Virtual Machines

Architecting with GCP Fundamentals: Infrastructure

COMPUTE ENGINE, KUBERNETES ENGINE, APP ENGINE, CLOUD FUNCTIONS

QWIKLABS CREATING VIRTUAL MACHINES, WORKING WITH VIRTUAL MACHINES

Google Cloud

Last modified 2017-11-27

Agenda

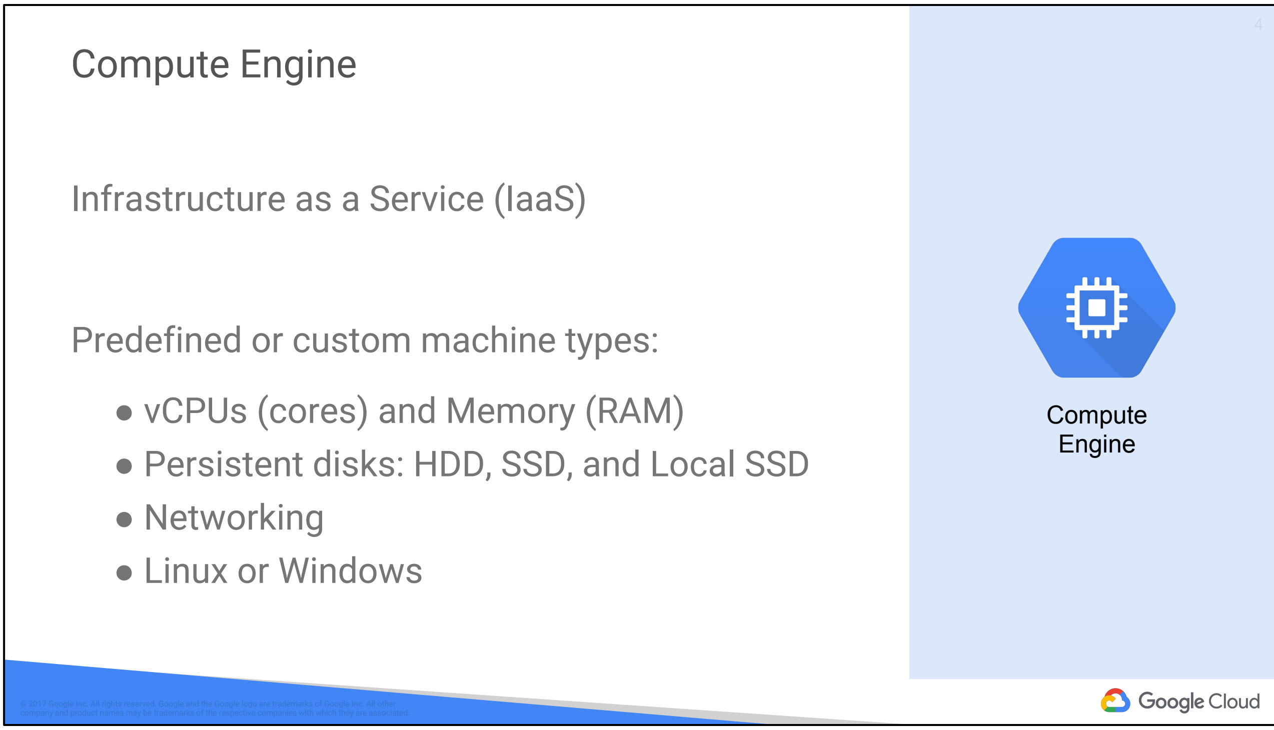
- **Compute Engine**
- Lab
- Compute options (vCPU and Memory)
- Images
- Disk options
- Common Compute Engine actions
- Lab
- Quiz

GCP compute and processing options

	Compute Engine	Kubernetes Engine	App Engine Standard	App Engine Flexible	Cloud Functions
Language support	Any	Any	Python Java PHP Go	Python Java Node.js Go Ruby PHP .NET Custom Runtimes	Node.js
Usage model	IaaS	IaaS PaaS	PaaS	PaaS	Microservices Architecture
Scaling	Server Autoscaling	Cluster	Autoscaling managed servers		Serverless
Primary use case	General Workloads	Container Workloads	Scalable web applications Mobile backend applications		Lightweight Event Actions

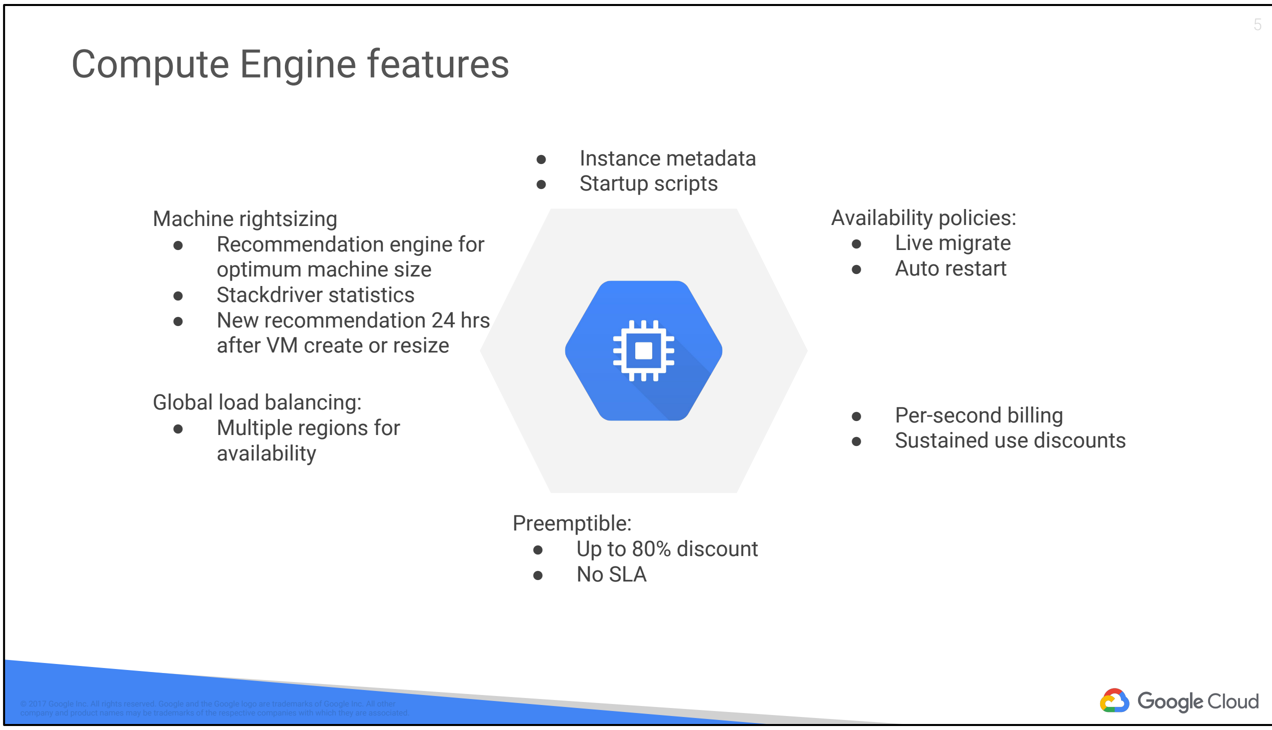
Google Cloud

SLA for Compute Engine: https://cloud.google.com/compute/sla



Compute Engine lets you create and run virtual machines on Google infrastructure. Compute Engine offers scale, performance, and value that allows you to easily launch large compute clusters on Google's infrastructure. There are no upfront investments, and you can run thousands of virtual CPUs on a system that has been designed to be fast and to offer strong consistency of performance.

An instance is a virtual machine (VM) hosted on Google's infrastructure. You can create an instance by using the Google Cloud Platform Console or the gcloud command-line tool. A Compute Engine instance can run Linux and Windows Server images provided by Google or any customized versions of these images. You can also build and run images of other operating systems. 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.



If the VM availability policy is set to the default, *live migrate*, during regular system maintenance, your VM will be migrated to different hardware so there is no downtime due to maintenance activities. Auto restart refers to what behavior the VM should take after a hardware failure or a system event. If marked *auto restart*, the system will try to launch a replacement VM. Auto restart does not restart the VM if it was terminated due to a user event, such as shutting down and terminating the VM.

Compute Engine monitors the CPU and memory utilization of running virtual machines and makes sizing recommendations using the last 8 days of data. The recommendation is designed for minimizing expenses by optimizing machine type However, you may have other goals:

- Run overcapacity, for example, to handle spike in traffic
- Run overcapacity for resilience
- Run undercapacity for guaranteed utilization, queue up work
- Run undercapacity to cap spend

https://cloud.google.com/compute/docs/instances/viewing-sizing-recommendations-for-instances

Compute

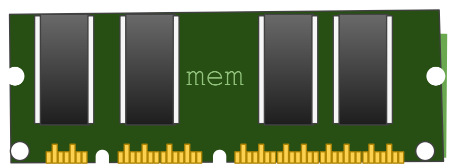
High CPU, high memory, standard, and shared-core machine types

- Network throughput scales at 2 Gb per vCPU
- Max throughput of 16 Gb or 8 vCPU

A vCPU is equal to 1 hyperthreaded core

2 vCPU is equal to 1 physical core

TAS



TAS

Google Cloud

3 machine types

- Higher proportion of memory to CPU
- Higher proportion of CPU to memory
- A blend of both in Google's standard configuration

Software-defined networking allows you to scale network and disk I/O performance. 2 Gb of throughput per vCPU up to 16 Gb. This also shares throughput with disk throughput.

Storage

Persistent disks

- Standard, SSD, or local SSD
- Standard and SSD PDs scale in performance for each GB of space allocated

Resize disks, migrate instances with no downtime

Google Cloud

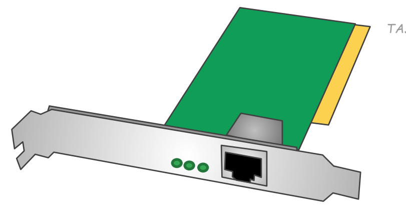
- Software-defined networking allows standard and SSD PD to scale per GB allocated up to their maximum allowed I/O defined at https://cloud.google.com/compute/docs/disks/performance
- Local SSD is only allocated in fixed predefined disk sizes
- Live migration of disk only applies to Standard and SSD PD. Local SSD is considered ephemeral and should not be relied upon to be available during a live migration
-

By default, most Compute Engine-provided Linux images will automatically run an optimization script that configures the instance for peak local SSD performance.
For more information, see https://cloud.google.com/compute/docs/disks/performance

Networking

Robust networking features

- Default, custom networks
- Inbound/outbound firewall rules
  - IP based
  - Instance/group tags
- Regional HTTPS load balancing
- Network load balancing
  - Does not require pre-warming
- Global and multi-regional subnetworks



Google Cloud

Your networks connect your instances to each other and to the internet. You can segment your networks, use firewall rules to restrict access to instances, and create static routes to forward traffic to specific destinations.

Scale your applications on Compute Engine from zero to full throttle with Cloud Load Balancing, with no pre-warming needed. Distribute your load-balanced compute resources in single or multiple regions close to your users and to meet your high availability requirements. Cloud Load Balancing can put your resources behind a single anycast IP and scale your resources up or down with intelligent autoscaling. Cloud Load Balancing comes in a variety of flavors and is integrated with Cloud CDN for optimal application and content delivery.

Subnetworks segments your Cloud network IP space into subnetworks. Subnetwork prefixes can be automatically allocated, or you can create a custom topology. For more information about subnetworks, see: https://cloud.google.com/compute/docs/subnetworks.

Pricing

Per-second billing, sustained use discounts

- 1 minute minimum

Preemptible instances

- Live at most 24 hours
- Can be pre-empted with a 30-second notification via API
- Up to 80% discount

Custom machine types

- Customize amount of memory and CPU

Recommendation Engine

- Notifies you of underutilized instances

Google Cloud

For more information about preemptible instances, see https://cloud.google.com/compute/docs/instances/preemptible.
For more information about using custom machine types, see https://cloud.google.com/compute/docs/instances/creating-instance-with-custom-machine-type.

VM charges and discounts

- All machines are billed for a minimum of 1 minute
- Per-second charge
- Lower price for preemptible instances
- Scaled discounts for sustained use
- Committed use discounts
- Always Free usage quotas

Custom-type sustained use discount

% of usage	% of base rate
First 25	100
Next 25	80
Last 25	60

Combined resources

Predefined-type sustained use discount

% of month usage	% of base rate
0–25	100
25–50	80
50–75	60
75–100	40

Inferred instances

Google Cloud

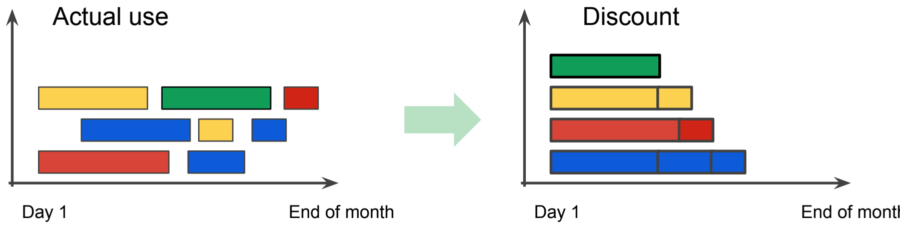
All machines are charged for 1 minute at boot time. This is the minimum charge for a VM. After that, per-second pricing begins.
Discounts are complicated. Predefined machine types are discounted based on the percent of monthly use. And custom-type is discounted on a percent of total use.

A new discount called Committed Use Discounts was announced in March of 2017. Customers can received up to a 57% discount in exchange for a one- or three-year commitment paid monthly with no up-front costs.
https://cloud.google.com/compute/docs/instances/signing-up-committed-use-discounts

Since March 2017, Google Cloud Platform offers Always Free Usage Limits, which provides free usage of one f1-micro instance per month (US regions only, excluding Northern Virginia). For more information, see
https://cloud.google.com/free/docs/always-free-usage-limits#compute_name

Inferred instances means that for billing purposes, the same type of machine used in the same zone will be combined into a single charge so that you get the most discount—as if it were one machine in use the whole time. And *combined resources* means that memory and vCPU of the same type are combined so that you get the discount on the greatest resource consumption in custom types.

Inferred instances discount



Usage of VMs of the same machine type in the same zone is combined as if they were one machine, to give you the best discount.

Google Cloud

The example shows a customer's usage that comprises eight distinct instances. The instances are combined to find the smallest number of simultaneous running instances, which are called "inferred instances". In this example, Compute Engine combines the instances to make four inferred instances with the longest possible duration. Compute Engine then calculates sustained use discounts based on the percentage of time that each of these inferred instances were running.

VM access

Linux: SSH

- SSH from console, SSH from CloudShell via Cloud SDK
- SSH from computer, third-party client and generate key pair
- Requires firewall rule to allow tcp:22

Windows: RDP

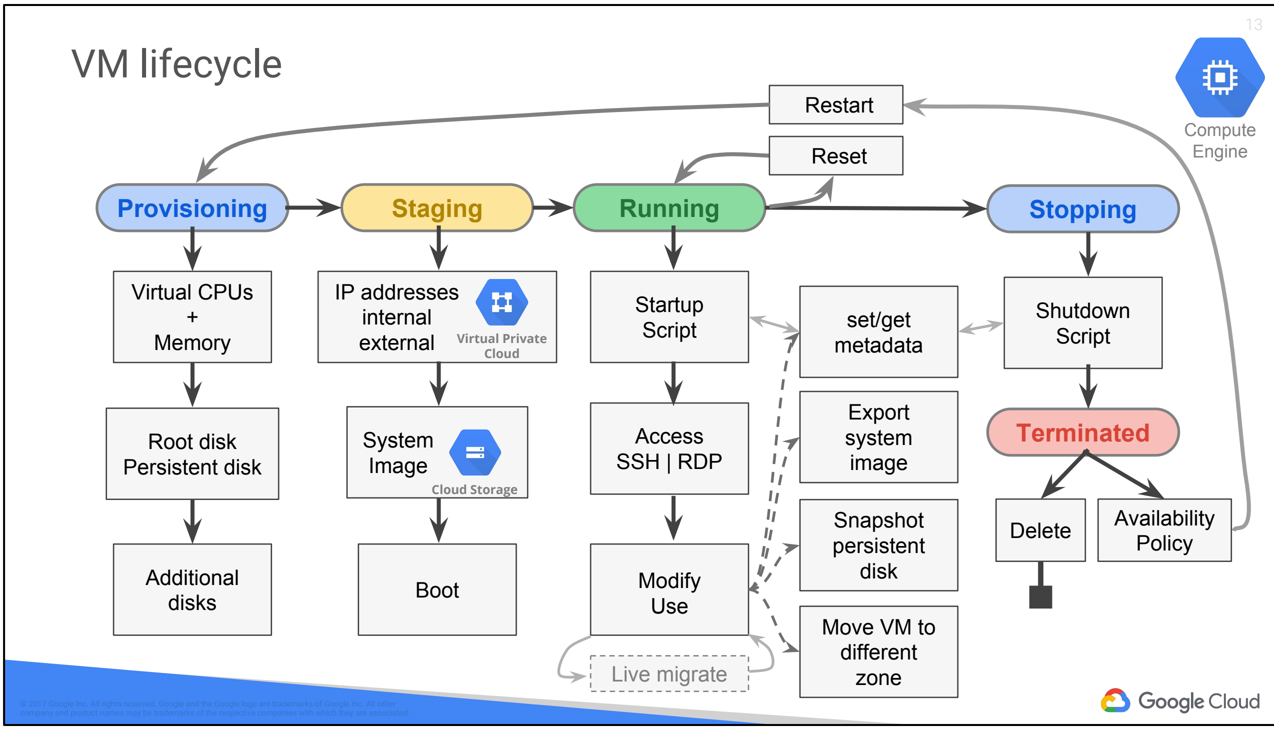
- Requires setting the Windows password
- Requires firewall rule to allow tcp:3389
- RDP clients
  - Chrome extension, third-party apps, MS Windows RDP client
  - Linux freerdp, remmina
- Powershell terminal

Google Cloud

Owners/Editors can use console, CloudShell, or SDK gcloud compute. Users that are not members can use third-party SSH.
In addition to SSH, you can protect access to instances with external IP addresses using port forwarding and SOCKS proxies. For more information, see https://cloud.google.com/solutions/connecting-securely#https-and-ssl

For more information on SSH key management, see:
https://cloud.google.com/compute/docs/instances/adding-removing-ssh-keys

Some RDP clients are incompatible because they don't support Network Level Authentication (NLA).



Create a new VM from the GCP Console, CloudShell gcloud command, or API. The rest of this module covers the subjects on this slide.

- PROVISIONING: Resources being reserved for the instance
- STAGING: Resources acquired; instance being prepared
- RUNNING: Instance is booting up or running
- STOPPING: Instance is being stopped due to failure or shutdown
  - Temporary status; instance will move to TERMINATED
  - When documentation says "STOPPED" it means "TERMINATED" state.
- TERMINATED: Instance is shut down or encountered failure (through API or inside the guest)
- Check instance state in gcloud using:
  - gcloud compute instances describe <instance>
  -

Difference between resetting and restarting:
https://cloud.google.com/compute/docs/instances/restarting-an-instance

Changing VM state from running

	methods	Shutdown Script time	state
`reset`	console, `gcloud`, API, OS	no	remains running
`restart`	console, `gcloud`, API, OS	no	terminated➡running
`reboot`	OS: `sudo reboot`	~90 sec	running➡running
`stop`	console, `gcloud`, API	~90 sec	running➡terminated
`shutdown`	OS: `sudo shutdown`	~90 sec	running➡terminated
`delete`	console, `gcloud`, API	~90 sec	running➡N/A
preemption	*automatic*	~30 sec	N/A

"ACPI Power Off"

Google Cloud

Stopping an instance moves it into a Terminated state. So "stopped" and "terminated" are the same state.
Stopping sends an Advanced Configuration Power Interface (ACPI) "ACPI Power Off" command.
Script normally has about 90 seconds to run; preempted VMs have 30 seconds.
Running and time allowed for shutdown script is not guaranteed.
When a VM is stopped, it loses ephemeral external IPs and the contents of RAM memory.
Terminated VMs still exist and can be restarted until they are deleted.

Availability policy: Automatic changes

Called "scheduling options" in SDK/API

Automatic restart

- Automatic VM restart due to crash or maintenance event
    - Not preemption or a user-initiated terminate

On host maintenance

- Determines whether host is live-migrated or terminated due to a maintenance event. Live migration is the default.

Live migration

- During maintenance event, VM is migrated to different hardware without interruption.
- Metadata indicates occurrence of live migration.

Google Cloud

Automatic restart due to maintenance event, hardware failure, or software failure.

For more information on live migration, see https://cloud.google.com/compute/docs/instances/live-migration

Stopped (Terminated) VM

Note: You don't *have* to stop a VM to make many of these changes. However, these are the actions that you *can* make to a stopped VM.

No charge for stopped VM

- Still charged for attached disks and IPs

Actions

- Change the machine type
- Add or remove attached disks; change auto-delete settings
- Modify instance tags
- Modify custom VM or project-wide metadata
- Remove or set a new static IP
- Modify VM availability policy
- ***Can't*** change the image of a stopped VM

Google Cloud

Can't change the image of a stopped VM
You can create a new VM from the stopped disk, even if the original VM was deleted
Machine type can increase or decrease in size, but the disk must be big enough to hold the image.

Agenda

- Compute Engine
- **Lab**
- Compute options (vCPU and Memory)
- Images
- Disk options
- Common Compute Engine actions
- Lab
- Quiz

Google Cloud

Lab: Creating virtual machines

Objectives

In this lab, you learn how to perform the following tasks:

- Create several standard VMs
- Create advanced VMs

Completion: 30 minutes

Access: 60 minutes

Compute Engine

Lab review

In this lab, you:

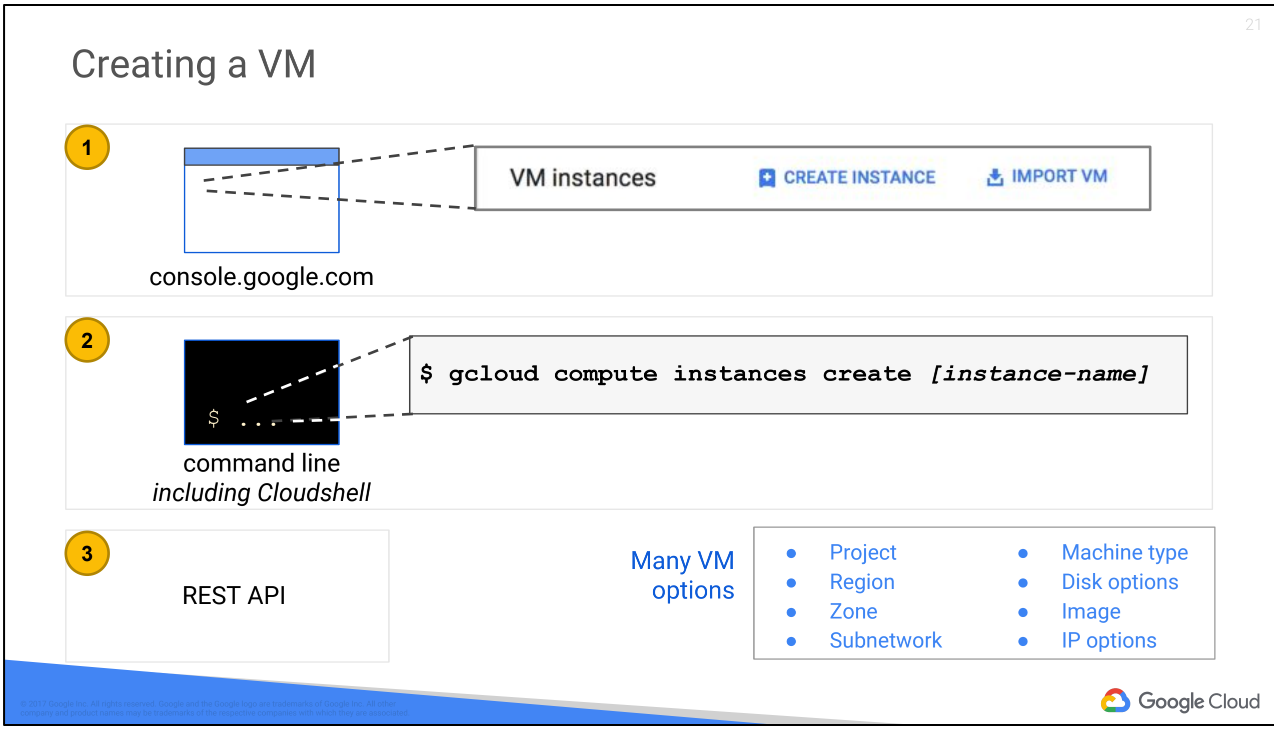
- Created several virtual machine instances of different types with different characteristics:
  - One was a small utility VM for administration purposes.
  - You also created a standard VM and a custom VM.
- You launched both Windows and Linux VMs and deleted VMs.

Google Cloud

Agenda

- Compute Engine
- Lab
- **Compute options (vCPU and Memory)**
- Images
- Disk options
- Common Compute Engine actions
- Lab
- Quiz

Google Cloud



The GCP Console does more validation and checking for you. Note that the VM choices are constrained by the network structure. There are inherent networking dependencies. Subnetwork determines DHCP internal IP address on the VM, and subnetwork is homed in one region but may cross multiple zones in that region. The GCP Console interface analyzes GCP and presents you with only pull-down selection options that are *possible* to provision and stage. When you use the `gcloud compute` command or the API, you are expected to know which options are possible and which are not. If you submit the request with impossible options, the command and API can't check them first. So they will instead attempt to start the VM, and the command may result in a "fail" after a few minutes.

REST API:
https://cloud.google.com/compute/docs/api/how-tos/api-requests-responses

Machine types

n1-standard-*vCPUs*					n1-highmem-*vCPUs*		n1-highcpu-*vCPUs*	
vCPUs	Mem GB	# Disks/Beta	Max Storage		High Mem		High CPU vCPUs	High CPU Mem
1	3.75 GB	16 / 32	64 TB					
2	7.5 GB	16 / 64	64 TB		13 GB		2	1.80
4	15 GB	16 / 64	64 TB		26 GB		4	3.60
8	30 GB	16 / 128	64 TB		52 GB		8	7.20
16	60 GB	16 / 128	64 TB		104 GB		16	14.4 GB
32	120 GB	16 / 128	64 TB		208 GB		32	28.8 GB
64	240	16/ 128	64 TB					

1	0.6 GB	4 / 16	3 TB	**f1-micro**
1	1.7 GB	4 / 16	3 TB	**g1-small**

shared core
micro-bursting

Google Cloud

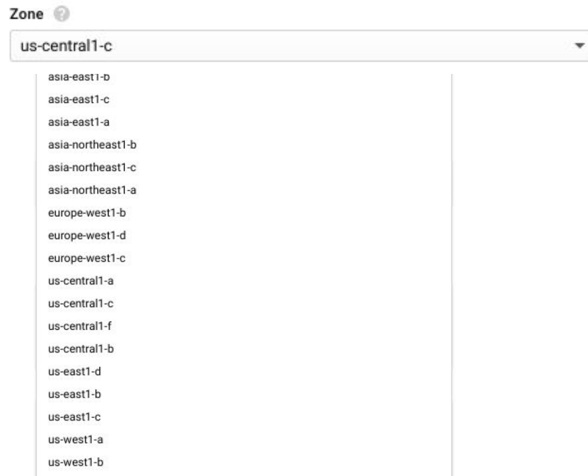
https://cloud.google.com/compute/docs/machine-types
High CPU doesn't mean more CPU power. It means less memory for the same CPU power.
Micro-bursting: The shared CPU can occasionally use more than its specified share of the CPU for a brief period.
Some zones do not offer 32-core or 64-core VMs. (Currently: us-central1-a and europe-west1-b. *Check documentation for latest information*.)

GPUs are now available in limited locations. NVIDIA® Tesla® K80 GPUs.
https://cloud.google.com/compute/docs/gpus/

Choosing region and zone

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

- Different zones support different CPU architectures
  - Broadwell = 64 CPU max
  - Sandy Bridge = 16 CPU max
  - Ivy Bridge = 32 CPU max
  - Haswell = 64 CPU max
  - Skylake = 96 CPU max



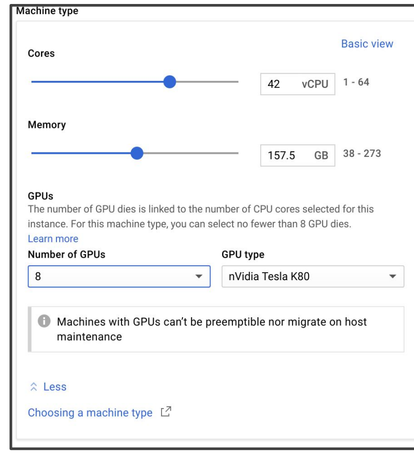
Google Cloud

https://cloud.google.com/compute/docs/regions-zones/regions-zones#available
Distributing multiple VMs across multiple zones is recommended for higher availability.

Examples:
us-west1-a/us-west1-b
(Broadwell processors)
us-central-1-a
(Sandy Bridge processors)
Limits the number of CPUs

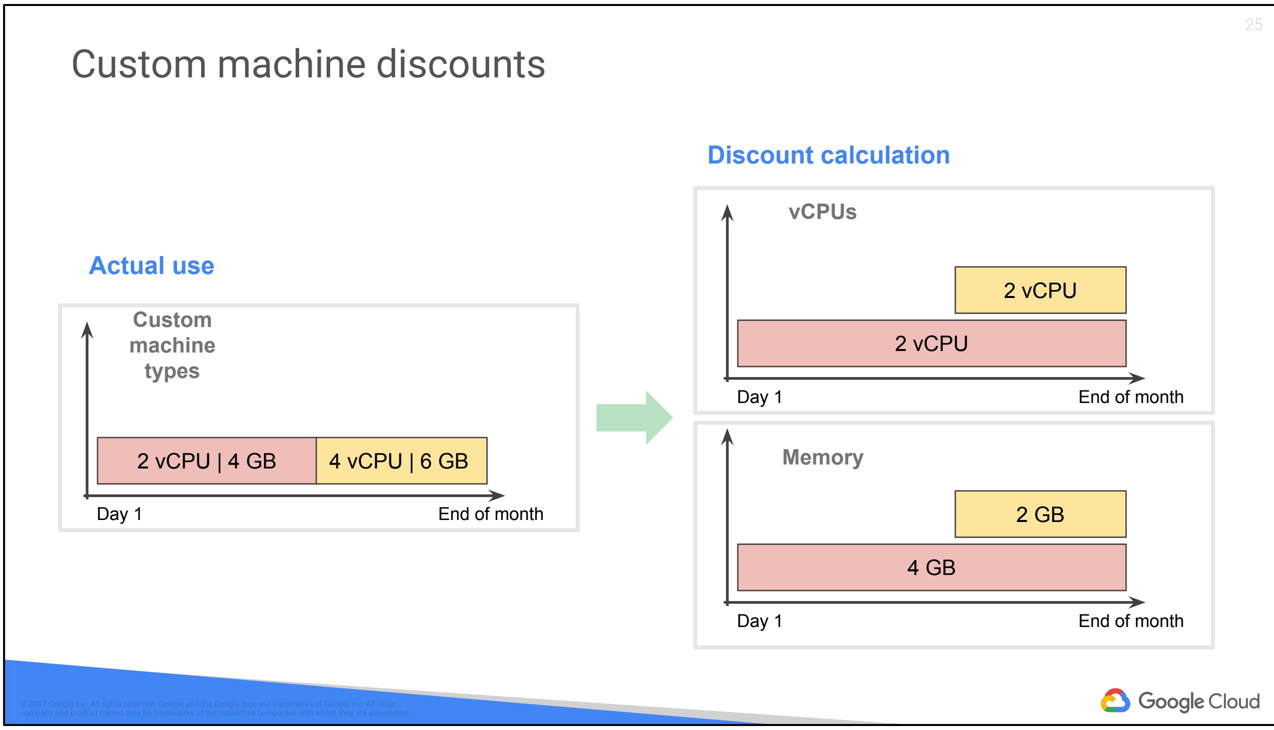
Creating custom machine types

- Number of vCPUs per instance
  - Only 1 vCPU or an even number of vCPUs
- Memory
  - 0.9 GB to 6.5 GB per vCPU
- When to select custom:
  - Requirements fit between the predefined types
  - Need more memory or more CPU
  - Need GPUs
- Customize the amount of memory and CPU for your machine
- Get recommendations for a predefined match
  - Custom VM will generally be more expensive than an identical predefined VM



Google Cloud

Charges based on number of vCPUs and memory hours consumed. Sustained use discounts are calculated differently.
Currently, SUSE OS is not supported in custom types.
Multiple network interfaces (NICs) on a single instance is currently not supported.



Accumulating inferred instances by common machine type in a zone isn't possible with custom machine types.
Instead, usage is broken down by resource (vCPU, memory), and the total accumulated resource usage is discounted.

Preemptible

- Lower price for interruptible service (up to 80%)
- VM might be terminated at any time
  - No charge if terminated in the first 10 minutes
  - 24 hours max
  - 30-second terminate warning, but not guaranteed
    - *Time for a shutdown script*
- No live migrate; no auto restart
- You *can* request that CPU quota for a region to be split between regular and preemption
  - Default is preemptible VMs count against region CPU quota

Google Cloud

Regular instances usually have a 90-second shutdown notice. Preemptible instances have a 30-second shutdown notice.
Pricing is not bid-based as with other cloud providers. This is intentional to provide you with more reliable cost estimation.
Preemption is based on most recently launched first (not longest running, as with other cloud providers). Avoids preempting too many instances from a single customer.
In one case, preemption rate averages were measured between 5% and 15% over a 7-day project.
Preemptible instances complete your batch processing tasks without placing additional workload on your existing instances, and without requiring you to pay full price for additional normal instances. Therefore, using preemptible instances avoids running out of VMs for mission-critical work

Agenda

- Compute Engine
- Lab
- Compute options (vCPU and Memory)
- **Images**
- Disk options
- Common Compute Engine actions
- Lab
- Quiz

Google Cloud

What's in an image?

- Contents of an image
  - Boot loader
  - Operating system
  - File system structure
  - Software
  - Customizations
- Image storage
  - A tar and gzip'd file
  - In a private area of Cloud Storage managed by the image service

Google Cloud

https://cloud.google.com/compute/docs/images

Images

- Public base images
  - Google, third-party vendors, and community; Premium images (p)
  - Linux
    - CentOS, CoreOS, Debian, RHEL(p), SUSE,(p) Ubuntu, openSUSE and FreeBSD
  - Windows
    - Windows Server 2016(p), 2012-r2(p), 2008-r2(p)
    - SQL Server pre-installed on Windows(p)
- Custom images
  - Create new image from VM: pre-configured and installed SW
  - Import from on-prem, workstation, or another cloud
  - Management features: image sharing, image family, deprecation

Google Cloud

Premium images (p) incur additional per-second charges with the exception of SQL Server images, which are charged per minute. Premium image prices vary with the machine type. Prices are global and do not vary by region or zone.
https://cloud.google.com/compute/pricing#premiumimages
openSUSE and FreeBSD are the community-supported OSs

Virtualbox or AWS
Image can be stored in Cloud Storage

Image family: points to most recent image so scripts and templates don't need to reference a specific version

Image management best practices:
https://cloud.google.com/solutions/image-management-best-practices

Compute Engine uses operating system images to create the root persistent disks for your instances. You specify an image when you create an instance. Images contain a boot loader, an operating system, and a root file system.

For more information about importing Linux images, see https://cloud.google.com/compute/docs/images/import-existing-image.

Agenda

- Compute Engine
- Lab
- Compute options (vCPU and Memory)
- Images
- **Disk options**
- Common Compute Engine actions
- Lab
- Quiz

Google Cloud

Boot disk

- VM comes with a single root persistent disk
- Image is loaded onto root disk during first boot
  - Bootable: you can attach to a VM and boot from it
  - Durable: can survive VM terminate
- Some OS images are customized for Compute Engine
- Can survive VM deletion if “Delete boot disk when instance is deleted” is disabled

Google Cloud

https://cloud.google.com/compute/docs/images

The image name contains the version of the OS. Generally, the latest patches were applied prior to creating the image.
In most cases the image has some level of "auto update" enabled, at least for security patches. Check the documentation of the particular OS.
This behavior might not meet your requirements. In some cases, stability is preferred over security, or applying security patches is under change management to minimize risks.

Persistent disks

Network storage appearing as a block device

- Attached to a VM through the network interface
- Durable storage: *can* survive VM terminate
- Bounded to zone: cannot be moved between zones
- Bootable: you can attach to a VM and boot from it
- Snapshots: incremental backups

Features

- HDD (magnetic) or SSD (faster, solid-state) options
- Disk resizing: even running and attached!
- Can be attached in read-only mode to multiple VMs
- Checksums built-in; Automatic encryption—your key optional
- Performance: https://cloud.google.com/compute/docs/disks/performance

Google Cloud

- Persistent disk: use SSD instead of HDD for high-rate random IOPS
- Persistent disks can be attached to multiple instances in read-only mode for data sharing
- Disk resizing: you can make them bigger, but never smaller
- Economical for storing large amounts of local data
- Performs best with sustained read/write of large files
- Virtual machines can save up I/O capability and burst for an I/O spike well above the average
- Allows smaller disks to be used in cases where typical I/O is low but periodic bursting is required over the maximum allowed sustained I/O
- Performance: https://cloud.google.com/compute/docs/disks/performance

Local SSD disks

Physically attached to VM—*not available on shared core*

- More IOPS, lower latency, higher throughput, than persistent disk
- SCSI or NVMe interface
  - NVMe requires an NVMe-enabled image
  - NVMe is faster than SCSI

375-GB disk up to eight, total of 3 TB

Up to 680K read IOPS and up to 360K write IOPS

Data survives a reset, but not a VM stop or terminate

- VM-specific, cannot be reattached to a different VM

Can use your own encryption keys

Google Cloud

- Limited to 4 partitions total of 1.5 TB in us-central1-a and europe-west1-b due to hardware differences in those zones.
- Local SSDs are not available to shared core machines.

RAM disk

- `tmpfs`
- Faster than local disk, slower than memory
  - Use when your application expects a file system structure and cannot directly store its data in memory
  - Fast scratch disk, or fast cache
- Very volatile—erase on stop or restart
- May need a larger machine type if RAM was sized for the application
- Consider using a persistent disk to back up RAM disk data

Google Cloud

https://cloud.google.com/compute/docs/disks/mount-ram-disks
https://www.kernel.org/doc/Documentation/filesystems/tmpfs.txt

Summary of disk options

	Persistent disk HDD	Persistent disk SSD	Local SSD disk	RAM disk
Data redundancy	Yes	Yes	No	No
Encryption at rest	Yes	Yes	Yes	N/A
Snapshotting	Yes	Yes	No	No
Bootable	Yes	Yes	No	Not
Use case	General, bulk file storage	Very random IOPS	High IOPS and low latency	low latency and risk of data loss

Google Cloud

HDD = Standard Hard Disk Drive
SSD = Solid State Drive

Cloud Storage versus Persistent disk
Both persistent disks and Cloud Storage buckets are *network storage that is accessible to the VM*. Notice that the persistent disk is created via the Compute Engine API—part of Compute Engine, not part of Cloud Storage. You use `gcloud` for disks, not `gsutil`. They are ***not*** using the same hardware.

An open-source method exists to mount a Cloud Storage bucket as part of the VM file system. It is recommended that you avoid this technique. To use Cloud Storage, use the appropriate tools or go through the API (covered in the Storage module). Cloud Storage is an object store, and some of the basic assumptions contradict the behavior of a file system. For example, an object store provides no serialization lock, so if multiple writers are writing to a file at once, the last one "wins." At several critical points, the simulation of a file system on top of an object store breaks down and can produce random errors that you would not expect of a file system.

Maximum disks

Number of cores	Disk number limit
Shared core	16
1 core	32
2–4 cores	64
8 or more cores	128

Google Cloud

This is a Beta release of increased persistent disk limits.
https://cloud.google.com/compute/docs/disks/

Persistent disk management differences

Cloud Persistent Disk

- Single file system is best
- Resize (grow) disks
- Resize filesystem
- Built-in redundancy
  - Built-in snapshot service
- Automatic encryption prior to write—use your keys

Computer Hardware Disk

- Partitioning
- Repartition disk
- Reformat
- Redundant disk arrays
- Subvolume management and snapshots
- Encrypt files before write to disk

Google Cloud

There are many differences between a physical hard disk in a computer and a persistent disk, which is essentially a virtual networked device.

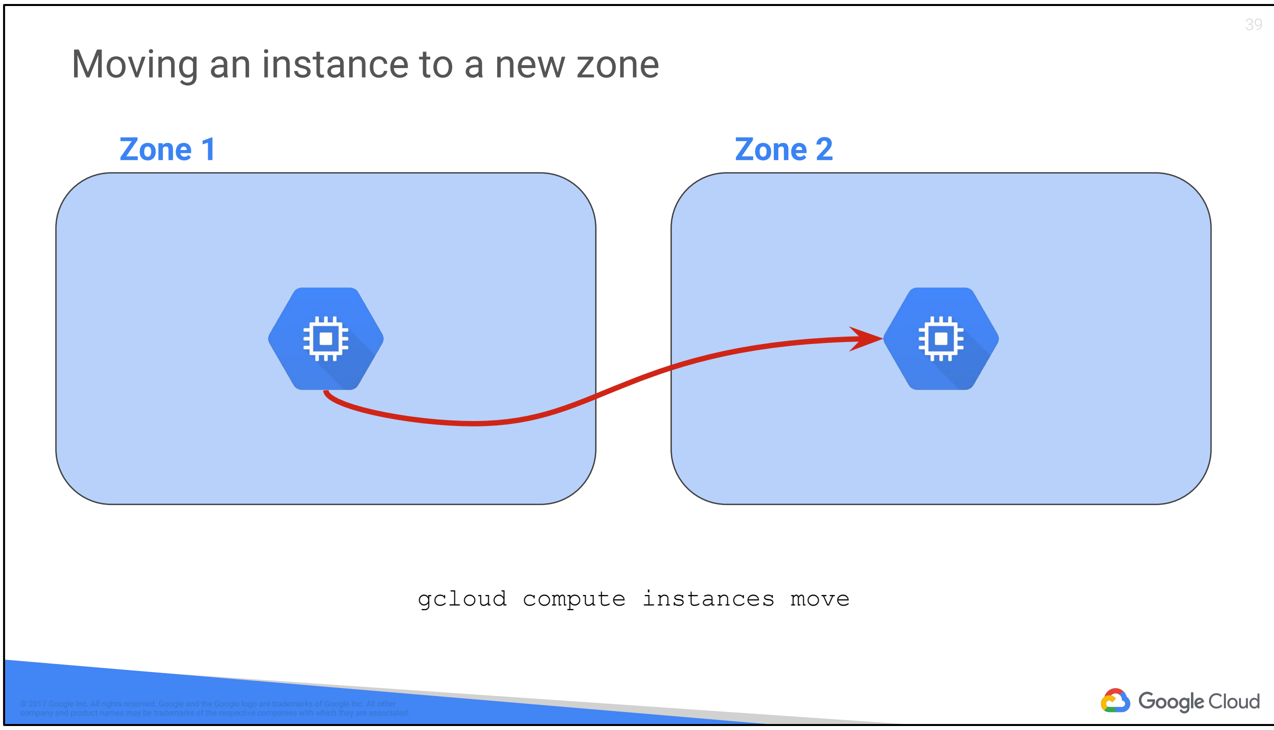
You CAN perform all of the hardware disk actions listed on the right. They just won't have the benefit you anticipate, and might be a lot of work for something that already comes standard with persistent disks.

A single file system gives the best performance on a persistent disk. If you need more space, IOPS, or throughput, you can resize a persistent disk and resize the file system that it contains. To do something similar on a computer disk, you would partition, and then when you need more space, you might re-partition and reformat the partitions. You don't need to build redundant disk arrays to get durability: the system already has built-in redundancy. You don't need to implement striping for performance improvement: that's done automatically and transparently.
You don't need to use something like subvolume management for snapshotting: there is an automatic built-in snapshotting service. Finally, you don't have to encrypt files before writing them to disk, because all files are encrypted before they are transmitted to the persistent disk.

Agenda

- Compute Engine
- Lab
- Compute options (vCPU and Memory)
- Images
- Disk options
- **Common Compute Engine actions**
- Lab
- Quiz

Google Cloud



- Move using gcloud command-line utility:
    - `gcloud compute instances move`
- Use API `moveInstance` method
    - Make a POST request to the `moveInstance` method with a request body that contains the `targetInstance` and the `destinationZone`
- Use disk snapshots (manual procedure)

Requirements

- Quotas to:
    - Create new snapshots
    - Promote any ephemeral external IP addresses
    - Create the new instance and disks in the destination zone
- Persistent disks attached to the instance you want to move cannot be attached to more than one instance
- Source instance cannot contain a local SSD
- Must update any existing references you have to the original resource (not done automatically)

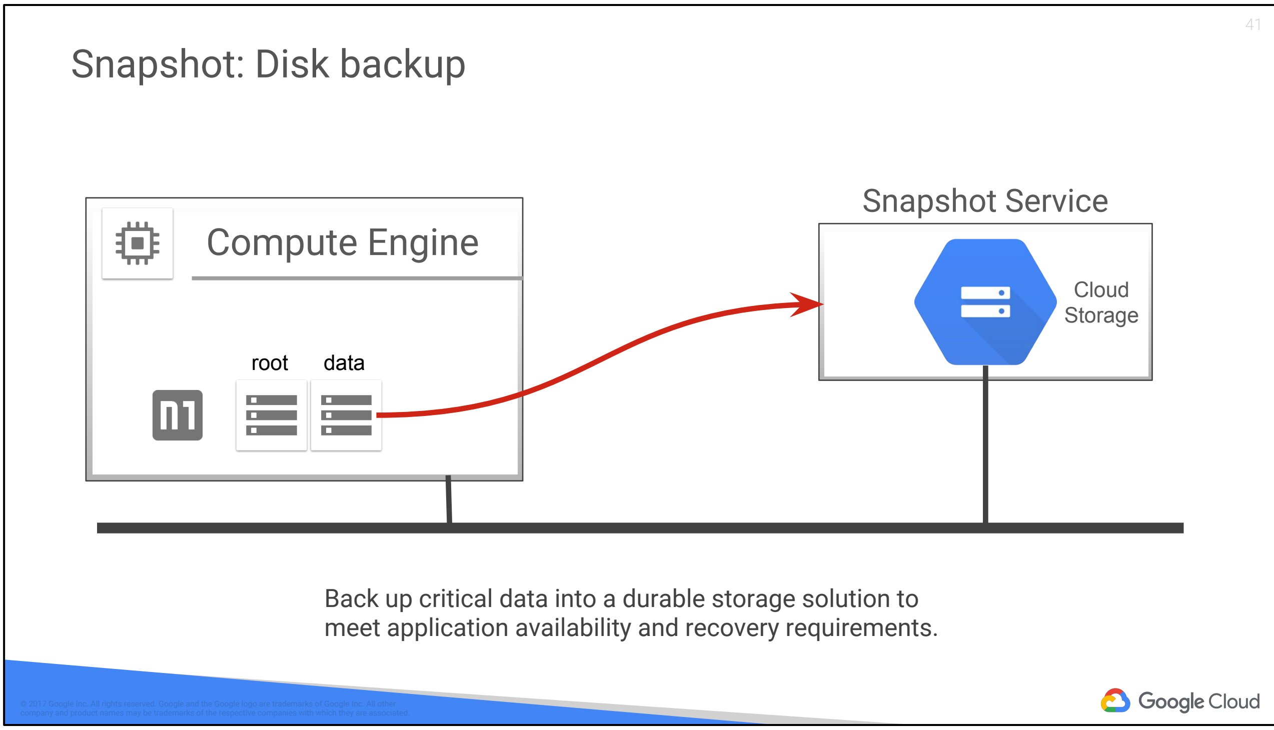
For more information about the requirements for moving instances between zones, see https://cloud.google.com/compute/docs/instances/moving-instance-across-zones#requirements

Moving an instance to a new zone

- Automated process (moving within region):
  - `gcloud compute instances move`
  - Update references to VM, not automatic
- Manual process (moving between regions):
  - Snapshot all persistent disks on the source VM
  - Create new persistent disks in destination zone restored from snapshots
  - Create new VM in the destination zone and attach new persistent disks
  - Assign static IP to new VM
  - Update references to VM
  - Delete the snapshots, original disks, and original VM

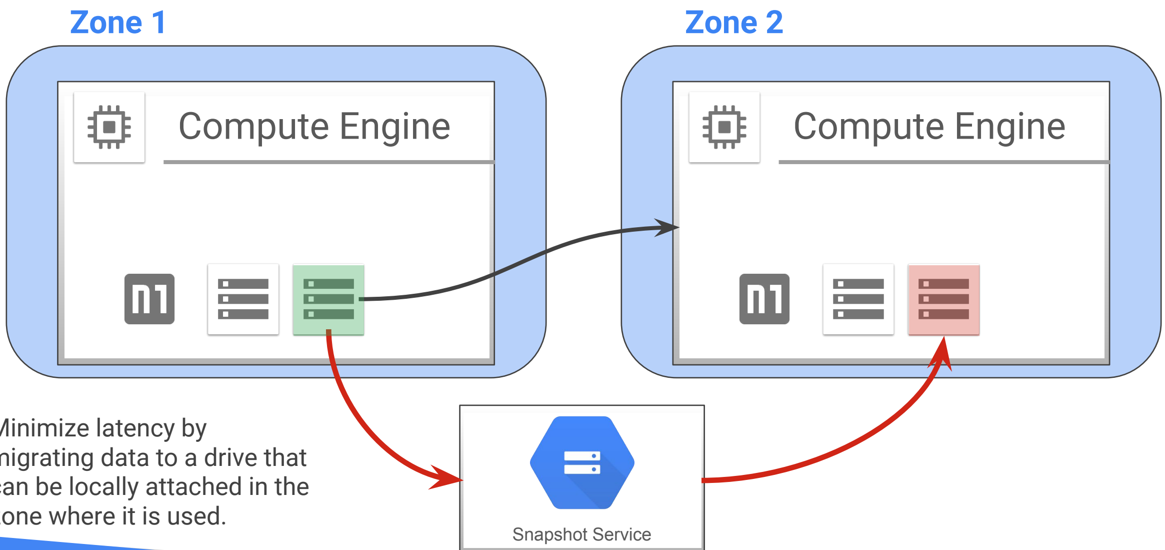
Google Cloud

Don't use on a VM with a local SSD. The local SSD data cannot be backed up and will just be discarded.
Persistent disks must be attached to only the VM you are going to move, not to multiple VMs.
Sufficient quota must exist for all the resources copied during duplication, or the process will fail.
Why would you move a VM to a different zone? To support availability policy. Also, if a zone is deprecated, you can use this process to preserve your application.

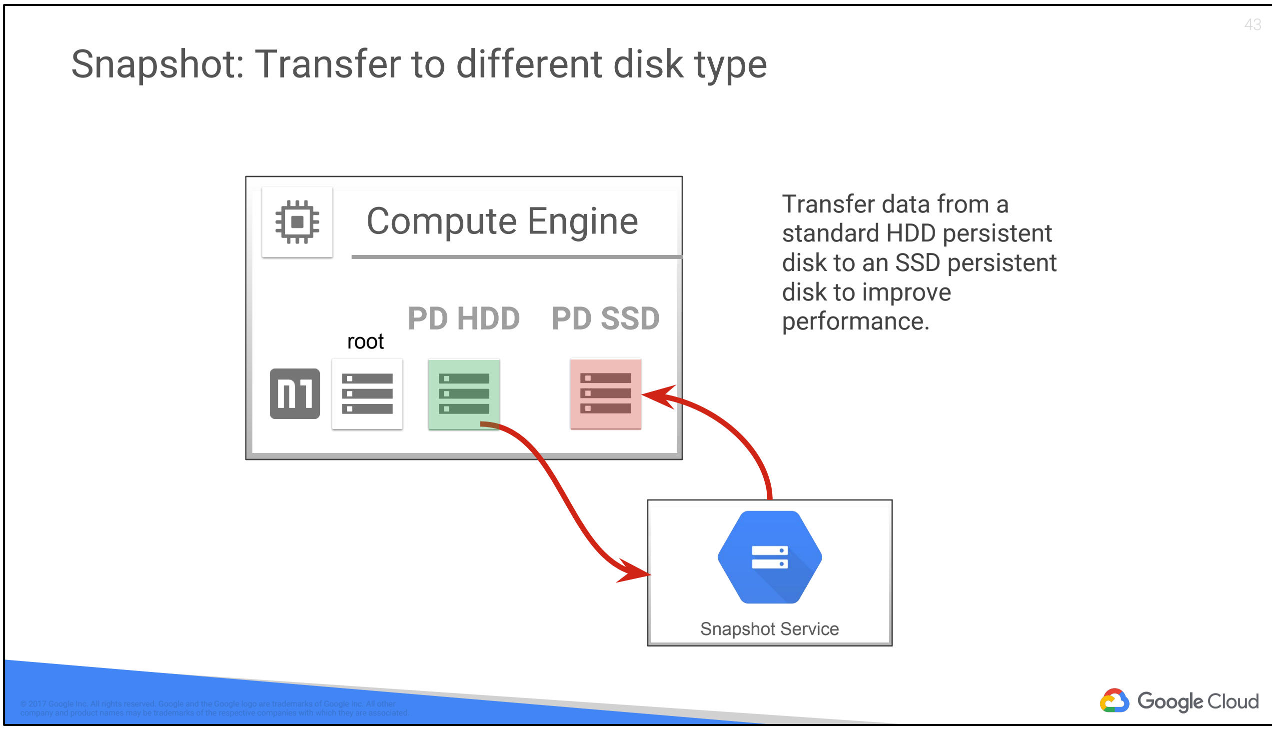


Cloud Storage is covered in the next part of this class.

Snapshot: Migrate data between zones



Minimize latency by migrating data to a drive that can be locally attached in the zone where it is used.



Standard HDD persistent disks are efficient and economical for handling sequential read/write operations, but are not optimized to handle high rates of random input/output operations per second (IOPS). If your applications require high rates of random IOPS, use SSD persistent disks.

Persistent disk snapshots

- Snapshot is not available for local SSD
- Creates an *incremental* backup to Cloud Storage
  - Not visible in *your* buckets—managed by the snapshot service
  - Consider cron jobs for periodic incremental backup
- Snapshots can be restored to a new persistent disk
  - New disk can be in another region or zone in the same project
  - Basis of VM migration: "moving" a VM to a new zone
    - Snapshot doesn't backup VM metadata, tags, etc.
  - Cannot restore to other disk types
- Don't use for database migration across zones

Google Cloud

https://cloud.google.com/compute/docs/disks/create-snapshots

Snapshots are the size of the existing disk. They cannot be restored to a smaller disk. So "shrinking" a disk would be an OS copy process of data from one attached disk to a second smaller attached disk. Not through Compute Engine services.

Multiple copies of each snapshot are redundantly stored across multiple locations with automatic checksums to ensure the integrity of your data.
You cannot share snapshots across projects.

Some people use the term "differential" (ie from "diff") backup, thinking about the differences that have occurred to the disk since the last backup. However, in storage discussions it is common to use "incremental" backup, because each backup file itself is incremental within the lineage, usually date-time-stamped. Both terms are technically correct.

Databases have existing replication technologies that are better suited. When you snapshot a database, you have to write-lock the database, which affects (violates) the SLA. Behind the scenes, you are using compression and decompression on highly structured and already compressed data, so when you restore, it can take hours.

Snapshot disk preparation

- Boot disk, halt the system
  - sudo shutdown -h now
- Additional disk, unmount the file system
  - sudo umount </mount/point>

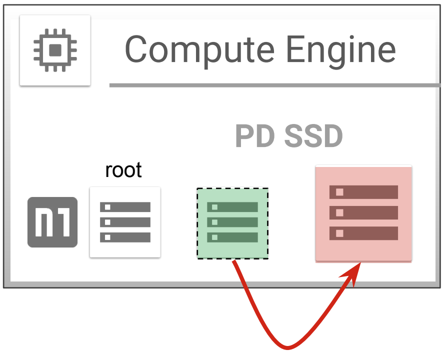
  If unmount isn't possible

  - Stop applications from writing to the persistent disk
  - Complete pending writes and flush cache
    - sudo sync
  - Suspend writing to the disk device
    - sudo fsfreeze -f </mount/point>

Google Cloud

On Windows, you can enable Volume Shadow Service (VSS), which enables a disk to be backed up without having to be shut down.
https://cloud.google.com/compute/docs/instances/windows/creating-windows-persistent-disk-snapshot?hl=en_US&_ga=1.38057187.1394348339.1478206226

Resize persistent disk



- Improve I/O performance by increasing storage capacity of a persistent disk
- You can edit a persistent disk, increase size, and increase IOPS capacity
- A disk can be resized even when it is attached to a VM and while it is running
- You don't need to use a snapshot to accomplish this
- You can grow disks, but never shrink them

	Sustained Random IOPS Limit (Read or Write)	**Sustained throughput limit (mb/s)**
10 GB SSD	300	4.8
100 GB SSD	3000	48
500 GB SSD	15000	240

Google Cloud

Note: A snapshot cannot be restored to a smaller disk. So not only can you not shrink an existing disk, you can't restore to a disk to shrink it. If you need to do something like that, use OS copy tools. So, it is better to be conservative with the size of the disk and to later expand it as necessary.

Agenda

- Compute Engine
- Lab
- Compute options (vCPU and Memory)
- Images
- Disk options
- Common Compute Engine actions
- **Lab**
- Quiz

Google Cloud

Lab: Working with Virtual Machines

Objectives

In this lab, you learn how to perform the following tasks:

- Customize an application server
- Install and configure necessary software
- Configure network access
- Schedule regular backups

Completion: 45 minutes

Access: 90 minutes

Compute Engine

Google Cloud

Lab review

In this lab, you:

- Created a customized virtual machine instance by installing base software (a headless JRE) and application software (a Minecraft game server).
- Customized the VM by attaching and preparing a high-speed SSD data disk, and reserved a static external IP so the address would remain consistent.
- Verified availability of the gaming server online.
- Set up a backup system to back up the server's data to a Cloud Storage bucket, and tested the backup system.
- Automated backups using cron.
- Set up maintenance scripts using metadata for graceful startup and shutdown of the server.

Agenda

- Compute Engine
- Lab
- Compute options (vCPU and Memory)
- Images
- Disk options
- Common Compute Engine actions
- Lab
- **Quiz**

Google Cloud

Quiz

Which statement is true of virtual machine instances in Compute Engine (Compute Engine)?

1. A VM in Compute Engine always maps to a single hardware computer in a rack.
2. Compute Engine uses VMware to create virtual machine instances.
3. In Compute Engine, a VM is a networked service that simulates the features of a computer.
4. All Compute Engine VMs are single tenancy and do not share CPU hardware.

Google Cloud

Quiz

Which statement is true of virtual machine instances in Compute Engine (Compute Engine)?

1. A VM in Compute Engine always maps to a single hardware computer in a rack.
2. Compute Engine uses VMware to create virtual machine instances.
3. In Compute Engine, a VM is a networked service that simulates the features of a computer. *
4. All Compute Engine VMs are single tenancy and do not share CPU hardware.

Google Cloud

Explanation:
VMs in Compute Engine are a collection of networked services. This includes disks (persistent disks) which are network-attached. In some cases, the GCP VM behaves unlike hardware or other kinds of virtual machines; for example, when a multi-tenant virtual CPU "bursts," using excess capacity beyond the VM spec.

Quiz

What is inferred instance discounting?

1. Usage of VMs for a partial month is extrapolated out to the entire month to give the best discount.
2. Usage of VMs of the same machine types across all zones in a region is combined as if they were one machine to give the best discount.
3. Usage of VMs of similar machine types in the same zone is combined as if they were one machine to give the best discount.
4. Usage of VMs of the same machine type in the same zone is combined as if they were one machine to give the best discount.

Quiz

What is inferred instance discounting?

1. Usage of VMs for a partial month is extrapolated out to the entire month to give the best discount.
2. Usage of VMs of the same machine types across all zones in a region is combined as if they were one machine to give the best discount.
3. Usage of VMs of similar machine types in the same zone is combined as if they were one machine to give the best discount.
4. Usage of VMs of the same machine type in the same zone is combined as if they were one machine to give the best discount. *

Google Cloud

Explanation:
Compute Engine combines usage of the same machine type in the same zone as if it were one virtual machine being used for a continuous period. This provides sustained use discounts without penalizing the user for terminating VMs when they are not needed and starting more on demand.

Quiz

Which statement is true of persistent disks?

1. Persistent disks are encrypted by default.
2. Persistent disks are physical hardware devices connected directly to VMs.
3. Persistent disks are always HDDs (magnetic spinning disks).
4. Once created, a persistent disk cannot be resized.

Google Cloud

Quiz

Which statement is true of persistent disks?

1. Persistent disks are encrypted by default. *
2. Persistent disks are physical hardware devices connected directly to VMs.
3. Persistent disks are always HDDs (magnetic spinning disks).
4. Once created, a persistent disk cannot be resized.

Google Cloud

Explanation:
Persistent disks are not physical disks; they are a virtual networked service. For example, persistent disks already include encryption as a default service. So if an administrator configures encryption as they might on a hardware disk, it will be redundant and might provide little or no additional benefit.

More resources

Instances
https://cloud.google.com/compute/docs/instances/

Regions and zones
https://cloud.google.com/compute/docs/regions-zones/regions-zones

Preemptible instances
https://cloud.google.com/compute/docs/instances/preemptible

Optimizing disk performance
https://cloud.google.com/compute/docs/disks/performance

Google Cloud

Google Cloud

Create and boot an instance

Select a public image
```
gcloud compute images list
```

Create the instance (automatically uses the latest version)
```
gcloud compute instances create [INSTANCE_NAME]--image-family [IMAGE_FAMILY] --image-project [IMAGE_PROJECT]
```

Create an instance from a SPECIFIC image (version)
```
gcloud compute instances create [INSTANCE_NAME]--image [IMAGE_ID] --image-project [IMAGE_PROJECT]
```

Create an instance from a CUSTOM image
```
gcloud compute instances create [EXAMPLE_INSTANCE]--image [IMAGE_NAME]
```

Create an instance from a SHARED image
```
gcloud compute instances create [INSTANCE_NAME]--image [IMAGE] --image-project [IMAGE_PROJECT]
```

Create an instance from a SNAPSHOT (backup image)

Step 1: Create the persistent disk from the snapshot
```
gcloud compute disks create [DISK_NAME] --source-snapshot [SNAPSHOT_NAME]
```
Step 2: Start the instance by booting from the persistent disk
```
gcloud compute instances create [INSTANCE_NAME] --disk name=[DISK_NAME],boot=yes
```

Preemptible instance add:
```
--preemptible
```


Google Cloud

Create and attach a persistent disk

Create the disk
```
gcloud compute disks create [DISK_NAME] --size [DISK_SIZE] --type [DISK_TYPE]
```

Attach the disk to the instance
```
gcloud compute instances attach-disk [INSTANCE_NAME] --disk [DISK_NAME]
```

Identify disk
```
ls /dev/disk/by-id
```

Format the disk (in this case with ext4)
```
sudo mkfs.ext4 -F -E lazy_itable_init=0,lazy_journal_init=0,discard /dev/disk/by-id/google-[DISK_NAME]
```

Create a mount point
```
sudo mkdir -p /mnt/disks/[MNT_DIR]
```

Mount the disk
```
sudo mount -o discard,defaults /dev/disk/by-id/google-[DISK_NAME] /mnt/disks/[MNT_DIR]
```

Configure access
```
sudo chmod a+w /mnt/disks/[MNT_DIR]
```


Google Cloud

Just reviewing the process and commands that you will be performing in the lab.