Google Cloud

Partner Certification Academy

Associate Cloud Engineer

The information in this presentation is classified:

Google confidential & proprietary

⚠ This presentation is shared with you under NDA.

- Do **not** record or take screenshots of this presentation.
- Do **not** share or otherwise distribute the information in this presentation with anyone **inside** or **outside** of your organization.

Thank you!

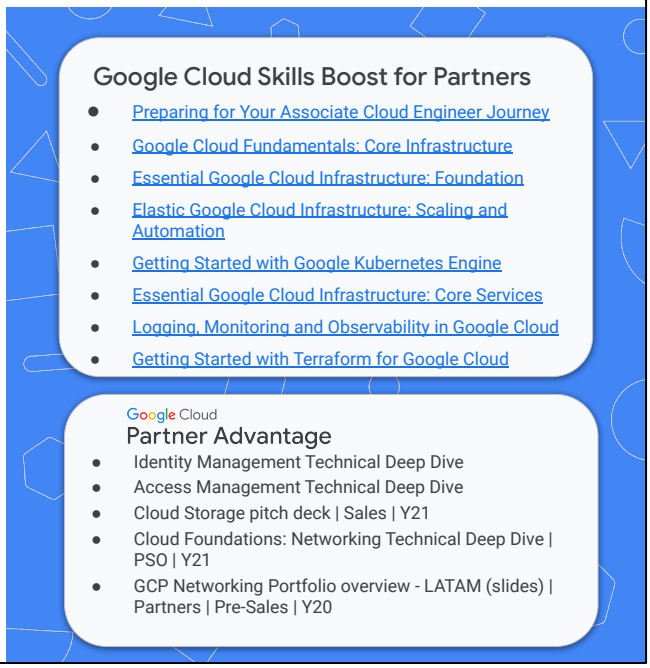
Google Cloud

Source Materials

Some of this program's content has been sourced from the following resources:

- Google Cloud certification site
- Google Cloud documentation
- Google Cloud console
- Google Cloud courses and workshops
- Google Cloud white papers
- Google Cloud Blog
- Google Cloud YouTube channel
- Google Cloud samples
- Google codelabs
- Google Cloud partner-exclusive resources

Google Cloud



Other sources used in the preparation of these materials

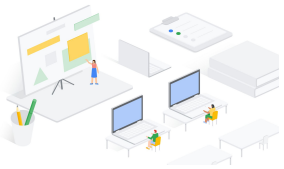
Innovating with Data and Google Cloud
Managing Security in Google Cloud
Application Development with Cloud Run
Tech Refresh Professional Cloud Architect
Architecting with Google Kubernetes Engine: Foundations

Session logistics

- When you have a question, please:
  - Click the Raise hand button in Google Meet.
  - Or add your question to the Q&A section of Google Meet.
  - Please note that answers may be deferred until the end of the session.
- These slides are available in the Student Lecture section of your Qwiklabs classroom.
- The session is **not recorded.**
- Google Meet does not have persistent chat.
  - If you get disconnected, you will lose the chat history.
  - Please copy any important URLs to a local text file as they appear in the chat.

Path to Service Excellence

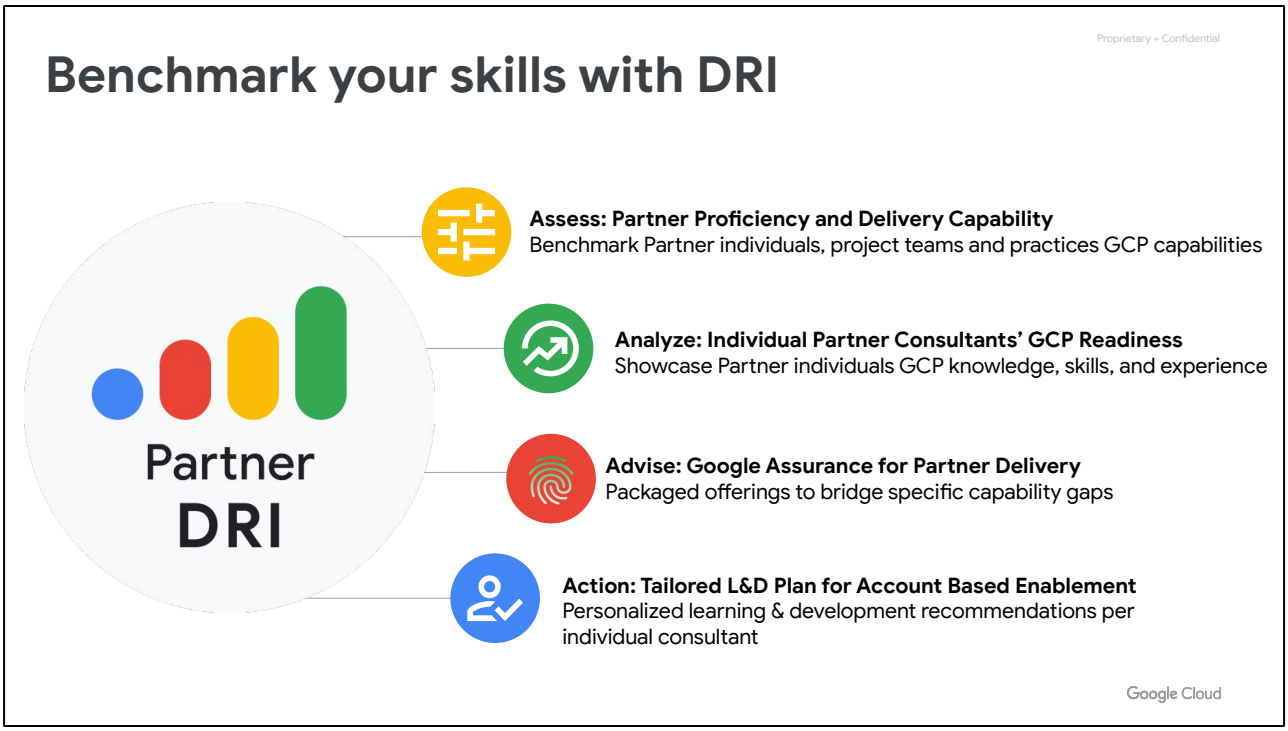
Certification



Advanced Solutions Training

Delivery Readiness Index

Certification is just one step on your professional journey. Google Cloud also offers our partners access to advanced solutions training, and a new quality-focused program called Delivery Readiness Index (DRI) to help you achieve service excellence with your customers.



DRI helps to benchmark partner proficiency and capability at any point during the customer journey however should be used primarily as a lead measure to predict and prepare for partner delivery success.

DRI assesses and analyzes Partner Consultant GCP proficiency by creating a DRI Profile inclusive of their GCP knowledge, skills, and experience.

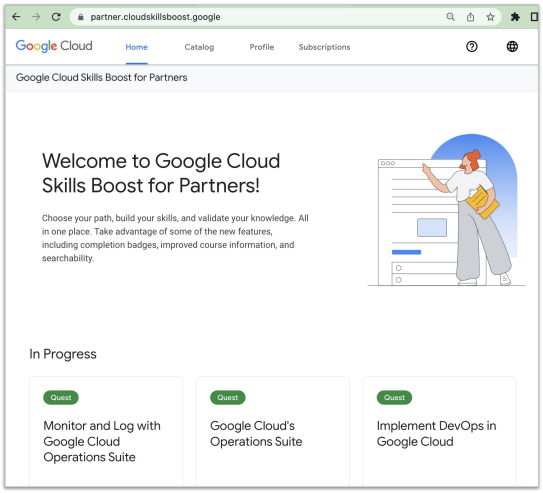
With the DRI insights, we can prescriptively advise the partner project team on the ground and bridge niche capability gaps.

DRI also takes action. For partner consultants, DRI generates a tailored L&D plan that prescribes personalized learning, training, and skill development to build GCP proficiency.

Google Cloud Skills Boost for Partners

https://partner.cloudskillsboost.google/

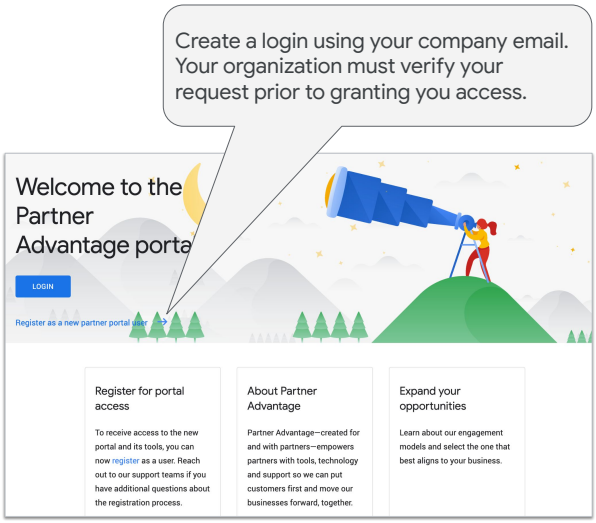
- On-demand course content
- Hands-on labs
- Skill Badges
- **FREE** to Google Cloud Partners!



Google Cloud

Partner Advantage

- Resources for Google Cloud partner organizations:
  - Recent announcements
  - Solutions/role-based training
  - Live/pre-recorded webinars on various topics
    - Partner Advantage Live Webinars
- Complements the certification self-study material presented on Google Cloud Skills Boost for Partners
- Helpful Links:
  - Getting started on Partner Advantage
  - Join Partner Advantage
  - Get help accessing Partner Advantage



https://www.partneradvantage.goog

Google Cloud

The getting started link:
https://support.google.com/googlecloud/topic/9198654#zippy=

Note the top section, **"Getting Started & User Guides"** and two key documents → Direct Partners to this if they need to enroll into Partner Advantage

1. Logging in to the Partner Advantage Portal - Quick Reference Guide
2. Enrolling in the Partner Advantage Program - Quick Reference Guide

Some context on enrolling in PA:
Access to Partner Portal is given in 2 ways

- Partner Admin Led: Partner Administrator at Partner Company can set up users
- User Led: User can go through Self Registration
  - https://www.partneradvantage.goog/GCPPRM/s/partneradvantageportallogin?language=en_US
  - Or directly to the User Registration Form, https://www.partneradvantage.goog/GCPPRM/s/partnerselfregistration?language=en_US

Please Note

- After a user self-registers, they receive an email that essentially states:
  - "Hi {Partner Name}, you are one step away from joining the Google Cloud Partner Advantage Community. Please click to continue with the

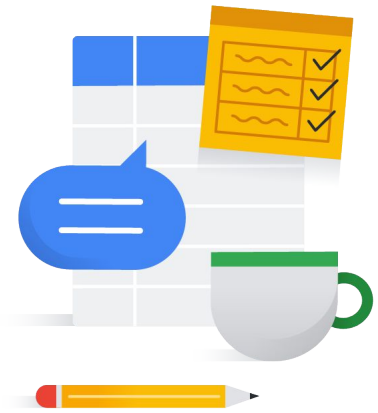
    - user registration process. See you in the cloud, The Partner Advantage Team
- Once registered, they can access limited content until their **Partner Administrator approves the user**
- Their Partner Administrator also receive an email notifying them that a member of their organization has registered themselves on their organization's Google Cloud Partner Advantage account.
    - It also states that this user has limited access to the portal
    - They are provided instructions on how to review and provision the appropriate access for the user that has registered
- Once their admin approves the user, they receive an email that states:
    - Hi {User Name}, Your Partner Administrator has updated your access to the Google Cloud Partner Advantage portal. You have been granted edit access to additional account information on the portal on behalf of your organization to help build your business. For additional access needs, please work with your Partner Administrator. See you in the cloud, The Partner Advantage Team

The net takeaway is, on the Support Page (the first link on this slide) Google Cloud Partner Advantage Support, there's a section **"Issue accessing Partner Advantage Portal? Click here for troubleshooting steps"**

- The source of their issue can be related to the different items shown
- Additionally, there's a Partner Administrator / Partner Adminstrator Team at their partner organization that has to approve their access.. Until that step is completed, they will have access issues/limitation. They will need to identify who this person or team is at their organization

Program issues or concerns?

- Problems with **accessing** Cloud Skills Boost for Partners
  - partner-training@google.com
- Problems with **a lab** (locked out, etc.)
  - support@qwiklabs.com
- Problems with accessing Partner Advantage
  - https://support.google.com/googlecloud/topic/9198654



Today's agenda

01 Program Overview

02 Accessing Course Content

03 Begin module 1 technical content review

Program Overview

Partner Certification Academy

A differentiated learning experience for the busy professional

Our goal is to help you prepare for Google Cloud certification exams

These programs may include:
- On-demand learning
- Self-paced labs
- Mentor-led workshops
- A voucher for the exam

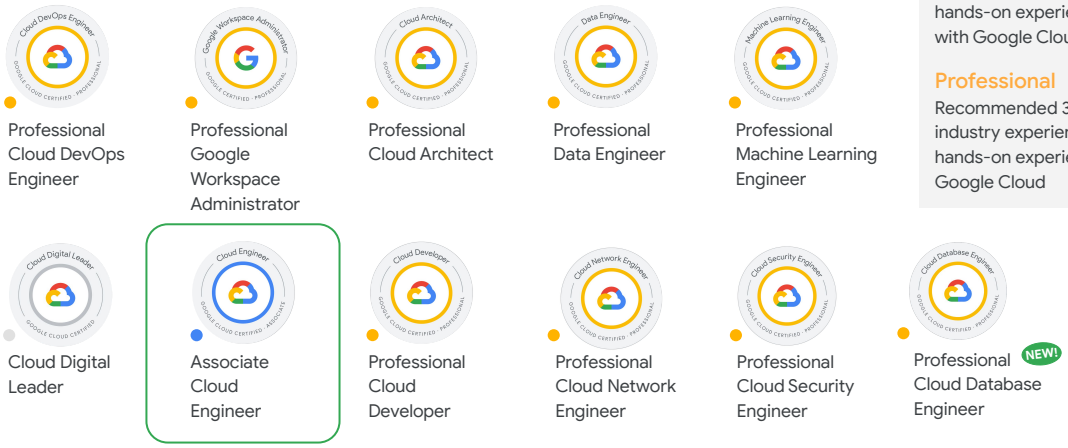
The workshop sessions:
- Are NOT training sessions - that's the purpose of the on-demand content.
- Help you review key concepts on the exam guide.
- Will NOT discuss actual exam questions.

Google Cloud is working with our partners to provide a differentiated learning experience - Partner Certification Academy (PCA)

These programs may include:
On-demand learning
Self-paced labs
Mentor-led workshops
A voucher for the exam

The workshop sessions:
Are NOT training sessions - that's the purpose of the on-demand content.
Help you review key concepts on the exam guide.
Will NOT discuss actual exam questions.

Google Cloud Certifications



Foundational
No cloud experience needed

Associate
Recommended 6+ months hands-on experience with Google Cloud

Professional
Recommended 3+ years industry experience & 1 year hands-on experience with Google Cloud

Google Cloud

More information:
https://cloud.google.com/certification#certification_paths

Associate Cloud Engineer

The Google Cloud Certified

Associate Cloud Engineer exam assesses your ability to:

- Setup a cloud solution environment
- Plan and configure a cloud solution
- Deploy and implement a cloud solution
- Ensure successful operation of a cloud solution
- Configure access and security

For more information:

https://cloud.google.com/certification/cloud-engineer

Associate Cloud Engineer
https://cloud.google.com/certification/cloud-engineer

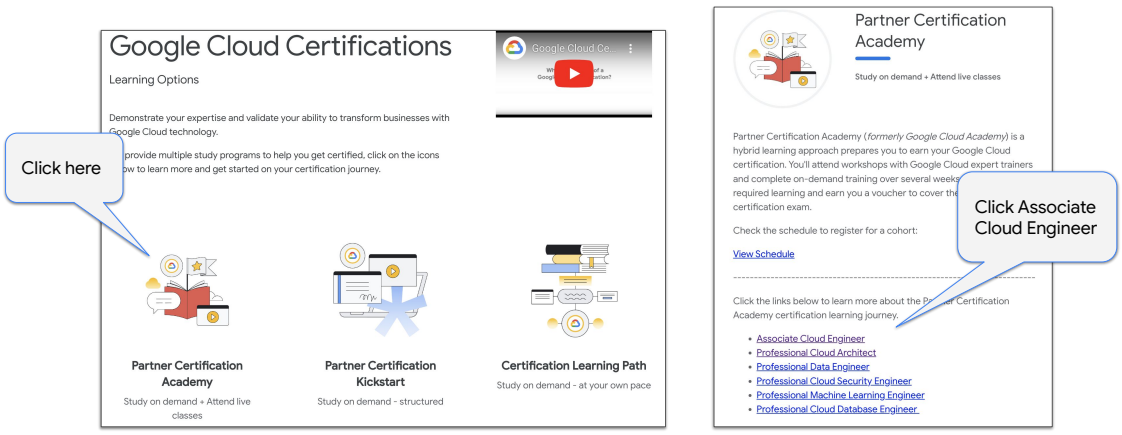
Exam Guide
https://cloud.google.com/certification/guides/cloud-engineer

Sample Questions
https://docs.google.com/forms/d/e/1FAIpQLSfexWKtXT2OSFJ-obA4iT3GmzgiOCGvjrT9OfxilWC1yPtmfQ/viewform

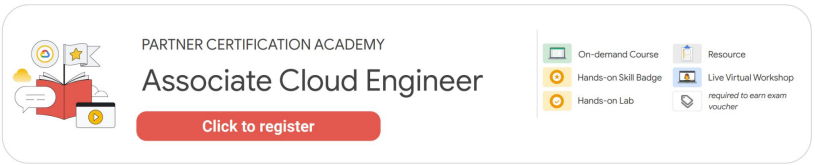
Accessing Partner Certification Academy Content

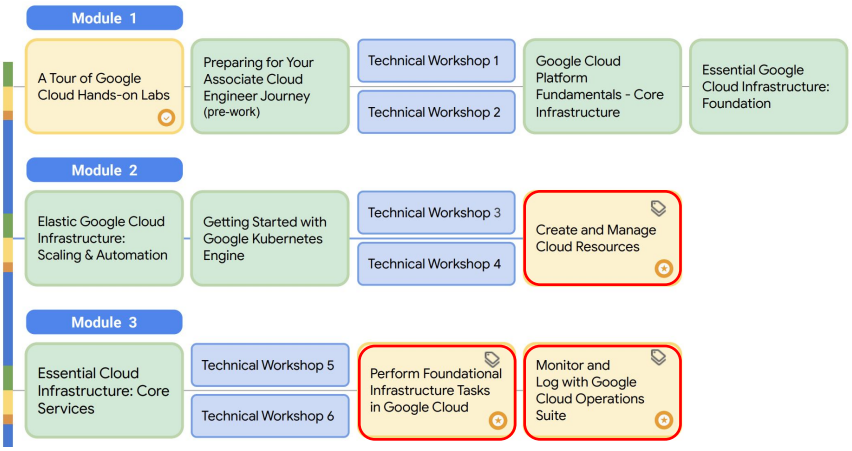
Learning Path - Partner Certification Academy Website

Go to: https://rsvp.withgoogle.com/events/partner-learning/google-cloud-certifications

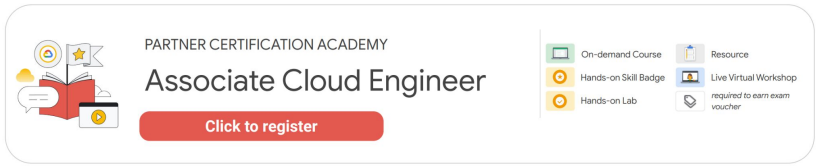


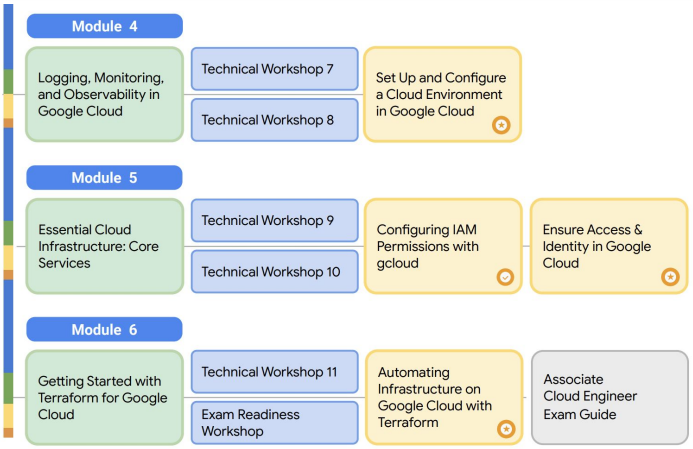
https://rsvp.withgoogle.com/events/partner-learning/google-cloud-certifications



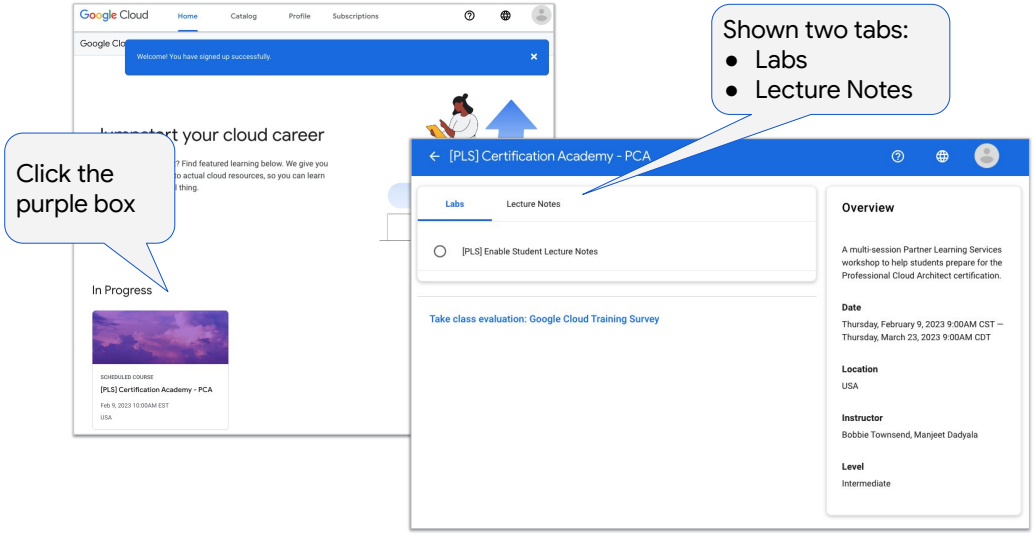


Needed for Exam Voucher





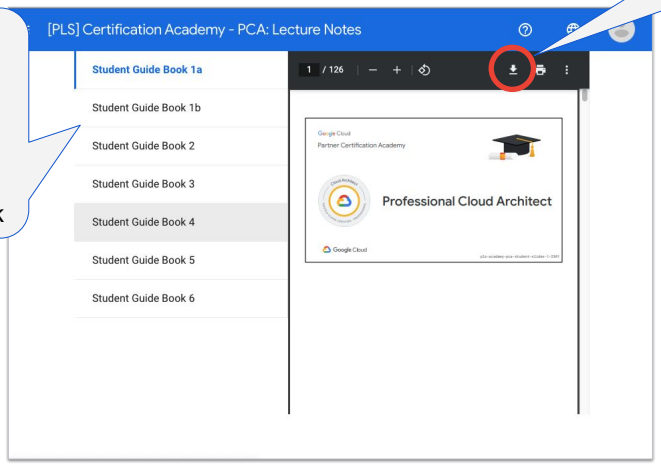
Sign into partner.cloudskillsboost.google



Downloading the lecture notes

Select the Lecture Notes tab

Select a Student Guide Book

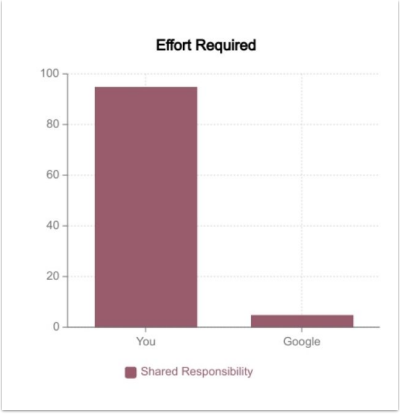


Click the download icon to download the PDF

Issues? Email: partner-training@google.com

Your Responsibilities

- **Workshop Day:** Meet for the cohort's weekly workshops (optional)
- **During the week:** Review material covered in the week's workshop, complete any course(s) as needed, perform hands-on labs, review additional suggested material.
- **Any time:** Reach out to your Mentor with questions



Important: You must allocate time between each weekly session to study and familiarize yourself with any prerequisite knowledge that will be covered in the workshops.
You will not pass the exam if you don't put in the work.

Experienced with AWS or Azure?

Speed your learning journey with:

- Google Cloud Fundamentals for Azure Professionals
- Google Cloud Fundamentals for AWS Professionals
- Compare AWS and Azure services to Google Cloud

Additional learning resources

- The Cloud Girl
  - https://github.com/priyankavergadia/GCPSketchnote
- Developer Cheat Sheet
  - https://googlecloudcheatsheet.withgoogle.com/
- Google Cloud product list
  - https://cloud.google.com/terms/services
- 21 products explained in under 2 minutes
  - https://cloud.google.com/blog/topics/inside-google-cloud/21-google-cloud-tools-each-explained-under-2-minutes

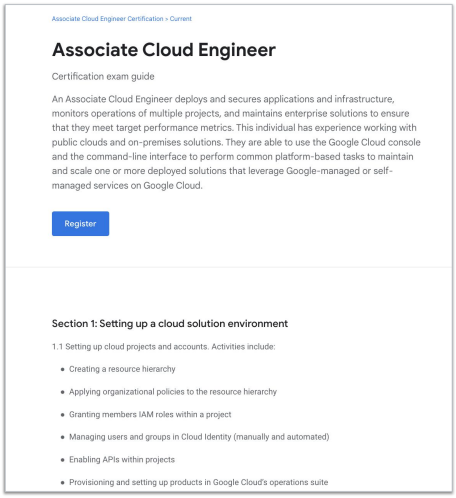
Associate Cloud Engineer (ACE) Exam Guide

Each module of this course covers Google Cloud services based on the topics in the ACE Exam Guide

The primary topics are:

- Compute Engine
- VPC Networks
- Google Kubernetes Engine
- Cloud Run, Cloud Functions and App Engine
- Cloud Storage and database options
- Resource Hierarchy/Identity and Access Management (IAM)
- Logging and Monitoring

Next discussion



Associate Cloud Engineer Certification > Current

Associate Cloud Engineer

Certification exam guide

An Associate Cloud Engineer deploys and secures applications and infrastructure, monitors operations of multiple projects, and maintains enterprise solutions to ensure that they meet target performance metrics. This individual has experience working with public clouds and on-premises solutions. They are able to use the Google Cloud console and the command-line interface to perform common platform-based tasks to maintain and scale one or more deployed solutions that leverage Google-managed or self-managed services on Google Cloud.

Register

Section 1: Setting up a cloud solution environment

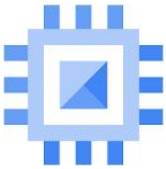
1.1 Setting up cloud projects and accounts. Activities include:

- Creating a resource hierarchy
- Applying organizational policies to the resource hierarchy
- Granting members IAM roles within a project
- Managing users and groups in Cloud Identity (manually and automated)
- Enabling APIs within projects
- Provisioning and setting up products in Google Cloud's operations suite

https://cloud.google.com/certification/guides/cloud-engineer/

Compute Engine

Exam Guide Overview - Compute Engine



Compute Engine

3.1 Deploying and implementing Compute Engine resources. Tasks include:

- 3.1.1 Launching a compute instance using the Google Cloud console and Cloud SDK (gcloud) (e.g., assign disks, availability policy, SSH keys)
- 3.1.2 Creating an autoscaled managed instance group using an instance template
- 3.1.3 Generating/uploading a custom SSH key for instances
- 3.1.4 Installing and configuring the Cloud Monitoring and Logging Agent
- 3.1.5 Assessing compute quotas and requesting increases

2.2 Planning and configuring compute resources. Considerations include:

- 2.2.1 Selecting appropriate compute choices for a given workload (e.g., Compute Engine, Google Kubernetes Engine, Cloud Run, Cloud Functions)
- 2.2.2 Using preemptible VMs and custom machine types as appropriate

2.3 Planning and configuring data storage options. Considerations include:

- 2.3.1 Product choice (e.g., Cloud SQL, BigQuery, Firestore, Cloud Spanner, Cloud Bigtable)
- 2.3.2 Choosing storage options (e.g., Zonal persistent disk, Regional balanced persistent disk, Standard, Nearline, Coldline, Archive)

4.1 Managing Compute Engine resources. Tasks include:

- 4.1.1 Managing a single VM instance (e.g., start, stop, edit configuration, or delete an instance)
- 4.1.2 Remotely connecting to the instance
- 4.1.3 Attaching a GPU to a new instance and installing necessary dependencies
- 4.1.4 Viewing current running VM inventory (instance IDs, details)
- 4.1.5 Working with snapshots (e.g., create a snapshot from a VM, view snapshots, delete a snapshot)
- 4.1.6 Working with images (e.g., create an image from a VM or a snapshot, view images, delete an image)
- 4.1.7 Working with instance groups (e.g., set autoscaling parameters, assign instance template, create an instance template, remove instance group)
- 4.1.8 Working with management interfaces (e.g., Google Cloud console, Cloud Shell, Cloud SDK)

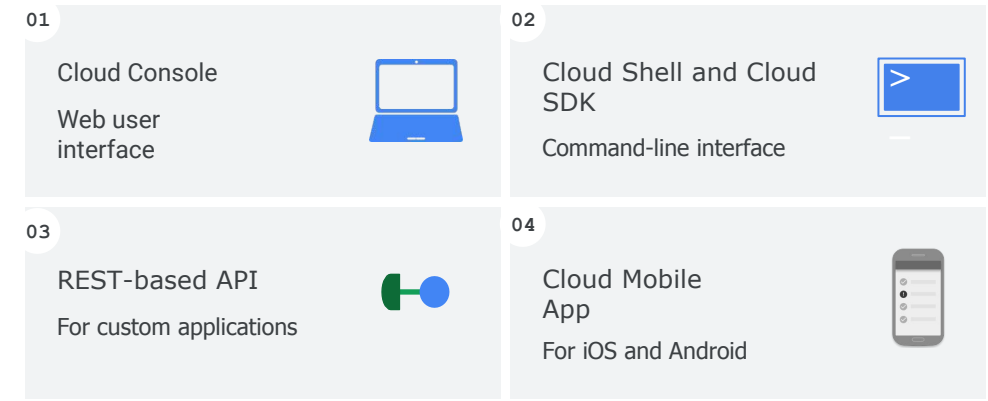
Exam Guide - Compute Engine

4.1 Managing Compute Engine resources. Tasks include:

- 4.1.1 Managing a single VM instance (e.g., start, stop, edit configuration, or delete an instance)
- 4.1.2 Remotely connecting to the instance
- 4.1.3 Attaching a GPU to a new instance and installing necessary dependencies
- 4.1.4 Viewing current running VM inventory (instance IDs, details)
- 4.1.5 Working with snapshots (e.g., create a snapshot from a VM, view snapshots, delete a snapshot)
- 4.1.6 Working with images (e.g., create an image from a VM or a snapshot, view images, delete an image)
- 4.1.7 Working with instance groups (e.g., set autoscaling parameters, assign instance template, create an instance template, remove instance group)
- 4.1.8 Working with management interfaces (e.g., Google Cloud console, Cloud Shell, Cloud SDK)

Next discussion

There are four ways to interact with Google Cloud



First is the Google **Cloud Console**, which is Google Cloud's Graphical User Interface (GUI) which helps you deploy, scale, and diagnose production issues in a simple web-based interface. With the Cloud Console, you can easily find your resources, check their health, have full management control over them, and set budgets to control how much you spend on them. The Cloud Console also provides a search facility to quickly find resources and connect to instances via SSH in the browser.

Cloud Console provides web-based interaction

- Simple web-based graphical user interface
- Easily find resources, check their health, have full management control over them, and set budgets
- Provides a search facility to quickly find resources and connect to instances via SSH in the browser

First is the Google **Cloud Console**, which is Google Cloud's Graphical User Interface (GUI) which helps you deploy, scale, and diagnose production issues in a simple web-based interface. With the Cloud Console, you can easily find your resources, check their health, have full management control over them, and set budgets to control how much you spend on them. The Cloud Console also provides a search facility to quickly find resources and connect to instances via SSH in the browser.

Cloud SDK is a collection of command line tools

- Set of tools to manage resources and applications hosted on Google Cloud
- Includes:
  - **gcloud tool** - Provides the main command-line interface for Google Cloud products and services
  - **gsutil** - Provides access to Cloud Storage from the command line
  - **bq** - A command-line tool for BigQuery

Cloud SDK: https://cloud.google.com/sdk/docs/

The **Cloud SDK** is a set of tools that you can use to manage resources and applications hosted on Google Cloud. These include the gcloud tool, which provides the main command-line interface for Google Cloud products and services, as well as gsutil, which lets you access Cloud Storage from the command line, and bq, a command-line tool for BigQuery. When installed, all of the tools within the Cloud SDK are located under the bin directory.

`gcloud components list` shows what is currently installed

Cloud Shell provides command line access to resources

- Provides command-line access to cloud resources directly from a browser
- Linux-based virtual machine with a persistent 5-GB home directory
- The Cloud SDK gcloud command and other utilities are always installed, available, up to date, and fully authenticated

```
$ gcloud compute instances list
```

```
NAME          ZONE            INTERNAL_IP   EXTERNAL_IP
nginxstack-1  us-central1-f   10.128.0.3    35.238.84.245
nginxstack-2  Us-central1-f   10.128.0.4    35.225.177.18
nginxstack-3  Us-central1-f   10.128.0.2    35.239.250.238
```


How Cloud Shell works:
https://cloud.google.com/shell/docs/how-cloud-shell-works

Cloud Shell provides command-line access to cloud resources directly from a browser. Cloud Shell is a Debian-based virtual machine with a persistent 5-GB home directory, which makes it easy to manage Google Cloud projects and resources. With Cloud Shell, the Cloud SDK gcloud command and other utilities are always installed, available, up to date, and fully authenticated.

APIs allow code to control your Cloud resources

- Google Cloud services offer APIs that allow code to be written to control them
- The Google APIs Explorer shows what APIs are available, and in what versions
- Google provides Cloud Client and Google API Client libraries

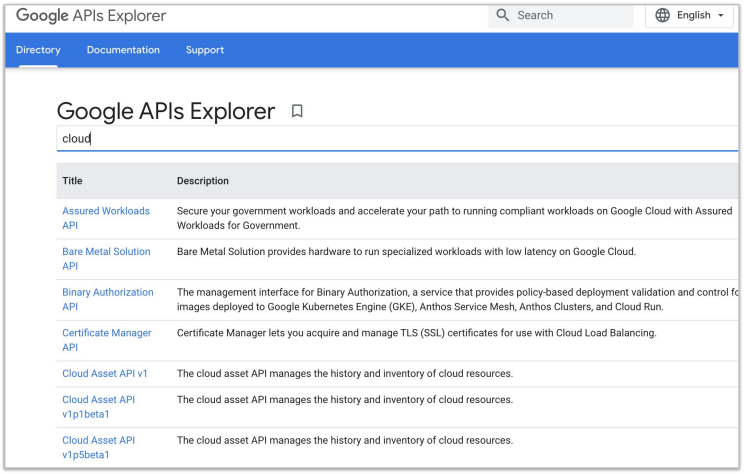
The services that make up Google Cloud offer **APIs**, so that code you write can control them. The Cloud Console includes a tool called the Google APIs Explorer that shows what APIs are available, and in what versions. You can try these APIs interactively, even those that require user authentication.

Suppose you've explored an API, and you're ready to build an application that uses it. Do you have to start coding from scratch? No. Google provides Cloud Client and Google API Client libraries in many popular languages to take a lot of the drudgery out of the task of calling Google Cloud from your code. Languages currently represented in these libraries are: Java, Python, PHP, C#, Go, Node.js, Ruby and C++.

API Explorer provides documentation on Google APIs

Interactive tool to try APIs in a browser

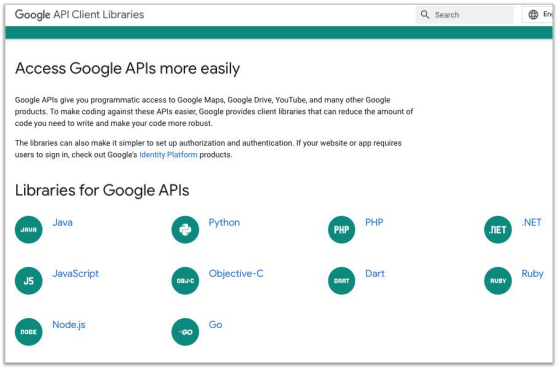
- Find methods for each API and what parameters they support
- Execute requests and see real-time response
- API calls can be authenticated



https://developers.google.com/apis-explorer

Use client libraries to control Google Cloud resources from code

- **Cloud Client Libraries**
  - Community-owned, hand-crafted client libraries
- **Google API Client Libraries**
  - Open-source, generated
  - Support various languages



Cloud Client Libraries

https://developers.google.com/api-client-library

Example: Using the Node.js client library to list VM instances

`gcloud` command which makes the same API call

```
gcloud compute instances list
```

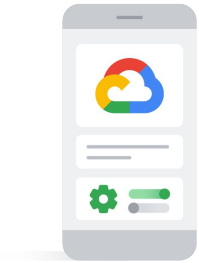
```
/**
 * TODO(developer): Uncomment and replace these variables before running the sample.
 */
// const projectId = 'YOUR_PROJECT_ID';
// const zone = 'europe-central2-b'

const compute = require('@google-cloud/compute');

// List all instances in the given zone in the specified project.
async function listInstances() {
  const instancesClient = new compute.InstancesClient();

  const [instanceList] = await instancesClient.list({
    project: projectId,
    zone,
  });

  console.log(`Instances found in zone ${zone}:`);

  for (const instance of instanceList) {
    console.log(` - ${instance.name} (${instance.machineType})`);
  }
}

listInstances();
```

Retrieving the list of instances

Sending the instance names and machine types to the log

From
https://cloud.google.com/compute/docs/api/libraries

Manage your resources with Cloud Console Mobile App



cloud.google.com/console-app

- Start, stop, and use SSH to connect into Compute Engine instances, and see logs
- Stop and start Cloud SQL instances
- Up-to-date billing information for projects and alerts for those going over budget
- Customizable graphs showing key metrics
- Alerts and incident management

Google Cloud app
https://cloud.google.com/app

The **Cloud Console Mobile App** can be used to start, stop, and use ssh to connect to Compute Engine instances, and to see logs from each instance. It also lets you stop and start Cloud SQL instances. Additionally, you can administer applications deployed on App Engine, by viewing errors, rolling back deployments, and changing traffic splitting.

The Cloud Console Mobile App provides up-to-date billing information for your projects, and billing alerts for projects that are going over budget.

You can set up customizable graphs showing key metrics such as CPU usage, network usage, requests per second, and server errors.

The mobile app also offers alerts and incident management.

Exam Guide - Compute Engine

3.1 Deploying and implementing Compute Engine resources. Tasks include:

- 3.1.1 Launching a compute instance using the Google Cloud console and Cloud SDK (gcloud) (e.g., assign disks, availability policy, SSH keys)
- 3.1.2 Creating an autoscaled managed instance group using an instance template
- 3.1.3 Generating/uploading a custom SSH key for instances
- 3.1.4 Installing and configuring the Cloud Monitoring and Logging Agent
- 3.1.5 Assessing compute quotas and requesting increases

2.2 Planning and configuring compute resources. Considerations include:

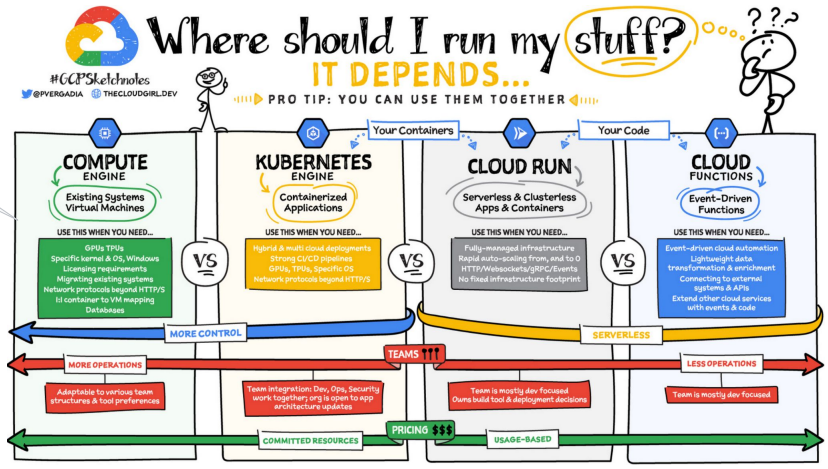
- 2.2.1 Selecting appropriate compute choices for a given workload (e.g., Compute Engine, Google Kubernetes Engine, Cloud Run, Cloud Functions)
- 2.2.2 Using preemptible VMs and custom machine types as appropriate

2.3 Planning and configuring data storage options. Considerations include:

- 2.3.1 Product choice (e.g., Cloud SQL, BigQuery, Firestore, Cloud Spanner, Cloud Bigtable)
- 2.3.2 Choosing storage options (e.g., Zonal persistent disk, Regional balanced persistent disk, Standard, Nearline, Coldline, Archive)

Choosing a Google Cloud compute option

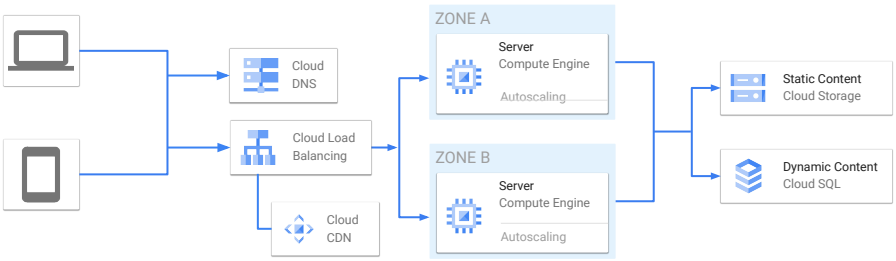
Focus of this section



Where should I run my stuff? Choosing a Google Cloud compute option

Compute Options - Where should I run my stuff?
https://cloud.google.com/blog/topics/developers-practitioners/where-should-i-run-my-stuff-choosing-google-cloud-compute-option

Use Compute Engine when you need control over operating systems or lift & shift situations



Tip: Look through all the links in What is Compute Engine? Use cases, security, pricing and more

What is Compute Engine? Use cases, security, pricing and more
https://cloud.google.com/blog/topics/developers-practitioners/what-compute-engine-use-cases-security-pricing-and-more

Compute Engine is a great solution when you need complete control over your operating systems, or an application that is a database.

Instance groups and autoscaling as shown on this slide allow you to meet variations in demand on your application. These will be discussed later in this module

Compute Engine Machine Types

Tip: More detail and use cases found here: https://cloud.google.com/compute#section-6

Google Cloud

Choosing the right VM type
https://cloud.google.com/compute#section-6

Tau VMs - Scale-out Optimized Machine Types

Best price / performance and full x86 compatibility

No application porting required !

- **Tau**, a new family of virtual machines, delivers **42% better price-performance among leading cloud providers** for scale-out apps
- First instances under the Tau VM family are based on 3rd gen AMD EPYC processors, **preserving full x86 compatibility.**
- Also supported in **GKE** (Google Kubernetes Engine), just add T2D to your GKE node pools

Ideal for scale-out workloads:

- Web servers
- Containerized microservices
- Data-logging processing
- Media transcoding
- Large-scale Java applications

Tau VMs ideal for scale-out workloads including web servers, containerized microservices, data-logging processing, media transcoding, and large-scale Java applications.

Compute Optimized Machine Types

Performance sensitive for CPU workloads.
Or, licensed applications that may benefit from more powerful cores.

- **High-performance Web Servers**
- **Gaming (AAA Game Servers)**
- **High Performance Computing (HPC)**
  - Simulations (Finite Element Analysis, Oil & Gas, CFD, Monte Carlo, Product Simulation, Weather, Physics, Chemistry)
  - Financial Services (Financial analysis & simulation workloads)
  - Genomic Analysis
- **Media Transcoding**
- **Electronic Design Automation**

Memory Optimized Machine Types

Designed for high-end business critical applications *Optimized for SAP*

- In-memory databases
  SAP HANA
- In-memory Analytics
- In memory persistent cache

Accelerator Optimized (GPUs) Machine Types

A2 VM Family: Newest NVIDIA GPUs, Optimized for ML, HPC and other parallelized CUDA Compute Workloads

Optimized Hardware

- Newest, Highest Performance NVIDIA A100 GPUs with 40 GB Memory
- Up to 96 Cascade Lake vCPUs and 1.3T memory
- *a2-highgpu* fixed VM shapes w/ 1,2,4 or 8 GPUs for scale-out workloads
- *a2-megagpu* VM shape w/**16 GPUs** for scale-up workloads
- NVLINK, optional Local SSD and optimized 100 Gbps Networking

Optimized Virtualization

- New transparent Numa topology, enabling maximum system throughput
- Memory speed improvements via 1G pages

Giant boost for all GPU accelerated Workloads

- ML Training - Create NexGen AI and maximize data scientist productivity
- ML Inference - Performance and versatility for running AI at Scale
- HPC - Faster & larger scale simulation to unlock new discoveries
- Data Analytics - Faster analytics with larger data set

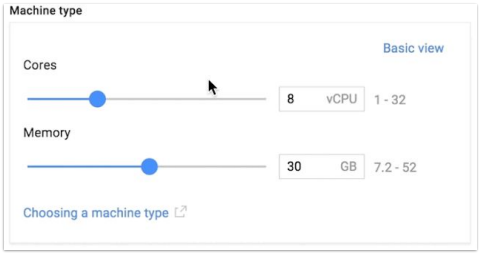
Google Cloud

CUDA:
Nvidia calls its parallel processing platform CUDA. CUDA Cores are the processing units inside a GPU just like AMD's Stream Processors.

CUDA is an abbreviation for Compute Unified Device Architecture. It is a name given to the parallel processing platform and API which is used to access the Nvidia GPUs instruction set directly.

Custom Machine Types

- Specify number of vCPU cores the memory
- Optimize resources
- Manage costs



Custom machine types:
https://cloud.google.com/custom-machine-types

A machine type specifies a particular collection of virtualized hardware resources available to a VM instance, including the system memory size, vCPU count, and maximum persistent disk capability.

Predefined machine types:

- Have a fixed collection of resources, are managed by Compute Engine and are available in multiple different classes.
- Each class has a predefined ratio of GB of memory per vCPU

Custom machine types:

- These let you specify the number of vCPUs and the amount of memory for your instance.

Exam Guide - Compute Engine

3.1 Deploying and implementing Compute Engine resources. Tasks include:

- 3.1.1 Launching a compute instance using the Google Cloud console and Cloud SDK (gcloud) (e.g., assign disks, availability policy, SSH keys)
- 3.1.2 Creating an autoscaled managed instance group using an instance template
- 3.1.3 Generating/uploading a custom SSH key for instances
- 3.1.4 Installing and configuring the Cloud Monitoring and Logging Agent
- 3.1.5 Assessing compute quotas and requesting increases

2.2 Planning and configuring compute resources. Considerations include:

- 2.2.1 Selecting appropriate compute choices for a given workload (e.g., Compute Engine, Google Kubernetes Engine, Cloud Run, Cloud Functions)
- 2.2.2 Using preemptible VMs and custom machine types as appropriate

2.3 Planning and configuring data storage options. Considerations include:

- 2.3.1 Product choice (e.g., Cloud SQL, BigQuery, Firestore, Cloud Spanner, Cloud Bigtable)
- 2.3.2 Choosing storage options (e.g., Zonal persistent disk, Regional balanced persistent disk, Standard, Nearline, Coldline, Archive)

Preemptible and Spot VMs

- A highly discounted VM compared to the price of standard VMs
  - Discount of 60-91% discount
  - Availability depends on having excess compute capacity in a zone
    - May or may not have availability in a given zone at a given time
    - Will have to try another zone or wait for the resource to be available
- Compute Engine might stop preemptible/spot instances at any time due to system events
  - Preemptible VMs - always stopped after they run for 24 hours.
    - May be stopped before the 24 hour time period
    - When restarted, the 24 hour clock resets
  - Spot VMs - stopped/deleted when Google needs the resource elsewhere
    - Spot VMs are the latest version of preemptible VMs
    - Can specify termination or deletion when creating the VM

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

Creating a preemptible VM:
https://cloud.google.com/compute/docs/instances/create-use-preemptible#creating_a_preemptible_vm

Kubernetes use case:
https://cloud.google.com/kubernetes-engine/docs/how-to/preemptible-vms

Spot VM Instances:
https://cloud.google.com/compute/docs/instances/spot

Using Spot instances with GKE:
https://cloud.google.com/kubernetes-engine/docs/concepts/spot-vms

Top 5 use cases for Google Cloud Spot VMs explained:
https://cloud.google.com/blog/products/compute/google-cloud-spot-vm-use-cases-and-best-practices

Preemptible/Spot VMs - additional details

- Offer the same machine types, options, and performance as regular compute instances
- Use cases
  - Stateless and scalable workloads that can be stopped and checkpointed in less than 30 seconds, or is location and hardware flexible
- Provides no live migration or automatic restart during maintenance events
- Not covered by Service Level Agreement due to the preceding limitations,
- No free tier

Tip: Look through all the links in Top 5 use cases for Google Cloud Spot VMs explained + best practices

Exam Guide - Compute Engine

3.1 Deploying and implementing Compute Engine resources. Tasks include:

- 3.1.1 Launching a compute instance using the Google Cloud console and Cloud SDK (gcloud) (e.g., assign disks, availability policy, SSH keys)
- 3.1.2 Creating an autoscaled managed instance group using an instance template
- 3.1.3 Generating/uploading a custom SSH key for instances
- 3.1.4 Installing and configuring the Cloud Monitoring and Logging Agent
- 3.1.5 Assessing compute quotas and requesting increases

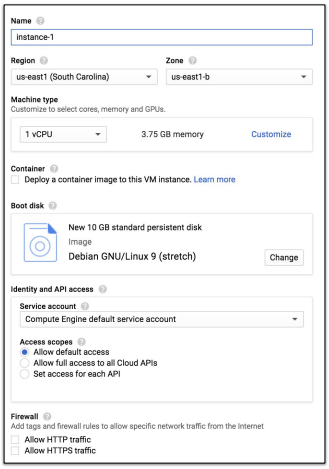
2.2 Planning and configuring compute resources. Considerations include:

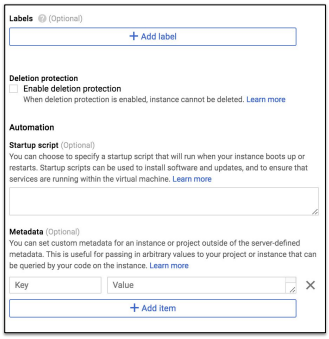
- 2.2.1 Selecting appropriate compute choices for a given workload (e.g., Compute Engine, Google Kubernetes Engine, Cloud Run, Cloud Functions)
- 2.2.2 Using preemptible VMs and custom machine types as appropriate

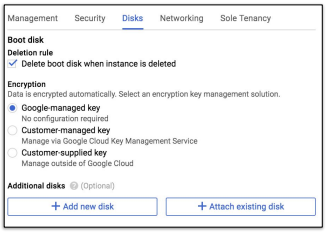
2.3 Planning and configuring data storage options. Considerations include:

- 2.3.1 Product choice (e.g., Cloud SQL, BigQuery, Firestore, Cloud Spanner, Cloud Bigtable)
- 2.3.2 Choosing storage options (e.g., Zonal persistent disk, Regional balanced persistent disk, Standard, Nearline, Coldline, Archive)

Creating VMs using Cloud Console







These screenshots represent the windows you will see when creating a virtual machine using the Google Cloud Console.

The screen on the left lets you set some basic details about your server, including the name, size, operating system, and permissions.

The middle screen is used to set labels, deletion protection, and bootstrapping options.

The right screen sets up your virtual disk parameters, including encryption options.

Creating VMs with the CLI

Partial syntax for `gcloud compute instances create` command

```
gcloud compute instances create INSTANCE_NAMES [INSTANCE_NAMES ...]
  [--accelerator=[count=COUNT],[type=TYPE]] [--async]
  [--no-boot-disk-auto-delete]
  [--boot-disk-device-name=BOOT_DISK_DEVICE_NAME]
  [--boot-disk-provisioned-iops=BOOT_DISK_PROVISIONED_IOPS]
  [--boot-disk-size=BOOT_DISK_SIZE] [--boot-disk-type=BOOT_DISK_TYPE]
  [--can-ip-forward] [--confidential-compute]
  [--create-disk=[PROPERTY=VALUE,...]] [--csek-key-file=FILE]
  [--deletion-protection] [--description=DESCRIPTION]
  [--disk=[auto-delete=AUTO-DELETE],[boot=BOOT],
  [device-name=DEVICE-NAME],[mode=MODE],[name=NAME],[scope=SCOPE]]
  [--enable-display-device] [--[no-]enable-nested-virtualization]
  [--[no-]enable-uefi-networking] [--hostname=HOSTNAME]
  [--instance-termination-action=INSTANCE_TERMINATION_ACTION]
  [--ipv6-network-tier=IPV6_NETWORK_TIER]
  [--ipv6-public-ptr-domain=IPV6_PUBLIC_PTR_DOMAIN]
  [--labels=[KEY=VALUE,...]] [--local-ssd=[device-name=DEVICE-NAME],
  [interface=INTERFACE]] [--machine-type=MACHINE_TYPE]
  [--maintenance-policy=MAINTENANCE_POLICY] [--metadata=KEY=VALUE,
  [KEY=VALUE,...]] [--metadata-from-file=KEY=LOCAL_FILE_PATH,[...]]
  [--min-cpu-platform=PLATFORM] [--min-node-cpu=MIN_NODE_CPU]
  [--network=NETWORK] [--network-interface=[PROPERTY=VALUE,...]]
  [--network-performance-configs=[PROPERTY=VALUE,...]]
  [--network-tier=NETWORK_TIER] [--preemptible]
  [--private-ipv6-google-access-type=PRIVATE_IPV6_GOOGLE_ACCESS_TYPE]
  [--private-network-ip=PRIVATE_NETWORK_IP]
  [--provisioning-model=PROVISIONING_MODEL]
```


Here is the command line version:
https://cloud.google.com/sdk/gcloud/reference/compute/instances/create

Creating VMs with the CLI - examples

- Standard VM
  - `gcloud compute instances create vm1 --zone=us-central1-a --machine-type=e2-medium`
- Preemptible VM
  - `gcloud compute instances create vm1 --zone us-central1-b --machine-type=e2-medium` **`--preemptible`**
- Spot VM
  - `gcloud compute instances create vm1 --zone us-central1-b --machine-type=e2-medium` **`--provisioning-model=SPOT`**
- Custom VM
  - `gcloud compute instances create my-vm --custom-cpu 4 --custom-memory 5`

Exam Guide - Compute Engine

3.1 Deploying and implementing Compute Engine resources. Tasks include:

- 3.1.1 Launching a compute instance using the Google Cloud console and Cloud SDK (gcloud) (e.g., **assign disks**, availability policy, SSH keys)
- 3.1.2 Creating an autoscaled managed instance group using an instance template
- 3.1.3 Generating/uploading a custom SSH key for instances
- 3.1.4 Installing and configuring the Cloud Monitoring and Logging Agent
- 3.1.5 Assessing compute quotas and requesting increases

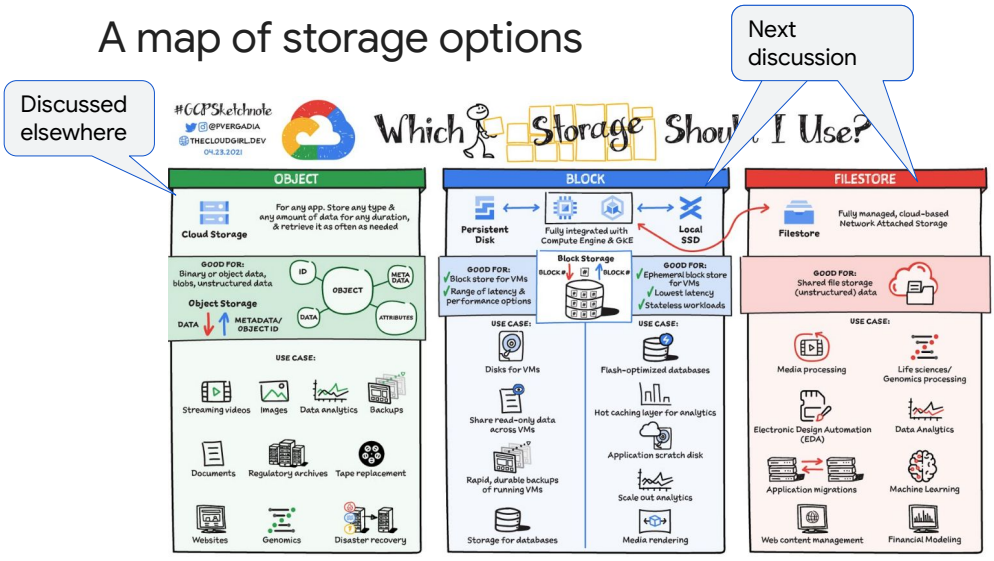
2.2 Planning and configuring compute resources. Considerations include:

- 2.2.1 Selecting appropriate compute choices for a given workload (e.g., Compute Engine, Google Kubernetes Engine, Cloud Run, Cloud Functions)
- 2.2.2 Using preemptible VMs and custom machine types as appropriate

2.3 Planning and configuring data storage options. Considerations include:

- 2.3.1 Product choice (e.g., Cloud SQL, BigQuery, Firestore, Cloud Spanner, Cloud Bigtable)
- 2.3.2 Choosing storage options (e.g., **Zonal persistent disk, Regional balanced persistent disk,** Standard, Nearline, Coldline, Archive)

A map of storage options



A map of storage options in Google Cloud

A map of storage options in Google Cloud
https://cloud.google.com/blog/topics/developers-practitioners/map-storage-options-google-cloud

Compute Engine storage options

- Each VM has a single boot persistent disk (PD) that contains the operating system
  - Not best practice to place non-system data on the boot disk
- When apps require additional storage space add one or more additional storage options to your instance
  - **Zonal persistent disk**: Efficient, reliable block storage.
  - **Regional persistent disk**: Regional block storage replicated in two zones.
  - **Local SSD:** High performance, transient, local block storage.
  - **Cloud Storage buckets**: Affordable object storage (discussed later)
  - **Filestore**: High performance NFSv3 file storage for Google Cloud users.

Link to disk types: https://cloud.google.com/compute/docs/disks#disk-types
Link to extreme disk:
https://cloud.google.com/compute/docs/disks/extreme-persistent-disk

The first disk that we create is what we call a persistent disk. That means it's going to be attached to the VM through the network interface. Even though it's persistent, it's not physically attached to the machine. This separation of disk and compute allows the disk to survive if the VM terminates. You can also perform snapshots of these disks, which are incremental backups that we'll discuss later.

The choice between HDD and SSD disks comes down to cost and performance. To learn more about disk performance and how it scales with disk size, please refer to the documentation page.

Another cool feature of persistent disks is that you can dynamically resize them, even while they are running and attached to a VM.

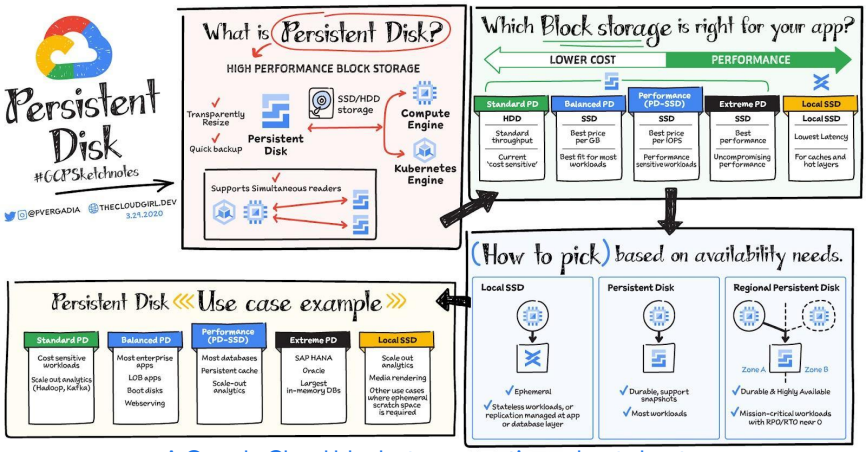
You can also attach a disk in read-only mode to multiple VMs. This allows you to share static data between multiple instances, which is cheaper than replicating your data to unique disks for individual instances.

Zonal persistent disks offer efficient, reliable block storage. Regional persistent disks provide active-active disk replication across two zones in the same region. Regional persistent disks deliver durable storage that is synchronously replicated across zones and are a great option for high-performance databases and enterprise applications that also require high availability. When you configure a zonal or regional persistent disk, you can select one of the following disk types.

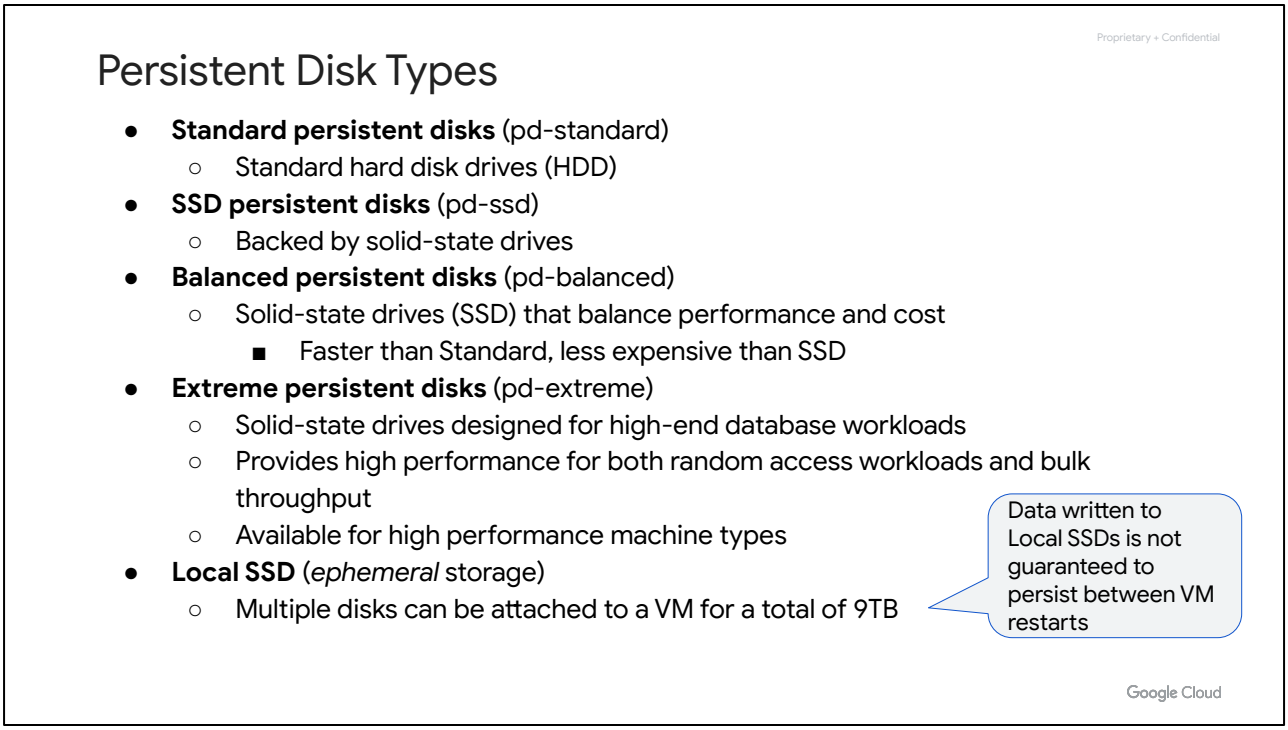
- Standard persistent disks (`pd-standard`). These types of disks are back by standard hard disk drives (HDD).
- Balanced persistent disks (`pd-balanced`). These types of disks are backed by solid state drives (SSD). They are an alternative to SSD persistent disks that balance performance and cost.
- SSD persistent disks (`pd-ssd`). These types of disks are backed by by solid state drives (SSD).

By default, Compute Engine encrypts all data at rest. Google Cloud handles and manages this encryption for you without any additional actions on your part. However, if you wanted to control and manage this encryption yourself, you can either use Cloud Key Management Service to create and manage key encryption keys (which is known as customer-managed encryption keys) or create and manage your own key encryption keys (known as customer-supplied encryption keys).

Persistent Disk



[A Google Cloud block storage options cheat sheet]



Link to disk types: https://cloud.google.com/compute/docs/disks#disk-types
Link to extreme disk:
https://cloud.google.com/compute/docs/disks/extreme-persistent-disk
Local ssd: https://cloud.google.com/compute/docs/disks/local-ssd

The first disk that we create is what we call a persistent disk. That means it's going to be attached to the VM through the network interface. Even though it's persistent, it's not physically attached to the machine. This separation of disk and compute allows the disk to survive if the VM terminates. You can also perform snapshots of these disks, which are incremental backups that we'll discuss later.

The choice between HDD and SSD disks comes down to cost and performance. To learn more about disk performance and how it scales with disk size, please refer to the documentation page.

Another cool feature of persistent disks is that you can dynamically resize them, even while they are running and attached to a VM.

You can also attach a disk in read-only mode to multiple VMs. This allows you to share static data between multiple instances, which is cheaper than replicating your

data to unique disks for individual instances.

Zonal persistent disks offer efficient, reliable block storage. Regional persistent disks provide active-active disk replication across two zones in the same region. Regional persistent disks deliver durable storage that is synchronously replicated across zones and are a great option for high-performance databases and enterprise applications that also require high availability. When you configure a zonal or regional persistent disk, you can select one of the following disk types.

- Standard persistent disks (`pd-standard`). These types of disks are back by standard hard disk drives (HDD).
- Balanced persistent disks (`pd-balanced`). These types of disks are backed by solid state drives (SSD). They are an alternative to SSD persistent disks that balance performance and cost.
- SSD persistent disks (`pd-ssd`). These types of disks are backed by by solid state drives (SSD).

By default, Compute Engine encrypts all data at rest. Google Cloud handles and manages this encryption for you without any additional actions on your part. However, if you wanted to control and manage this encryption yourself, you can either use Cloud Key Management Service to create and manage key encryption keys (which is known as customer-managed encryption keys) or create and manage your own key encryption keys (known as customer-supplied encryption keys).

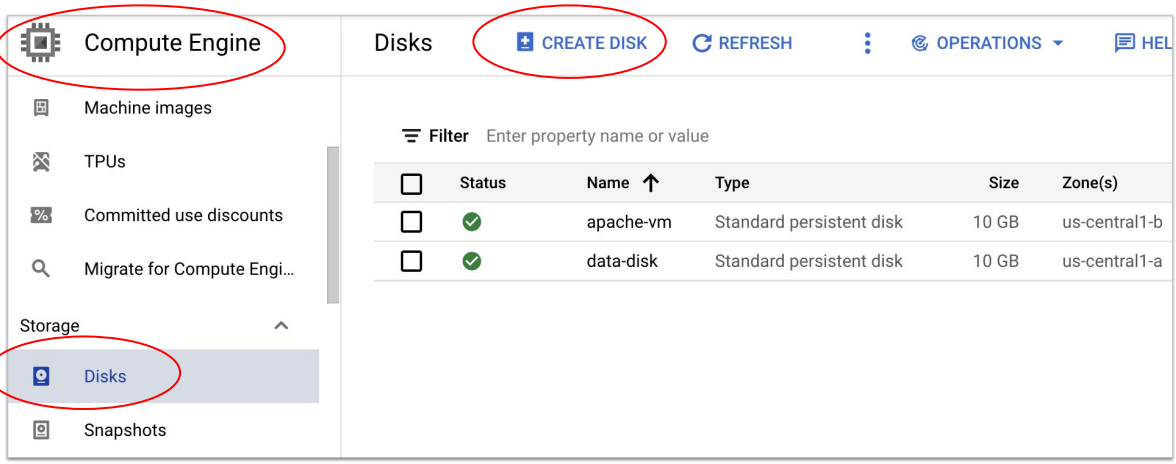
Zonal vs Regional disks

- Balanced, SSD and Standard disk types can be zonal or regional
  - Regional disks provide high availability
    - Synchronously replicates data between two zones in a region
    - Can be used in the event of a zonal outage where VM becomes unavailable
      - Spin up a VM in the secondary zone and force attach the disk
        - Time to recover - time to create VM (several minutes) + time to force attach disk (~1 minute)

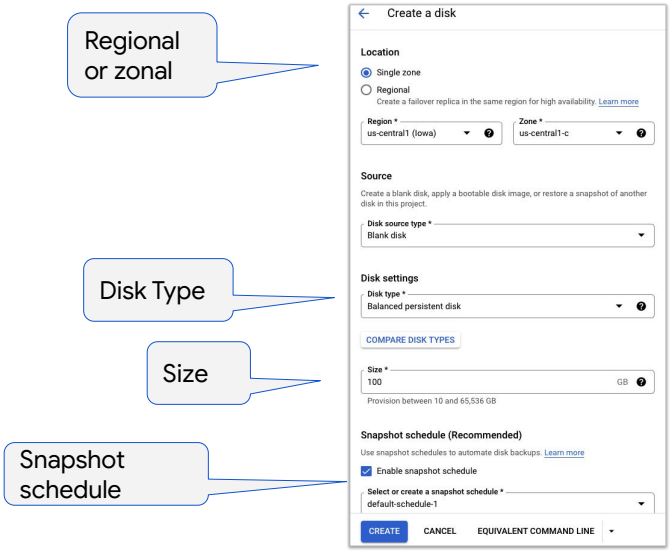
```
gcloud compute instances attach-disk myvm2 \
     --disk data-disk --disk-scope=regional \
     --force-attach
```


High availability options using regional PDs
https://cloud.google.com/compute/docs/disks/high-availability-regional-persistent-disk

Creating disks with the Console



Creating disks with the Console (continued)



Creating disks with the CLI

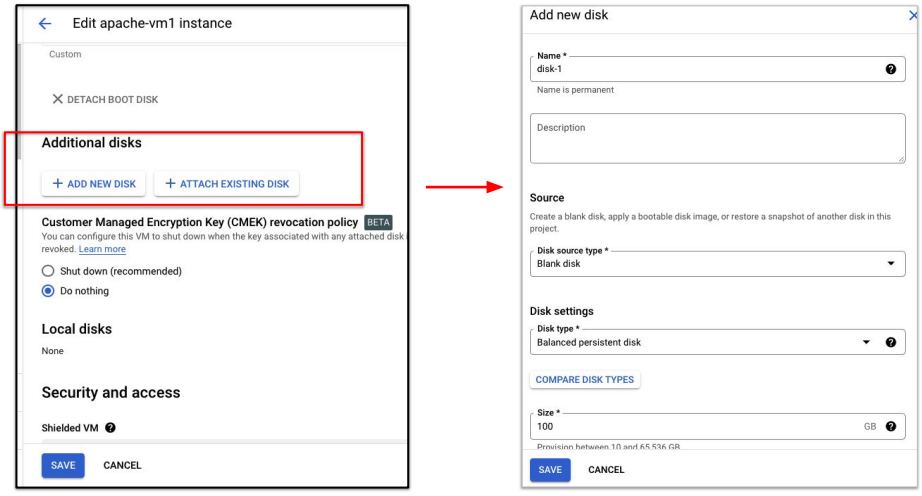
- Creating a zonal disk

```
gcloud compute disks create data-disk \
      --zone=us-central1-a \
      --size=10GB \
      --type=pd-standard
```

- Creating a regional disk

```
gcloud compute disks create disk-1 \
    --type=pd-balanced \
    --size=10GB \
     --region=us-central1 \
     --replica-zones=us-central1-c,us-central1-a
```

Adding additional disks with the Console



Add a persistent disk to your VM
https://cloud.google.com/compute/docs/disks/add-persistent-disk

Share disks between VMs:
https://cloud.google.com/compute/docs/disks/sharing-disks-between-vms

Adding additional disk using the CLI

- Attaching a zonal disk to a VM

```
gcloud compute instances attach-disk myvm \
     --disk data-disk \
     --zone=us-central1-a
```

- Attaching a regional disk to a VM

```
gcloud compute instances attach-disk myvm \
     --disk data-disk \
     --disk-scope=regional
```

- Must also format the disk to be used by the given file system or operating system

Attach disk:
https://cloud.google.com/sdk/gcloud/reference/compute/instances/attach-disk

Formatting and mounting a non-boot disk on a Linux VM
https://cloud.google.com/compute/docs/disks/add-persistent-disk#format_and_mount_linux

Formatting and mounting a non-boot disk on a Windows VM
https://cloud.google.com/compute/docs/disks/add-persistent-disk#format_and_mount_windows

Summary: Persistent Disk Options

Network storage appearing as a block device

- Attached to a VM through the network interface
- Durable storage: can survive VM terminate
- Bootable: you can attach to a VM and boot from it
- Snapshots: incremental backups
- Performance: Scales with size
- Disk resizing: even running and attached!
- Can be attached in read-only mode to multiple VMs
- Local SSD available for fast caching
- Zonal or Regional
- Encrypted by Google by default
  - Customers can do their own encryption using
    - Customer managed encryption keys
    - Customer supplied encryption keys

Filestore...

What is Filestore?

- Cloud-based managed file storage service for the Unix file system (POSIX)
- Provides native experience for standing up Network Attached Storage (NAS) for Compute Engine and Kubernetes Engine
- High-performance, fully managed network attached storage
  - Mount as file shares on Compute Engine instances
  - Used to store and serve files such as documents, images, videos, audio files, and other data
- Pay for what you use
- Capacity scales automatically scale based on demand
- Use cases:
  - Enterprise application migrations (SAP)
  - Media rendering where file shares are needed
  - Web content management

YouTube video: https://www.youtube.com/watch?v=CUwpXqEitA0

Filestore
https://cloud.google.com/filestore

Filestore use cases

Rendering

Mount Filestore volumes on Compute Engine instances, enabling visual effects artists to collaborate on the same file share. Burst compute to meet rendering demands.

Application migrations

Filestore can support a broad range of enterprise applications that need a shared filesystem interface to data. Ex. SAP

Web content management

Web developers creating websites and blogs that serve file content to their audience will find it easy to integrate Filestore with web software like Wordpress.

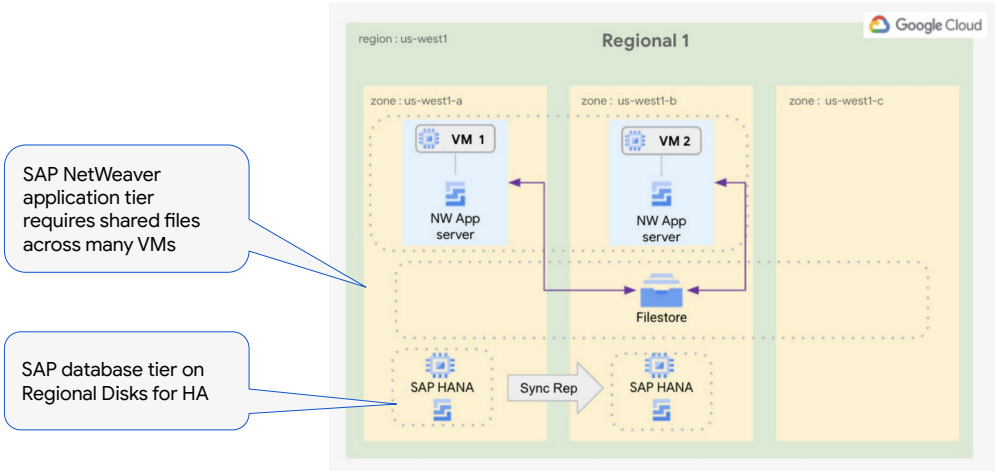
Media processing

Graphic design, video and image editing and other media workflows use files as an input and files as the output. Filestore helps creators access shared storage to manipulate and produce large files.

Home directories

Users across your organization probably need to access and share common data sets. You can host file content in Filestore and enable shared access to that data.

Filestore example - SAP NetWeaver



Announcing Filestore Enterprise, for your most demanding apps (2021)

Announcing Filestore Enterprise, for your most demanding apps(2021)
https://cloud.google.com/blog/products/storage-data-transfer/google-cloud-announces-filestore-enterprise-for-business-critical-apps

Note that this blog is from 2021, and the “public preview” items mentioned in the blog are now generally available (GA)

Exam Guide - Compute Engine

3.1 Deploying and implementing Compute Engine resources. Tasks include:

- 3.1.1 Launching a compute instance using the Google Cloud console and Cloud SDK (gcloud) (e.g., assign disks, availability policy, **SSH keys)**
- 3.1.2 Creating an autoscaled managed instance group using an instance template
- 3.1.3 **Generating/uploading a custom SSH key for instances**
- 3.1.4 Installing and configuring the Cloud Monitoring and Logging Agent
- 3.1.5 Assessing compute quotas and requesting increases

2.2 Planning and configuring compute resources. Considerations include:

- 2.2.1 Selecting appropriate compute choices for a given workload (e.g., Compute Engine, Google Kubernetes Engine, Cloud Run, Cloud Functions)
- 2.2.2 Using preemptible VMs and custom machine types as appropriate

2.3 Planning and configuring data storage options. Considerations include:

- 2.3.1 Product choice (e.g., Cloud SQL, BigQuery, Firestore, Cloud Spanner, Cloud Bigtable)
- 2.3.2 Choosing storage options (e.g., Zonal persistent disk, Regional balanced persistent disk, Standard, Nearline, Coldline, Archive)

Connecting to Linux Machines

- Use SSH to connect to and administrate Linux machines
  - Requires an SSH key which by default is automatically generated if accessing from within Google Cloud
- Simplest way is to use the Web Console and just click the SSH button next to the VM you want to connect to
  - Opens a terminal window in a new browser tab
  - Requires a firewall rule allowing SSH

A nice feature of the Console is the ability to open an SSH session to a virtual machine with a press of a button.

From the Console, you can use SSH to connect and administer various types of Linux instances.

A project-wide SSH is used by default. Alternatively, you can use your own keys.

Simply press the SSH button associated with the instance:

- This will open a terminal window in a new browser tab

If you are using the default network, a firewall rule for SSH is already created. For auto and custom networks, you will need to create one.

SSH with `gcloud` - first step

- Useful for developers who are not given access to the Web Console
- SSH keys are exchanged automatically when `gcloud` is initialized
  - Stored in the `~/.ssh` folder
- Creating an SSH key pair
  - `ssh-keygen -t rsa -f ~/.ssh/my-ssh-key -C myusername -b 2048`
  - Public key file: `~/.ssh/my-ssh-key.pub`
  - Private key file: `~/.ssh/my-ssh-key`

About SSH connections:
https://cloud.google.com/compute/docs/instances/ssh

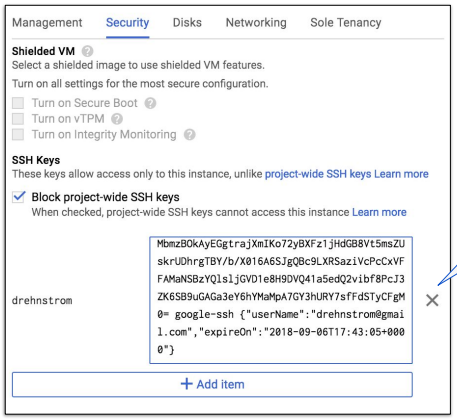
Create SSH keys:
https://cloud.google.com/compute/docs/connect/create-ssh-keys

Add SSH keys to VMs:
https://cloud.google.com/compute/docs/connect/add-ssh-keys

Youtube video that demonstrates SSH key creation and access:
https://www.youtube.com/watch?v=8QGpHQ2SyG8

SSH with gcloud - second step

When checked, project-wide keys cannot be used to access this VM



Enter public ssh key here.

Restricts access to users with the corresponding private key

SSH with gcloud - third step

To connect from local command line

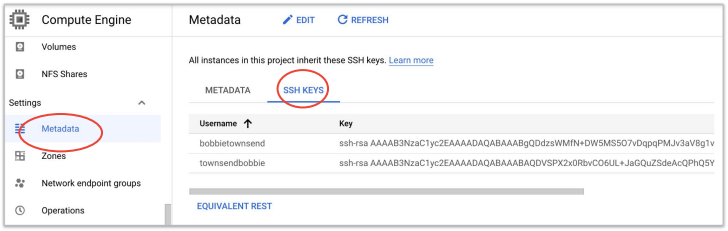
- `gcloud` automatically exchanges keys
  - Looks in the `~/.ssh` folder for the private key

```
gcloud compute ssh myvm --zone=us-east1-b
```

Project-Wide SSH Keys

Can be added in Compute Engine metadata

- These can be used to SSH into **any machine** within a project unless otherwise specified
- Less secure than machine level ssh keys



Project-wide SSH keys can be added in Compute Engine metadata. Once added, they can be used to SSH into any machine in a project.

Exam Guide - Compute Engine

3.1 Deploying and implementing Compute Engine resources. Tasks include:

- 3.1.1 Launching a compute instance using the Google Cloud console and Cloud SDK (gcloud) (e.g., assign disks, **availability policy,** SSH keys)
- 3.1.2 Creating an autoscaled managed instance group using an instance template
- 3.1.3 Generating/uploading a custom SSH key for instances
- 3.1.4 Installing and configuring the Cloud Monitoring and Logging Agent
- 3.1.5 Assessing compute quotas and requesting increases

2.2 Planning and configuring compute resources. Considerations include:

- 2.2.1 Selecting appropriate compute choices for a given workload (e.g., Compute Engine, Google Kubernetes Engine, Cloud Run, Cloud Functions)
- 2.2.2 Using preemptible VMs and custom machine types as appropriate

2.3 Planning and configuring data storage options. Considerations include:

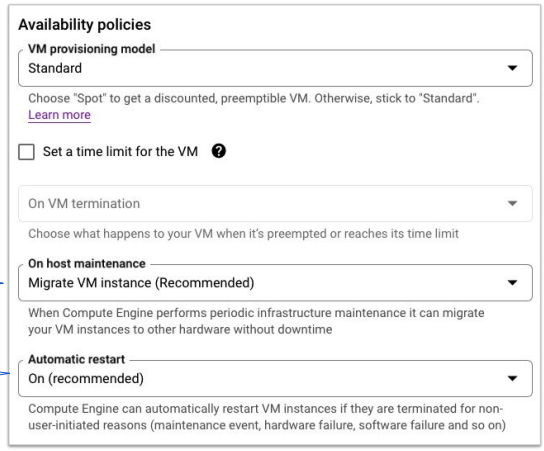
- 2.3.1 Product choice (e.g., Cloud SQL, BigQuery, Firestore, Cloud Spanner, Cloud Bigtable)
- 2.3.2 Choosing storage options (e.g., Zonal persistent disk, Regional balanced persistent disk, Standard, Nearline, Coldline, Archive)

Availability policies

- Determines what happens to VMs when Google does maintenance on the underlying infrastructure or if terminated due to a non-user-initiated reason

Maintenance options are
- Migrate to another hardware
- Terminate the instance

Restart options are
- On
- Off



Live migration during maintenance events:
https://cloud.google.com/compute/docs/instances/live-migration

Set VM host maintenance policy:
https://cloud.google.com/compute/docs/instances/setting-instance-scheduling-options

Exam Guide - Compute Engine

4.1 Managing Compute Engine resources. Tasks include:

- 4.1.1 Managing a single VM instance (e.g., start, stop, edit configuration, or delete an instance)
- 4.1.2 Remotely connecting to the instance
- 4.1.3 Attaching a GPU to a new instance and installing necessary dependencies
- 4.1.4 Viewing current running VM inventory (instance IDs, details)
- 4.1.5 Working with snapshots (e.g., create a snapshot from a VM, view snapshots, delete a snapshot)
- 4.1.6 Working with images (e.g., create an image from a VM or a snapshot, view images, delete an image)
- 4.1.7 Working with instance groups (e.g., set autoscaling parameters, assign instance template, create an instance template, remove instance group)
- 4.1.8 Working with management interfaces (e.g., Google Cloud console, Cloud Shell, Cloud SDK)

Starting, stopping or deleting an instance

- Start a VM

```
gcloud compute instances start example-instance
--zone=us-central1-a
```

- Stop

```
gcloud compute instances stop test-instance
--zone=us-central1-a
```

- Delete

```
gcloud compute instances delete instance-1
--zone=us-central2-a
```


Start: https://cloud.google.com/sdk/gcloud/reference/compute/instances/start

Stop: https://cloud.google.com/sdk/gcloud/reference/compute/instances/stop

Delete: https://cloud.google.com/sdk/gcloud/reference/compute/instances/delete

Editing configuration, or updating an instance

- Some VM properties can be changed without stopping the VM

```
gcloud compute instances update example-instance
--zone=us-central1-a
--update-labels=k0=value1,k1=value2 --remove-labels=k3
```

- Other properties are updateable only when a VM is stopped
  - Updating these is a 3 step process
    - Stop the vm
    - Use the instances export command to export the VM's properties to a file
    - Edit the file and make the appropriate changes
    - Use the instances update-from-file command to update the target instance.

Partial example output

```
creationTimestamp: '2022-08-09T04:13:25.743-07:00'
deletionProtection: false
description: ''
disks:
- autoDelete: true
  boot: true
  deviceName: persistent-disk-0
  diskSizeGb: 10
  guestOsFeatures:
  - type: UEFI_COMPATIBLE
  - type: VIRTIO_SCSI_MULTIQUEUE
  - type: GVNIC
  index: 0
  interface: SCSI
  kind: compute#attachedDisk
  licenses:
  - https://www.googleapis.com/compute/v1/projects/debian-
  mode: READ_WRITE
  source: https://www.googleapis.com/compute/v1/projects/b
  type: PERSISTENT
displayDevice:
  enableDisplay: false
fingerprint: jA07n2WbFIs=
id: 7993863297061355450
keyRevocationActionType: NONE
kind: compute#instance
labelFingerprint: 42WmSpB8rSM=
lastStartTimestamp: '2022-09-19T05:45:57.391-07:00'
lastStopTimestamp: '2022-09-20T03:15:34.068-07:00'
```


Update:
https://cloud.google.com/sdk/gcloud/reference/compute/instances/update

Update instance properties:
https://cloud.google.com/compute/docs/instances/update-instance-properties

Exam Guide - Compute Engine

4.1 Managing Compute Engine resources. Tasks include:

- 4.1.1 Managing a single VM instance (e.g., start, stop, edit configuration, or delete an instance)
- 4.1.2 Remotely connecting to the instance
- 4.1.3 Attaching a GPU to a new instance and installing necessary dependencies
- 4.1.4 Viewing current running VM inventory (instance IDs, details)
- 4.1.5 Working with snapshots (e.g., create a snapshot from a VM, view snapshots, delete a snapshot)
- 4.1.6 Working with images (e.g., create an image from a VM or a snapshot, view images, delete an image)
- 4.1.7 Working with instance groups (e.g., set autoscaling parameters, assign instance template, create an instance template, remove instance group)
- 4.1.8 Working with management interfaces (e.g., Google Cloud console, Cloud Shell, Cloud SDK)

Connecting to Machines without an External IP

- Use a bastion host
  - Create a machine in the same network that has a public IP
  - Connect to it, then connect to the private machine from there
  - Turn the machine off when you don't need it
- Use Identity Aware Proxy (IAP) - TCP Forwarding
  - Uses IAM to control access to services like SSH and RDP on VM instances
    - Avoids openly exposing these services to the internet
  - Requests must pass authentication and authorization checks
  - Can be used as an alternative to a Bastion host
- Site to site: Use Cloud VPN/Interconnect
  - Connect from your network to the Google Cloud network via a private IP address

Topics are covered in Module 5 along with IAM

In most cases, you may not want to give an instance an external IP address for security purposes, yet you will still need to connect to them. What are your options?

You can also leverage a bastion host, or jumpbox.
This would be one virtual machine with a public IP address that you can connect to, that can then give you access to other instances in the same network via a private IP address.
The bastion host can be turned off when not in use.

You can also leverage a VPN connection.
VPN allows you to connect to your network via an encrypted tunnel over the internet using a private IP address.

Another way is to add an “IAP-secured Tunnel User” role to enable SSH via Identity Away Proxy, or IAP.
This method will work for both the Console and gcloud.
You will have to set up a firewall rule to allow SSH traffic from 35.235.240.0/20.

Exam Guide - Compute Engine

4.1 Managing Compute Engine resources. Tasks include:

- 4.1.1 Managing a single VM instance (e.g., start, stop, edit configuration, or delete an instance)
- 4.1.2 Remotely connecting to the instance
- 4.1.3 Attaching a GPU to a new instance and installing necessary dependencies
- 4.1.4 Viewing current running VM inventory (instance IDs, details)
- 4.1.5 Working with snapshots (e.g., create a snapshot from a VM, view snapshots, delete a snapshot)
- 4.1.6 Working with images (e.g., create an image from a VM or a snapshot, view images, delete an image)
- 4.1.7 Working with instance groups (e.g., set autoscaling parameters, assign instance template, create an instance template, remove instance group)
- 4.1.8 Working with management interfaces (e.g., Google Cloud console, Cloud Shell, Cloud SDK)

What are GPUs?

- GPUs were originally designed to accelerate the rendering of 3D graphics
- Over time, they became more flexible and programmable, enhancing their capabilities
  - Allowed graphics programmers to create more interesting visual effects and realistic scenes with advanced lighting and shadowing techniques
  - Other developers also began to tap the power of GPUs to dramatically accelerate additional workloads in
    - High performance computing (HPC)
    - Deep learning
    - Video Processing
    - and more

Using GPUs with Compute Engine

- Use GPUs to accelerate specific workloads on your VMs such as machine learning and data processing
  - Provides NVIDIA GPUs in passthrough mode allowing VMs to have direct control over the GPUs and their associated memory
- Supported by specific machine types (N1 or A2 machine series)
  - Not supported by shared-core or memory-optimized machine types
- Google provides Deep Learning VM images with GPU drivers pre-installed
  - Can also use public/custom images and install drivers yourself

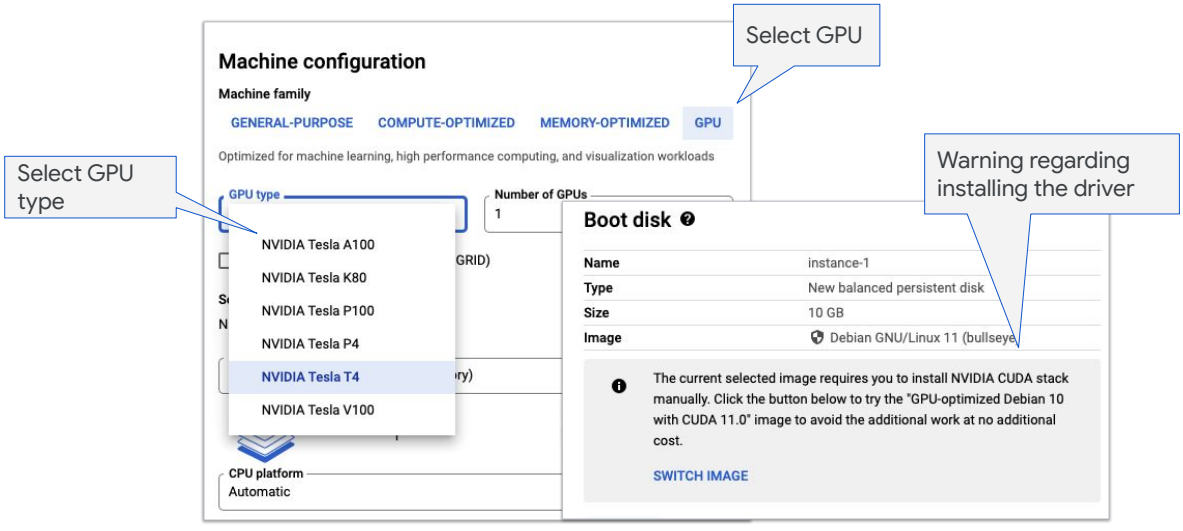
GPU platforms
https://cloud.google.com/compute/docs/gpus

About GPUs:
https://cloud.google.com/compute/docs/gpus/about-gpus

Create a VM with attached GPUs:
https://cloud.google.com/compute/docs/gpus/create-vm-with-gpus

Installing GPU drivers:
https://cloud.google.com/compute/docs/gpus/install-drivers-gpu

Creating a VM with a GPU in the Console



Creating a VM with a GPU in CLI

- First check for GPU availability in the desired region

```
gcloud compute regions describe us-central1
```

Partial output:

```
  metric: NVIDIA_P4_GPUS
  usage: 0.0
- limit: 1.0
  metric: NVIDIA_P4_VWS_GPUS
  usage: 0.0
- limit: 100.0
```

Limit = Max quota for this region

Not a guarantee that a resource is always available

See: Quotas and limits

- Next create the VM the the appropriate machine type and image
  - Install a driver if needed

Exam Guide - Compute Engine

4.1 Managing Compute Engine resources. Tasks include:

- 4.1.1 Managing a single VM instance (e.g., start, stop, edit configuration, or delete an instance)
- 4.1.2 Remotely connecting to the instance
- 4.1.3 Attaching a GPU to a new instance and installing necessary dependencies
- 4.1.4 Viewing current running VM inventory (instance IDs, details)
- 4.1.5 Working with snapshots (e.g., create a snapshot from a VM, view snapshots, delete a snapshot)
- 4.1.6 Working with images (e.g., create an image from a VM or a snapshot, view images, delete an image)
- 4.1.7 Working with instance groups (e.g., set autoscaling parameters, assign instance template, create an instance template, remove instance group)
- 4.1.8 Working with management interfaces (e.g., Google Cloud console, Cloud Shell, Cloud SDK)

VM Manager

- Suite of tools to manage operating systems for large VM fleets running Windows and Linux on Compute Engine
- Includes
  - OS patch management
    - Apply on-demand and scheduled patches.
    - Provides patch compliance reporting for your environment.
  - OS inventory management*
    - Collect and review operating system information
  - OS configuration management
    - Install, remove, and auto-update software packages.

*This is the only one mentioned in the ACE Exam Guide and discussed here. Engineers will find the other tools useful as well.

VM Manager:
https://cloud.google.com/compute/docs/manage-os

Set up VM Manager:
https://cloud.google.com/compute/docs/manage-os#gcloud

The exam guide mentions only OS inventory management, and that's what is discussed here. Engineers may find the other tools useful as well

OS Inventory Management

- Identify VMs that are running a specific version of an operating system.
- View operating system packages that are installed on a VM.
- Generate a list of operating system package updates that are available for each VM
- Identify missing operating system packages, updates, or patches for a VM
- View vulnerability reports for a VM

OS inventory management:
https://cloud.google.com/compute/docs/instances/os-inventory-management

View operating system details:
https://cloud.google.com/compute/docs/instances/view-os-details

Create an OS policy assignment:
https://cloud.google.com/compute/docs/os-configuration-management/create-os-policy-assignment

Enable OS Inventory Management - Code

#Enable the API (one time activity)

```
gcloud services enable osconfig.googleapis.com
```

#Set enable-osconfig in project-wide metadata, so that it applies to all of the instances (alternatively, can set it at the VM level)

```
gcloud compute project-info add-metadata --project bt-managed-instance-grp \
  --metadata=enable-osconfig=TRUE
```

#create a VM with the OS Config agent installed

```
 gcloud compute instances create vm-os-config-demo \
   --machine-type=e2-micro --zone=us-east1-b \
   --metadata=startup-script='#! /bin/bash
   apt update
   apt -y install google-osconfig-agent'
```

#get an inventory

```
gcloud compute os-config inventories list --location=us-east1-b
```

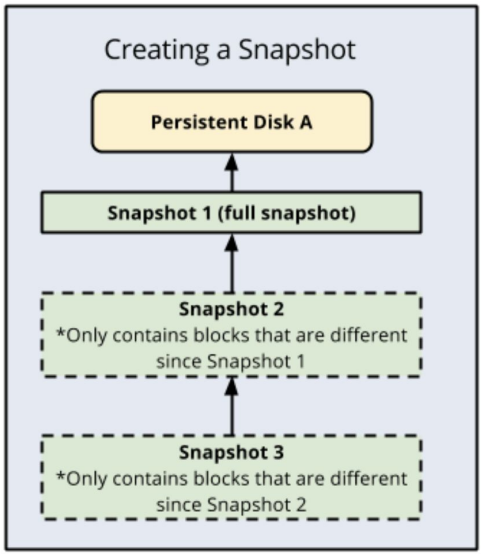
Exam Guide - Compute Engine

4.1 Managing Compute Engine resources. Tasks include:

- 4.1.1 Managing a single VM instance (e.g., start, stop, edit configuration, or delete an instance)
- 4.1.2 Remotely connecting to the instance
- 4.1.3 Attaching a GPU to a new instance and installing necessary dependencies
- 4.1.4 Viewing current running VM inventory (instance IDs, details)
- 4.1.5 Working with snapshots (e.g., create a snapshot from a VM, view snapshots, delete a snapshot)
- 4.1.6 Working with images (e.g., create an image from a VM or a snapshot, view images, delete an image)
- 4.1.7 Working with instance groups (e.g., set autoscaling parameters, assign instance template, create an instance template, remove instance group)
- 4.1.8 Working with management interfaces (e.g., Google Cloud console, Cloud Shell, Cloud SDK)

Snapshots

- Are incremental backups of data from persistent disks
  - No need to stop VM to take a snapshot
- Multiple copies are stored across multiple locations automatically
  - Snapshots can be shared across projects
- Create snapshot schedules to make backups of disks on a predetermined schedule



Persistent disk snapshots:
https://cloud.google.com/compute/docs/disks/snapshots

Create and manage disk snapshots (includes Create, List, View, Delete):
https://cloud.google.com/compute/docs/disks/create-snapshots

Move a VM instance between zones or regions:
https://cloud.google.com/compute/docs/instances/moving-instance-across-zones

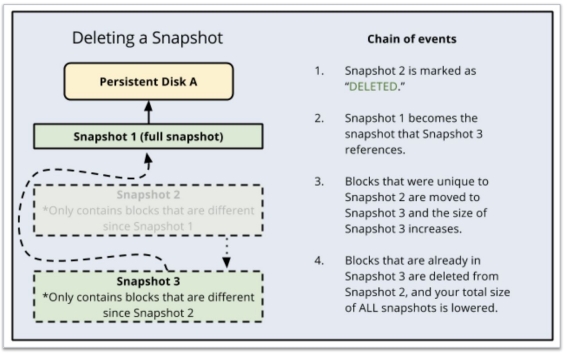
Best practices for persistent disk snapshots
https://cloud.google.com/compute/docs/disks/snapshot-best-practices

Introducing Google Cloud Backup and DR (new 2022):
https://cloud.google.com/blog/products/storage-data-transfer/introducing-google-cloud-backup-and-dr

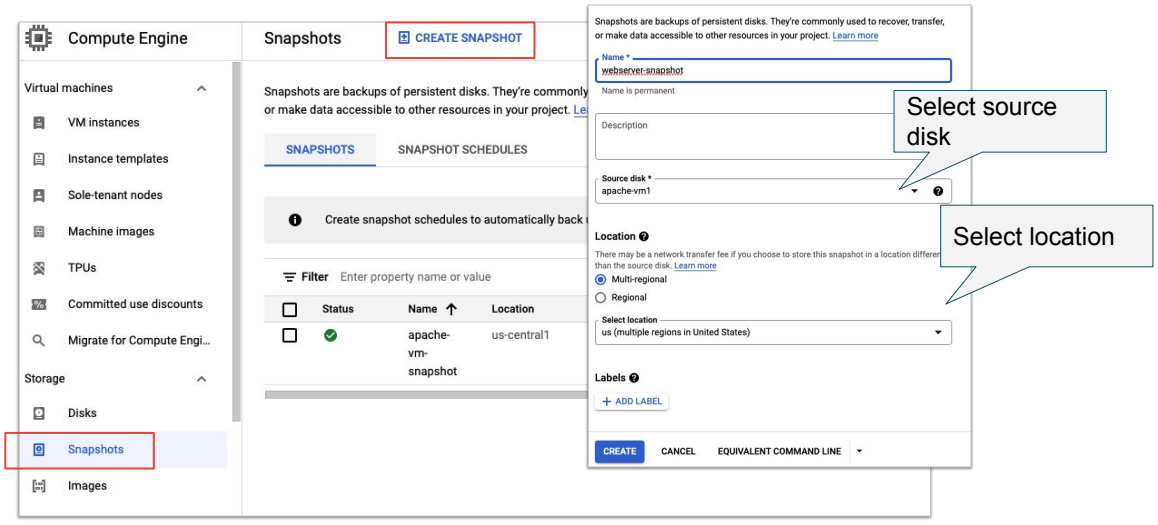
Backup and DR Service (new 2022):
https://cloud.google.com/backup-disaster-recovery

Deleting Snapshots

- A deleted snapshot is immediately marked as DELETED in the system.
  - Is deleted outright If no dependent snapshots
- If dependant snapshots exist
  - Data required for restoring other snapshots is moved into the next snapshot
  - Data not required for restoring is deleted
  - The next snapshot no longer references the snapshot marked for deletion, and instead references the snapshot before it



Create a Snapshot




```
gcloud compute snapshots create webserver
--source-disk ws-disk
--source-disk-region=us-central1
```

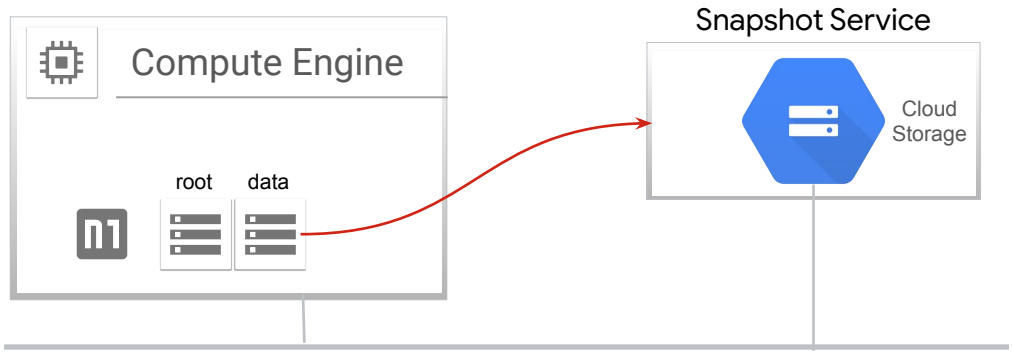
Snapshot storage

- Stored in Cloud Storage and have a choice of
  - Multi-regional location, such as Asia
    - Provides higher availability (99.95% SLA vs 99.9% for regional)
    - Potentially slower snapshot restoration performance
  - Regional location, such as asia-south1
    - Use for compliance, e.g., GDPR
    - Use when all resources created from the snapshot will be in the same region - provides fastest restoration performance
- Network costs may be incurred when creating a disk from a snapshot
  - Multi-regional location storage
    - No network costs as long as the new persistent disk is created in one of the regions of the multi-regional group
  - Regional storage
    - *Will* incur network costs if the new disk is created in another region

Snapshot use cases

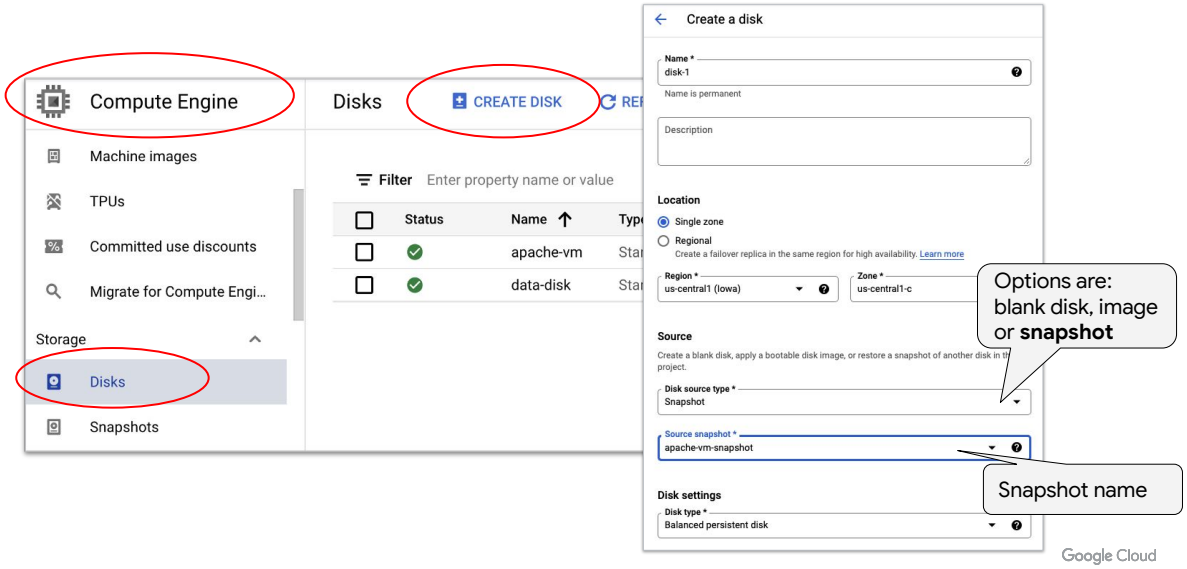
- Snapshots have many use cases
  - Can be used as source for a new disk
  - Can play a part in a disaster recovery plan
  - Can backup data
  - Can be used to move a VM to another zone/region
  - Can be used to migrate data from one disk type to another
- Examples are shown on the following slides

Example - using a snapshot to backup data



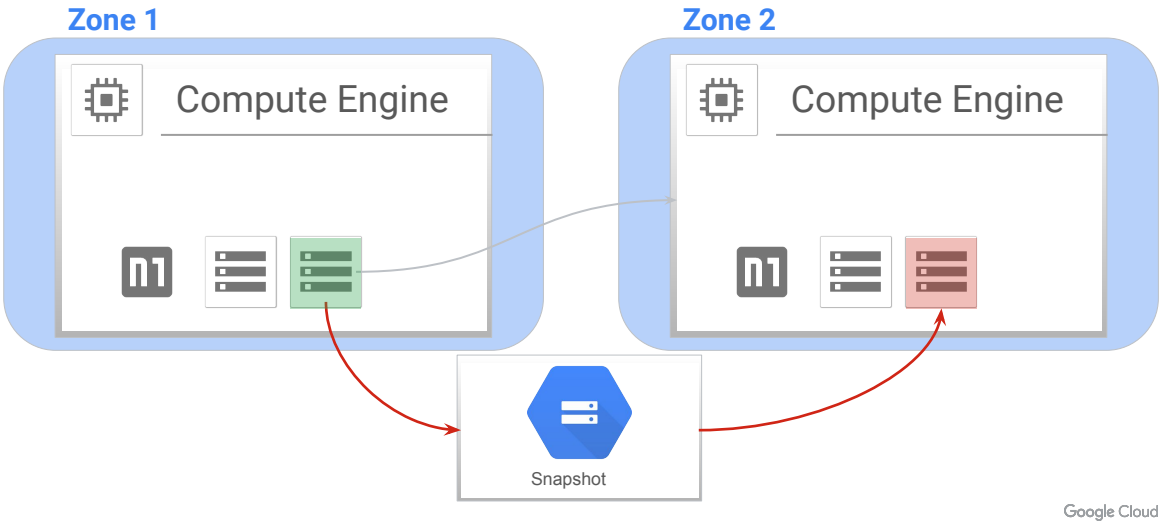
Snapshots have many use cases. For example, they can be used to backup critical data into a durable storage solution to meet application, availability, and recovery requirements.

Example - create disk from snapshot



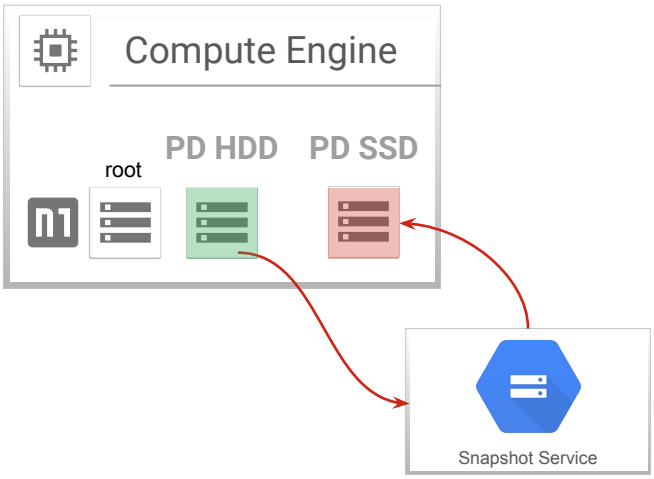
Another use case is the ability to create a new disk from a snapshot. Afterwards, this disk can be attached to a VM.

Example - using a snapshot migrate data between zones



Snapshots can also be used to migrate data between zones. For example, you might want to minimize latency by migrating data to a drive that can be attached to a VM in another zone.

Example - Transferring data to SSD to improve performance



Another snapshot use case of transferring data to a different disk type. For example, if you want to improve disk performance, you could use a snapshot to transfer data from a standard HDD persistent disk to a SSD persistent disk.

Exam Guide - Compute Engine

4.1 Managing Compute Engine resources. Tasks include:

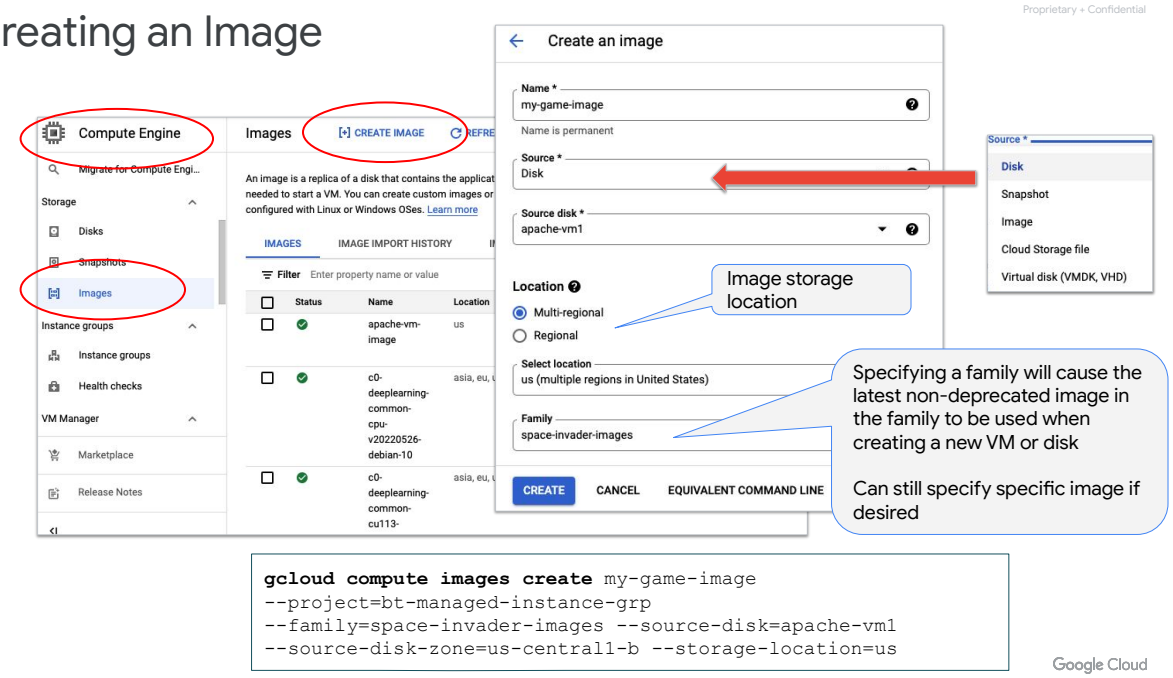
- 4.1.1 Managing a single VM instance (e.g., start, stop, edit configuration, or delete an instance)
- 4.1.2 Remotely connecting to the instance
- 4.1.3 Attaching a GPU to a new instance and installing necessary dependencies
- 4.1.4 Viewing current running VM inventory (instance IDs, details)
- 4.1.5 Working with snapshots (e.g., create a snapshot from a VM, view snapshots, delete a snapshot)
- 4.1.6 Working with images (e.g., create an image from a VM or a snapshot, view images, delete an image)
- 4.1.7 Working with instance groups (e.g., set autoscaling parameters, assign instance template, create an instance template, remove instance group)
- 4.1.8 Working with management interfaces (e.g., Google Cloud console, Cloud Shell, Cloud SDK)

Compute Engine 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
    - SQL Server pre-installed on Windows
  - The complete list of images is shown in the Web Console > Images or `gcloud compute images list`
- 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

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

Creating an Image



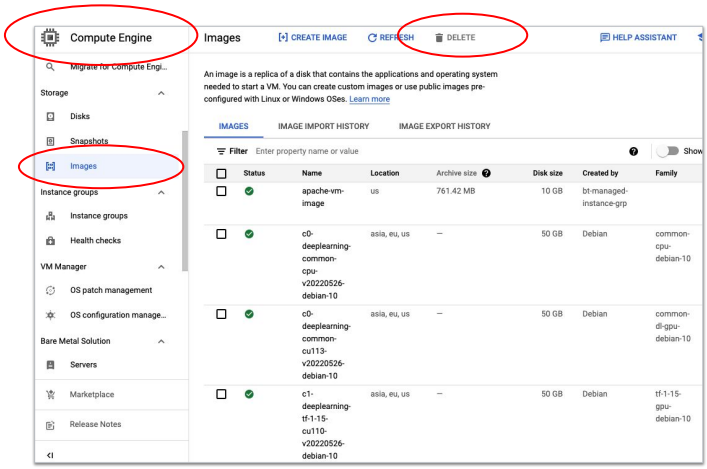
Create an image from a VM disk, other image or a snapshot:
https://cloud.google.com/sdk/gcloud/reference/compute/images/create

View: https://cloud.google.com/sdk/gcloud/reference/compute/images/list

Delete: https://cloud.google.com/sdk/gcloud/reference/compute/images/delete

Enable guest operating system features on custom images:
https://cloud.google.com/compute/docs/images/create-delete-deprecate-private-images#guest-os-features

View & Delete Images




```
gcloud compute images list
```

```
gcloud compute images delete \
apache-vm-image
```

Snapshots vs Images

Both can be used as the basis for a new VM

- Takes longer to spin up a VM from a snapshot (data needs to be restored) versus an image (which is already in the state to be booted)

Snapshots	Images
● Best for disk backups ● Can be scheduled ● Good for the use cases mentioned on the prior slides ● Lower storage cost than images ● Can be created while VM is running	● Best for infrastructure re-use ○ Boot disk/data disk images for new VMs ○ Managed Instance Group templates require images (not snapshots) ● Can be versioned and deprecated

Exam Guide - Compute Engine

4.1 Managing Compute Engine resources. Tasks include:

- 4.1.1 Managing a single VM instance (e.g., start, stop, edit configuration, or delete an instance)
- 4.1.2 Remotely connecting to the instance
- 4.1.3 Attaching a GPU to a new instance and installing necessary dependencies
- 4.1.4 Viewing current running VM inventory (instance IDs, details)
- 4.1.5 Working with snapshots (e.g., create a snapshot from a VM, view snapshots, delete a snapshot)
- 4.1.6 Working with images (e.g., create an image from a VM or a snapshot, view images, delete an image)
- 4.1.7 Working with instance groups (e.g., set autoscaling parameters, assign instance template, create an instance template, remove instance group)
- 4.1.8 Working with management interfaces (e.g., Google Cloud console, Cloud Shell, Cloud SDK)

Exam Guide - Compute Engine

3.1 Deploying and implementing Compute Engine resources. Tasks include:

- 3.1.1 Launching a compute instance using the Google Cloud console and Cloud SDK (gcloud) (e.g., assign disks, availability policy, SSH keys)
- 3.1.2 Creating an autoscaled managed instance group using an instance template
- 3.1.3 Generating/uploading a custom SSH key for instances
- 3.1.4 Installing and configuring the Cloud Monitoring and Logging Agent
- 3.1.5 Assessing compute quotas and requesting increases

2.2 Planning and configuring compute resources. Considerations include:

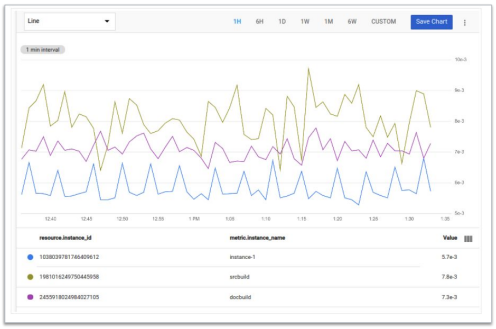
- 2.2.1 Selecting appropriate compute choices for a given workload (e.g., Compute Engine, Google Kubernetes Engine, Cloud Run, Cloud Functions)
- 2.2.2 Using preemptible VMs and custom machine types as appropriate

2.3 Planning and configuring data storage options. Considerations include:

- 2.3.1 Product choice (e.g., Cloud SQL, BigQuery, Firestore, Cloud Spanner, Cloud Bigtable)
- 2.3.2 Choosing storage options (e.g., Zonal persistent disk, Regional balanced persistent disk, Standard, Nearline, Coldline, Archive)

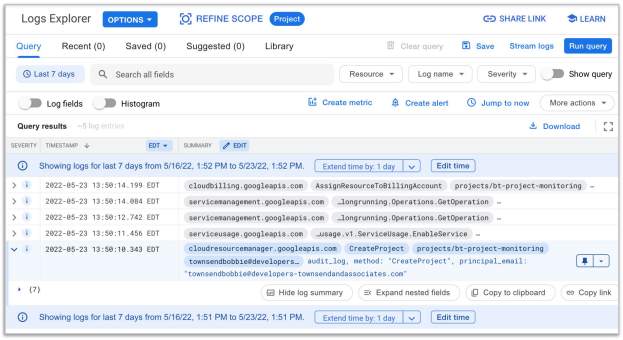
Cloud Monitoring

- Collects metrics that measure the performance of the service infrastructure
  - Cloud Run, disks, buckets, virtual machines, and many more
- AWS resources can be monitored as well
- Example metrics include:
  - Request count: for example, the number of HTTP requests per minute that result in 2xx or 5xx responses.
  - Response latencies: for example, the latency for HTTP 2xx responses
- Custom metrics can be created



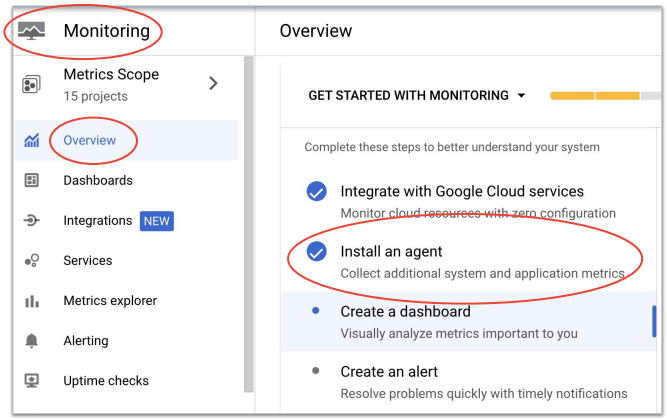
Cloud Logging

Store, search, analyze, monitor, and alert on logging data and events from Google Cloud and Amazon Web Services



What are the Cloud Monitoring and Logging Agents?

- To capture metrics and logs from within a VM, need to install an agent
- In the past there were two separate agents for Monitoring and Logging
- A single agent was introduced in 2021
- Both methods are shown as the date of the last update to the ACE exam is unknown



Legacy method: Installing the Cloud Logging agent on individual VMs:
https://cloud.google.com/logging/docs/agent/logging/installation

Legacy method:Configuring the Logging Agent:
https://cloud.google.com/logging/docs/agent/logging/configuration

Legacy method: Installing the Cloud Monitoring agent on individual VMs:
https://cloud.google.com/monitoring/agent/monitoring/installation

Legacy method: Configuring the Cloud Monitoring agent:
https://cloud.google.com/monitoring/agent/monitoring/configuration

New method: Ops Agent overview:
https://cloud.google.com/stackdriver/docs/solutions/agents/ops-agent

New method: Configuring the Ops Agent:
https://cloud.google.com/stackdriver/docs/solutions/agents/ops-agent/configuration

Monitoring Agent (Legacy)

- Installed on VMs to provide additional metrics not available externally
  - Memory usage and uptime, for example
- Also provides metrics on common applications
  - Apache, Cassandra, CouchDB, HBase, IIS, JVM, Kafka, etc.
- Add the following code to your startup script

```
curl -sSO
https://dl.google.com/cloudagents/install-monitoring-agent.sh
sudo bash install-monitoring-agent.sh
```


When using Compute Engine VMs, sometimes Operations needs to get metrics from within the VM. Memory utilization might be the most common example of this.

To solve this, you can install the Operations Monitoring Agent on the VM. This agent will then report VM metrics to the Operations system.

In addition, the monitoring agent will detect commonly used applications, and if they are running on that machine, it will automatically report metrics on those applications as well. For example, if Apache Web Server was running on a machine, you could then add its utilization to your dashboard.

Just install the monitoring agent in your virtual machine's startup script.

Many common applications are supported. See the documentation for details and installation instructions.

Logging Agent (Legacy)

- Automatically stream logs from third-party applications, system logs, and your code to the Operations logs
- Install on VMs running Linux or Windows
  - AWS EC2 VMs also supported
  - Not required for App Engine or Kubernetes clusters
- Add the following code to your startup script

```
curl -sSO https://dl.google.com/cloudagents/install-logging-agent.sh
sudo bash install-logging-agent.sh
```


If you're using Kubernetes, Cloud Functions, or App Engine to deploy your applications, then all messages sent to standard out end up in the Operations logs automatically. When using virtual machines though, you will need to add the Logging Agent when you create the VM.

Once the Logging Agent is added, any messages you log from your application will go to Operations. Also, the logs for third-party applications like Apache web server will be aggregated into the Operations logs.

The Logging Agent can also be used on VMs running in AWS EC2.

Just add a couple lines of code to your startup script.

Ops Agent (New 2021)

- Automatically capture monitoring metrics and stream logs from third-party applications, system logs, and your code to the Operations logs
- Install on VMs running Linux or Windows
  - AWS EC2 VMs also supported
  - Not required for App Engine or Kubernetes clusters
- Add the following code to your startup script

```
curl -sSO
https://dl.google.com/cloudagents/add-google-cloud-ops-agent-repo.sh
sudo bash add-google-cloud-ops-agent-repo.sh --also-install
```

Exam Guide - Compute Engine

3.1 Deploying and implementing Compute Engine resources. Tasks include:

- 3.1.1 Launching a compute instance using the Google Cloud console and Cloud SDK (gcloud) (e.g., assign disks, availability policy, SSH keys)
- 3.1.2 Creating an autoscaled managed instance group using an instance template
- 3.1.3 Generating/uploading a custom SSH key for instances
- 3.1.4 Installing and configuring the Cloud Monitoring and Logging Agent
- 3.1.5 Assessing compute quotas and requesting increases

2.2 Planning and configuring compute resources. Considerations include:

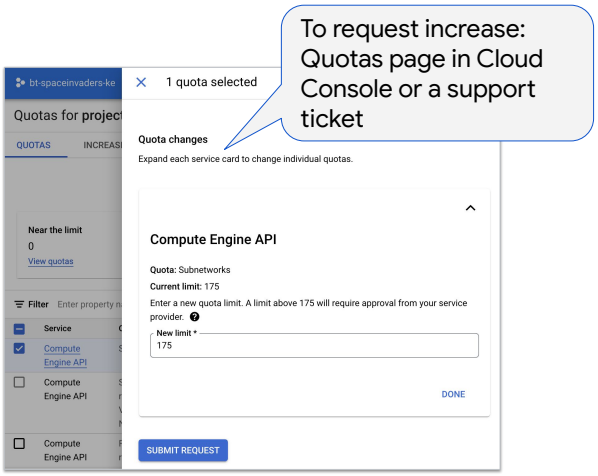
- 2.2.1 Selecting appropriate compute choices for a given workload (e.g., Compute Engine, Google Kubernetes Engine, Cloud Run, Cloud Functions)
- 2.2.2 Using preemptible VMs and custom machine types as appropriate

2.3 Planning and configuring data storage options. Considerations include:

- 2.3.1 Product choice (e.g., Cloud SQL, BigQuery, Firestore, Cloud Spanner, Cloud Bigtable)
- 2.3.2 Choosing storage options (e.g., Zonal persistent disk, Regional balanced persistent disk, Standard, Nearline, Coldline, Archive)

All resources are subject to project quotas or limits

- Examples
  - How many resources you can create per project
    - 300 VPC subnets/VPC
  - How quickly you can make API requests in a project: rate limits
    - 300 admin actions/minute (Cloud Spanner)
  - How many resources you can create per region
    - 24 CPUs region/project



Note: The numbers shown here vary by customer

Work with quotas
https://cloud.google.com/docs/quota

Resource quotas:
https://cloud.google.com/compute/quotas

Requesting an increase in quota:
https://cloud.google.com/compute/quotas#requesting_additional_quota

Compute Engine quotas:
https://cloud.google.com/compute/quotas

BigQuery quotas:
https://cloud.google.com/bigquery/quotas

All resources in Google Cloud are subject to project quotas or limits. These typically fall into one of the three categories shown here:

- How many resources you can create per project, e.g., number of VPC networks
- How quickly you can make API requests in a project or rate limits.

- There also regional quotas, e.g., the number of CPUs allowed per region

Given these quotas, you may be wondering, how do I spin up one of those 96-core VMs?

As your use of Google Cloud expands over time, your quotas may increase accordingly. If you expect a notable upcoming increase in usage, you can proactively request quota adjustments from the Quotas page in the Cloud Console. This page will also display your current quotas.

If quotas can be changed, why do they exist?

Quota benefits

- Prevent runaway consumption in case of an error or malicious attack
- Prevent billing spikes or surprises
- Forces sizing consideration and periodic review

Project quotas prevent runaway consumption in case of an error or malicious attack. For example, imagine you accidentally create 100 instead of 10 Compute Engine instances using the gcloud command line.

Quotas also prevent billing spikes or surprises. Quotas are related to billing, but we will go through how to set up budgets and alerts later, which will really help you manage billing.

Finally, quotas force sizing consideration and periodic review. For example, do you really need a 96-core instance, or can you go with a smaller and cheaper alternative?

It is also important to mention that quotas are the maximum amount of resources you can create for that resource type *as long as those resources are available*. Quotas do not guarantee that resources will be available at all times. For example, if a region is out of local SSDs, you cannot create local SSDs in that region, even if you still had quota for local SSDs.

Exam Guide - Compute Engine

3.1 Deploying and implementing Compute Engine resources. Tasks include:

- 3.1.1 Launching a compute instance using the Google Cloud console and Cloud SDK (gcloud) (e.g., assign disks, availability policy, SSH keys)
- 3.1.2 Creating an autoscaled managed instance group using an instance template
- 3.1.3 Generating/uploading a custom SSH key for instances
- 3.1.4 Installing and configuring the Cloud Monitoring and Logging Agent
- 3.1.5 Assessing compute quotas and requesting increases

2.2 Planning and configuring compute resources. Considerations include:

- 2.2.1 Selecting appropriate compute choices for a given workload (e.g., Compute Engine, Google Kubernetes Engine, Cloud Run, Cloud Functions)
- 2.2.2 Using preemptible VMs and custom machine types as appropriate

2.3 Planning and configuring data storage options. Considerations include:

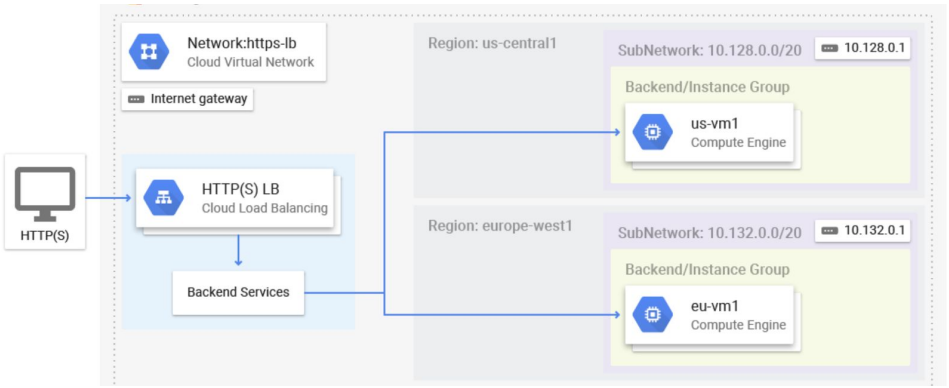
- 2.3.1 Product choice (e.g., Cloud SQL, BigQuery, Firestore, Cloud Spanner, Cloud Bigtable)
- 2.3.2 Choosing storage options (e.g., Zonal persistent disk, Regional balanced persistent disk, Standard, Nearline, Coldline, Archive)

Exam Guide - Compute Engine

4.1 Managing Compute Engine resources. Tasks include:

- 4.1.1 Managing a single VM instance (e.g., start, stop, edit configuration, or delete an instance)
- 4.1.2 Remotely connecting to the instance
- 4.1.3 Attaching a GPU to a new instance and installing necessary dependencies
- 4.1.4 Viewing current running VM inventory (instance IDs, details)
- 4.1.5 Working with snapshots (e.g., create a snapshot from a VM, view snapshots, delete a snapshot)
- 4.1.6 Working with images (e.g., create an image from a VM or a snapshot, view images, delete an image)
- 4.1.7 Working with instance groups (e.g., set autoscaling parameters, assign instance template, create an instance template, remove instance group)
- 4.1.8 Working with management interfaces (e.g., Google Cloud console, Cloud Shell, Cloud SDK)

Scalable, Fault Tolerant Architecture Illustrated



This illustration shows a fault tolerant architecture using load balancing and autoscaling.

In this example, HTTP(S) traffic is being delivered to an Global Load Balancer. The load balancer is a managed service, so performance and high availability is built into the service.

The backend service is responsible for distributing traffic to the appropriate region based on latency.

Each region, in this example, has a managed instance group of instances that will be increased or decreased based on a predefined metric you choose based on the application.

The infrastructure shown can survive a region outage and still function.

Scalable, Fault-Tolerant Architecture

To set up what is diagrammed on the previous slide, you need to create:

An instance template	One or more instance groups	A load balancer	A backend service

In order to build the infrastructure in the previous slide, you would need to:

- Build an instance template
- Create one or more managed instance groups(MIG) at the regions required
- Build a global load balancer with the MIG as the backend service

Compute Engine elasticity - Instance Groups

- Two types
  - **Managed instance groups**
    - Group of **identical** machines created from a template
    - Use case: When need VM elasticity based on demand
  - **Unmanaged instance groups**
    - Group of VMs with **different configurations**
    - Use case: Lift and shift of on-premise workloads that need a load balancer to serve traffic
      - No automatic elasticity or automatic healing

https://cloud.google.com/compute/docs/instance-groups

Instance groups:
https://cloud.google.com/compute/docs/instance-groups/

Basic scenarios for creating managed instance groups (MIGs):
https://cloud.google.com/compute/docs/instance-groups/creating-groups-of-managed-instances#create_managed_group

Managed instance groups create VMs based on instance templates

- Instance templates define the VMs: image, machine type, etc.
  - Test to find the smallest machine type that will run your program
- Instance group manager creates the machines.
- Set up auto scaling to optimize cost and meet varying user workloads.
- Add a health check to enable auto healing.
- Use multiple zones for high availability.

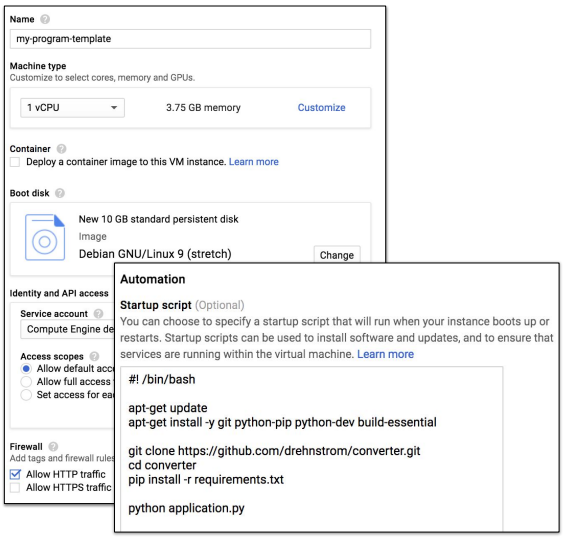
Managed instance groups create VMs based on instance templates. Instance templates are just a resource used to define VMs and managed instance groups. The templates define the boot disk image or container image to be used, the machine type, labels, and other instance properties like a startup script to install software from a Git repository.

The virtual machines in a managed instance group are created by an instance group manager. Using a managed instance group offers many advantages, such as autohealing to re-create instances that don’t respond and creating instances in multiple zones for high availability.

Creating an Instance Template

Contains the information required to build a VM for a deployment

- Like building a VM, but don't create the instance, create the template
- Include a startup script to automate code deployment
- Can also use a custom image



Instance template:
https://cloud.google.com/compute/docs/instance-templates/

Create instance templates:
https://cloud.google.com/compute/docs/instance-templates/create-instance-templates

Creating an instance template is similar to creating a virtual machine.

The template will include all information required to build the instance.

The information provided to build the template can include startup scripts to automate code deployment if needed.

Creating instance template with gcloud

- Creating a template with a specified machine type and image

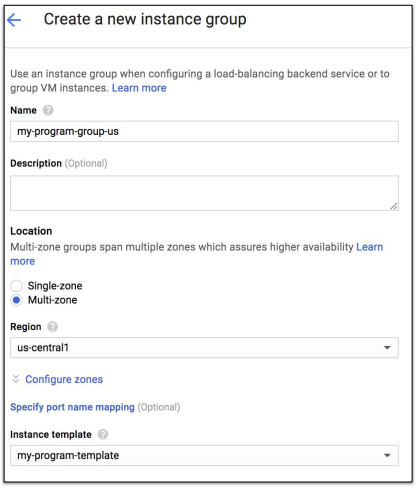
```
gcloud compute instance-templates create example-template \
    --machine-type=e2-standard-4 \
    --image-family=debian-10 \
    --image-project=debian-cloud \
    --boot-disk-size=250GB
```

- Refer to the documentation for other syntax examples

Create a new instance template
https://cloud.google.com/compute/docs/instance-templates/create-instance-templates#create_a_new_instance_template

Instance Groups

- Instance groups create machines based on instance templates
- Specify how many machines to create
  - In which region
- Based on what template
- Autoscaler adds and removes machines based on demand
- Health check ensures machines are working



An instance group enables you to manage a group of virtual machines as a single entity.

All virtual machines in a managed instance group are built from the same template.

When building the group, you specify:

- Number and location of instances
- Template to build from
- Metric or metrics to use to scale instances
- Health check to verify machines are working properly

Creating instance group with `gcloud`

- Create a regional MIG with 3 instances within the us-east1 region

```
gcloud compute instance-groups managed create example-rmig \
    --template example-template  \
    --size 3 \
    --region us-east1
```

- Create a zonal MIG with 3 instances within the us-east1 region

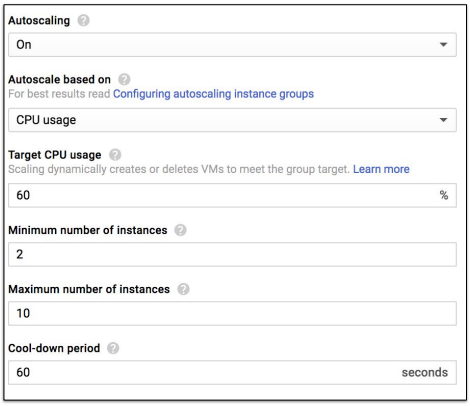
```
gcloud compute instance-groups managed create example-zmig \
    --template example-template \
    --size 3 \
  --zones us-east1-b,us-east1-c
```


Create a MIG in a single zone:
https://cloud.google.com/compute/docs/instance-groups/create-zonal-mig

Create a MIG with VMs in multiple zones in a region:
https://cloud.google.com/compute/docs/instance-groups/distributing-instances-with-regional-instance-groups

Specifying an Autoscaling Configuration

- Defines when to create or destroy machines
- Pick a metric and a threshold
  - CPU utilization
  - Load balancing capacity
  - Monitoring metrics
  - Predictions based on history
- Specify minimum and maximum number of machines



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

Autoscaling groups of instances:
https://cloud.google.com/compute/docs/autoscaler/#managed_instance_groups

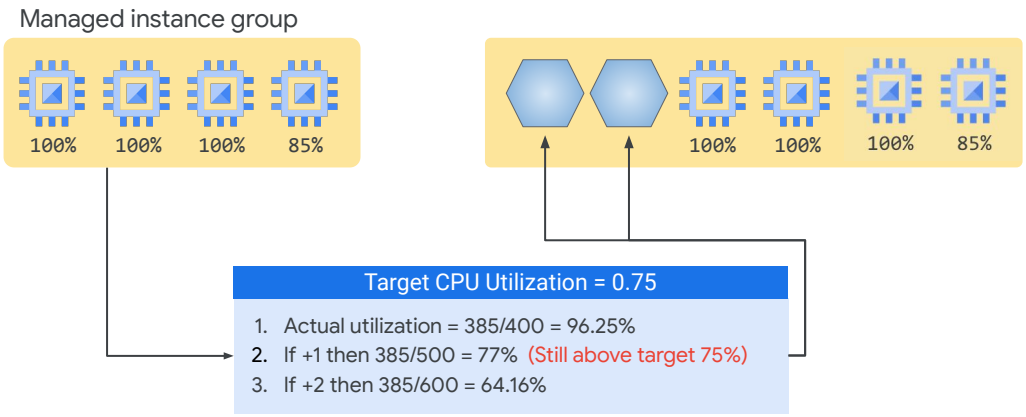
Create a MIG with autoscaling enabled:
https://cloud.google.com/compute/docs/instance-groups/create-mig-with-basic-autoscaling

Part of the configuration of the managed instance group includes how to configure autoscaling. This configuration defines when to create and destroy instances.

Pick the metrics and thresholds to use.

Specify a minimum and maximum number of instances to run.

Example Scale-out policy decision



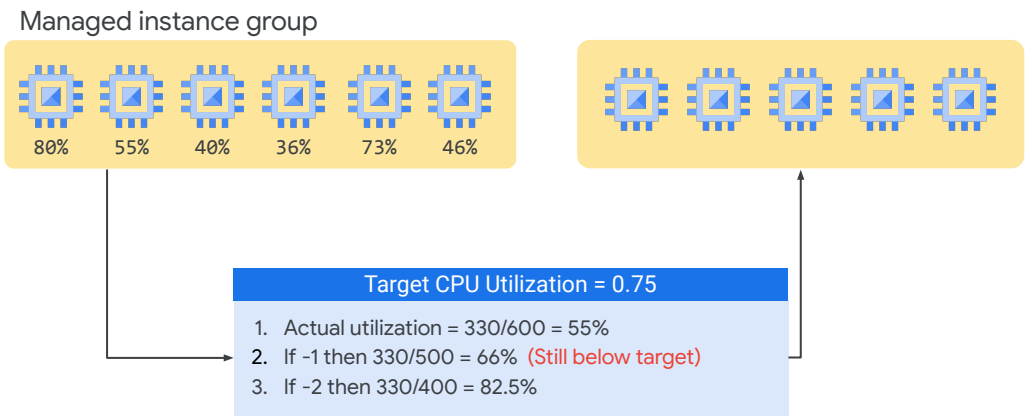
The percentage utilization that an additional VM contributes depends on the size of the group. The 4th VM added to a group offers 25% increase in capacity to the group. The 10th VM added to a group only offers 10% more capacity, even though the VMs are the same size.

In this case shown in the diagram Autoscaler is conservative and rounds up. In other words, it would prefer to start an extra VM that isn't really needed than to possibly run out of capacity.

https://cloud.google.com/compute/docs/autoscaler/understanding-autoscaler-decisions

https://cloud.google.com/compute/docs/autoscaler/multiple-policies

Example Scale-in policy decision



In this example, removing one VM doesn't get close enough to the target of 75%. Removing a second VM would exceed the target. Autoscaler behaves conservatively. So it will shut down one VM rather than two VMs. It would prefer underutilization over running out of resource when it is needed.

Autoscaling instance group with gcloud

- Set autoscaling based on target CPU utilization

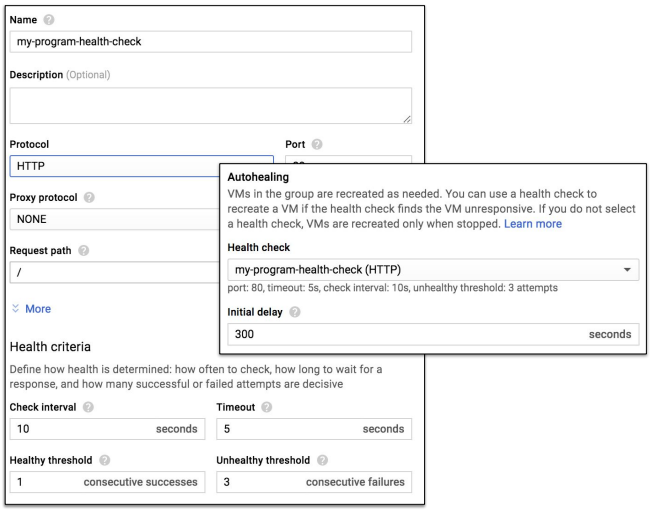
```
gcloud compute instance-groups managed set-autoscaling example-rmig \
    --max-num-replicas 20 \
    --target-cpu-utilization 0.60 \
    --cool-down-period 90
```

Cool down period: Specifies how long it takes for your application to initialize

When scaling out, the autoscaler ignores data from VMs that are still initializing because those VMs might not yet represent normal usage of your application

Health Checks

- Makes requests to the machines in the instance groups
- If no response, they are shut off and new ones created
- Parameters control:
  - Where to make request
  - Via what port
  - How often



Setting up health checking and autohealing: https://cloud.google.com/compute/docs/instance-groups/autohealing-instances-in-migs

Health checks ensure that only fully operable instances are being used from the group. With health checks, instances that fail the health check can be excluded and new instances can take their place.

Parameters required to configure the health check include:

- Where to make the request
- What port to use
- And how often to run the check

Health checks need a firewall rule to allow incoming probes from certain ports. See https://cloud.google.com/compute/docs/instance-groups/autohealing-instances-in-migs

Creating health checks with gcloud

- Create a regional MIG with 3 instances within the us-east1 region

```
gcloud compute health-checks create http example-check --port 80 \
       --check-interval 30s \
       --healthy-threshold 1 \
       --timeout 10s \
       --unhealthy-threshold 3
```

- Looks for response on port 80
- Checks every 30 seconds
- Considered “healthy” if it responds within 1st attempt
- Waits 10 seconds to get a response
- Considered “unhealthy” if it fails to respond within 3 attempts

Exam Guide - Compute Engine

3.1 Deploying and implementing Compute Engine resources. Tasks include:

- 3.1.1 Launching a compute instance using the Google Cloud console and Cloud SDK (gcloud) (e.g., assign disks, availability policy, SSH keys)
- 3.1.2 Creating an autoscaled managed instance group using an instance template
- 3.1.3 Generating/uploading a custom SSH key for instances
- 3.1.4 Installing and configuring the Cloud Monitoring and Logging Agent
- 3.1.5 Assessing compute quotas and requesting increases

2.2 Planning and configuring compute resources. Considerations include:

- 2.2.1 Selecting appropriate compute choices for a given workload (e.g., Compute Engine, Google Kubernetes Engine, Cloud Run, Cloud Functions)
- 2.2.2 Using preemptible VMs and custom machine types as appropriate

2.3 Planning and configuring data storage options. Considerations include:

- 2.3.1 Product choice (e.g., Cloud SQL, BigQuery, Firestore, Cloud Spanner, Cloud Bigtable)
- 2.3.2 Choosing storage options (e.g., Zonal persistent disk, Regional balanced persistent disk, Standard, Nearline, Coldline, Archive)

Exam Guide - Compute Engine

4.1 Managing Compute Engine resources. Tasks include:

- 4.1.1 Managing a single VM instance (e.g., start, stop, edit configuration, or delete an instance)
- 4.1.2 Remotely connecting to the instance
- 4.1.3 Attaching a GPU to a new instance and installing necessary dependencies
- 4.1.4 Viewing current running VM inventory (instance IDs, details)
- 4.1.5 Working with snapshots (e.g., create a snapshot from a VM, view snapshots, delete a snapshot)
- 4.1.6 Working with images (e.g., create an image from a VM or a snapshot, view images, delete an image)
- 4.1.7 Working with instance groups (e.g., set autoscaling parameters, assign instance template, create an instance template, remove instance group)
- 4.1.8 Working with management interfaces (e.g., Google Cloud console, Cloud Shell, Cloud SDK)


Google Cloud