to MB, 1 GB = 1024 MB.

In the API, you must always provide memory in megabytes. If you use the Google Cloud CLI, you can provide the total memory for a VM in gigabytes or megabytes. However, the gcloud CLI expects the memory value to be an integer, so you cannot provide a float value. For example, to express 5.75 GB, convert 5.75 GB into MB instead. In this case, 5.75 GB is 5888 MB.

Compute Engine – Method 1

❑ Create a VM with a custom machine type:

Console

gcloud

API

Go

Java

Node.js

Python

1. In the Google Cloud console, go to the **Create an instance** page.

Go to Create an instance

2. In the **Zone** list, select the zone where you want to host this VM. The **Series** list is filtered to include only the machine type families available in the selected zone.
3. In the **Machine configuration** section, select **General-purpose**.
 - a. In the **Series** list, click **N1** in **First Generation** for N1 custom machine types or **E2**, **N2**, or **N2D** for **Second Generation** custom machine types.
 - b. From the **Machine type** list, select **Custom**.
 - c. To specify the number of vCPUs and the amount of memory for the VM instance, drag the sliders or enter the values in the text boxes. The console displays an estimated cost for the VM as you change the number of vCPUs and memory.
4. Continue to create the VM.

Compute Engine – Method 1

❑ **Create a VM with a custom machine type:** [Add extended memory to a machine type](#)

Depending on the machine, each machine type has a specific amount of memory by default. For example, if you create a custom N1 VM, it can have up to 6.5 GB of memory per vCPU. For custom N2 VMs, this number increases up to 8 GB of memory per vCPU.

With extended memory, you can add memory to a machine type with no limitations per vCPU. You can add extended memory up to certain limits based on the machine type:

- For N1 machine types, you can add up to a total of 624 GB of memory per VM.
- For N2 machine types, you can add up to a total of 640 GB of memory per VM.
- For N2D machine types, you can add up to a total of 768 GB of memory per VM.
- Extended memory is not available for E2 machine types.

Compute Engine – Method 1

❑ Add extended memory during instance creation:

[Console](#) [gcloud](#) [API](#) [Go](#) [Java](#) [Node.js](#) [Python](#)

1. In the Google Cloud console, go to the **Create an instance** page.

[Go to Create an instance](#)
2. In the **Zone** list, select the zone where you want to host this VM. The **Series** list is filtered to include only the machine type families available in the selected zone.
3. In the **Machine configuration** section, select **General-purpose**.
 - a. In the **Series** list, click **N1** in **First Generation** for N1 custom machine types or **N2** or **N2D** for **Second Generation** custom machine types.
 - b. In the **Machine type** list, select **Custom**.
 - c. To specify the number of vCPUs drag the **Cores** slider or enter the value in the text box.
 - d. To add extended memory, select **Extend memory**, and do the following:
 - i. To specify the amount of memory for the VM instance, drag the **Memory** slider or enter the value in the text box.

The console displays an estimated cost for the VM as you change the number of vCPUs and memory.
4. Continue to create the VM.

Compute Engine – Method 1

❑ Add extended memory to an existing VM instance:

Console

gcloud

API

Go

Java

Node.js

Python

1. In the Google Cloud console, go to the **VM instances** page.

Go to VM instances

2. Choose the stopped VM you want to modify from the VM list.

3. Click **Edit** at the top of the page.

4. In **Machine configuration**, select **General-purpose**.

5. From the **Machine type** list, select **Custom**.

6. Select the number of vCPUs you want.

7. To add extended memory, select **Extend memory** and then specify the amount of memory you want.

8. Save your changes.

To add more memory to an existing instance, you must first stop the instance.

After the instance stops, follow the instructions below to add more memory to the VM.

Compute Engine – Method 2

❑ Create a VM instance from an image:

This section explains how to create a VM from a public OS image or a custom image. A VM contains a bootloader, a boot file system, and an OS image.

View a list of public images available on Compute Engine

Before you create a VM by using a public image, review the list of public images that are available on Compute Engine.

For more information about the features available with each public image, see [Feature support by operating system](#).

[Console](#) [gcloud](#) [API](#) [C#](#) [Go](#) [Java](#) [Node.js](#) [PHP](#) [Python](#) [Ruby](#)

1. In the Google Cloud console, go to the **Images** page.

Go to Images

Compute Engine – Method 2

❑ Create a VM instance from an image: [Create a VM from a custom image](#)

A custom image belongs only to your project.

To create a VM with a custom image, you must first create a custom image if you don't already have one.

[Console](#)

[gcloud](#)

[API](#)

[Go](#)

[Java](#)

[Node.js](#)

[Python](#)

1. In the Google Cloud console, go to the **VM instances** page.

[Go to VM instances](#)

2. Select your project and click **Continue**.

3. Click **Create instance**.

4. Specify a **Name** for your VM. For more information, see [Resource naming convention](#).

5. Optional: Change the **Zone** for this VM. Compute Engine randomizes the list of zones within each region to encourage use across multiple zones.

6. Select a **Machine configuration** for your VM.

7. In the **Boot disk** section, click **Change**, and then do the following:

- Select the **Custom Images** tab.
- To select the image project, click **Select a project**, and then do the following:
 - Select the project that contains the image.
 - Click **Open**.
- In the **Image** list, click the image that you want to import.
- Select the type and size of your boot disk.
- Optional: For advanced configuration options, click **Show advanced configuration**.
- To confirm your boot disk options, click **Select**.

8. In the **Firewall** section, to permit HTTP or HTTPS traffic to the VM, select **Allow HTTP traffic** or **Allow HTTPS traffic**.

The Google Cloud console adds a network tag to your VM and creates the corresponding ingress firewall rule that allows all incoming traffic on `tcp:80` (HTTP) or `tcp:443` (HTTPS). The network tag associates the firewall rule with the VM. For more information, see [Firewall rules overview](#) in the Virtual Private Cloud documentation.

9. To create and start the VM, click **Create**.

Compute Engine – Method 2

❑ Create custom images:

You can create custom images from source disks, images, snapshots, or images stored in Cloud Storage and use these images to create virtual machine (VM) instances. Custom images are ideal for situations where you have created and modified a persistent boot disk or specific image to a certain state and need to save that state for creating VMs.

Create the image

You can create disk images from the following sources:

- A persistent disk, even while that disk is attached to a VM
- A snapshot of a persistent disk
- Another image in your project
- An image that is shared from another project
- A [compressed RAW image](#) in Cloud Storage

Compute Engine – Method 2

❑ Create an image:

1. In the Google Cloud console, go to the **Create an image** page.

[Go to Create an image](#)

2. Specify the **Name** of your image.
3. Specify the **Source** from which you want to create an image. This can be a persistent disk, a snapshot, another image, or a disk.raw file in Cloud Storage.
4. If you are creating an image from a disk attached to a running VM, check **Keep instance running** to confirm that you want to create the image while the VM is running. You can [prepare your VM](#) before creating the image.
5. In the **Based on source disk location (default)** drop-down list, specify the location to store the image. For example, specify `us` to store the image in the `us` multi-region, or `us-central1` to store it in the `us-central1` region. If you don't make a selection, Compute Engine stores the image in the multi-region closest to your image's source location.
6. Optional: specify the properties for your image.
 - **Family**: the [image family](#) this new image belongs to.
 - **Description**: a description for your custom image.
 - **Label**: a [label](#) to group together resources.
7. Specify the encryption key. You can choose between a Google-managed key, a [Cloud Key Management Service \(Cloud KMS\)](#) key or a [customer-supplied encryption \(CSEK\)](#) key. If no encryption key is specified, images are encrypted using a Google-managed key.
8. Click **Create** to create the image.

Compute Engine – Method 3

❑ Create a VM similar to an existing VM:

If you have a VM with a specific configuration that you would like to reuse, you can use the Google Cloud console to create a VM that is similar to an existing VM.

You can only use the Google Cloud console to duplicate a VM's configuration.

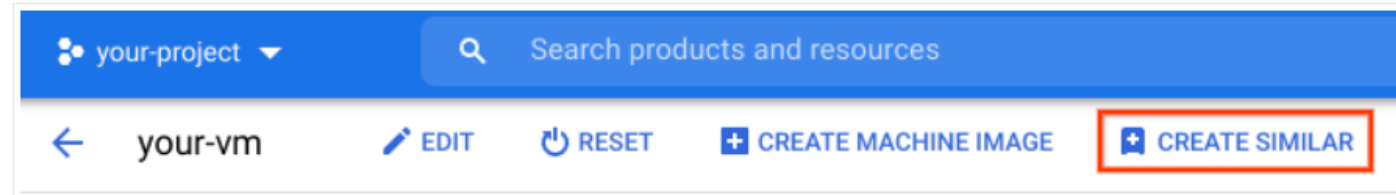
You cannot duplicate a VM's configuration by using the gcloud CLI or the Compute Engine API.

Create a VM that is similar to an existing VM

1. Go to the **VM instances** page.

[Go to the VM instances page](#)

2. Click the name of the VM. The **VM instance details** page opens.
3. In the toolbar at the top of the **VM instance details** page, click **Create similar**.



The Google Cloud console copies the configuration and opens the **Create an instance** page.

4. On the **Create an instance** page, customize the new VM.
5. Click **Create**.

Compute Engine – Method 4

❑ Create a VM with attached GPUs:

Compute Engine provides graphics processing units (GPUs) that you can add to your virtual machines (VMs). You can use these GPUs to accelerate specific workloads on your VMs such as machine learning and data processing.

This page explains how to create a VM with attached GPUs. If you want to add GPUs to existing VMs, see Add or remove GPUs (<https://cloud.google.com/compute/docs/gpus/add-remove-gpus>).

Before you begin

- If you want to use the command-line examples in this guide, do the following:
 1. Install or update to the latest version of the [Google Cloud CLI](#).
 2. [Set a default region and zone](#).
- If you want to use the API examples in this guide, [set up API access](#).
- Read about [GPU pricing on Compute Engine](#) to understand the cost to use GPUs on your VMs.
- Read about [restrictions for VMs with GPUs](#).
- Check your [GPU quota](#).
- Choose an operating system image:


Compute Engine – Method 4

❏ Create a VM with attached GPUs:

Checking GPU quota

To protect Compute Engine systems and users, new projects have a global GPU quota, which limits the total number of GPUs you can create in any supported zone.

Use the `regions describe` command to ensure that you have sufficient GPU quota in the region where you want to create VMs with GPUs.

```
gcloud compute regions describe REGION 
```

Replace *REGION* with the `region` that you want to check for GPU quota.

Compute Engine – Method 4

❑ Create VMs with attached GPUs:

If you have a VM with a specific configuration that you would like to reuse, you can use the Google Cloud console to create a VM that is similar to an existing VM.

To create a VM with attached GPUs, complete the following steps:

1. Create the VM. The method used to create a VM depends on the GPU model.
 - To create a VM with A100 GPUs, see [Create a VM with attached GPUs \(A100 GPUs\)](#).
 - To create a VM with any other available model, see [Create a VM with attached GPUs \(other GPU types\)](#).
2. For the VM to use the GPU, you need to [install the GPU driver on your VM](#).
3. If you enabled an NVIDIA RTX virtual workstation(formerly known as NVIDIA GRID), [install a driver for virtual workstation](#).

Compute Engine

❑ Connect to VMs by SSH:

Compute Engine uses key-based SSH authentication to establish connections to all Linux virtual machine (VM) instances. You can optionally enable SSH for Windows VMs. By default, passwords aren't configured for local users on Linux VMs.

Connect to Linux VMs using Google tools: To connect to Linux instances through the Google Cloud console or the Google Cloud CLI, complete the steps in one of the following tabs

Console

gcloud

1. In the Google Cloud console, go to the **VM instances** page.

Go to VM instances

2. In the list of virtual machine instances, click **SSH** in the row of the instance that you want to connect to.

<input type="checkbox"/>	Name ^	Zone	Recommendation	Internal IP	External IP	Connect
<input type="checkbox"/>	instance-1	us-east1-b		10.142.0.2 (nic0)	35.231.114.114 ↗	<div>SSH</div>

Compute Engine

❏ **Manage VMs:** [VM instance lifecycle](#)

A VM instance can transition through many states as part of its lifecycle. When you create a VM, Compute Engine provisions resources to start the VM. Next, the VM moves into staging, where it prepares for first boot. During and after start-up, a VM is considered running. During its lifetime, a running VM can be repeatedly stopped and restarted or suspended and resumed.

A VM can be in one of the following states:

- PROVISIONING
- STAGING
- RUNNING
- STOPPING
- REPAIRING
- TERMINATED
- SUSPENDING
- SUSPENDED

Compute Engine - Report

TASK I: Choosing the right virtual machine type

Phân tích cách sử dụng các loại hình virtual machine workloads on CLOUD:

- Scale out workloads (T2A, T2D)
- General purpose workloads (E2, N2, N2D, N1)
- Ultra-high memory (M2, M1)
- Compute-intensive workloads (C3, C2, C2D)
- Most demanding applications and workloads (A2)

→ bằng Tiếng Việt !!!

Compute Engine - Report

TASK II: GPU platforms

<https://cloud.google.com/compute/docs/gpus>

<https://cloud.google.com/compute/docs/gpus/gpu-regions-zones>

Nếu hệ thống “compute engine with GPU” yêu cầu cấu hình:

☺ 3 GPUs, 100GB memory, 50 vCPUs, 200 GB capacity

Để đảm bảo triển khai hệ thống, ta phải sử dụng loại “machine type” tối thiểu nào với:

NVIDIA GPUs A100-40GB

NVIDIA GPUs A100-80GB

NVIDIA V100 GPUs

☺ Liệt kê tất cả các “zone” ở US có thể cung cấp 2 loại GPU platforms: A100 80GB và A100 40GB

Compute Engine - Report

TASK III: Compute Engine API and Pricing Estimation

<https://cloud.google.com/compute/all-pricing>

☺ Enable Compute Engine API

Trình tự các bước thực hiện, ảnh chụp minh chứng

☺ Ước tính chi phí (*monthly*) để “create compute engine” với các cấu hình:

- E2 machine, on-demand price: 3 vCPUs + 20 GB memory at Singapore (asia-southeast1)
- E2 high-memory machine: 4 vCPUs + 32 GB memory at Toronto (northamerica-northeast2)
- N2D machine, on-demand price: 5 vCPUs + 24 GB memory at Paris (euro-west9)
- N1 machine, on-demand price: 2 vCPUs + 64 GB memory at Hong Kong (asia-east2)
- M1 memory-optimized machine: 40 vCPUs + 961 GB memory at Tokyo (asia-northeast1)

TASK VI: Create a VM with a custom machine type

Tạo virtual machine bằng tùy chỉnh machine type

- 😊 Đặt tên VM của mỗi nhóm: VM-dh20hm-group#, (# là số thứ tự nhóm)
- 😊 Region: thuộc Hong Kong, zone: thuộc asia-east2-b
- 😊 Machine config: E2 (e2-small, 2 vCPU, 2GB memory)
- 😊 Boot disk: public image – Debian GNU/Linux 11 – Standard persistent disk – size: 20GB

Trình tự các bước thực hiện, ảnh chụp minh chứng

Compute Engine - Report

TASK V: Compute Engine Application – Create Website Wordpress

☺ Mỗi nhóm khởi tạo loại google engine:

“Google Click to Deploy” dành cho ứng dụng “wordpress”

☺ Đặt tên cho ứng dụng VM: wordpress-dh20hm-group#, (# là số thự tự nhóm)

☺ Region: thuộc Hong Kong, zone: thuộc asia-east2-b

☺ Machine config: E2 (e2-small, 2 vCPU, 2GB memory)

☺ Boot disk: default

☺ Hiện thị: “Deployment Manager” của ứng dụng VM Wordpress vừa tạo

☺ Active và show website đã khởi tạo từ ứng dụng Compute Engine

“Hello World, this is group # in the class of DH20HM, Nong Lam University”