GCP Compute Service

Cloud Compute Engine

Google Compute Engine is Google's IAAS for running Virtual Machine on Google Cloud Platform (GCP)

Google Cloud Platform

GCE: Features

GCE

- ✓ Scalable, High-Performance Virtual Machines
- ✓ GCE VM's come with persistent disk storage, and deliver consistent performance
- ✓ GCE VMs also supports Local SSD and RAM disk
- ✓ GCE VM's are backed by Google's private global fiber network.
- ✓ Flexible for different work loads.
- ✓ Linux & Windows Support
- ✓ Predefined and Custom Machine Types
- ✓ Supports Global Load Balancing
- ✓ Supports Auto Scaling

GCE: Features

GCE

- ✓ Sub Min Billing
- ✓ Charges : Min level Increment beyond 1 min.
- ✓ Global Scoped Images
- ✓ Supports CDN
- ✓ Disk Performance – Local SSD
- ✓ High Network Performance
- ✓ **Live Migration – Transparent Maintenance**
- ✓ **Auto Restart**
- ✓ Machine Right Sizing
- ✓ Automatic Discount

  Committed Saving, Sustained Use, Inferred Instance, Custom Machine Discount, Preemptible VM

GCE: Resources

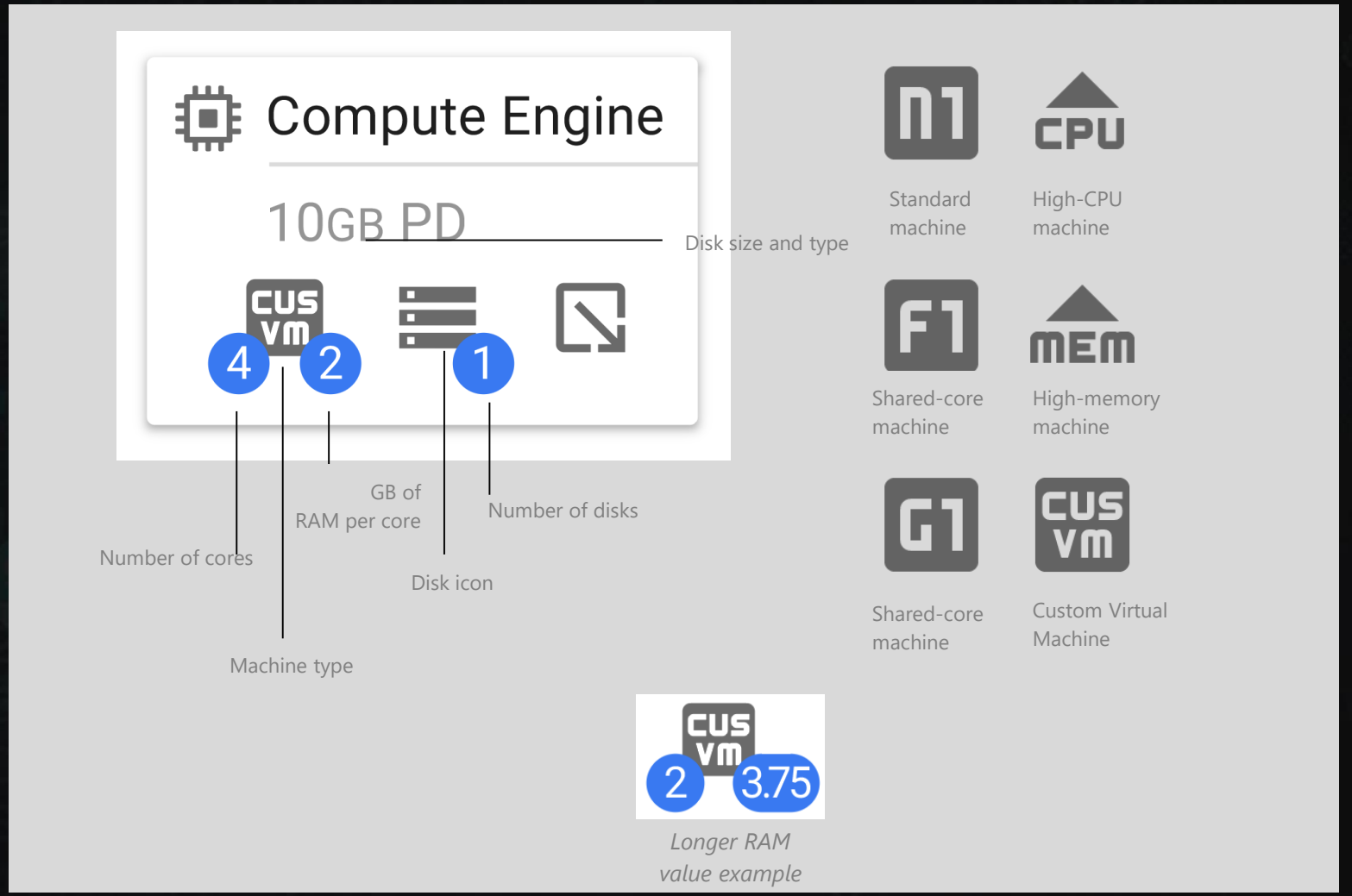
GCE

- ✓ Disks
- ✓ Images
- ✓ Networks
- ✓ Firewalls
- ✓ Address
- ✓ Regions and Zones
- ✓ Instances
- ✓ Load Balancer , Auto Scaling
- ✓ Instance Templates
- ✓ Instance Groups

Google Cloud Platform

GCE: Representation

GCE



Compute Engine Performance

- Processor Family - 2ghz, 2.2 ghz, 2.5ghz …

Haswell, Broadwell, Ivy Bridge, Sandy Bridge processors

- What do we mean by vCPU
  - 2vCPU - -> 1 CPU
- Network throughput
  - Per vCPU -> 2Gbps
  - Max 16gbps for 8 cores
- Disk IO
  - Based on Type and size of disk.

Accessing VM

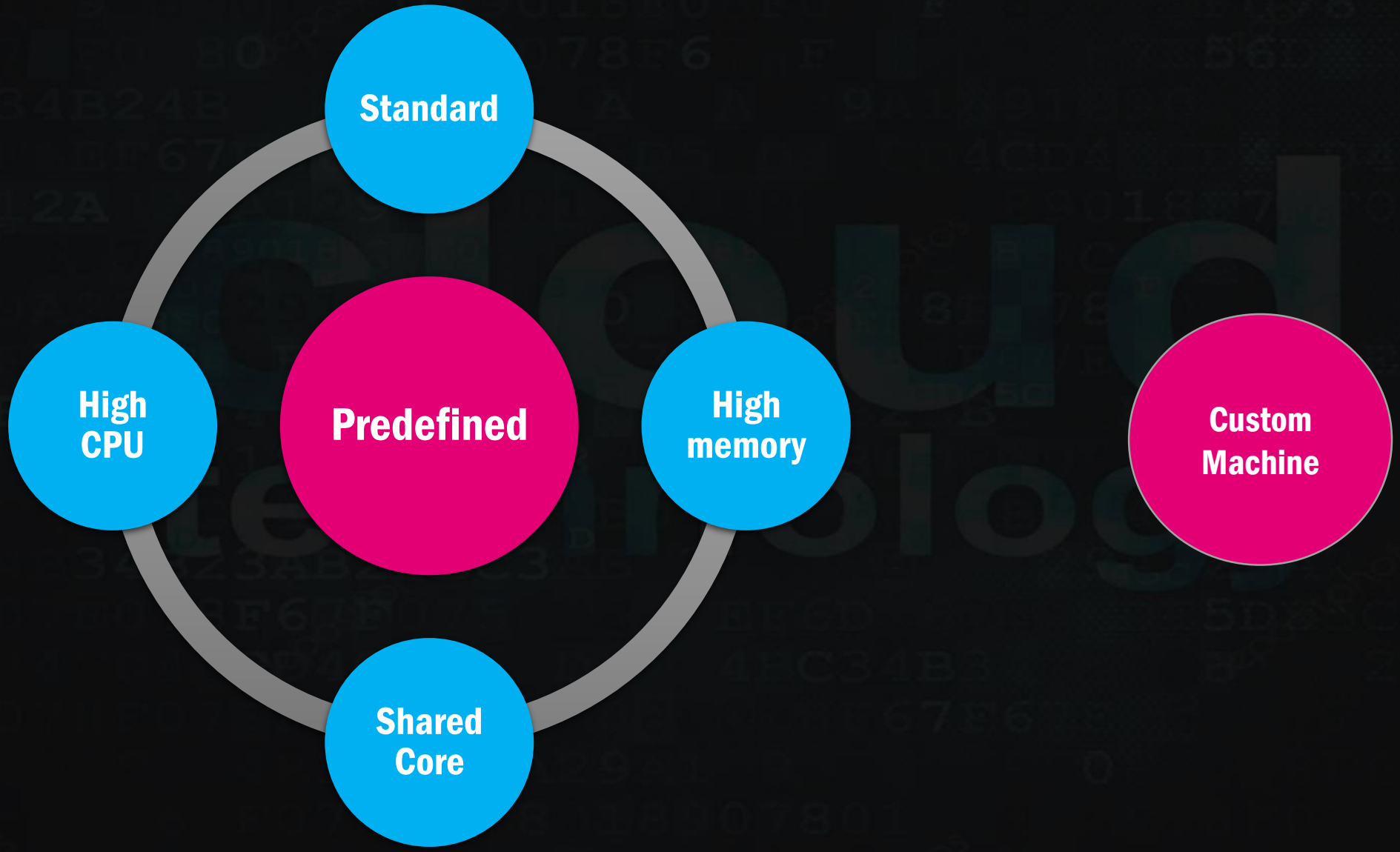
- Windows
  - rdp
  - Rdp clients
  - Powershell
  - Needs to set password
  - Firewall rule should allow : tcp:3389

- Linux
  - SSH from Console - key is not required.
  - SSH from CloudShell
  - SSH via Cloud SDK
  - Desktop, Third party client – Based on key
  - Need firewall opened to allow tcp:22

Compute Engine - Machine Types

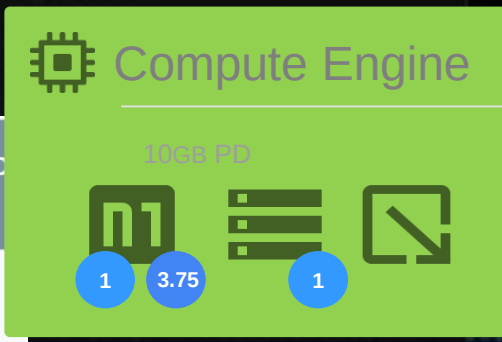
Google Cloud Platform

GCE : Machine Type



Google Cloud Platform

GCE : Predefined - Standard

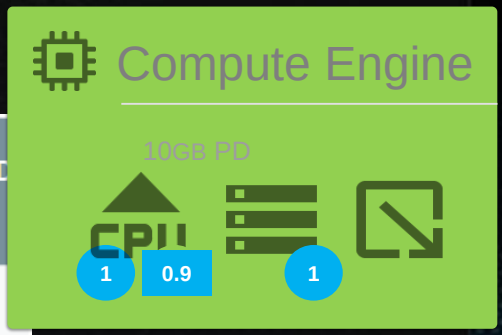


Machine name	Description	Virtual CPUs[1]	Memory (GB)	Max number of persistent disks (PDs)[2]	Max total PD size (TB)
n1-standard-1	Standard machine type with 1 virtual CPU and 3.75 GB of memory.	1	3.75	16 (32 in Beta)	64
n1-standard-2	Standard machine type with 2 virtual CPUs and 7.5 GB of memory.	2	7.50	16 (64 in Beta)	64
n1-standard-4	Standard machine type with 4 virtual CPUs and 15 GB of memory.	4	15	16 (64 in Beta)	64
n1-standard-8	Standard machine type with 8 virtual CPUs and 30 GB of memory.	8	30	16 (128 in Beta)	64
n1-standard-16	Standard machine type with 16 virtual CPUs and 60 GB of memory.	16	60	16 (128 in Beta)	64
n1-standard-32	Standard machine type with 32 virtual CPUs and 120 GB of memory.	32	120	16 (128 in Beta)	64
n1-standard-64	Standard machine type with 64 virtual CPUs and 240 GB of memory.	64	240	16 (128 in Beta)	64

Standard machine types are suitable for tasks that have a balance of CPU and memory needs. Standard machine types have 3.75 GB of RAM per virtual CPU.

Google Cloud Platform

GCE : Predefined - High CPU

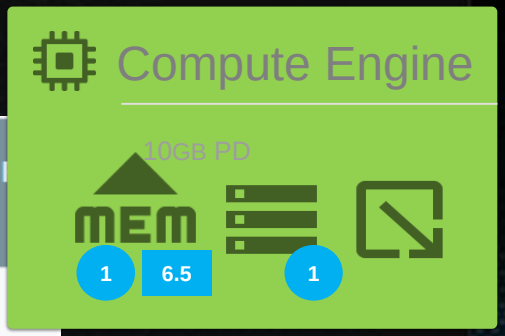


Machine name	Description	Virtual CPUs[1]	Memory (GB)	Max number of persistent disks (PDs)[2]	Max total PD size (TB)
n1-highcpu-2	High-CPU machine type with 2 virtual CPUs and 1.80 GB of memory.	2	1.80	16 (64 in Beta)	64
n1-highcpu-4	High-CPU machine type with 4 virtual CPUs and 3.60 GB of memory.	4	3.60	16 (64 in Beta)	64
n1-highcpu-8	High-CPU machine type with 8 virtual CPUs and 7.20 GB of memory.	8	7.20	16 (128 in Beta)	64
n1-highcpu-16	High-CPU machine type with 16 virtual CPUs and 14.4 GB of memory.	16	14.4	16 (128 in Beta)	64
n1-highcpu-32	High-CPU machine type with 32 virtual CPUs and 28.8 GB of memory.	32	28.8	16 (128 in Beta)	64
n1-highcpu-64	High-CPU machine type with 64 virtual CPUs and 57.6 GB of memory.	64	57.6	16 (128 in Beta)	64

High-CPU machine types are ideal for tasks that require more virtual CPUs relative to memory. High-CPU machine types have 0.90 GB of RAM per virtual CPU

Google Cloud Platform

GCE : Predefined – High Memory



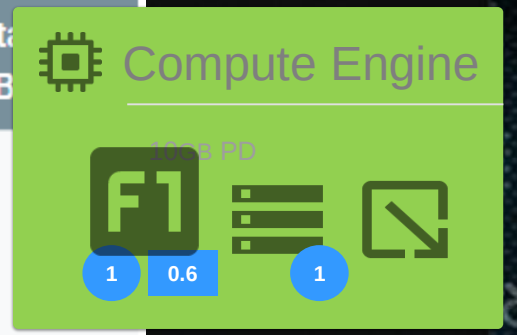
Machine Name	Description	Virtual CPUs[1]	Memory (GB)	Max number of persistent disks (PDs)[2]	Max total size (TB)
n1-highmem-2	High memory machine type with 2 virtual CPUs and 13 GB of memory.	2	13	16 (64 in Beta)	64
n1-highmem-4	High memory machine type with 4 virtual CPUs, and 26 GB of memory.	4	26	16 (64 in Beta)	64
n1-highmem-8	High memory machine type with 8 virtual CPUs and 52 GB of memory.	8	52	16 (128 in Beta)	64
n1-highmem-16	High memory machine type with 16 virtual CPUs and 104 GB of memory.	16	104	16 (128 in Beta)	64
n1-highmem-32	High memory machine type with 32 virtual CPUs and 208 GB of memory.	32	208	16 (128 in Beta)	64
n1-highmem-64	High memory machine type with 64 virtual CPUs and 416 GB of memory.	64	416	16 (128 in Beta)	64

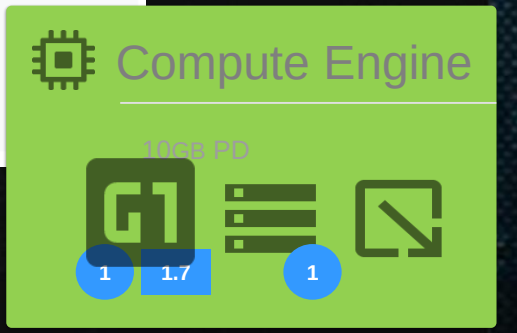
High-memory machine types are ideal for tasks that require more memory relative to virtual CPUs. High-memory machine types have 6.50GB of RAM per virtual CPU

Google Cloud Platform

GCE : Predefined – Shared Core

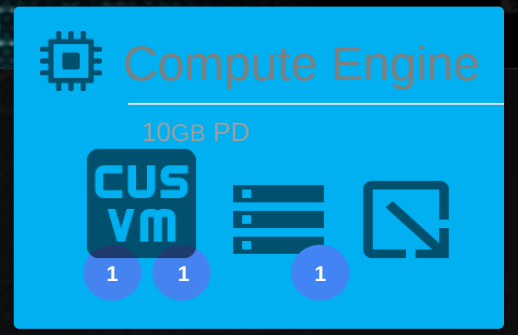
Machine name	Description	Virtual CPUs	Memory (GB)	Max number of persistent disks (PDs)[1]	Max tota size (TB
`f1-micro`	Micro machine type with 0.2 virtual CPU, 0.60 GB of memory, backed by a shared physical core.	0.2	0.60	4 (16 in Beta)	3
`g1-small`	Shared-core machine type with 0.5 virtual CPU, 1.70 GB of memory, backed by a shared physical core.	0.5	1.70	4 (16 in Beta)	3





Allows – F1 Micro bursting
Persistent disk usage is charged separately from machine type pricing.

GCE : Custom Machine Types



Google Cloud Platform

Custom machine types are ideal for the following scenarios

- Workloads that are **not a good fit** for the predefined machine types that are available to you.
- Workloads that require more processing power or more memory, but don't need all of the upgrades that are provided by the next larger predefined machine type.

Custom Machine Types costs slightly more to use than an equivalent predefined machine type

GCE : Custom Machine Types : Limitations

- Only machine types with **1 vCPU or an even number of vCPUs** can be created.
- Memory must be between **0.9 GB per vCPU, up to 6.5 GB** per vCPU.
- The total memory of the instance must be a multiple of **256 MB.**
- The maximum number of vCPUs allowed depends on the zone.

Supporting Zones

Haswell and Broadwell processors	Support up to 64vCPUs
Ivy Bridge processors	Support up to 32vCPUs
Sandy Bridge processors	Support up to 16vCPUs

https://cloud.google.com/compute/docs/regions-zones/regions-zones#available

GCE : Custom Machine Types

Examples of invalid machine types

- **1 vCPU, 0.6 GB of total memory** — Invalid because the total memory is less than the minimum 0.9 GB.
- **1 vCPU, 0.9 GB of total memory** — Invalid because the total memory must be a multiple of 256 MB. For 1 vCPU, use a minimum of 1024 MB memory.

Examples of valid machine types

- **32 vCPUs, 29 GB of total memory** — Valid because the total number of vCPUs is even and the total memory is a multiple of 256 MB. The amount of memory per vCPU is 0.9 GB, which satisfies the minimum requirement.
- **1 vCPU, 1 GB of total memory** — Valid because it has one vCPU, which is the minimum value, and the total memory is a multiple of 256 MB. The amount of memory per vCPU is also between the acceptable range of 0.9 GB to 6.5 GB per vCPU.

GCE：GPUs and Machine types

Google Compute Engine provides graphics processing units (GPUs) that you can attach to your virtual machine instances.

However, instances with lower numbers of GPU dies are limited to a maximum number of vCPUs. In general, higher numbers of GPU dies allow you to create instances with higher numbers of vCPUs and system memory.

GPU model	GPUs	GPU boards	GPU memory	Available vCPUs	Available memory	Available zones
NVIDIA® Tesla® K80	1 GPU	1/2 board	12 GB GDDR5	1 - 8 vCPUs	1 - 52 GB	• us-west1-b • us-east1-c • us-east1-d • europe-west1-b • europe-west1-d • asia-east1-a
	2 GPUs	1 board	24 GB GDDR5	1 - 16 vCPUs	1 - 104 GB	
	4 GPUs	2 boards	48 GB GDDR5	1 - 32 vCPUs	1 - 208 GB	
	8 GPUs	4 boards	96 GB GDDR5	1 - 64 vCPUs	1 - 416 GB	

GCE : GPUs and Machine types

Restrictions

- ✓ You must have **GPU quota** before you can create instances with GPUs..
- ✓ Instances with **one or more GPUs** have a maximum number of vCPUs for each GPU that you attach to the instance.
- ✓ You **cannot** attach GPUs to instances with **shared-core machine** types.
- ~~✓ You cannot attach GPUs to **preemptible** instances.~~
- ✓ GPU instances cannot **live migrate**.
- ✓ GPUs require device **drivers** in order to function properly. You can use any driver that you like, but you must ensure that these drivers are installed and configured properly.

Compute Engine – Disks, VM Lifecycle

Google Cloud Platform

GCE : Compute Engine and Persistent Disks

Persistent Disk

- ✓ Network Storage & Attached VM through network Interface
- ✓ Persistent and independent of compute(instance)
- ✓ Zonal
- ✓ Used as Bootable, Snapshots
- ✓ Use SSD as well as magnetic
- ✓ Resize dynamically (even when instance is running)
- ✓ Attached to multiple VM for read only data
- ✓ Automatic Encryption – You can choose your own key
- ✓ Lower performance with corresponding Local SSD/Ram disk

GCE : Compute Engine and Local Disks

Local Disks

- ✓ Local Disk can be attached to VM
- ✓ Ephemeral in nature –
  - ✓ Data stays on Restart but not on Instance stopped /terminate
- ✓ Provided high IOPS based on size of disk
  - ✓ Upto 680K read and 360 write
- ✓ Predefined size (375G) and upto 8 -> max 3TB
- ✓ can not live migrate
- ✓ SCSI or NVMe Interface
- ✓ Not available for Shared Core

GCE : Compute Engine and RAM Disks

RAM Disk

- ✓ Faster then any disk option available
- ✓ Ephemeral - goes away on stop, restart, terminate
- ✓ Size of Ram is dependent on many factors.. While planning
- ✓ tmpfs
- ✓ Always choose persistence disk at storage – but you need to catch machine life cycle events to store latest data..

GCE : Boot disk & Persistent Disk

VM Comes with single root boot disk – Persistent Disk (not local Disk)

Images is loaded into root persistent disk at the time of boot process

Different types

- Bootable – for running VM
- Snapshots - incremental backup

Durable : Can survive if you delete VM instance – You have to choose that option

VM Live Migrate

- During Maintenance , VM is migrated to another hardware without stopping or restarting
- You can check metadata to check if VM is in live migration
- May see some impact if vm is very busy (working 90%+ CPU)

Automatic Restart

- VM starts automatically if crash or maintenance events
- Not – for manually stop

GCE : Compute Engine moving to another Zone

- ✓ Manual option using GCP Console
- ✓ Using gcloud Command

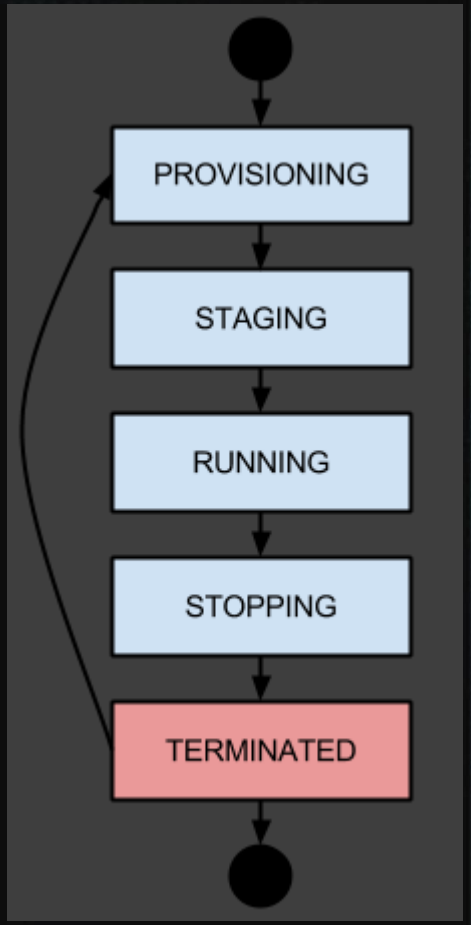
GCE : CPU Platforms

CPUs

- ✓ 2.6 GHz Intel Xeon E5 (Sandy Bridge)
- ✓ 2.5 GHz Intel Xeon E5 v2 (Ivy Bridge)
- ✓ 2.3 GHz Intel Xeon E5 v3 (Haswell)
- ✓ 2.2 GHz Intel Xeon E5 v4 (Broadwell)
- ✓ 2.0 GHz Intel Xeon (Skylake)

GCE : Instance Lifecycle

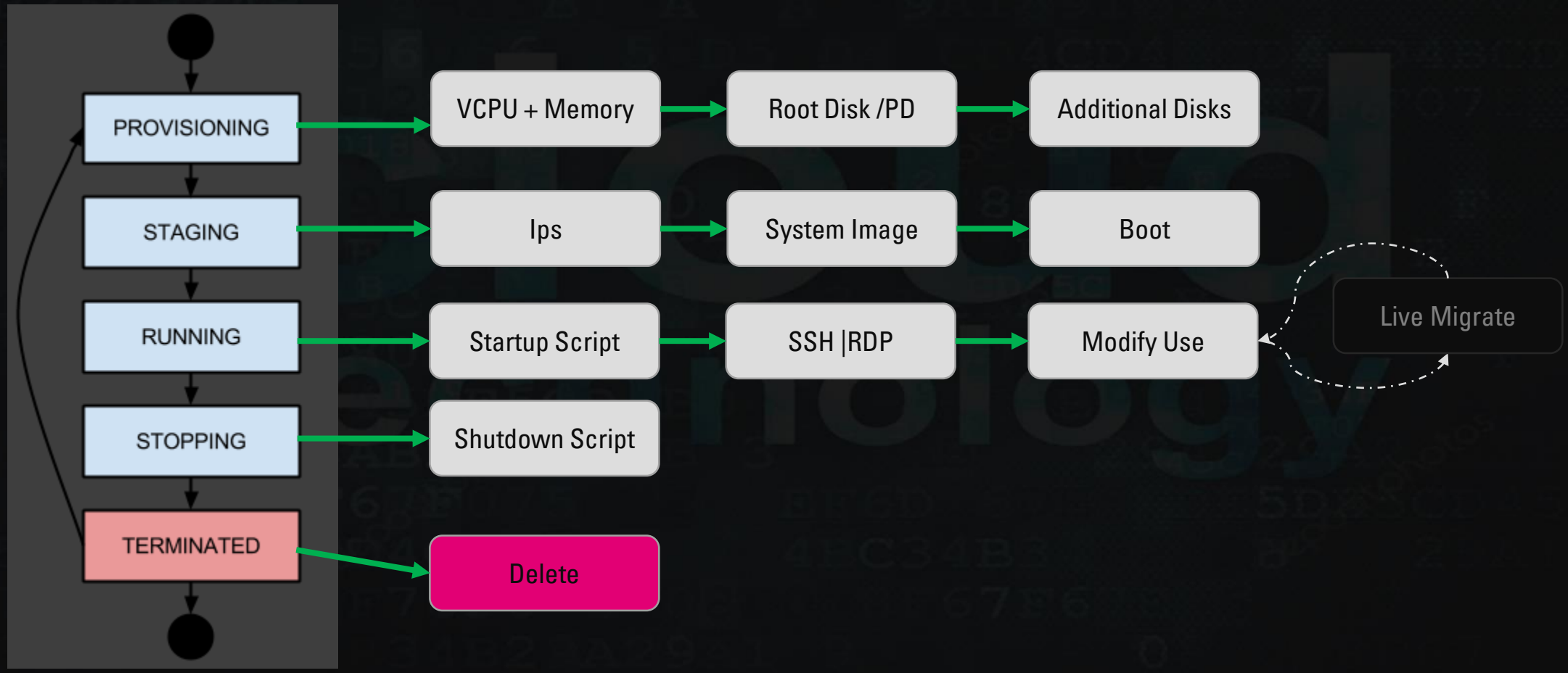
- **PROVISIONING** - Resources are being reserved for the instance. The instance isn't running yet.
- **STAGING** - Resources have been acquired and the instance is being prepared for launch.
- **RUNNING** - The instance is booting up or running. You should be able to ssh into the instance soon, though not immediately, after it enters this state.
- **STOPPING** - The instance is being stopped either due to a failure, or the instance being shut down. This is a temporary status and the instance will move to TERMINATED.
- **TERMINATED** - The instance was shut down or encountered a failure, either through the API or from inside the guest. You can choose to restart the instance or delete it.



$gcloud compute instances list

$gcloud compute instances describe example-instance

VM Lifecycle



Compute Engine Discount / pricing innovation

Google Cloud Platform

Pricing

- All machine types are charged a minimum of **1 minute**.

For example, if you run your virtual machine for 30 seconds, you will be billed for 1 minute of usage.

- After 1 minute, instances are charged in **1 second increments.**

- **Always Free tier**
  - 1 f1-micro VM instance per month (US regions, excluding Northern Virginia).
  - 30 GB of Standard persistent disk storage per month.
  - 5 GB of snapshot storage per month.
  - 1 GB egress from North America to other destinations per month (excluding Australia and China).
  - All usage aggregated

Pricing – Other Consideration

- Disk Pricing – Prorated to second
- Snapshots – only charged for amount used .
- Unused static IP are charged per hours (1 Cent currently)
- Custom Machine types charged for per vCPU, per RAM GB
- Predefined machines are billed based on instance you select
- Predefined machines are mostly cost less then equivalent custom Machin type
- Network egress based on type of egress used
- No ingress to vm or almost all GCP service

Pricing Innovation

- Preemptible VM - get upto 80% discount
- Sustained use discount
  - Inferred Instance discount - Predefined, Custom
- Committed use Discount
- Recommendation Engine – based on monthly usage

Preemptible VM

Get upto 80% discount

Things to Consider

- Can not live Migrate and auto Restart
- 24 hours max use and Not covered under SLA
- Charged if only started for 10 min , Less use will not be billed.
- When you attach GPU to preemptible – you quota will be used.
- Compute Engine sends signal for preemption to VM 30 sec

- Average preemption rate varies between 5% and 15% per seven days per project,

Sustained Use Discount - Predefined Machines

Usage Level (% of month)	% at which incremental is charged	Example incremental rate (USD/hour) for an n1-standard-1 instance
0%-25%	100% of base rate	$0.0475
25%-50%	80% of base rate	$0.0380
50%-75%	60% of base rate	$0.0285
75%-100%	40% of base rate	$0.0190

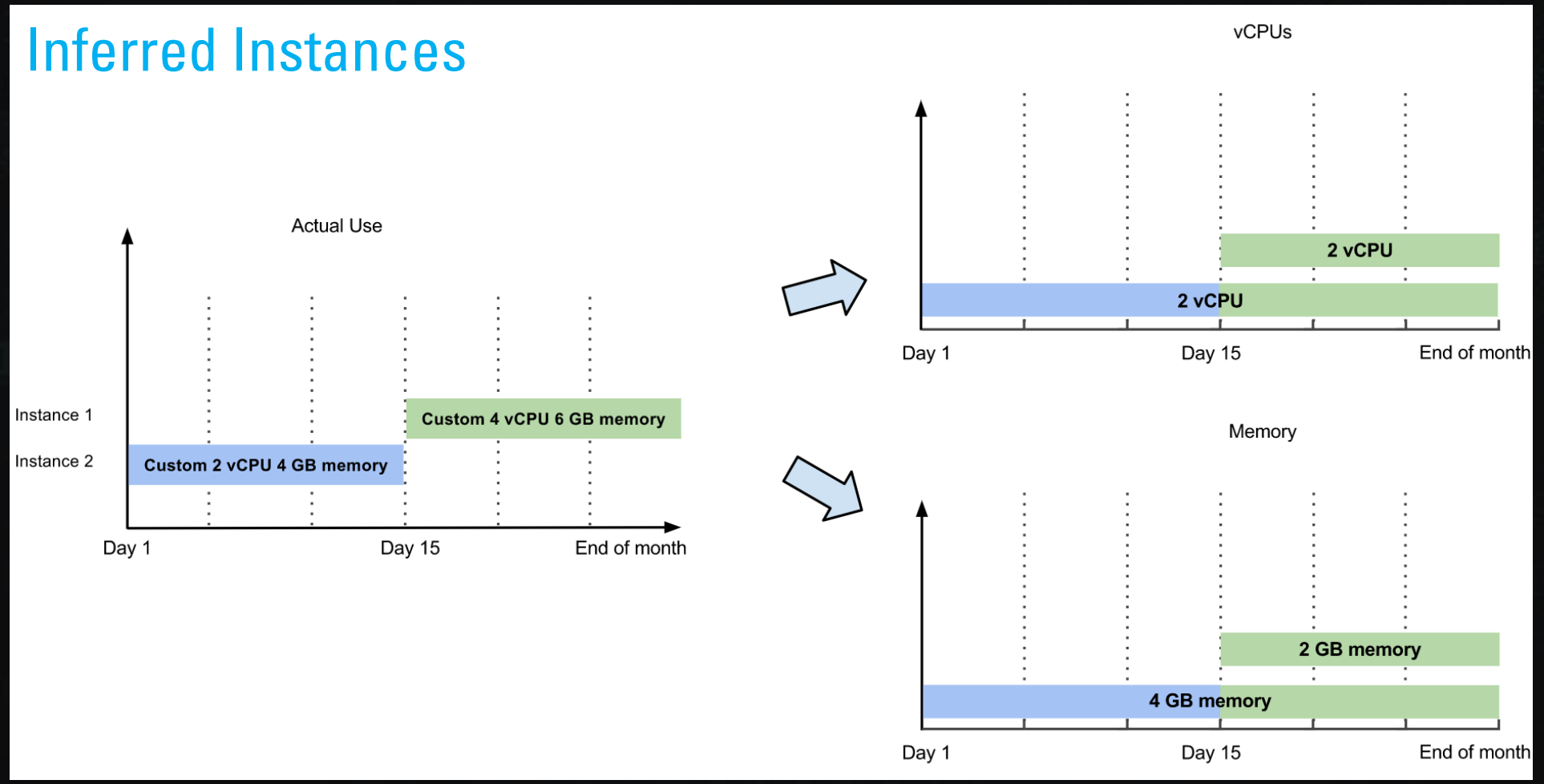
Inferred Instances

An inferred instance combines multiple, non-overlapping instances in the same zone into a single instance for billing.

With inferred usage, you are more likely to qualify for sustained use discounts.

Non-overlapping instances running the same predefined machine type in the same zone are combined to create one or more inferred instances.

Sustained Use Discount – Custom Machine



Committed use discount

- You can get up to 57% discount on normal prices on committed foe 1 year to 3 year.
- Upon purchase, you will be billed a monthly fee for the duration of the term you selected, whether or not you use the services.

- Committed use discounts work on all Compute Engine non-shared core machine types, including predefined and custom machine types. Committed use discounts are:
  - **Simple and flexible**: Discounts apply to the aggregate number of vCPUs or memory within a region so they are not affected by changes to your instance's machine type.
  - **No upfront costs**: There are no upfront costs for committed use discounts. Committed use discounts are applied to your bill every month.

Compute Engine Other Concepts

Google Cloud Platform

GCE : Instance & Images

- Google Compute Engine uses operating system images to create the root persistent disks for your instances

- Images contain a boot loader, an operating system, a root file system, Custom software

- You specify an image when you create an instance
- Stored in GCS Managed image service

Two types of images

Private Images – Maintained by you

Public Images – Maintained by google

Community supported images

Google Cloud Platform

GCE : Images

Public

- ✓ Compute Engine offers many preconfigured public images that have compatible Linux and Windows operating systems.
- ✓ To see the full list of public images with their image names, versions numbers, and image sizes, go to the Images page in the console.

Operating system	Support channel	Image family	Image project	Notes	Start an instance
CentOS	Compute Engine	centos-7 centos-6	centos-cloud		START
Container-Optimized OS from Google	Compute Engine	cos-stable cos-beta cos-dev	cos-cloud		START
CoreOS	CoreOS Support	coreos-stable coreos-beta coreos-alpha	coreos-cloud		START
Debian	Compute Engine	debian-9	debian-cloud		START

GCE : Images

Custom

A custom image is a boot disk image that you own and control access to Use custom images for the following tasks:

- **Import a boot disk image** to Compute Engine from your on-prem/local environment, or from your virtual machine instances running on another cloud
- **Create an image** from the root persistent disks of your existing Compute Engine instances. Then use that image to create new root persistent disks for your instances.
- **Copy one image to another image**. This feature is available in Beta using either the gcloud tool or the API. Use the same process that you use to create an image, but specify another image as the image source. You can also create an image from a custom image in a different project.

GCE : Instance Templates

Instance templates define the machine type, image, zone, and other instance properties for the instances in a managed instance group

Create an instance template once and can reuse it for multiple groups and configuration

Instance template is a global resource that is not bound to a zone or a region - However, you can still specify some zonal resources in an instance template, which restricts the template to the zone where that resource resides

GCE : Instance Groups

Instance groups are group of virtual machine (VM) instances so that you don't have to individually control each instance in your project

There are two types of Instance Groups

- **Managed Instance Groups**
  - Zonal Managed
  - Reginal Managed
- **Unmanaged Instance Groups**

Managed Instance Groups

- ✓ Use Identical instance resources
- ✓ Used for Auto-Scaling
- ✓ Rolling updates can be done
- ✓ Changes to instance will make change in all the instances
- ✓ Load balancing for only similar resources.

Un-Managed Instance Groups

- ✓ Use non- identical instance recourses.
- ✓ Can not be used for Auto-Scaling
- ✓ Rolling updates cant be used
- ✓ Can make arbitrary changes in any instance
- ✓ Can be used for load balancing pre-existing resources or Groups of dissimilar resources.

GCP Compute Service

GCP Compute Engine

Demo Next..

Google Cloud Platform

End