Linux Academy

Master Builder's Guide to Google Cloud

Choose a Section

Return to Table of Contents

Choose a Lesson

Return to Table of Contents

Choose a Lesson

Role of a Google Cloud Architect

Architect Exam and Course Overview

Role of a Google Cloud Architect

Next

Google's definition

- A Professional Cloud Architect enables organizations to leverage Google Cloud technologies. With a **thorough** understanding of cloud architecture and Google Cloud Platform, this individual can design, develop, and manage robust, secure, scalable, highly available, and dynamic solutions to drive business objectives.
- What does this mean in practice?

What does a traditional architect do?

- Plan, design, and review construction of buildings
- Poorly designed building does not withstand test of time and may collapse
- Must know which tools to build with and the best use of each one
- Manage business requirements
  - Budget, time line, redundancy, efficiency, special requirements

How a Cloud Architect is similar

- Plans, designs, builds, and manages **cloud** infrastructure
- Must be skilled in tools of the trade (and choose the best tool for the job)
- Infrastructure must stand the test of time
  - Future-proofing, redundancy, disaster recovery
- Manage business requirements
  - Authentication, costs, scalability, redundancy

How a Cloud Architect is different

- Not constrained by physical resources
  - Near infinite resources available
  - Scale, automate, and expand with no limits
  - Global reach
- Design with change in mind
  - "How will this scale?" "How will new features change this design?"
- Design with **scalability** and **automation** in mind
  - Common theme for the exam

As we move forward...

- Adopt the mindset of scalability and automation
- Start small, but start thinking at a larger scale
- Think like an architect! Consider how to design solutions for different requirements

Return to Table of Contents

Choose a Lesson

Role of a Google Cloud Architect

Architect Exam and Course Overview

Role of a Google Cloud Architect

Previous

Cloud Architect vs. Cloud Engineer

- Cloud Engineer = tactical
  - Operate, monitor, and modify jobs as needed
  - Example: Run jobs on a GKE cluster and measure performance
- Cloud Architect = strategic/design
  - Design and implement the solution to meet requirements
  - Example: Design and implement GKE cluster to meet customer requirements

Return to Table of Contents

Choose a Lesson

Role of a Google Cloud Architect

Architect Exam and Course Overview

Architect Exam and Course Overview

Next

Taking the exam

- 50 questions
- 2 hours
- Passing score not published by Google
- Pass/Fail only feedback at end of exam
  - No score
  - No breakdown of performance by topic
- Combination of case study and stand-alone questions

Exam scope

- VERY broad range of topics—a little bit of everything on GCP
- Some topics are conceptual, others quite detailed
- Mix of low-level (technical) and high-level (conceptual) topics
- Familiarity with technical topics help prepare for conceptual questions
- Heavy emphasis on meeting specific business requirements
  - Save costs, global scope, use managed services
- "Mile wide, inch deep, with a few deep holes here and there"

Case Studies

- Three possible studies (number on exam may vary)
- Available to view in advance
- Learn them inside and out
- Very representative of businesses who want to migrate to GCP
- Roughly 20-40% of exam questions (may vary per person)

Return to Table of Contents

Choose a Lesson

Role of a Google Cloud Architect

Architect Exam and Course Overview

Architect Exam and Course Overview

Previous

Two primary 'themes' in this course

- Become expert craftsmen using GCP's tools
- Use those tools to design solutions
  - In context of exam objectives
- Some topics will have deeper focus than others to match exam

Importance of hands-on practice

- PRACTICE, PRACTICE, PRACTICE
- Emphasis on hands-on approach
- Do the labs
- Do not be afraid to experiment and break something
- Create, experiment, break, learn, repeat
  - Take notes

Return to Table of Contents

Choose a Lesson

Return to Table of Contents

Choose a Lesson

Management Services

Resource Hierarchy

Labels

Quotas

Management Services

In these next few sections:

- Cloud Resource Manager
  - Quotas, IAM, Billing
- Cloud IAM (Identity and Access Management)
- Stackdriver (monitor all the things!)

To put it another way....

- How is our stuff organized and what are we paying for?
- Who has access to our stuff?
- What is happening to our resources?

Role of the Cloud Architect (i.e., "Why this matters")

- Not as fun as creating and designing infrastructure, but still vitally important
- If not properly managed:
  - Unauthorized access
  - Runaway/hidden costs
  - Project resources halted due to improper billing setup
  - Application/service errors that we don't know about and don't know how to fix

Return to Table of Contents

Choose a Lesson

- Management Services
- Resource Hierarchy
- Labels
- Quotas

Resource Hierarchy

Next

What is the GCP Resource Hierarchy?

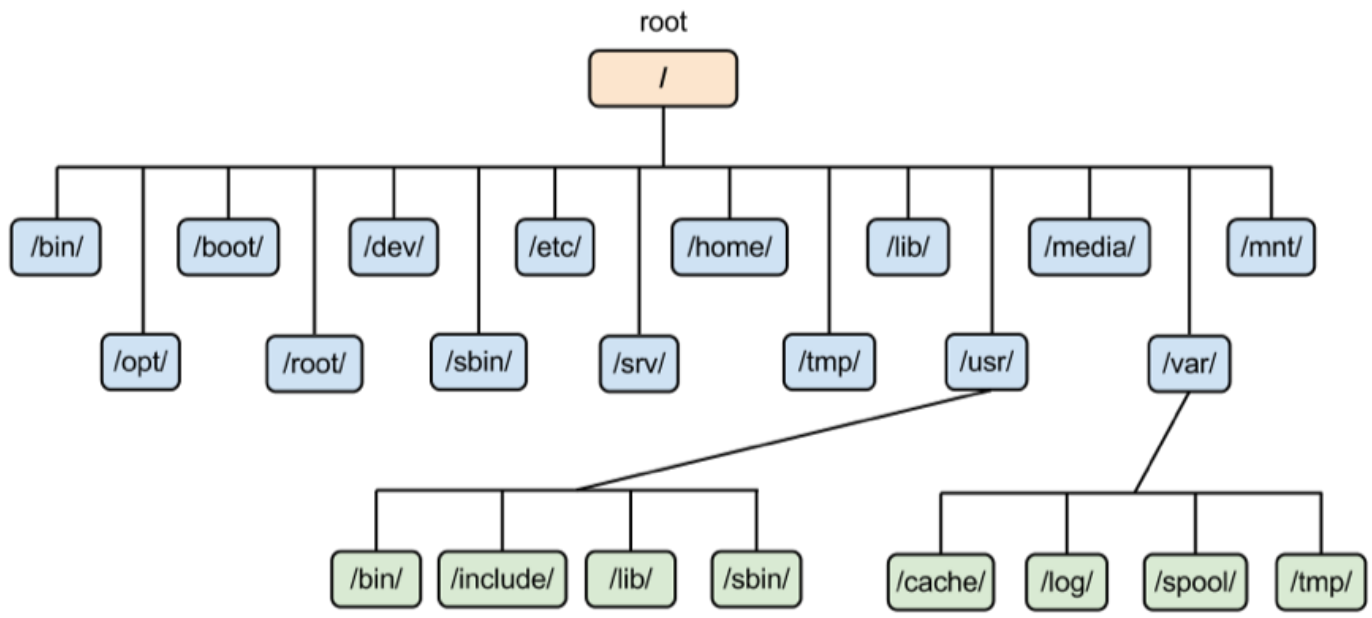
- Hierarchy of ownership via parent/child relationship
  - Identity and Access Management
  - Similar to a traditional file system
- Provide attach points and inheritance of access control and policies

Core principles

- Each child object has only one parent
- Permissions are inherited from top down
- More permissive parent policy always overrules more restrictive child policy

File System example

- Each child has only one parent
- Permissions are inherited from above



Return to Table of Contents

Choose a Lesson

- Management Services
- Resource Hierarchy
- Labels
- Quotas

Resource Hierarchy

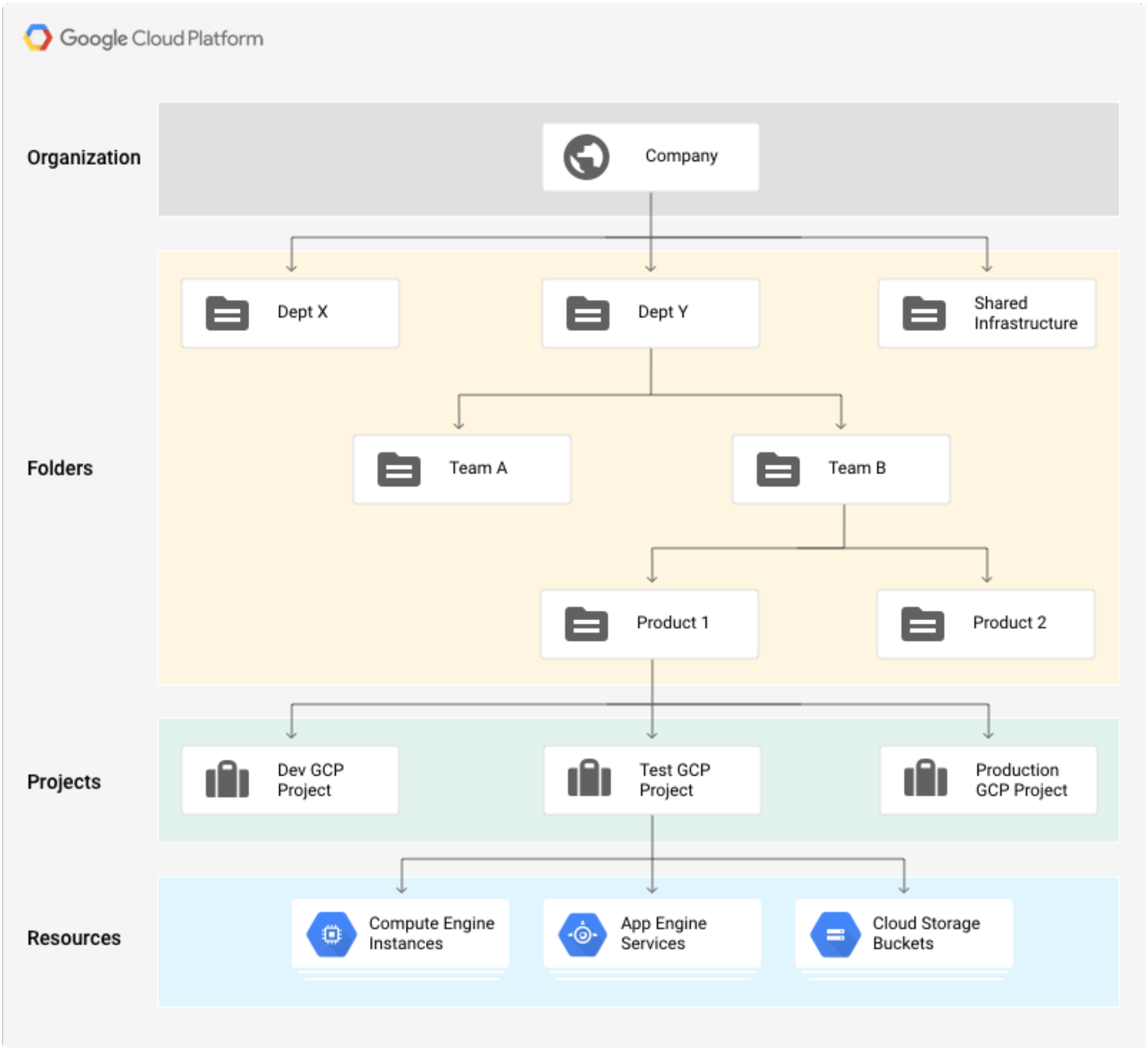
Previous

Next

Google Cloud Resource Hierarchy

- Organization (root node)
- Folders (optional)
- Projects
- Resources (inside projects)

Click each layer to learn more



Return to Table of Contents

Choose a Lesson

Management Services

Resource Hierarchy

Labels

Quotas

X

Organization Resource

- Represents an organization (e.g., a company) and is the **root** node in the GCP resource hierarchy.
- Hierarchical ancestor of project resources and Folders.
- IAM access control policies applied to the Organization resource are applied throughout the entire below hierarchy (through Folders, Projects, and Resources)
- Can grant access to different people in organization
- Not applicable to personal (e.g., Gmail) accounts

Key Roles

- Organization admin: Full power to edit all permissions
- Organization owner: Reserved for G Suite/Cloud Identity super admin

Return to Table of Contents

Choose a Lesson

Management Services

Resource Hierarchy

Labels

Quotas

X

Folders

- Additional (optional) grouping and isolation boundary between projects
- Collection of projects and other folders
- IAM roles applied to folder apply to all projects inside
- Useful for grouping by departments
- Useful for delegating administration rights
- Beware: Removing projects from folder will remove folder-applied IAM roles

Return to Table of Contents

Choose a Lesson

Management Services

Resource Hierarchy

Labels

Quotas

X

Projects

- **Core** organizational component of GCP
- Required to use any GCP resources
- Basis for creating, enabling, using, and paying for GCP services (i.e., everything)
- Exam tip: Become VERY familiar working with and managing projects

Identifiers

- Project ID (must be globally unique)
- Project number (automatically generated)
- Project name ("Friendly name")
    - Good practice to have identical Project name and ID

Return to Table of Contents

Choose a Lesson

X

Resources

- Everything that is created and used on GCP
  - Instances
  - Services
  - APIs
  - Cloud Storage buckets
  - Managed services
  - IAM policies

Return to Table of Contents

Choose a Lesson

Management Services

Resource Hierarchy

Labels

Quotas

Resource Hierarchy

Previous

Policy Inheritance

- Child nodes inherit parent permissions (all the way down)
  - Example: Project Editor role granted at Organization node applies to all folders, projects, and resources down the line
  - Example: Project Viewer role granted to folder applies to all projects and resources inside of folder
- More permissive parent policy will also overrule restrictive child policies
  - Example: If I grant Sue the Project Editor role at Organization node, I cannot remove Compute Admin role at project level.

Return to Table of Contents

Choose a Lesson

- Management Services
- Resource Hierarchy
- Labels
- Quotas

YOUR COMPANY NAME
Phone: 1-555-555-5555
Equipment Description
Serial Number: 1-2020321

Labels

Next

What are labels?

- Virtual 'sticky notes'
- Tool for organizing GCP resources
- Set in console, gcloud, or API
- No set rules on how to label—depends on your organization needs
- Search for and filter out resources by label
  - Useful for lookup tables and configuration files

How it works

- Key-value pair
- Key: Unique identifier
- Value: Identified data or pointer to data location
- Key cannot be empty, but Value can
- Up to 64 labels per resource

Examples

- Environment: env:prod/env:test
- Owner or point of contact: owner:matt, contact:devops
- Team or cost center: team:research, team:marketing
- App component: component:backend, component:frontend
- Resource state: state:readyfordeletion, state:inuse

Labels vs. Network Tags

- Labels:
  - Can be applied across all of GCP
  - Organization purposes only
  - Does not affect resource operation
- Network Tags:
  - Only for network/VPC resources
  - Affects resource operation (e.g., firewall rule application, network route)

Return to Table of Contents

Choose a Lesson

Management Services

Resource Hierarchy

Labels

Quotas

YOUR COMPANY NAME
Phone: 1-555-555-5555
Equipment Description
Serial Number: 1-2020321

Labels

Previous

Creating instance with labels

- `gcloud compute instances create [instance-name] --labels key=value,key=value`

List existing labels:

- `gcloud compute instances describe [instance-name] --format 'default(labels)'`

Update instance labels on existing resource

- `gcloud compute instances update [instance-name] --update-labels key=value,key=value`

Removing instance labels:

- `gcloud compute instances update [instance-name] --remove-labels key1,key2`
  - Only need to remove keys

Return to Table of Contents

Choose a Lesson

- Management Services
- Resource Hierarchy
- Labels
- Quotas

Quotas

What are quotas?

- Caps on resources you can create (typically per project)
  - Example: 24 total CPUs per region, 5 static IPs per project
- Prevent unexpected spikes in usage
- Generally one of three types:
  - Resources per project (global, not region-bound)
  - API rate limit requests per project
  - Per region
  - Be aware of overlap restrictions
- Free trial has additional quotas in place—trial should be primary for testing/evaluating

Why do we need quotas?

- Ability to instantly create infinite resources can lead to massive costs
- Protection from unexpected spikes in resource usage
- Prevent runaway consumption due to error or malicious intent
- Prevent unexpected spikes in billing
- Review sizing considerations
  - ”Do you really need a single 96 CPU VM?”

Increasing quota caps

- Most quotas are soft caps and can be raised by request
- Support ticket or self service form
- Quotas can be viewed in console
- Best practice—pro-actively request increase for anticipated demand

Service	Location	Current Usage	7 Day Peak Usage	Limit	
Compute Engine API CPUs	asia-east1	0	–	2,400	View hierarchy
Compute Engine API CPUs	asia-east2	0	–	24	View hierarchy
Compute Engine API CPUs	asia-northeast1	0	–	2,400	View hierarchy
Compute Engine API CPUs	asia-south1	0	–	24	View hierarchy

Return to Table of Contents

Choose a Lesson

Return to Table of Contents

Choose a Lesson

- IAM Overview
- Hands-On
- Service Accounts
- IAM Best Practices

IAM Overview

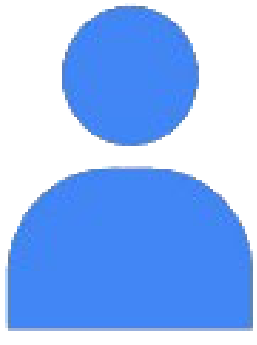
Next

Importance of IAM knowledge

- **Core**, **fundamental** skill required for both passing exam and performing role of cloud architect.

So what is Identity and Access Management (IAM)?

- Technical definition
  - With Cloud IAM, you can grant granular access to specific GCP resources and prevent unwanted access to other resources. Cloud IAM lets you adopt the security principle of least privilege, so you grant only the necessary access to your resources.
- Simple breakdown
  - Who = **Member**
  - Can do what = **Role**
  - On which resource = All resources on GCP



Who | can do what | on which resource

Click each item to learn more

Identity

Google Account

Service Account

Google Group

Google Apps Domain

Role

Owner

Viewer

Editor

compute.instanceAdmin

storage.objectAdmin

. . .

Resource

Cloud Platform

Projects

Compute Engine

App Engine

Cloud Storage

Cloud Logging

. . .

Return to Table of Contents

Choose a Lesson

IAM Overview

Hands-On

Service Accounts

IAM Best Practices

X

Who = Member = Identity

- Person or a service account
- Identity = email address
  - Log in with email address (joe@companyname.com)
- People authenticate via Google account (in form of email address)
- Service account = application/server account
  - Not associated with a person

Return to Table of Contents

Choose a Lesson

IAM Overview

Hands-On

Service Accounts

IAM Best Practices

X

Can do what = Role = collection of permissions

- Permissions: Determine what operations are allowed on resource
- Permissions format: <service>.<resource>.<verb>
  - compute.instances.delete
- Permissions are not directly assigned to **member** (the 'who')
- Permissions are bundled into **roles** — roles assigned directly to user
  - Members can be assigned multiple roles
- Example: **Role** of compute.instanceAdmin assigned to **Member** Joe (joe@company.com) includes multiple permissions

Return to Table of Contents

Choose a Lesson

IAM Overview

Hands-On

Service Accounts

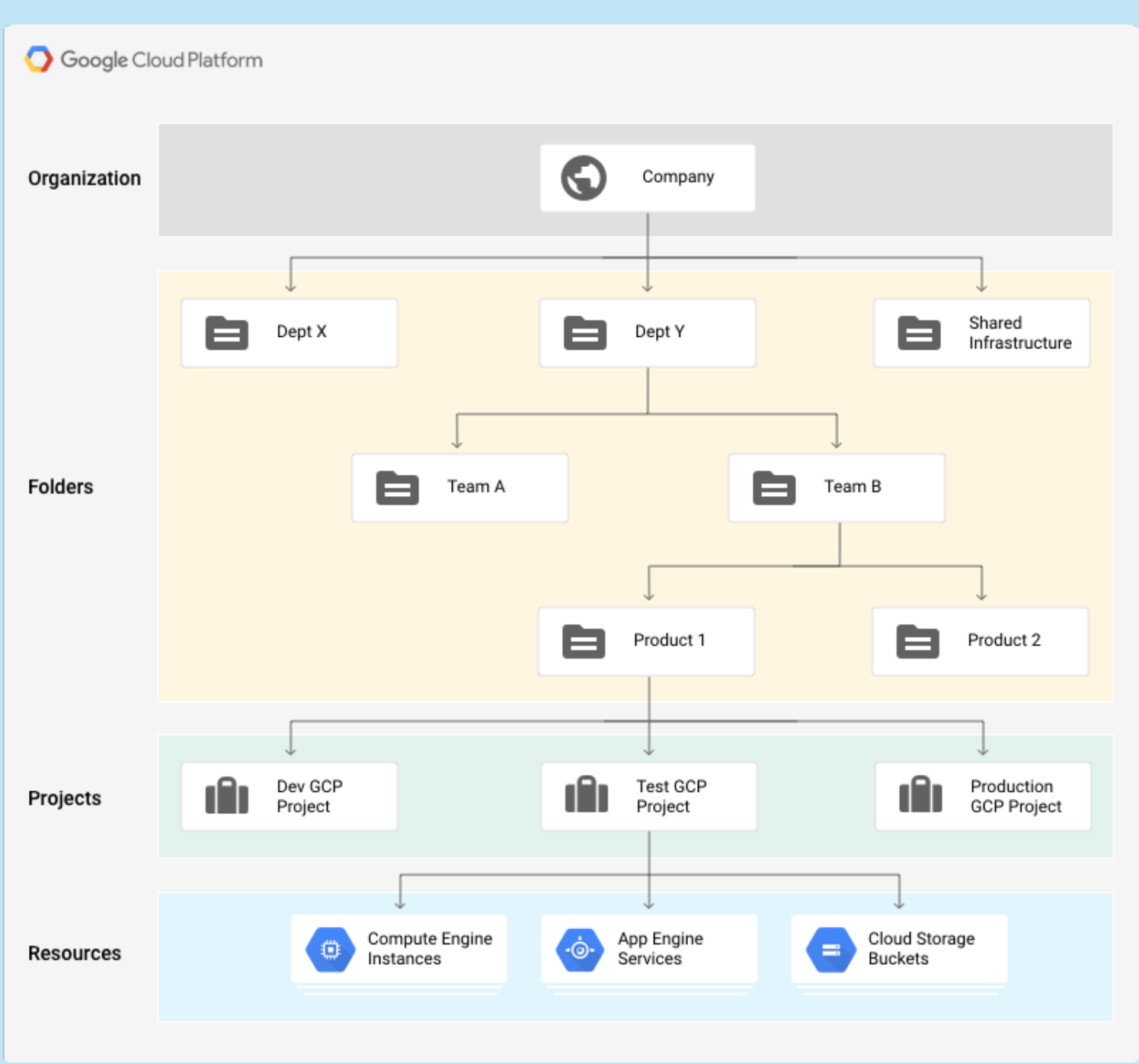
IAM Best Practices

X

Resource = What are we giving access to?

- All of the components of GCP
  - Compute Engine VMs
  - Cloud Storage buckets
- Includes Organizations, Folders, Projects, Services, and all resources inside

Resources



Return to Table of Contents

Choose a Lesson

- IAM Overview
- Hands-On
- Service Accounts
- IAM Best Practices

IAM Overview

Previous

Next

Members - Person = Google Account

- Google account = any person who interacts with GCP
- **Always** associated with email address
  - @gmail.com, @company.com
- NOT set up in GCP
  - Create, edit, delete Google accounts outside of GCP IAM (admin.google.com)
  - Give GCP access to Google accounts with GCP IAM

Google account types

- Personal - joe@gmail.com
  - Non-Gmail accounts associated with Google account (joe@yahoo.com)
- G Suite Domain
  - Represents company Internet domain name (companyname.com)
  - Uses company-wide Google apps (Gmail, Drive, etc.)
- Cloud Identity Domain
  - Similar to G Suite domain
  - Represents virtual group of all Google accounts in an organization
  - No access to G Suite applications (Gmail, Drive, etc)
  - Can sync existing directory service (e.g., Active Directory) to Cloud Identity
- Google Group - Collection of Google accounts
  - Also represented by email address (googlegroup@companyname.com)
  - Apply single policy to collection of users in group
  - No login credentials—not used to establish individual identity
  - Can include Service Accounts
- allAuthenticatedUsers
  - Special identifier to represent any GCP account
  - Not available for every service
  - Anonymous visitors not included
- allUsers (i.e., 'public')
  - Another special identifier
  - Not available for every service
  - Anyone and everyone (including anonymous visitors)

Service Accounts

- Belong to an application/server
- Not associated with a person
- Carry out application/server interactions
  - Example: Local server backup application writing data to Cloud Storage
  - Example: Web application subscribed to Pub/Sub topic
- Also identified by email address (e.g., <project_id>@developer.gserviceaccount.com)

Return to Table of Contents

Choose a Lesson

IAM Overview

Hands-On

Service Accounts

IAM Best Practices

IAM Overview

Previous

Next

Cloud Identity - sync with Active Directory

Why does this matter?

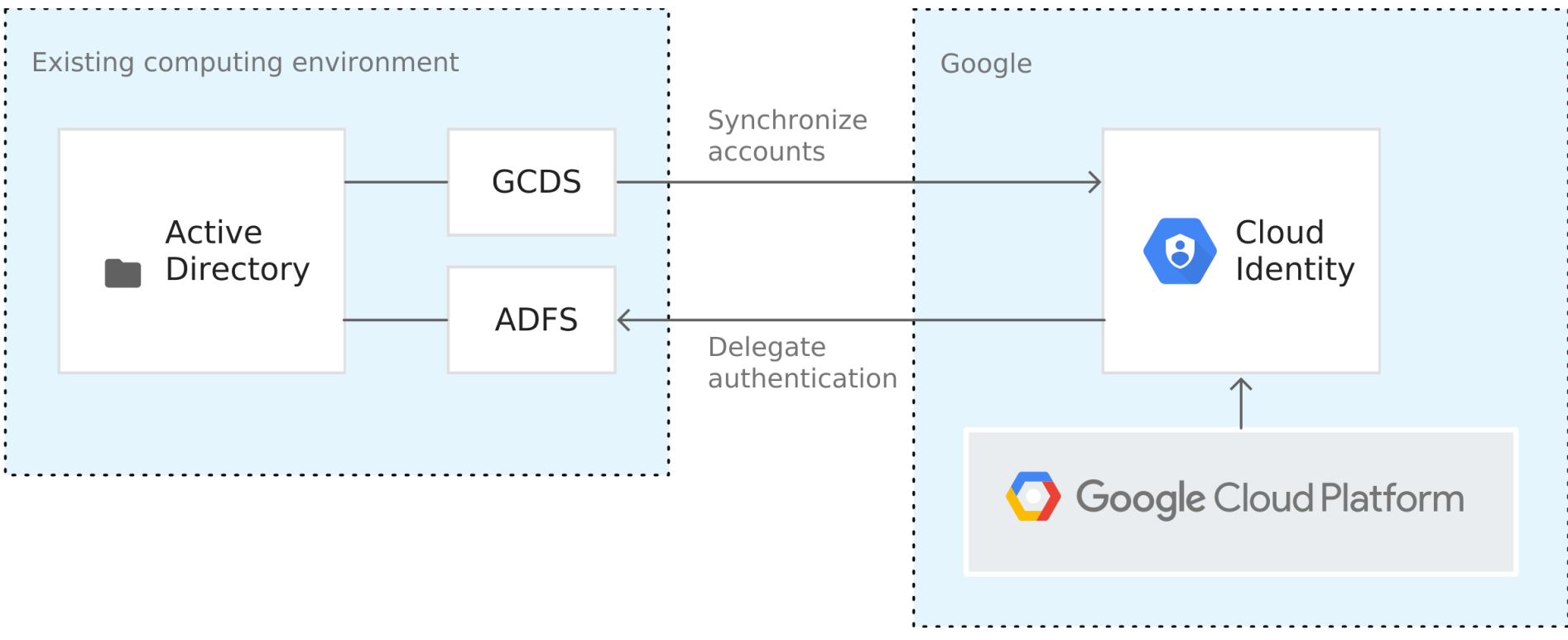
- Businesses not using G Suite need to maintain Google accounts under company domain for access
- Separately maintaining GCP accounts alongside Active Directory is cumbersome and error prone
- Need single source of identity management to manage both GCP and non-GCP identities

How it works

- Cloud Identity maps (or federates) AD accounts to Cloud Identity accounts
  - Mapping = Federation
- Active Directory is the 'single source of truth'.
- One-way sync from AD to Cloud Identity, not the other way around
- All user/password management done in AD
  - Create/modify/delete all in AD
  - Cloud Identity automatically creates Google account from AD accounts (or subset of accounts if desired)
- Provides single sign-on capability to GCP

How to synchronize - tools

- Google Cloud Directory Sync (GCDS)
  - Google-provided tool for synchronization process
  - Runs on AD server
- Active Directory Federation Services (ADFS)
  - Microsoft-provided tool



Return to Table of Contents

Choose a Lesson

IAM Overview

Hands-On

Service Accounts

IAM Best Practices

IAM Overview

Previous

Next

Role Types

- Primitive
- Predefined
- Custom

Primitive Roles

- Broad, original roles available on GCP (before current IAM environment)
- Applied across entire project
- Owner: Modify all resources and manage IAM and billing
- Editor: Modify all resources, no access to manage IAM and billing
- Viewer: View resources, cannot make changes

Predefined Roles

- More granular/specific access, not across entire project
- Applied to single service
- Example: compute.instanceAdmin allows access to modify instances, but does not affect any other service.

Custom Roles

- Even more granular than predefined roles
- Combine individual permissions when predefined roles are not specific enough

Return to Table of Contents

Choose a Lesson

- IAM Overview
- Hands-On
- Service Accounts
- IAM Best Practices

IAM Overview

Previous

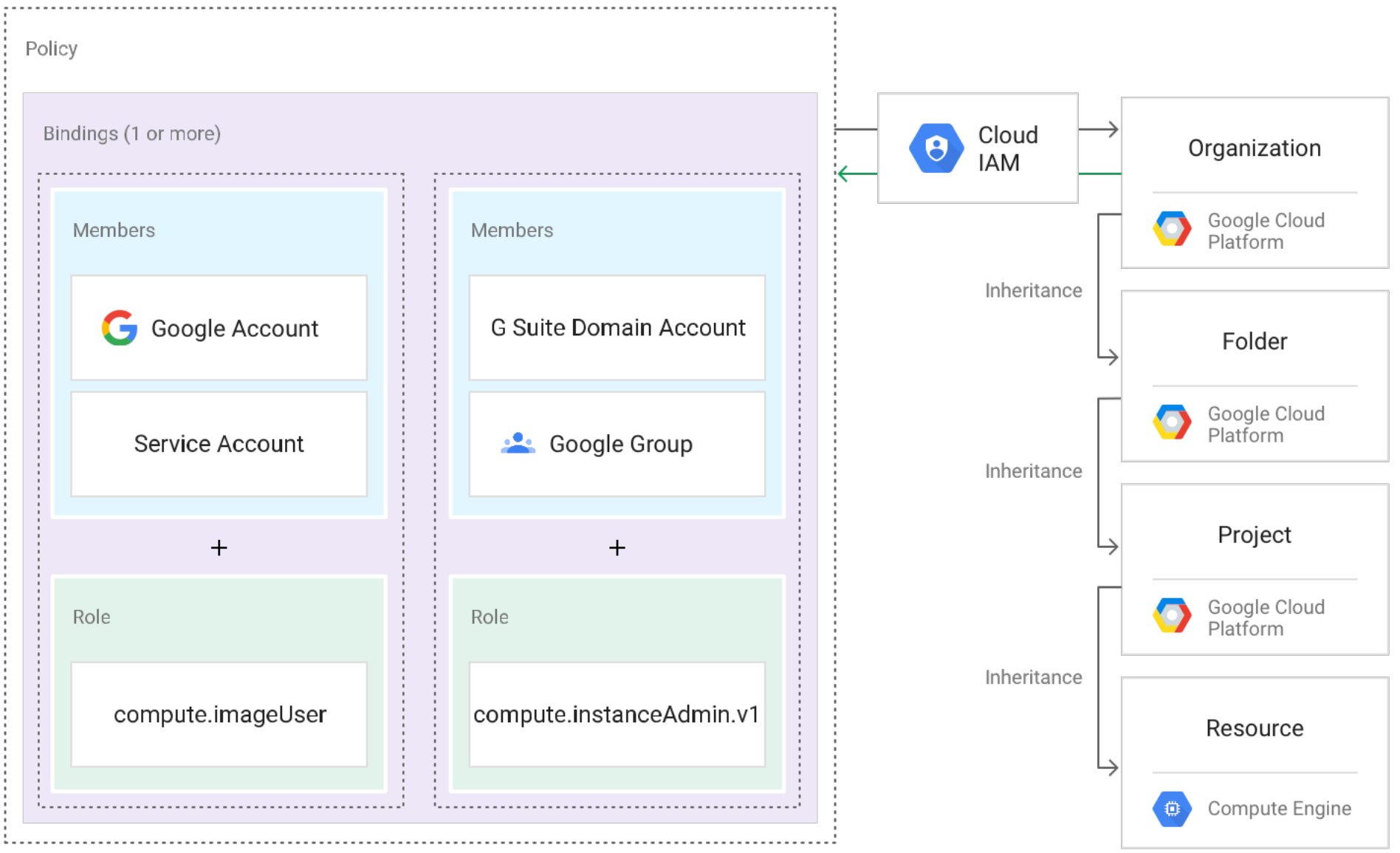
Next

Putting it together - IAM Policy

- How you grant roles to users
- Collection of statements that define who has what type of access
- Binding: Binds a list of members to a role

Policy Inheritance (or Hierarchy)

- Roles are enforced from the top down
- Organization > Folders > Project > Service/Resource
- Each child has exactly one parent
- Inherits roles from parent
- Permissive parent policy overrules child policies



Return to Table of Contents

Choose a Lesson

IAM Overview

Hands-On

Service Accounts

IAM Best Practices

IAM Overview

Previous

Putting EVERYTHING together:

IAM policies:

- **Grant members**... - (Users, groups, organizations, service accounts)
- ...**various roles**... - Primitive (broad), predefined/custom (granular)
- ...**in a hierarchal format**... - Parent overrules child
- ...**to GCP Resources**. - All layers of GCP

Examples:

- joe@linuxacademy.com is granted Owner role to Project 'Dev Environment'
- pw-dev-env@developer.gserviceaccount.com granted App Engine Service Admin role to App Engine

Return to Table of Contents

Choose a Lesson

IAM Overview

Hands-On

Service Accounts

IAM Best Practices

Service Accounts

Next

What is a Service Account?

- Special type of Google account
  - Not attached to a user
- Resources don't need end-user authentication
- Identity represented by email address
- Example: 126692436861-compute@developer.gserviceaccount.com
- Authenticate between applications and GCP services
- By default, GCE instances use service accounts for GCP access

Types of Service Accounts

- Google-managed
  - Represent different Google services and are automatically granted IAM roles
  - [PROJECT_NUMBER]@cloudservices.gserviceaccount.com
  - Generally invisible to end user
- User-managed
  - Created for/by you, based on enabled APIs in project
  - [PROJECT-NUMBER]-compute@developer.gserviceaccount.com
  - [PROJECT-ID]@appspot.gserviceaccount.com
  - Both automatically created and user-created

Both a Member and a Resource

- Service accounts are both a member (who) and a resource
  - Service accounts are granted permissions to a resource
  - Users (person) are granted role serviceAccountUser to a service account

Member	Role	Resource
bob@professionalwireless.net	Service Account User	Service Accounts
Service Account	Storage Admin	Cloud Storage Bucket

Return to Table of Contents

Choose a Lesson

- IAM Overview
- Hands-On
- Service Accounts
- IAM Best Practices

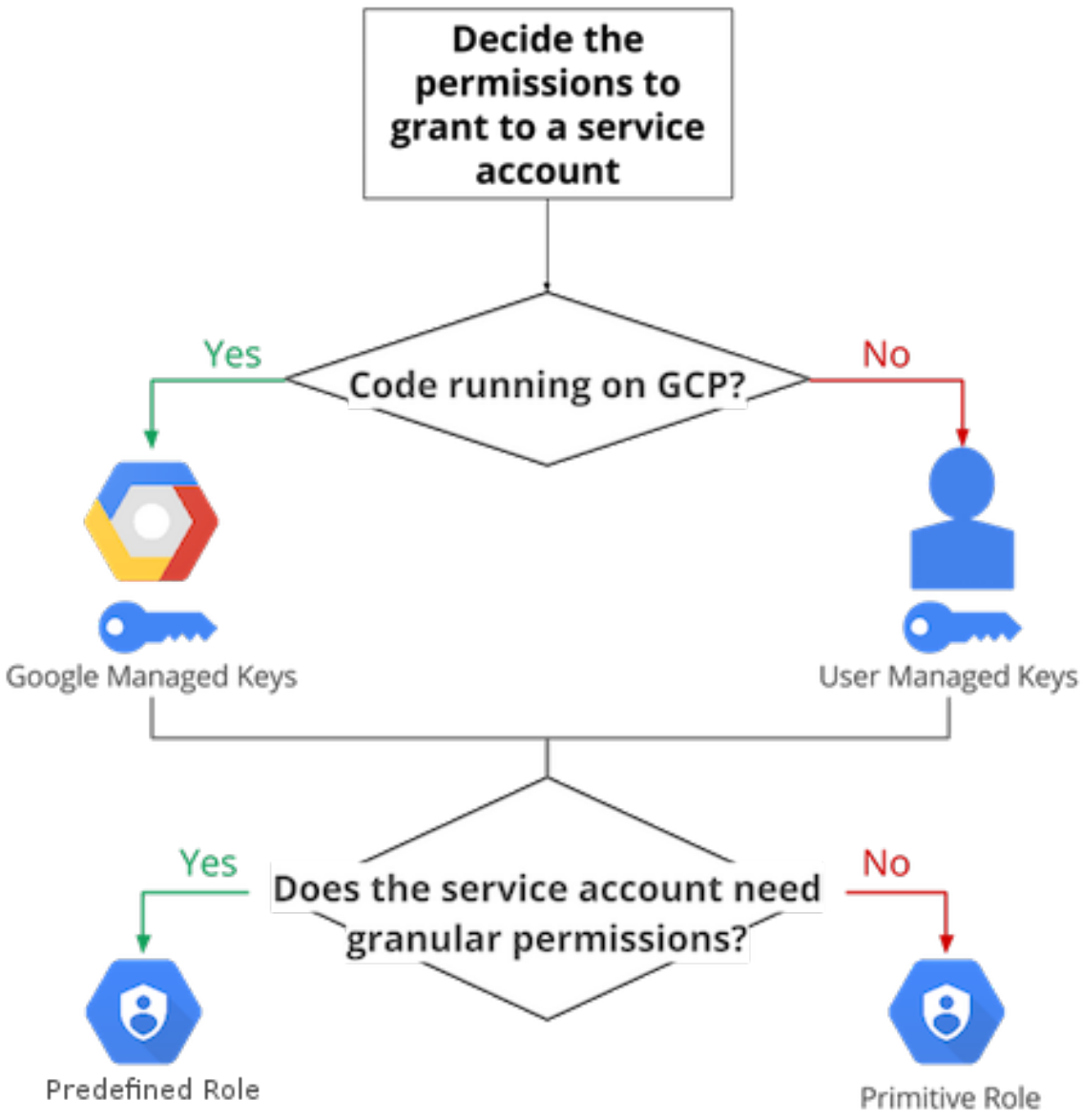
Service Accounts

Previous

Next

Service Account Keys

- Service account access managed by **account keys**
  - Think of it as the account's 'password'
- Default service account keys are managed by Google and can't be accessed/edited
- Custom service accounts can use user-managed (custom) keys, which you store/manage
  - Google maintains public copy for verification, but the public/private key pair is yours to manage
  - If you lose your private copy of the key, Google cannot retrieve it!
  - Can also use Google-managed keys



Return to Table of Contents

Choose a Lesson

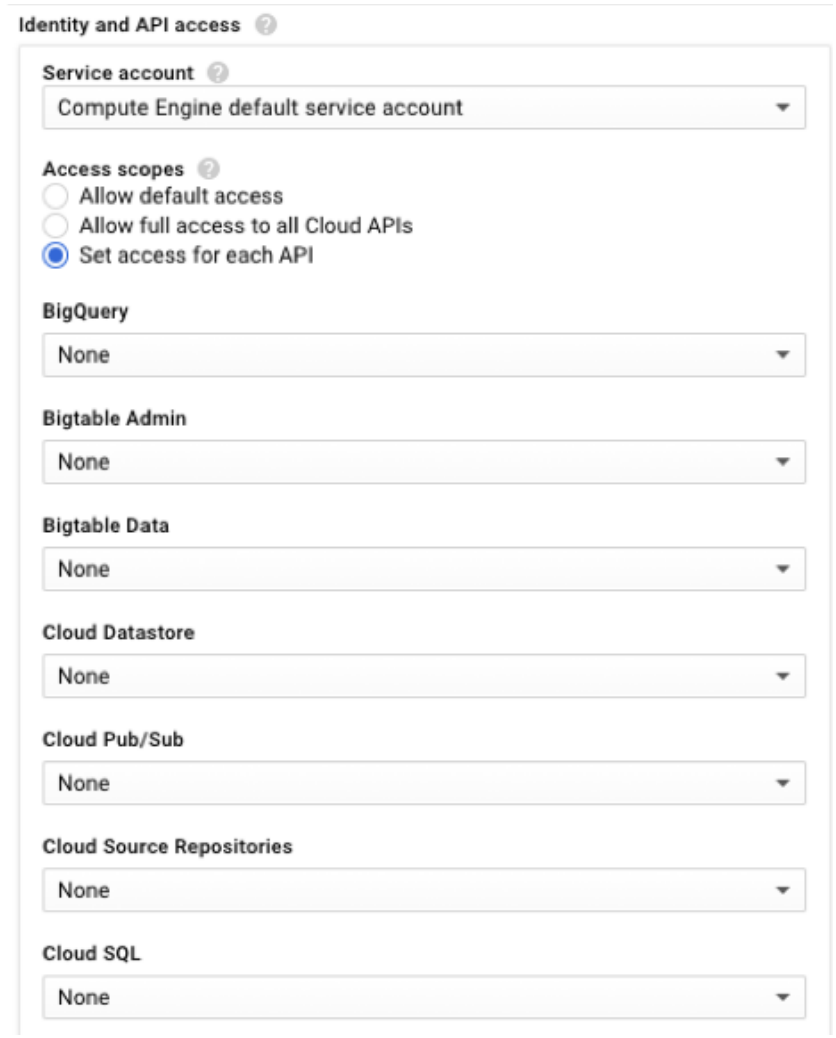
- IAM Overview
- Hands-On
- Service Accounts
- IAM Best Practices

Service Accounts

Previous

Service Account Scopes

- Legacy method of granting permissions for **default** service accounts for an individual instance
- Grant per-instance permissions to other GCP resources via the instance
- IAM roles or scopes determine service account permission for that instance



Putting it all together:

- Service accounts grant application/VMs access
- Users are granted access to act as a service account
  - Service accounts are granted access to resources.
- Service accounts use keys for access
- Service accounts granted access based on both scopes and IAM

Return to Table of Contents

Choose a Lesson

IAM Overview

Hands-On

Service Accounts

IAM Best Practices

IAM Best Practices

Principle of Least Privilege

- Give people just enough access to do what they need, and not any more
- Use predefined roles over primitive roles
- Grant roles at smallest scope necessary
  - e.g., Compute Instance Admin vs. Compute Admin
- (For service accounts) - Treat each app component as a separate trust boundary
  - Create separate service account for each service
- Restrict service account access
- Restrict who can create and manage service accounts (Service Account Admin)
- Careful with Owner roles (Editor might be better)
  - Owner can change IAM policy, Billing

Service Accounts

- Rotate service account keys (user managed)
  - **IMPORTANT**: Don't delete service in use by running resources
- Don't check in service account keys to source code or leave in downloads directory!
- Name service keys to reflect use and permissions

Auditing

- Use Cloud Audit logs to regularly audit IAM policy changes
- Export audit logs to Cloud Storage for long term retention
- Restrict log access with cloud logging roles
- We will discuss auditing later in this course

Other best practices

- When possible, use groups
- Have more than one Organization Administrator
- Separate production and development environments into separate projects

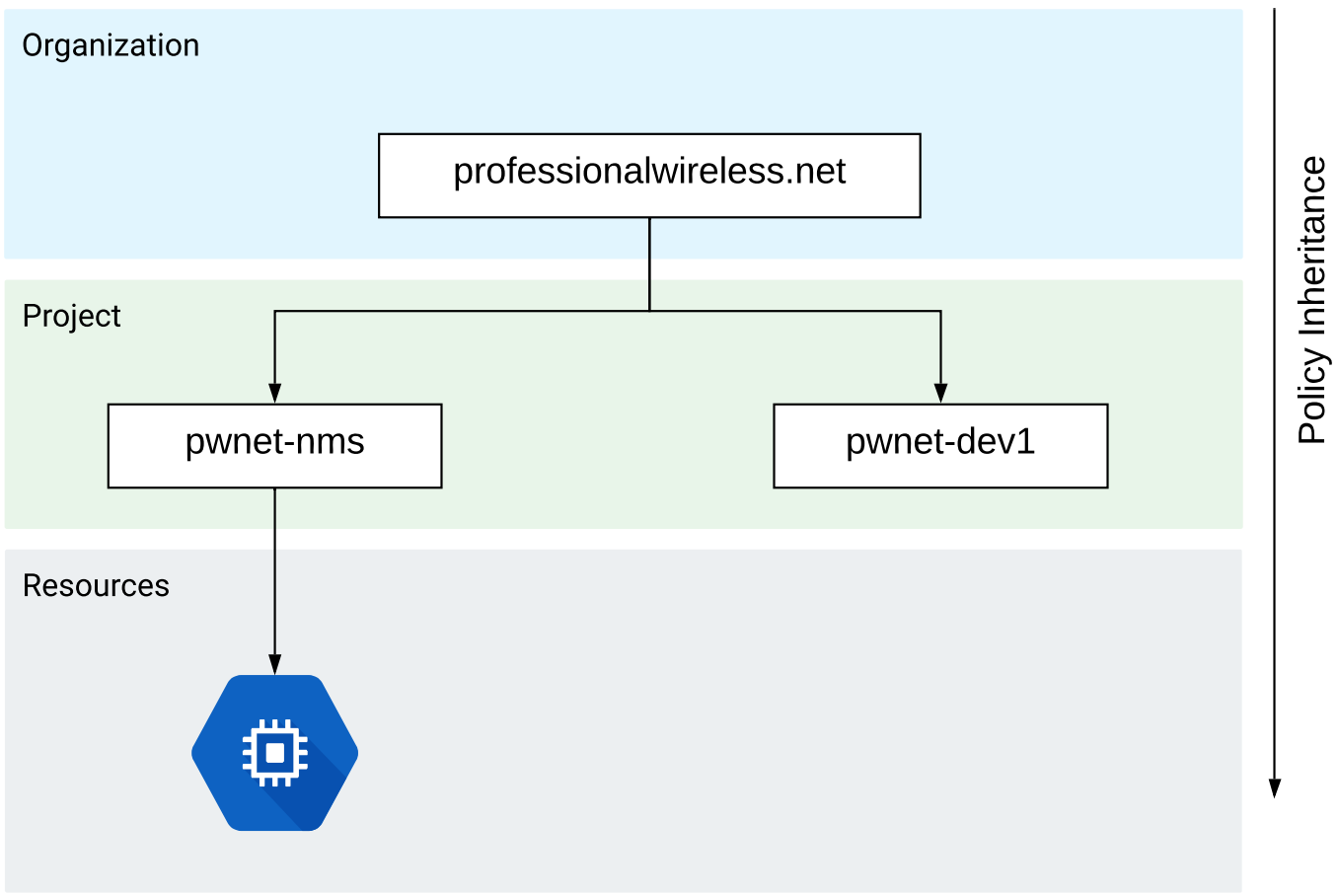
Return to Table of Contents

Choose a Lesson

- IAM Overview
- Hands-On
- Service Accounts
- IAM Best Practices

IAM Hands-On

Add domain as a member
Add single user as member
Creating projects
Attach to billing account
Enable viewing other projects



Basic Concepts
- MEMBERS
- assigned ROLES
- to RESOURCES

IAM policies inherited from top down

Resource Hierarchy
Organization
Folders
Projects
Resources/Services

View and edit IAM Policy

Retrieve IAM policy and download in YAML format

```
gcloud projects get-iam-policy (PROJECT_ID) >
(filename).yaml
```

Update IAM policy from updated file

```
gcloud projects set-iam-policy PROJECT_ID
(filename).yaml
```

Add single binding without downloading file

```
gcloud projects add-iam-policy-binding PROJECT_ID
--member user:(user's email)--role roles/editor
```

Return to Table of Contents

Choose a Lesson

Return to Table of Contents

Choose a Lesson

Billing Overview

Billing Hands-On

Billing Overview

Next

Why is this important?

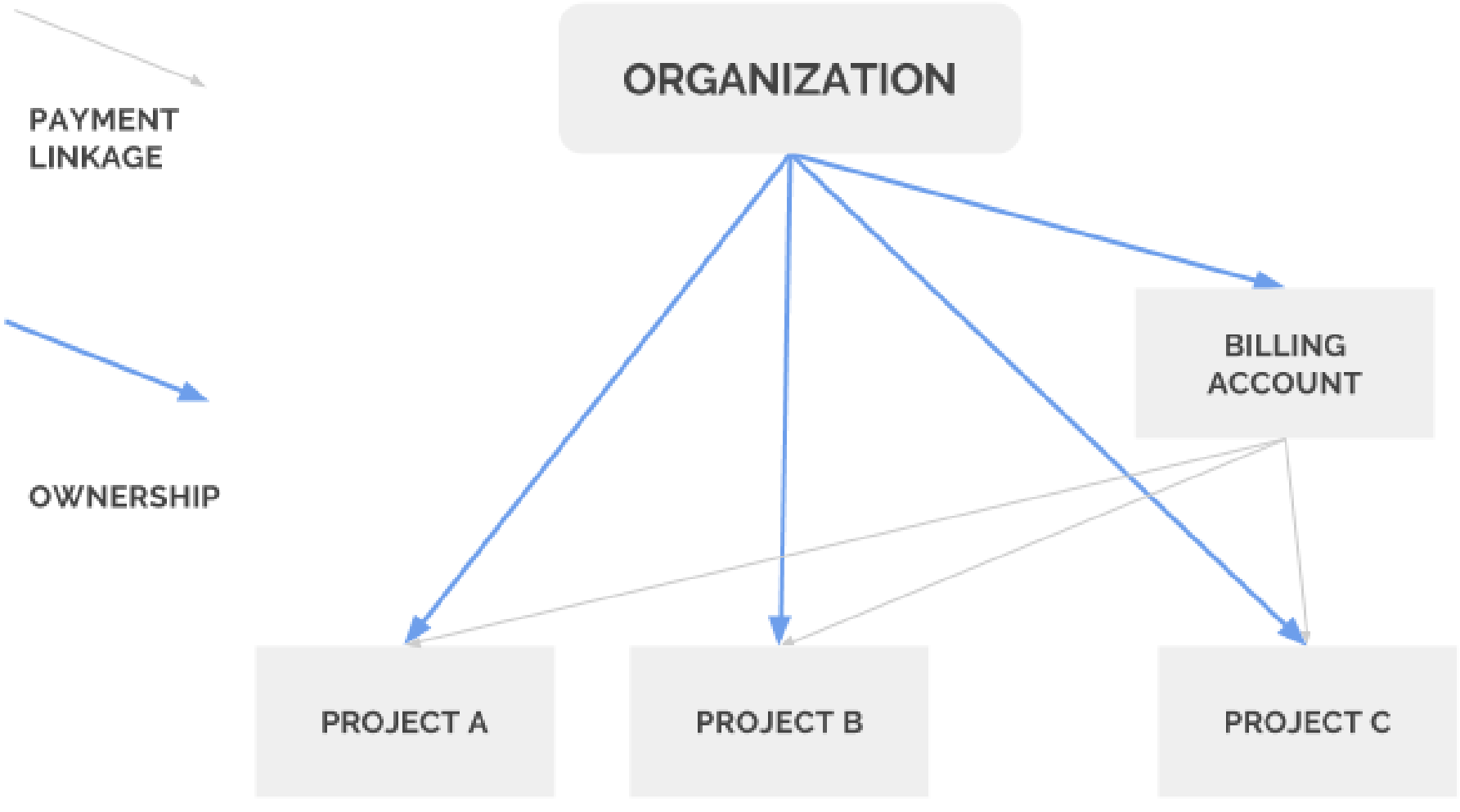
- Setting up and monitoring billing is an essential architect skill
  - Grant IAM roles to users to create, manage, and use billing accounts
  - Manage billing data for retention and analysis
  - Monitor costs, be alerted to runaway costs
- Two primary components we care about:
  - Assign billing IAM roles
  - Viewing and managing billing data

Billing + Cloud IAM

- Assigned mostly at organization level or within billing account
- Billing roles defined in Cloud IAM
  - Billing Account Creator create new billing accounts
    - Organization level only
  - Billing Account Administrator manage (but not create) accounts
    - Configure billing export
    - Link/unlink projects
    - Manage billing user roles
  - Billing Account User link project with billing account
    - Often paired with Project Creator
  - Billing Account Viewer view billing information
    - Usually granted to finance teams
  - Project Billing Manager link/unlink account to projects
    - Similar to Billing Account User, but with no access to project resources
    - Organization or project level

Billing relationships

- Ownership IAM inheritance
- Payment linkage Which billing account pays for which project



Return to Table of Contents

Choose a Lesson

Billing Overview

Billing Hands-On

Billing Overview

Previous

Viewing and managing billing data

- "What am I paying for?"
- View in web console
  - View trends, current resource billing
- Export to Cloud Storage and BigQuery
  - Forward exports to financial departments/auditors
- Set budgets and alerts
  - Does not stop billing/resource usage, just sends alerts at thresholds.

Return to Table of Contents

Choose a Lesson

Billing Overview

Billing Hands-On

Billing Hands-On

Exam perspective

- IAM roles
- Working with exports - when to use exports based on requirements
  - Cloud Storage - long term retention, "audit"
  - BigQuery - analysis

Guideposts for this demo

- IAM roles - Higher level role has lower level included
- Looking at current costs
- View trends
- Create export to Cloud Storage/BigQuery
- Run queries against sample public dataset

Public dataset queries:

Find all charges that were more than 3 dollars:

```
SELECT product, resource_type, start_time, end_time,
cost, project_id, project_name, project_labels_key, currency,
currency_conversion_rate,
usage_amount, usage_unit
FROM `cloud-training-prod-bucket.arch_infra.billing_data`
WHERE (cost > 3)
```

Find which product had the highest total number of records:

```
SELECT product, COUNT(*)
FROM `cloud-training-prod-bucket.arch_infra.billing_data`
GROUP BY product
LIMIT 200
```

Find which product most frequently cost more than a dollar:

```
SELECT product, cost, COUNT(*)
FROM `cloud-training-prod-bucket.arch_infra.billing_data`
WHERE (cost > 1)
GROUP BY cost, product
LIMIT 200
```

Return to Table of Contents

Choose a Lesson

Return to Table of Contents

Choose a Lesson

- Stackdriver Overview
- Stackdriver Logging
- Stackdriver Monitoring
- Trace, Debug, Error Reporting, and Profiler

Stackdriver Overview

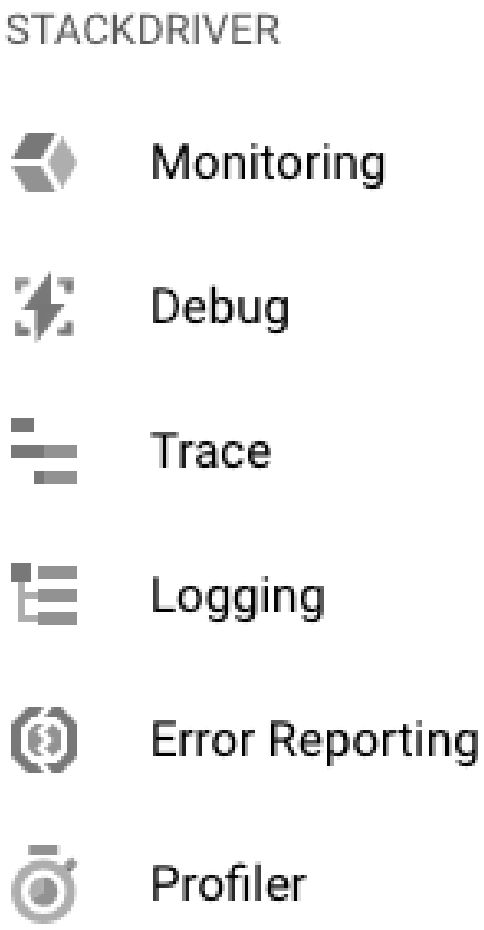
"HEY, what's going on?"

What is Stackdriver?

- Suite of tools for monitoring, logging, and tracking diagnostics for your applications
- Recent acquisition—previously exclusive to AWS
- Native monitoring of both GCP and AWS
  - Connect via service accounts/APIs on other platforms/on-premises
- Dynamically discover all GCP resources
  - Install Stackdriver client on VMs for even greater levels of monitoring

Five Different Products (with a sixth in Beta)

- Logging
  - Central collection for all activity logs
- Monitoring
  - Monitor metrics, health checks, dashboards, and alerts
- Error Reporting
  - Identify and understand application errors
- Trace
  - Find latency bottlenecks in applications
- Debug
  - Find/fix code errors in production
- Profiler (Beta)
  - Collect CPU/memory data—optimize performance

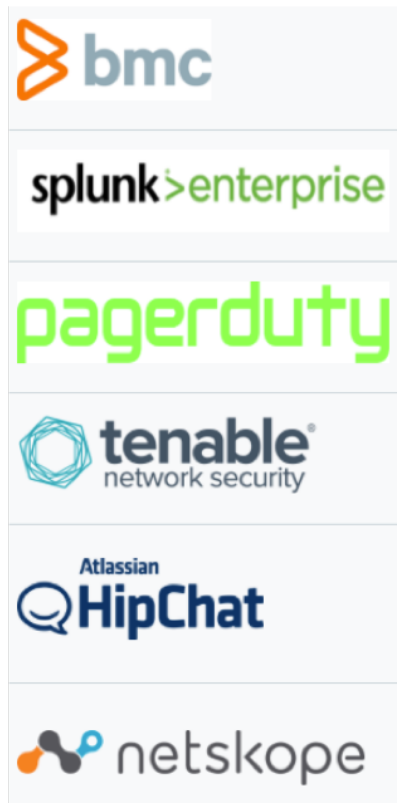


Stackdriver Benefits

- Multi-cloud monitoring—native to GCP and AWS
- Identify trends and prevent problems before they occur
- Centralized logging for all of GCP/AWS
  - One-stop shop
- Better signal-to-noise ratio
  - More relevant alerts
- Find and fix problems faster

Third-Party Integrations

- Site Reliability Engineer (SRE) vendors
- Centralized logging integrates with third-party products
- BMC, Splunk, Hipchat, PagerDuty, Netskope



Return to Table of Contents

Choose a Lesson

- Stackdriver Overview
- Stackdriver Logging
- Stackdriver Monitoring
- Trace, Debug, Error Reporting, and Profiler

Stackdriver Logging Concepts

Next

What is it?

- Single repository for log data and events from multiple sources
- Store, search, and analyze logs
- Collect platform, system, and application logs (with agent)
- Tight integration with Stackdriver Monitoring
- Real time and batch monitoring
- Export logs to other sources for long term storage/analysis

Exam Perspective

- IAM roles
- Exports
- How logging works with other Stackdriver products

Concepts and Terminology

- Associated by project
  - Log viewer only shows logs for one project
- Log entry: Records status or event
  - Includes log name (e.g., ‘syslog’, ‘compute.googleapis.com/activity’)
- Logs: Named collection of log entries
  - Only exist if there are log entries
- Retention period
  - Depends on log type

Log Types

- "Who did what, where, and when?"
- Audit Logs
  - Admin Activity, System Event, Access Transparency, Data Access
- Agent Logs
  - Agent installed on VMs
  - Records VM system and third-party app logs

Admin Activity/System Event Logs

- Always on, immutable, no charge
- Admin Activity: Administrative actions and API calls (e.g., Bob created an instance)
- System Events: GCE system event (e.g., live migration)

Data Access Logs

- Logs API calls that create, modify, or read user-provided data
- Not on by default, can become large
- Charge if beyond free limits

Agent Logs

- Agent installed on support VMs
- Logs data from third-party applications
- Also incurs charge beyond free limits

Other Types

- Access Transparency: Access logs when GCP support staff access data (with support package)

Return to Table of Contents

Choose a Lesson

- Stackdriver Overview
- Stackdriver Logging
- Stackdriver Monitoring
- Trace, Debug, Error Reporting, and Profiler

Stackdriver Logging Concepts

Previous
Next

Pricing

- First 50 GB/project/mo free—$0.50/GB after that
- Admin and system event logs exempt
- Formerly Standard and Premium tiers—gone as of July 2018

Retention

Log Type	Retention Period
Admin Activity	400 days
Data Access	30 days
System Event	400 days
Access Transparency	400 days
All other logs	30 days

Exporting logs

- After retention period, logs are deleted and cannot be recovered!
- Export logs for long term retention
- Long term storage (Cloud Storage), big data analysis (BigQuery), stream to other sources (Pub/Sub)

The basics:

- Requires a project and destination service
- Create **filter**—select log entries to export
- Choose **destination**—Cloud Storage, BigQuery, Pub/Sub
- Filter and destination held in a **sink**—direct what entries to copy to which destination
- Only new entries will be exported after sink creation

IAM Roles

- Logging Admin: Full control plus add others to Logging IAM
- Logs Viewer: View logs
- Logs Writer: Grant service accounts ability to write (create) logs
- Logs Configuration Writer: Create metrics and export sinks

Return to Table of Contents

Choose a Lesson

Stackdriver Overview

Stackdriver Logging

Stackdriver Monitoring

Trace, Debug, Error Reporting, and Profiler

Stackdriver Logging Concepts

Previous

Hands-On Guideposts

- View logs
- Filter (basic/advanced)
- Turn on real-time viewing
- Export logs to Cloud Storage and BigQuery
  - Cloud Storage: 'Audit', long term storage
  - BigQuery: Short term analysis
- Create a custom metric
- Enable data access logs

Return to Table of Contents

Choose a Lesson

- Stackdriver Overview
- Stackdriver Logging
- Stackdriver Monitoring
- Trace, Debug, Error Reporting, and Profiler

Stackdriver Monitoring

Next

What is it?

- Full-stack monitoring, powered by Google
- What is up? What is down? What is overloaded?
- Native monitoring of GCP, AWS, and third-party applications
- Monitor system and application metrics
- Interacts with Stackdriver Logging
- Easy to view insights with dashboards and alerts
- Uptime checks on external applications
  - Must be public and allow firewall traffic from uptime source IP addresses

Exam Perspective

- Troubleshoot reachability with external sources (e.g., uptime checks)
- Familiar with the logging/monitoring agent
- Know how to use Monitoring in conjunction with other Stackdriver products to troubleshoot problems—mostly conceptual
  - Work with alerts

PRICING

Feature	Price	Free allotment per month
Monitoring Data	$0.2580/MB: 150–100,000 MB $0.1510/MB: 100,000–250,000 MB $0.0610/MB: above 250,000 MB	All GCP Metrics First 150 MB per billing account
Monitoring API calls	$0.01/1000 API calls	First million API calls

Stackdriver Agent

- Software installed on VMs
  - Recommended but not required
- Without the agent, can still get CPU, disk/network traffic, and uptime info
- Agent accesses additional resource and application service info
- Monitors many third-party applications

General Best Practices

- Create single project for Stackdriver monitoring
  - ‘Single pane of glass’: Monitors all GCP/AWS resources across projects
  - IAM controls: Separate Stackdriver accounts for data and control isolation
- Determine monitoring needs in advance

Return to Table of Contents

Choose a Lesson

- Stackdriver Overview
- Stackdriver Logging
- Stackdriver Monitoring
- Trace, Debug, Error Reporting, and Profiler

Stackdriver Monitoring

Previous

Hands-On Guideposts

- Create project
- Sign up for account
- Go through initial setup process
- Install agent on instance
- Explore overview
- Create a dashboard
- Explore resources, alerting, uptime
- Misc. items

Exam tip: Firewall needs to allow traffic for communications (uptime checks/alerts), typically via port 80/443

Return to Table of Contents

Choose a Lesson

Stackdriver Overview

Stackdriver Logging

Stackdriver Monitoring

Trace, Debug, Error Reporting, and Profiler

Trace, Debug, Error Reporting, and Profiler

Next

Exam Perspective

- Conceptual knowledge on how each product is used to solve problems
- Interaction with Stackdriver Logging

Trace

- Find performance bottlenecks—latency (loading times)
- Collect data from Google App Engine (GAE), Google HTTP load balancers, or apps with Stackdriver Trace SDK
- Integrated into App Engine Standard—automatically enabled
- Available for GCE, GKE, and GAE (Flexible)
  - Requires enabling Stackdriver Trace API or SDK (depending on library)
- Can be installed on non-GCP resources

Trace answers questions:

- How long does it take my application to handle a given request?
  - Why is it taking my application so long to handle a request?
  - Why do some of my requests take longer than others?
- What is the overall latency of requests to my application?
- Which microservice is causing a bottleneck?
- Has latency for my application increased or decreased over time?
- What can I do to reduce application latency?

Error Reporting

- Real-time error monitoring and alerting in your application
- Quickly understand errors
- Write to Stackdriver Logging or Error Reporting API (beta)
- Automatic and real-time analysis
  - Alerts and dashboards
- Built in to App Engine Standard and Cloud Functions—automatically enabled
- In Beta for GAE Flexible, GCE, GKE, EC2 (AWS)
- GCE, GKE, and EC2 require Stackdriver logging agent to be installed
- Java, Python, JavaScript, Ruby, C#, PHP, and Go

Debug

- Debug application
- Inspect application state without stopping or slowing app
- Does not require adding log statements
- Automatically enabled in GAE Standard
- Available in GAE Flexible, GCE, and GKE with additional configuration
- Java, Python, Go, Node.js
- Can be installed on non-GCP resources

Profiler (beta)

- Profile resource intensive application components
- Collect CPU/RAM usage data --> associate with source code --> identify high resource usage components

Return to Table of Contents

Choose a Lesson

Stackdriver Overview

Stackdriver Logging

Stackdriver Monitoring

Trace, Debug, Error Reporting, and Profiler

Lesson Title

Previous

Demo Guideposts

- View Trace data on App Engine app
- Purposely break app and re-deploy
- Use Error Reporting and Debug to fix what we broke

Master Builder's Guide to Google Cloud

Return to Table of Contents

Choose a Lesson

Return to Table of Contents

Choose a Lesson

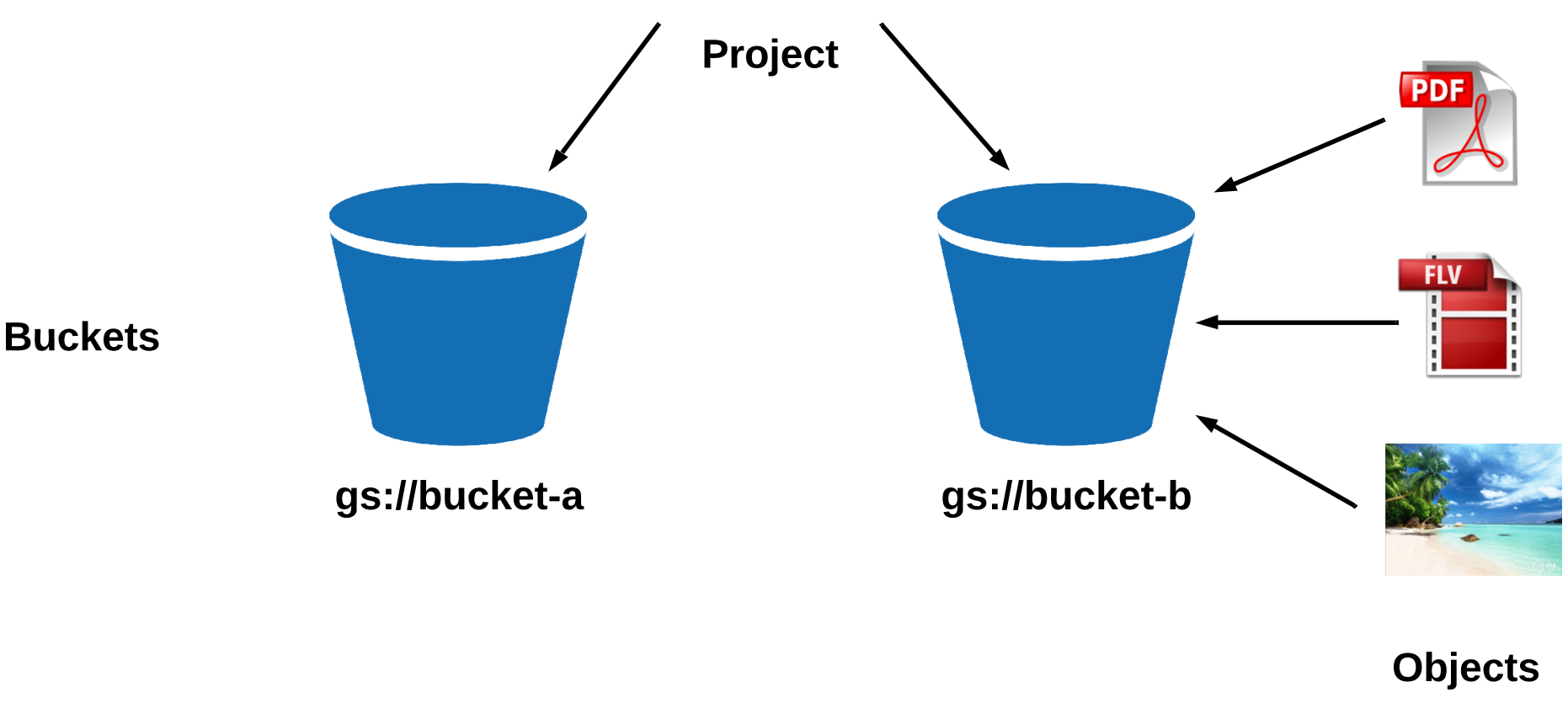
- Cloud Storage Concepts
- gsutil Command Basics
- Cloud Storage Security Concepts
- Object Versioning and Lifecycle Management
- Additional gsutil Hands-On

Cloud Storage Concepts

Next

What is it?

- Storage for unstructured data
  - What can a bucket hold? Everything!
  - Pictures, videos, files, scripts, etc.
  - Even databases for transfer/backup (or BigQuery read)
- Virtually infinite size
- Pay-per-use, not allocation (elastic)
- Primary unit is a bucket
  - Access managed via IAM
  - Can be arranged in file/folder format for organizational purposes
- Objects go inside of the bucket
  - Objects = files
  - Inherit bucket permissions and storage class
  - **Note**: Folders are also considered objects
- GCS is not a file system/block storage
  - Technically an object storage system
  - Different write/performance characteristics
  - Block storage = persistent disk (useful for SANs)



Return to Table of Contents

Choose a Lesson

Cloud Storage Concepts

gsutil Command Basics

Cloud Storage Security Concepts

Object Versioning and Lifecycle Management

Additional gsutil Hands-On

Cloud Storage Concepts

Previous

Next

Storage Classes

- Different pricing tiers for storage buckets
- Same performance for all classes
- Differences in geo-location, SLA, and operation restrictions
- Classes in brief:
  - Standard (Regional and Multi-regional)
  - Nearline/Coldline (Regional/Multi-regional)
- Class applied to bucket and/or objects

Multi-Regional (standard)	Span multiple regions Hot data from different geographical locations No retrieval costs 99.95% availability SLA	$0.026/GB/mo
Regional (standard)	Single region Hot data in single geographical location Highest regional performance No retrieval costs 99.9% availability SLA	$0.02/GB/mo
Nearline	Regional or multi-regional Archive data accessed once per month Lower monthly cost with some retrieval costs Minimum 30 day duration Ideal for backups 99.9% SLA multi-regional, 99% regional	$0.01/GB/mo $0.01 per GB retrieval
Coldline	Regional or multi-regional Archive data accessed once per year Even lower monthly costs with higher retrieval costs Minimum 90 day duration Ideal for 'cold' storage - legal compliance, disaster recovery 99.9% SLA multi-regional, 99% regional	$0.007/GB/mo $0.05 per GB retrieval

In Beta: Dual-region bucket

- Under multi-regional banner
- Data replicated in two regions with regional bucket performance
- More expensive than multi-regional

Return to Table of Contents

Choose a Lesson

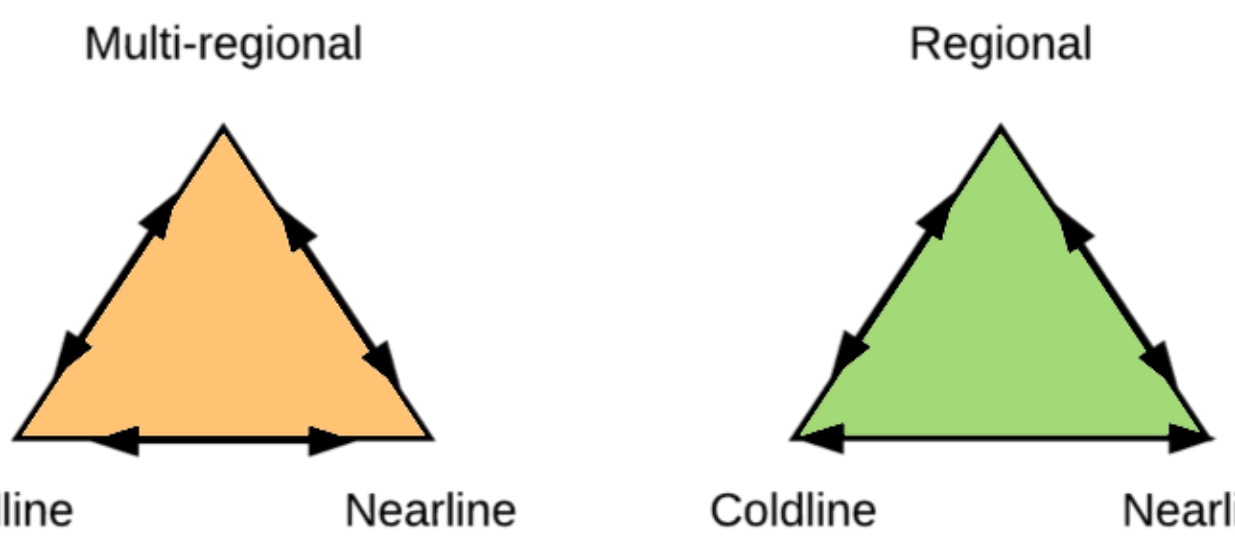
- Cloud Storage Concepts
- gsutil Command Basics
- Cloud Storage Security Concepts
- Object Versioning and Lifecycle Management
- Additional gsutil Hands-On

Cloud Storage Concepts

Previous

Changing Storage Classes

- Nearline/Coldline can be regional or multi-regional
- Cannot change from multi-regional to regional (and vice versa)
- Changing class only affects new objects
  - Previous objects are previous class (change with `gsutil`)
- Objects can be moved to another bucket
  - Same class: Can use web console
  - Different class: Must use `gsutil`



Return to Table of Contents

Choose a Lesson

- Cloud Storage Concepts
- gsutil Command Basics
- Cloud Storage Security Concepts
- Object Versioning and Lifecycle Management
- Additional gsutil Hands-On

gsutil Command Basics

What is gsutil?

- Command line format for working with Cloud Storage buckets/objects
- Different format compared to gcloud commands
- Useful for:
  - Creating and deleting buckets
  - Uploading, downloading, and deleting objects
  - Listing buckets and objects
  - Moving, copying, and renaming objects
  - Editing object and bucket ACLs

Accessing buckets

Bucket syntax = `gs://<BUCKET_NAME>/<OBJECT_NAME>`

gsutil Command Structure

- `gsutil <command> <options> <target>`
- `gsutil <command> <options> <source> <target>`
- Very similar to Linux commands
  - `cp`, `mv`, `rm`, `ls`; same for both Linux and `gsutil`
  - Almost identical structure
- Example: `gsutil cp file1.txt gs://mybucket`
  - `gsutil ls -a gs://mybucket`
- Top-level command options (before the command)
  - `gsutil **-m** cp -r gs://<BUCKET> <target>`

gsutil References

- GCP documentation: https://cloud.google.com/storage/docs/gsutil
- Command line help
  - `gsutil help <command> - gsutil help cp`
  - `gsutil help options` - help with top level command line options

Return to Table of Contents

Choose a Lesson

- Cloud Storage Concepts
- gsutil Command Basics
- Cloud Storage Security Concepts
- Object Versioning and Lifecycle Management
- Additional gsutil Hands-On

Cloud Storage Security Concepts

Next

Security is REALLY Important!

- When taken lightly, you can end up famous...in a **bad** way

Corporate tech giant leaves secret data exposed to public internet

There's a Hole in 1,951 Amazon S3 Buckets

Poorly configured cloud buckets strike again ·

Customer Data Left Publicly Exposed on Cloud Storage Server

Security is in YOUR control!

Cloud Storage access is as 'locked down' or 'wide open' as you choose

Return to Table of Contents

Choose a Lesson

- Cloud Storage Concepts
- gsutil Command Basics
- Cloud Storage Security Concepts
- Object Versioning and Lifecycle Management
- Additional gsutil Hands-On

Cloud Storage Security Concepts

Previous

Next

Access Management Principles

- Two methods: IAM and ACL
- ACL: Access Control Lists
- Some overlap between the two

Security via IAM

- Same IAM principles as the rest of GCP
- Hierarchical in nature
- Granted at project, resource (all of Cloud Storage), or individual bucket (but not objects)
- Possible to grant access to manage bucket but not view/read objects inside

IAM Roles

- Primitive project level roles (Owner/Editor/Viewer)
- Standard Storage Roles (work independently from ACLs)
  - Storage Admin
  - Storage Object Admin
  - Storage Object Creator
  - Storage Object Viewer
- Legacy roles – equivalent to ACL permissions - bucket level only
  - Storage Legacy Bucket Owner
  - Storage Legacy Bucket Reader
  - Storage Legacy Bucket Writer
  - Storage Legacy Object Owner
  - Storage Legacy Object Reader

or

or

Access Control Lists (ACLs)

- Define who has access to buckets/objects and what level of access
- Can be applied to bucket or individual objects
- Objects inherit ACL from default bucket ACL
- Can also be independent
- For non-legacy IAM roles, no overlap
- Sounds like a lot of overlap between IAM and ACL… because there is

Should you use ACLs?

- Best practice: Use IAM over ACL whenever possible
- IAM gives enterprise-grade control across all of GCP
- IAM leaves an audit trail for access
- When in doubt, use IAM
- However, use ACL to grant access to an object without granting access to bucket
- More fine-grained control

Signed URLs: Timed Access to Object Data

- Set a timer on access to a bucket or object
- Useful for temporarily giving access without need of signing in with Google account
- Gives user read, write, or delete access for limited time
- Anyone with the URL can access within the time period
  - URL is the 'key' to access the data

Return to Table of Contents

Choose a Lesson

- Cloud Storage Concepts
- gsutil Command Basics
- Cloud Storage Security Concepts
- Object Versioning and Lifecycle Management
- Additional gsutil Hands-On

Cloud Storage Security Concepts

Previous

Next

Customer-supplied encryption keys

- By default, GCP encrypts all GCS data with Google-supplied keys
- Customer can choose to manage their own keys
- Customer keys provided via customer-side .boto configuration file.

Best Practices

- NEVER share credentials!
- Remove application access when no longer needed
- Don’t make bucket names a target – e.g., ‘gs://confidential-customer-info’
- Use groups over individual users for IAM when possible
- Check default object ACL before adding objects
- Make sure publicly readable data is intended!
  - Once ‘out’, you can’t bring it back in
- Though possible, publicly writable buckets are a bad idea
- Use signed URLs for secure access without need of Google account

Return to Table of Contents

Choose a Lesson

- Cloud Storage Concepts
- gsutil Command Basics
- Cloud Storage Security Concepts
- Object Versioning and Lifecycle Management
- Additional gsutil Hands-On

Cloud Storage Security Concepts

Previous

Hands-On Guideposts

- Assign IAM roles to buckets
  - `gsutil iam ch user:<user_email>:<role1,role2> gs://<BUCKET>`
- Remove IAM role from bucket
  - `gsutil iam ch -d user:<user_email>:<role1,role2> gs://<BUCKET>`
- Remove all roles from bucket for given user
  - `gsutil iam ch -d user:<user_email> gs://<BUCKET>`
- Assign ACL roles to buckets and objects
  - `gsutil acl ch -u <user_email>:<O/R/W> gs://<BUCKET>`
- Remove ACL role
  - Delete all ACL's
    - `gsutil acl ch -d <user_email> gs://<BUCKET>`
- URL for direct access
- Mixed owner/read permissions
- Signed URL's
  - Create service account with key
  - Upload to cloud shell (or add to current CLI environment)
  - `gsutil signurl -d <time_period (10m)> <keyfile.json> gs://<BUCKET>/<object>`

Return to Table of Contents

Choose a Lesson

- Cloud Storage Concepts
- gsutil Command Basics
- Cloud Storage Security Concepts
- Object Versioning and Lifecycle Management
- Additional gsutil Hands-On

Object Versioning & Lifecycle Management

Next

Object Versioning

- Retrieve objects that are deleted or overwritten
- Applied to bucket level
- Disabled by default
- When enabled, deleted and overwritten objects are archived instead of deleted
- Object keeps same name but paired with unique identifier number
- Actions possible with versioning enabled:
  - List archived versions
  - Restore archived version to live state
  - Permanently delete older version
- If versioning is disabled, existing versions remain but new ones not created

Versioning Considerations

- No default limit on versions – can increase bucket size (and cost) greatly
  - Lifecycle Management helps with this
- Archive versions retain own ACL, which may not be same as live version

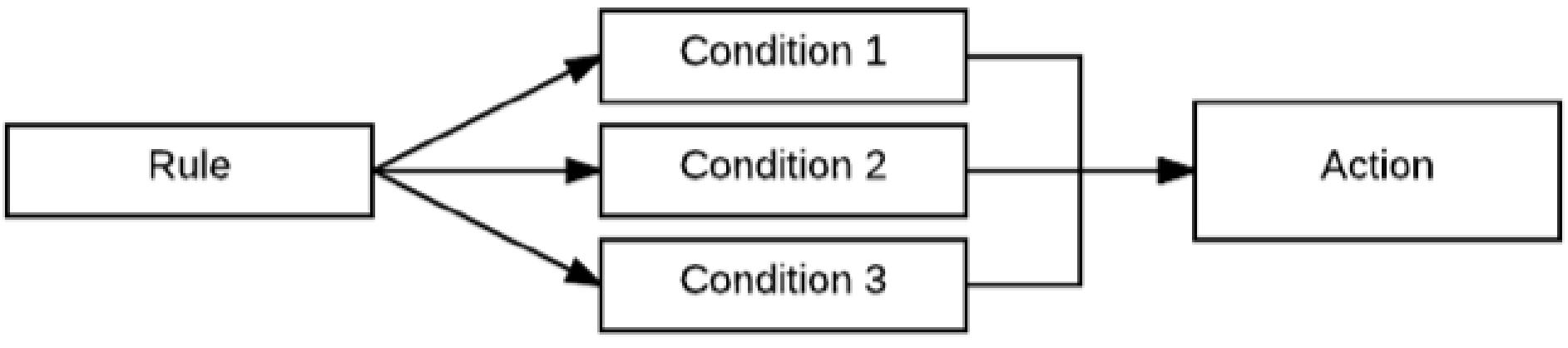
Versioning Terminology

- Generation: Update when object content overwritten (delete, overwrite file)
- Metageneration: Metadata generation/change
- No relationship between the two

Object Lifecycle Management

- Sets Time to Live (TTL) on an object
  - Archive/Delete older versions
  - Downgrade storage classes
- Applied to bucket level
- Often paired to object versioning, but not required
- Examples:
  - Downgrade the storage class of objects older than 365 days to Coldline Storage
  - Delete objects created before January 1, 2017
  - Keep only the 3 most recent versions of each object in a bucket with versioning enabled
- Implemented with combination of Rules, Conditions, and Actions

Click each item to learn more



Return to Table of Contents

Choose a Lesson

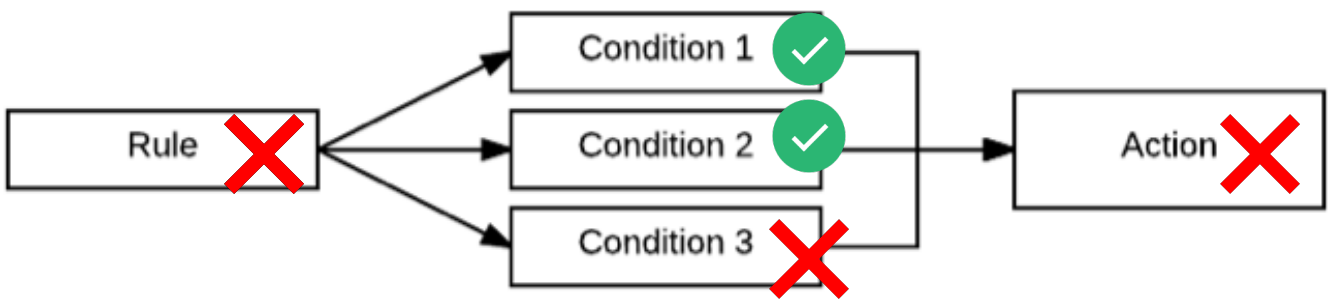
- Cloud Storage Concepts
- gsutil Command Basics
- Cloud Storage Security Concepts
- Object Versioning and Lifecycle Management
- Additional gsutil Hands-On

Object Versioning & Lifecycle Management

Previous

Next

Rules must have all conditions met before executing



Actions only need one Rule to be met before being executed

Return to Table of Contents

Choose a Lesson

Cloud Storage Concepts

gsutil Command Basics

Cloud Storage Security Concepts

Object Versioning and Lifecycle Management

Additional gsutil Hands-On

Object Versioning & Lifecycle Management

Previous

Hands-On Guideposts

- Versioning is command line/REST API only
- Check current policy
  - `gsutil versioning get gs://<BUCKET>`
- Enable Object Versioning - command line only
  - `gsutil versioning set on gs://<BUCKET>`
- 'Delete' an object (in reality, archive an object)
- Check full object details in bucket (also command line only)
  - `gsutil ls -a gs://<BUCKET>`
- Copy a few versions of files into bucket
  - `gsutil cp <file> gs://<BUCKET>`
- Create Lifecycle Management policy via web console
  - Command line version:
    - `gsutil lifecycle get gs://<BUCKET> > filename.json`
    - `gsutil lifecycle set filename.json gs://<BUCKET>`

Return to Table of Contents

Choose a Lesson

Cloud Storage Concepts

gsutil Command Basics

Cloud Storage Security Concepts

Object Versioning and Lifecycle Management

Additional gsutil Hands-On

X

Rules

- Specify set of conditions in order to take action
- If multiple conditions in a rule, all conditions must be met before action taken
- However, if multiple rules with same action, any met rule executes action

Return to Table of Contents

Choose a Lesson

Cloud Storage Concepts

gsutil Command Basics

Cloud Storage Security Concepts

Object Versioning and Lifecycle Management

Additional gsutil Hands-On

X

Condition

- Criteria to meet before taking action
- Age
  - Age in days (TTL)
  - Still valid if archived via versioning
- CreatedBefore
  - Object created before midnight of specified data (UTC)
- IsLive
  - Live or archived version of object
- MatchesStorageClass
  - Condition matches specified storage class
- NumberOfNewerVersions
  - Condition met when number of newer versions than target is reached

Return to Table of Contents

Choose a Lesson

Cloud Storage Concepts

gsutil Command Basics

Cloud Storage Security Concepts

Object Versioning and Lifecycle Management

Additional gsutil Hands-On

X

Actions

- Delete
  - Versioned or non-versioned
  - Delete live (versioned) objects creates archived object
  - Deleting archived object deletes permanently
  - Deleted objects cannot be recovered!
  - Test rules on development data
- SetStorageClass
  - Change storage class of affected object
  - Remember, cannot go from Regional to Multi-Regional, and vice versa
  - Regional Standard ▮ Regional Nearline/Coldline
  - Multi-Regional Standard ▮ Multi-Regional Nearline/Coldline

Return to Table of Contents

Choose a Lesson

Cloud Storage Concepts

gsutil Command Basics

Cloud Storage Security Concepts

Object Versioning and Lifecycle Management

Additional gsutil Hands-On

Additional gsutil Hands-On

Hands-On Guideposts

- Remove a bucket
  - `gsutil rm -r gs://<BUCKET>`
- Create a new bucket
  - `gsutil mb -l <location> -c <class> gs://<BUCKET>`
- Copy local files to bucket - use `-m` for parallel threading
  - `gsutil -m cp -r <files/directory> gs://<BUCKET>`
- Turn on versioning
  - `gsutil versioning get gs://<BUCKET>`
  - `gsutil versioning set on gs://<BUCKET>`
- View bucket folder contents
  - `gsutil ls gs://<BUCKET>/<folder>`
- View all subfolder contents
  - `gsutil ls -r gs://<BUCKET>`
- Change storage class in existing bucket (disable versioning first)
  - `gsutil versioning set off gs://<BUCKET>`
  - `gsutil -m rewrite -r -s NEARLINE gs://<BUCKET>/*`
    - (add -m for parallel threading)
- Give public read access to a few objects via ACL
  - View file structure
    - `gsutil ls gs://bucket`
  - Give public access to a couple files
    - `gsutil acl ch -u AllUsers:R gs://<BUCKET/<object>`
- Revoke public access
  - `gsutil acl ch -d AllUsers gs://<BUCKET/<object>`
- Delete bucket
  - `gsutil rm -r gs://<BUCKET>`

Return to Table of Contents

Choose a Lesson

Return to Table of Contents

Choose a Lesson

- Managed Databases Overview
- Managed Databases on Google Cloud
- Cloud SQL Closer Look
- BigQuery Closer Look

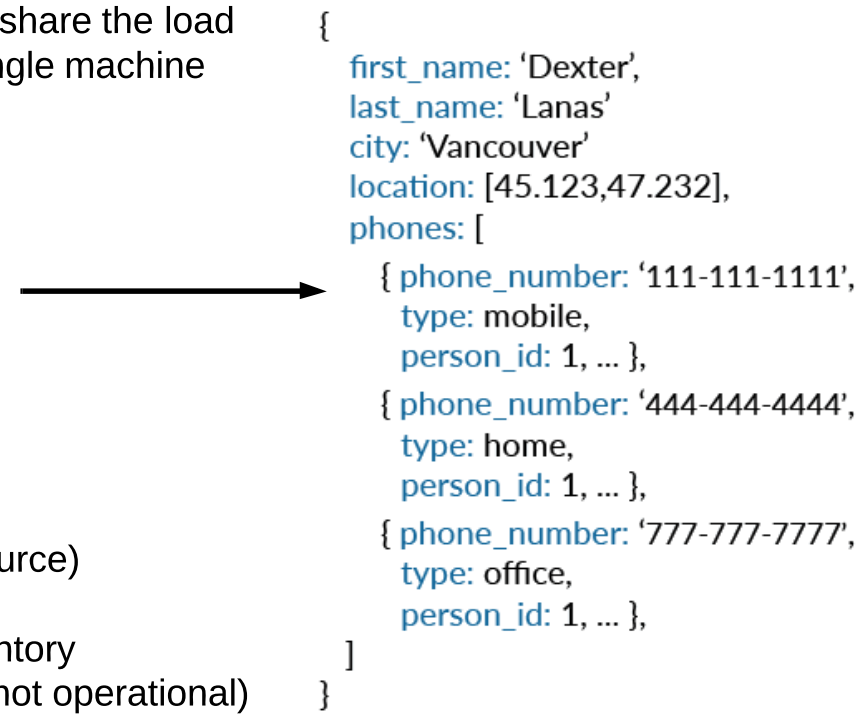
Managed Databases Overview

Managed DatabasesHigh Level Overview

- Multiple managed databases for ingesting, storing, and using structured data
- Each managed service has a specific purpose and use case
- Architect exam will test ability to select correct solution for a use case, with some deeper level knowledge required for a few services
- Additionally, different managed databases manage different 'stuff', which affects scaling, depending on the service
  - Range between completely 'no-ops' to managed with requirement to provision resources

Questions to ask ourselves

- Main factors:
  - Scalability/availability/performance
  - Relational (RDBMS/SQL) vs. non-relational (NoSQL)
  - Transactional vs. Analytical
- How does this solution scale (grow), and what are its limits?
  - Availability (global vs. regional), capacity, performance
  - Horizontal vs. vertical scaling (in compute)?
    - Horizontal: Scaling out; adding more machines to share the load
    - Vertical: Scaling up; adding more compute to a single machine
- What type of data?
  - Relational, non-relational
  - Relational: Tables and rows, spreadsheets
    - Strong consistency
  - Non-relational: Non-fixed relationships, JSON format
    - Eventual consistency
- What is our interaction method?
  - Transactional/Analytical
  - OLTP: Online Transaction Processing Transactional
  - OLAP: Online Analytical Processing Analytical
  - Transactional: Operational data (data is the original source)
    - Relatively simple queries
    - Example: Medical records database, product inventory
  - Analytical: Consolidation data (not the original source/not operational)
    - Complex queries
    - Example: Business insights, market trends, data warehousing



Exam Perspective

All about choosing the correct service to support requirements

- Storing and structuring data
- Retrieving and using data

Changes to the refreshed Architect exam:

- More in-depth than previous version
- BigQuery especially goes into much more details
- Also more on Datastore and Cloud SQL, specifically how to optimize performance and scale it
- Might be worth sampling Data Engineer course

Return to Table of Contents

Choose a Lesson

- Managed Databases Overview
- Managed Databases on Google Cloud
- Cloud SQL Closer Look
- BigQuery Closer Look

Managed Databases on Google Cloud

GCP resource for storage comparisons:
https://cloud.google.com/storage-options/

Next

Decision Tree - Storage Options

	Relational		Non-relational		Data Warehouse
Use Case	Cloud SQL	Cloud Spanner	Cloud Datastore (Firestore)	Cloud Bigtable	BigQuery
	Structured data Web framework	RDBMS+scale High transactions	Semi-structured Key-value data	High throughput analytics	Mission critical apps Scale+consistency
Example	Medical records Blogs	Global supply chain Retail	Product catalog Game state	Graphs IoT Finance	Large data analytics Processing using SQL

Return to Table of Contents

Choose a Lesson

- Managed Databases Overview
- Managed Databases on Google Cloud
- Cloud SQL Closer Look
- BigQuery Closer Look

Managed Databases on Google Cloud

Previous

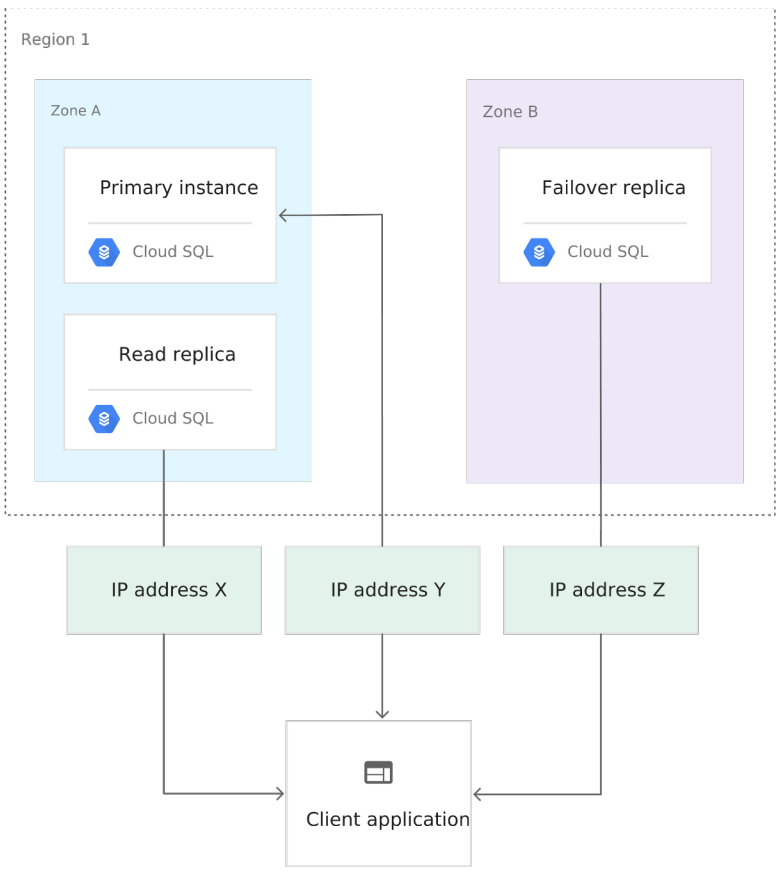
Next

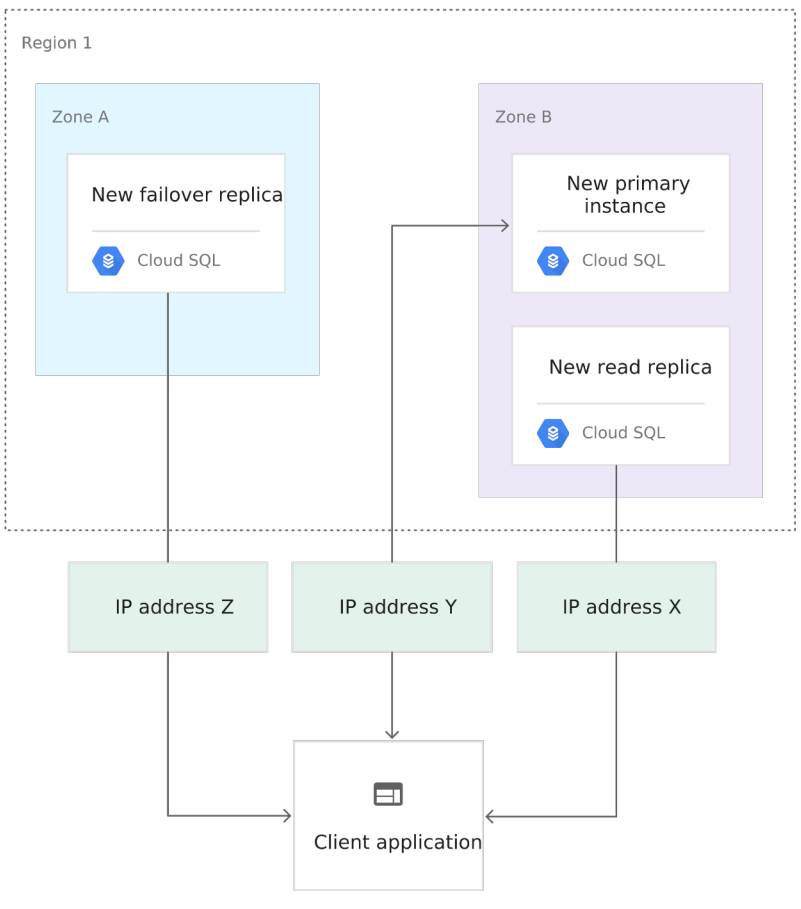
Mapping Storage Types – at a Glance

- Unstructured Data: Cloud Storage
- Relational Data (SQL): Cloud SQL and Spanner
- Non-relational (NoSQL) Data: BigTable and Datastore (Firestore)
- Big data analysis (SQL queries): Google BigQuery
- Newer options: Specific niche purposes
  - In-memory (REDIS) database: Cloud Memorystore
  - Managed file server: Cloud Filestore
- Unmanaged database: Compute Engine/Persistent Disk
  - Microsoft SQL Server, MongoDB, Cassandra, HBase, etc

Cloud SQL

- Hosted MySQL/PostgreSQL database
- Direct 'lift and shift' of MySQL, PostgreSQL databases
  - Minimum modification required
  - Exact same use cases and interaction as on-premises MySQL/PostgreSQL
- Traditional relational database (SQL)
- Often the 'first step' for on-premises to cloud migration
- Managed, but still need to provision individual machines
  - Managed single Compute Engine VM
  - Many similarities to configuring a Compute Engine instance
- What "stuff" is managed?
  - OS/database installation and updates
  - Storage (auto-scaling)
  - Backups
  - Failover, read replicas
- Limitations
  - Hard cap of 10 TB for single Cloud SQL instance
  - Limited scalability (more vertical than horizontal)
    - i.e., Add more compute to single machine vs. using multiple machines
  - Single region only (only single-region database in this list)





Return to Table of Contents

Choose a Lesson

- Managed Databases Overview
- Managed Databases on Google Cloud
- Cloud SQL Closer Look
- BigQuery Closer Look

Managed Databases on Google Cloud

Previous

Next

Cloud Spanner

- No compromises, horizontally scalable relational (RDBMS) database
- Typically, traditional RDBMS databases do not scale well - i.e., trade offs

Transactional Consistency vs. Scalability - Why not both?

- Cross-region availability
  - Global footprint, yet still strongly consistent
- Advantages of relational database, without the drawbacks
  - Horizontally scalable = near infinite capacity
- Not a direct lift and shift of MySQL, modification required
- More expensive than Cloud SQL (starting at $0.90/hr/node)

	Cloud Spanner	Traditional Relational	Traditional Non-relational
Schema	Yes	Yes	No
SQL	Yes	Yes	No
Consistency	Strong	Strong	Eventual
Availability	High	Failover	High
Scalability	Horizontal	Vertical	Horizontal
Replication	Automatic	Configurable	Configurable

Datastore

- Original App Engine database
- Non-relational (NoSQL) database
  - Semi-structured, ACID transactions
- Completely no-ops, no individual machines to interact with
- Structure in key/value pairs - flexible schema definition
- Scales from zero to terabytes
  - Grows with your application
  - Cost effective
- Ideal for web and mobile applications
- Will eventually be replaced by Cloud Firestore

Datastore Query Performance

- Improve query performance with indexes
- Like most databases, Datastore uses an index to speed up searching and queries
- By default, built-in index serves queries - fine for simple query types
- Composite index - necessary for complex queries
  - Defined in custom index configuration file (index.yaml)
  - Apply via gcloud command
  - `gcloud datastore create-indexes`

Return to Table of Contents

Choose a Lesson

Managed Databases Overview

Managed Databases on Google Cloud

Cloud SQL Closer Look

BigQuery Closer Look

Managed Databases on Google Cloud

Previous

Bigtable

- Non-relational (NoSQL) database
- Compared to Datastore:
  - More ideal for analytics
  - More expensive
  - Requires managing nodes
- HBase compatible
- The 'heavy hitter' option - performance over cost efficiency - pay whether using nodes or not
  - Store terabytes to petabytes
  - Very high volume of writes
  - Millisecond response time

BigQuery

- Fully managed, no-ops (serverless), data warehousing
- High capacity data warehouse/analytics
- Big data exploration and processing
- Not ideal for operational database
- Interact using familiar SQL queries
- Mountkirk Games and TerramEarth case studies

Newer options - very focused use cases

Memorystore

- Managed in-memory database - compatible with Redis
- Data stored in memory instead of standard storage (i.e. hard drive)
  - Results in very fast performance
  - Sub-millisecond data access
- Fully compatible with Redis database – easy lift and shift
- Regional presence – similar to Cloud SQL
- Up to 300 GB instance capacity
- Failover and high availability capabilities across two zones in single region
- Exam perspective - managed Redis database

Filestore (currently in beta)

- High performance file storage - NAS
- Managed file storage/sharing instance
- OS, updates, patches handled for you

Non-managed options - Microsoft SQL Server

- MS SQL server does not have a managed service available
- Can manually create high availability/failover configuration - use Compute Engine
  - Create High Availability - Always On Availability Groups
  - Must use Compute Engine instances - manual group configuration
  - https://cloud.google.com/compute/docs/instances/sql-server/configure-availability#configure_failover_cluster
  - Manually create high availability and failover cluster servers in different subnets, but same zone

Other non-managed databases can be hosted on Compute Engine

- Cassandra, MongoDB, HBase, Redis, etc.

Return to Table of Contents

Choose a Lesson

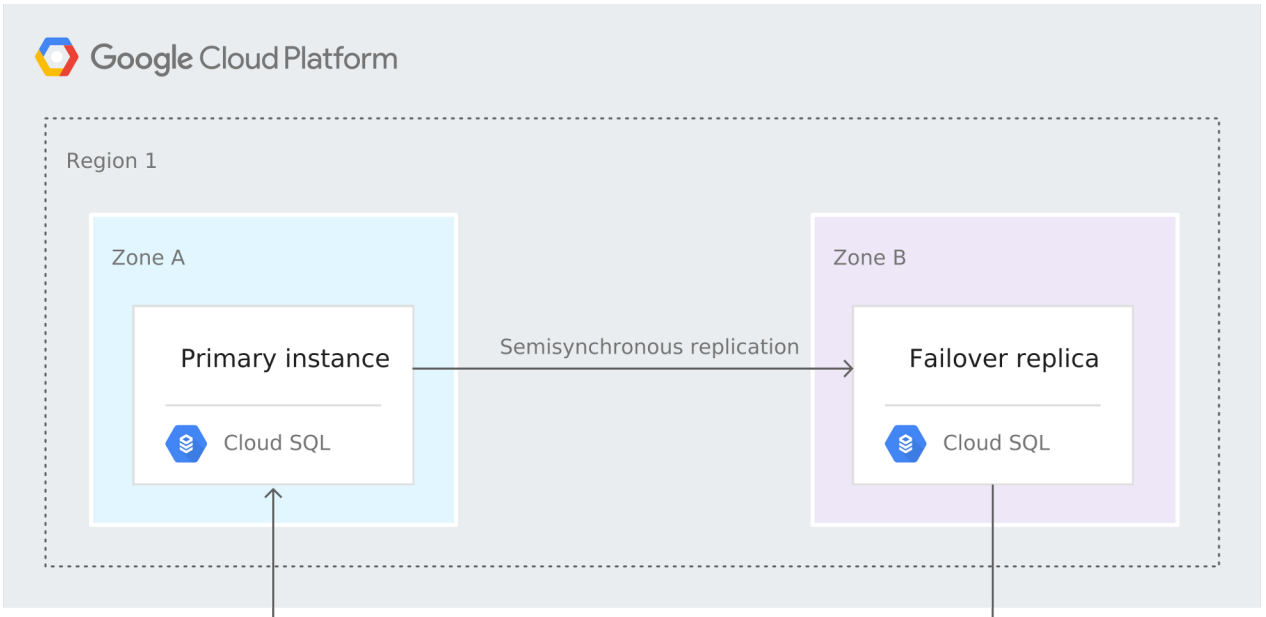
- Managed Databases Overview
- Managed Databases on Google Cloud
- Cloud SQL Closer Look
- BigQuery Closer Look

Cloud SQL - Closer Look

Next

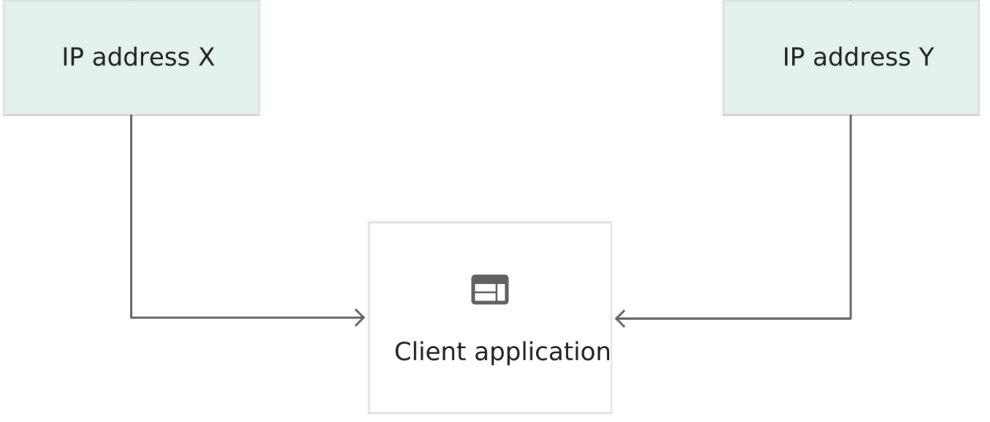
Improve availability and redundancy

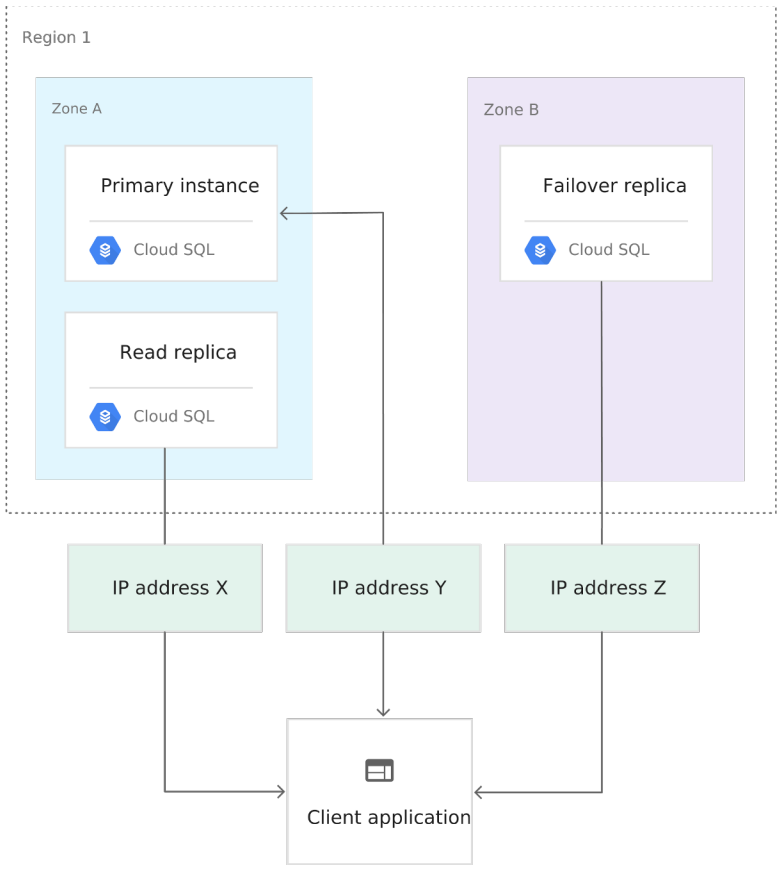
- By default, Cloud SQL = single instance
  - But with easy to configure read replica and failover instance
  - Note: Read replica and failover instance are separate entities
- Failover created in separate zone in same region
  - Data is automatically replicated to failover instance
- In case of primary instance failure, Cloud SQL will **automatically** fail over to the failover instance, which will become the new primary instance
  - New primary instance keeps same connection info (IP address)

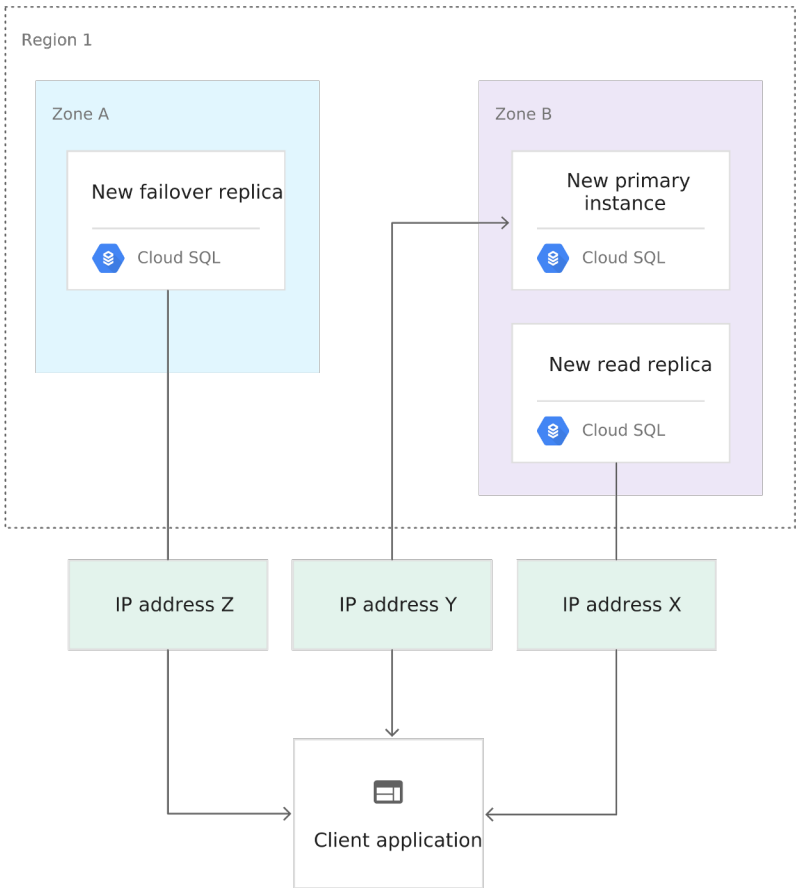


Failover Process

- Primary instance stops responding
- Cloud SQL automatically fails over to failover instance
- Failover instance becomes new primary instance with same connection info







Return to Table of Contents

Choose a Lesson

- Managed Databases Overview
- Managed Databases on Google Cloud
- Cloud SQL Closer Look
- BigQuery Closer Look

Cloud SQL - Closer Look

Previous

Failover Replication Lag

- Replication lag: Difference in time between primary instance makes an update and failover instance catches up
- Does not result in data loss
- If a result of infrequent spike in usage, delete, and recreate failover replica. Lag from unusually large updates usually resolve on their own
- Address by either increasing performance of failover instance (more RAM or disk size) or shard database so write operations are shared by multiple Cloud SQL instances

Scaling your Cloud SQL instance

- Other than read replica, Cloud SQL scales 'vertically'
  - Increase storage/compute for single instance
- Storage scales automatically, Compute does not
- Storage
  - Storage increases automatically, just enable automatic storage increase
  - No downtime required
  - Disk size directly tied to disk performance (similar for persistent disks in GCE)
- Compute (CPU/RAM)
  - Can increase compute, but requires restart
- How to address scaling compute needs?
  - Option A: Create alerts for high CPU utilization, then use maintenance window to manually increase
  - Option B: Plan compute needs in advance and provision as such (over-provision)
- Option A requires downtime to make adjustments, option B costs more but results in less downtime

Return to Table of Contents

Choose a Lesson

Managed Databases Overview

Managed Databases on Google Cloud

Cloud SQL Closer Look

BigQuery Closer Look

BigQuery - Closer Look

Next

What is BigQuery

- Serverless data warehousing for analytics
- Massive scale
- Interact via familiar SQL queries
- Structure is project - dataset - table
- IAM roles applied to project/dataset, not tables

Exam Perspectives

- Refreshed version of architect exam goes into much more detail on managing BigQuery resources and users
- By comparison: Data Engineer exam tests on creating and using BigQuery data with some permissions
- Architect focuses more on properly managing access to resources, especially cross project
- Some overlap between both

Viewing Job/Query History

- BigQuery's version of "HEY, what's going on?"
- View number of BigQuery jobs per person + details
- `bq ls -j -a (myproject)`
- Also view via web console - Query/Job history
- Job/query history persists for 6 months. If you want to delete job history sooner, contact support
- BigQuery Admin can view all jobs. BigQuery User and JobUser roles can only view their own.
- Can also view job/query data in Stackdriver - BigQuery
  - Export to GCS/BQ
  - Manage lifecycle in GCS (retention)
  - Configure expiration settings in BQ to set time limit on retention (explained later)

Detailed permissions to datasets and other projects

- Primary BQ permissions resource: https://cloud.google.com/bigquery/docs/access-control
- Common scenario - Run jobs and queries in Project A (billing project) while querying datasets in in Projects A and B
  - Grant BigQuery user role to users in billing project A, grant DataViewer role to same user(s) in projects containing data to be queried

Partitioning tables

- Why is this important?
- If you have a massive table, queries will be against the entire table (or all entries in a column), which equals increased costs
- Partitioning table = Dividing a large table into smaller logical segments, called partitions
- Results:
  - Improved query performance
  - Less costs
- Partitioning methods
  - Ingest time: When the data arrives to the BQ table
  - By Timestamp/Date: Timestamp in a certain column
  - You can only partition tables by one of the time factors above
- Alternate method: Sharded tables
- Completely separate tables divided by date
- Partitioning recommended over sharding for performance and overhead

Return to Table of Contents

Choose a Lesson

- Managed Databases Overview
- Managed Databases on Google Cloud
- Cloud SQL Closer Look
- BigQuery Closer Look

BigQuery - Closer Look

Previous

Set expiration date on table data...

- Not all data needs to remain forever - it does have a cost
- Can automatically remove data that is temporary in purpose
- Setting expiration date automatically removes data that is (x) days old
- If using partitioned tables, expiration is applied to individual partitions
- `bq mk --time_partitioning_type=DAY --time_partitioning_expiration=259200 [DATASET].[TABLE]`

....or just keep it

- Tables that are not edited for 90 days auto-convert to long term storage pricing
- Same rate as GCS Nearline - $0.01/GB/month
- Partitioned tables - each partitioned qualifies separately for long term storage pricing

Query external data sources (federation)

- BigQuery can run queries on external data in the following locations:
  - Google Cloud Storage
  - Bigtable
  - Google Drive
- Why do this?
  - Load and clean from external source, and write cleaned data into BigQuery
  - Frequently changing data to query in external source; no need to reload changing data into BigQuery, just query updated external source
  - Query performance will be impacted
- Why **not** do this?
  - Size concerns: BigQuery can handle nearly infinite amount of data; just load directly into BigQuery instead
  - Non-updating data: Data that is not being constantly updated (not appended) should just be loaded into BigQuery

Return to Table of Contents

Choose a Lesson

Return to Table of Contents

Choose a Lesson

- VPC Concepts
- Firewalls
- Shared VPC

VPC Concepts

Next

Prerequisites for this section

- Familiar with basic networking concepts
- Subnet (CIDR notation), IP address, routes, firewall, etc.
- Required skills for exam
- Go through **Subnetting Fundamentals** course for further study

What is a Virtual Private Cloud (VPC)?: "The basics"

- Software-Defined Network (SDN)
  - Virtual version of traditional physical networks
- Terms ‘VPC’ and ‘network’ are interchangeable
- Central foundation of all other networking functions on GCP
- **Global** (multi-regional) communications space, private communication among resources
- Create subnetworks from single VPC
  - Subnets are **regional**. Can have single subnet span multiple zones.
- Project based, but can share between projects with Shared VPC
- Traditional networking concepts apply
  - Firewalls, routes, load balancing, DNS, etc.
- Traffic to/from instances controlled with **firewall rules**.
- VPC instances can have both **external** and **internal** IP addresses
  - Assign to a network and subnet (if not default)
- Hybrid networking with on-premises networks with interconnect options
- Can configure private (internal-only) access to other GCP resources
- Incoming (**ingress**) traffic is free. Outgoing (**egress**) traffic has a cost

IAM roles

- Falls under Compute Engine
- Compute Admin: Full access to both instance and network admin functions
- Compute Network Admin: Network admin roles only

Quotas and Limits

- Current limit of 15,500 virtual machines per VPC (may change over time)
  - Cannot be raised via Quotas console
  - If more are needed, create a new VPC network or call customer sales engineer
  - No limit per subnet, just across entire VPC
- IPv4 unicast traffic only (within VPC)
  - No broadcast/multicast
  - No IPv6 internally, but global load balancer IPs and traditional App Engine do support
- Most other quotas can be increased by request

Return to Table of Contents

VPC Concepts

Choose a Lesson
VPC Concepts
Firewalls
Shared VPC

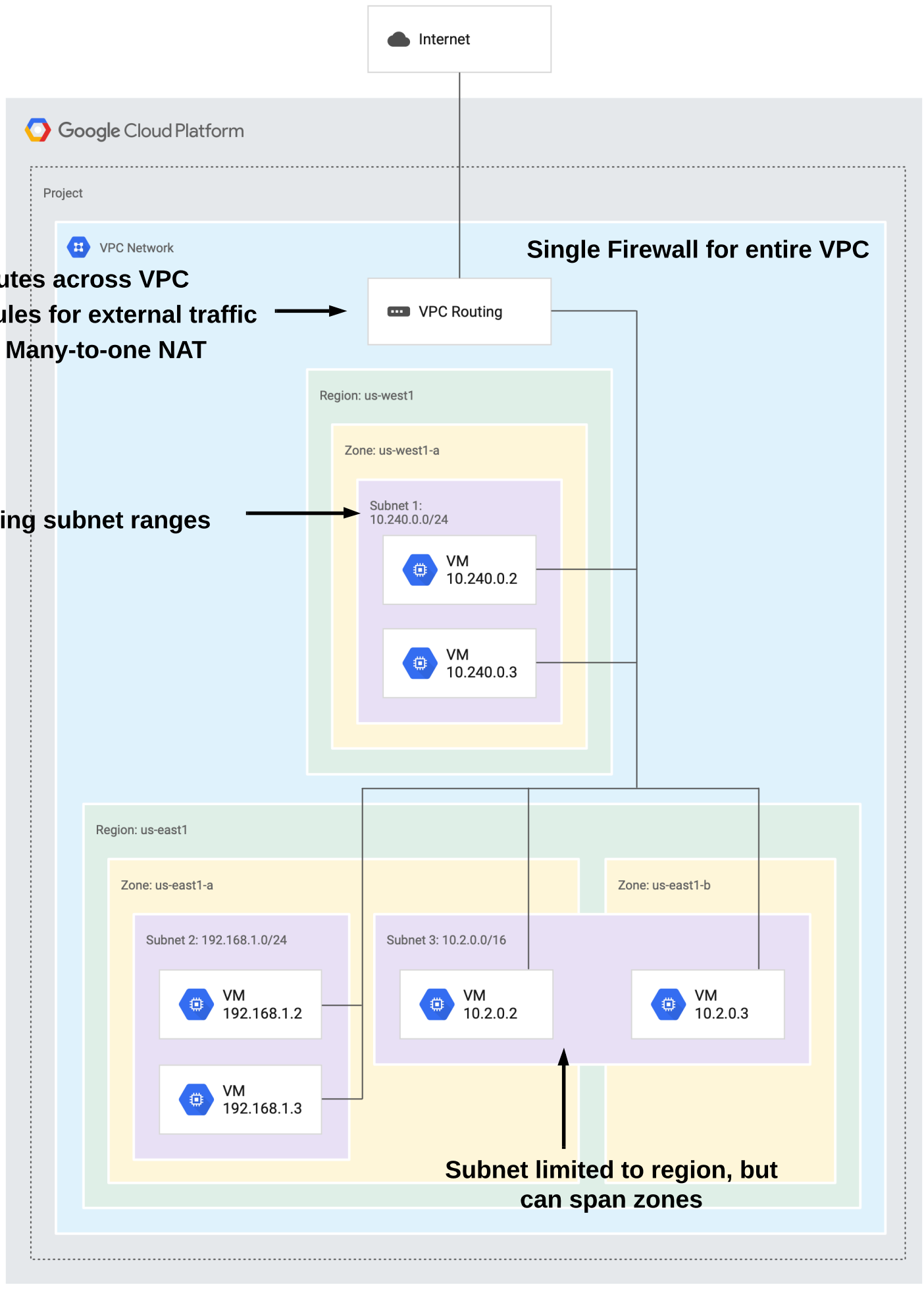
Previous

VPC Details

Dynamic routes across VPC
Custom routing rules for external traffic
Proxy server, Many-to-one NAT

Non-overlapping subnet ranges

Multiple regions on global, private network



Return to Table of Contents

Choose a Lesson

VPC Concepts

Firewalls

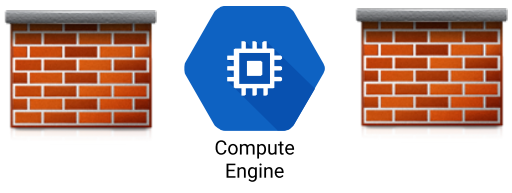
Shared VPC

Firewalls

Firewalls Basics

Next

- Allow/deny traffic to and from instances
  - Based on configuration
- Manage both **inbound (ingress)** and **outbound (egress)** traffic
- Defined at network (VPC) level, but enforced for each instance



- Rules manage both external access and also access between internal resources
- Implied ‘deny all’ ingress
- Implied ‘allow all’ egress
- Firewall components:
  - Directions: Ingress/egress
  - Source (traffic location source rule is applied to)
  - Target (traffic destination)
  - Protocol/port
  - Action: Allow/deny
  - Priority: Order rules evaluated—first matching rule applied

Conditions for determining access

- Source/target
- Protocols
  - Ports
- Network Tags on instances
  - Per-instance traffic management
  - Also applies to routing
- The above can be optionally combined for granular access

Return to Table of Contents

Choose a Lesson

- VPC Concepts
- Firewalls
- Shared VPC

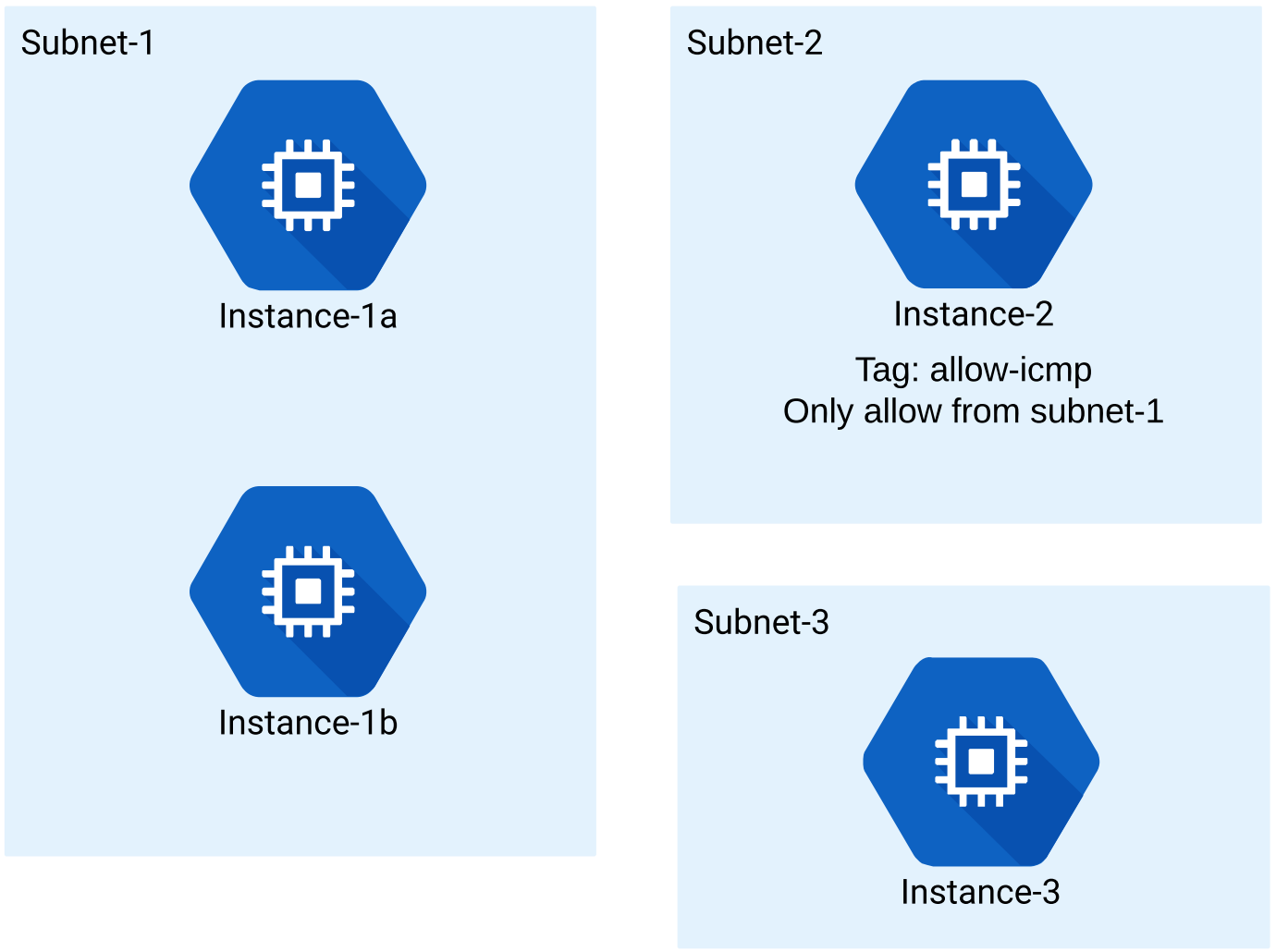
Firewalls

Previous

Hands-On Guideposts

- **Communicate with other instances using firewall rules and network tags**

- **Allow SSH access to all**
- **Allow ICMP (ping) to selected instances by network tag only from subnet-1**
- **View command line cross reference**



Return to Table of Contents

Choose a Lesson
VPC Concepts
Firewalls
Shared VPC

Shared VPC

Next

Shared VPC Overview

- By default, VPC is tied to a single project
- Need may exist to share network resources across projects
- Shared VPC shares VPC across projects within an organization
  - i.e., Cross-Project Networking – also its former name

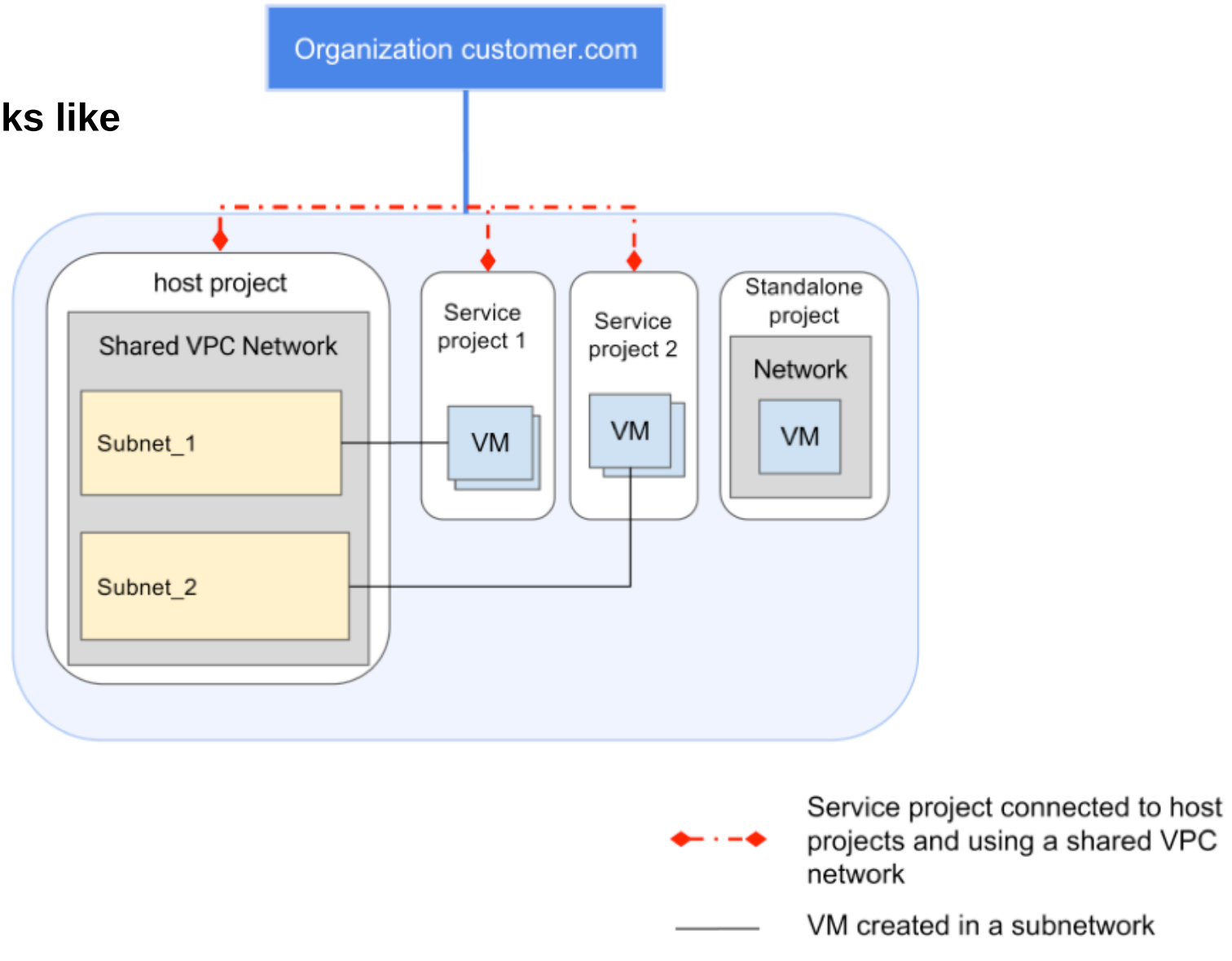
Concepts and Terminology

- **Host project**: Project hosting the shared VPC
- **Service project**: Project with permission to shared VPC
  - Shared VPC projects can be controlled by different departments
  - Ownership of resources in shared VPC maintained by project
- **Standalone project**: Project not using shared VPC (the default)
- **Shared VPC admin**: IAM role for administrator of shared VPC
- **Service project admin:** Project admin of shared VPC service project
  - More on IAM roles in a bit...

Why separate projects to begin with?

- Why not place everything in the same project?
- Separation of projects for access control and billing
  - ...but still need access to same VPC resources
- Projects are primary method of separating access

What it looks like



Return to Table of Contents

Choose a Lesson

VPC Concepts

Firewalls

Shared VPC

Shared VPC

Previous

Next

Considerations

- Only within **single** GCP organization
- Service project can only link to single host project
- Project cannot be both host and service project
- Existing projects can use shared VPC but existing instances cannot
- Reserved (static) IP addresses tied to project that reserved it

Resources that can use shared VPC

- Compute Engine Instances
- Compute Engine Instance Templates
- Compute Engine Instance Groups
- Google Kubernetes Engine clusters
- Internal IP Addresses
- Internal DNS
- Cloud DNS Private Zones
- Load Balancing

Use cases (diagrams to follow)

- Separate testing and production environments
- Two-tier web service
- Hybrid cloud scenarios

Return to Table of Contents

Choose a Lesson

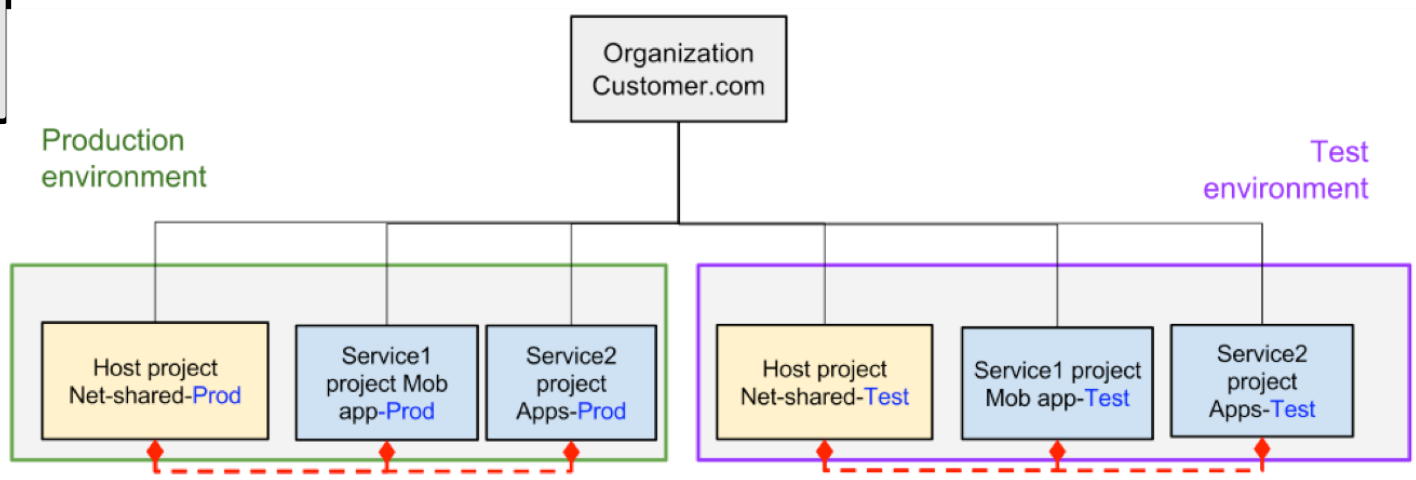
- VPC Concepts
- Firewalls
- Shared VPC

Shared VPC

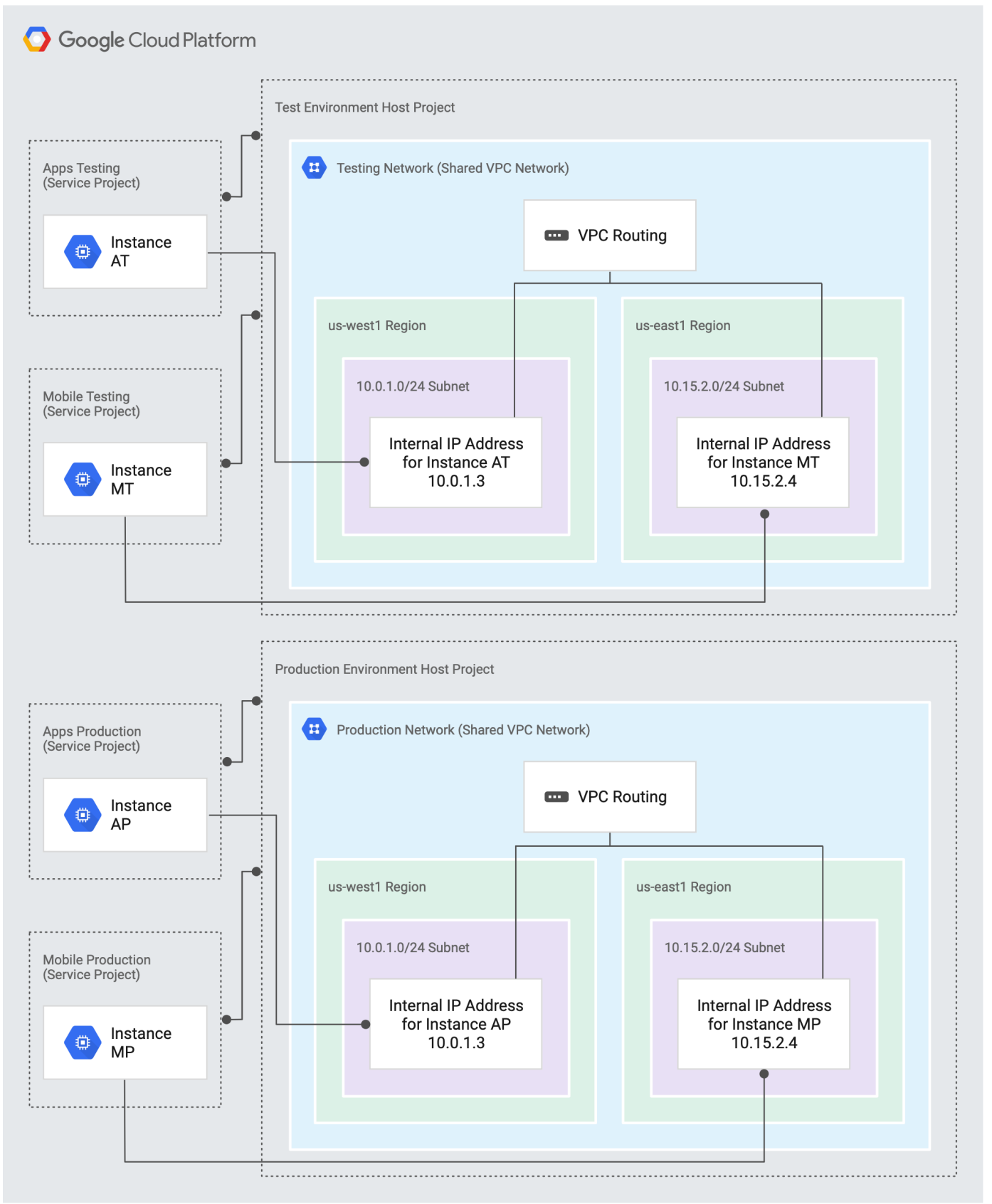
Previous

Next

Testing and Production Environments



Same subnet CIDR range and internal IP for testing and production equivalents



Return to Table of Contents

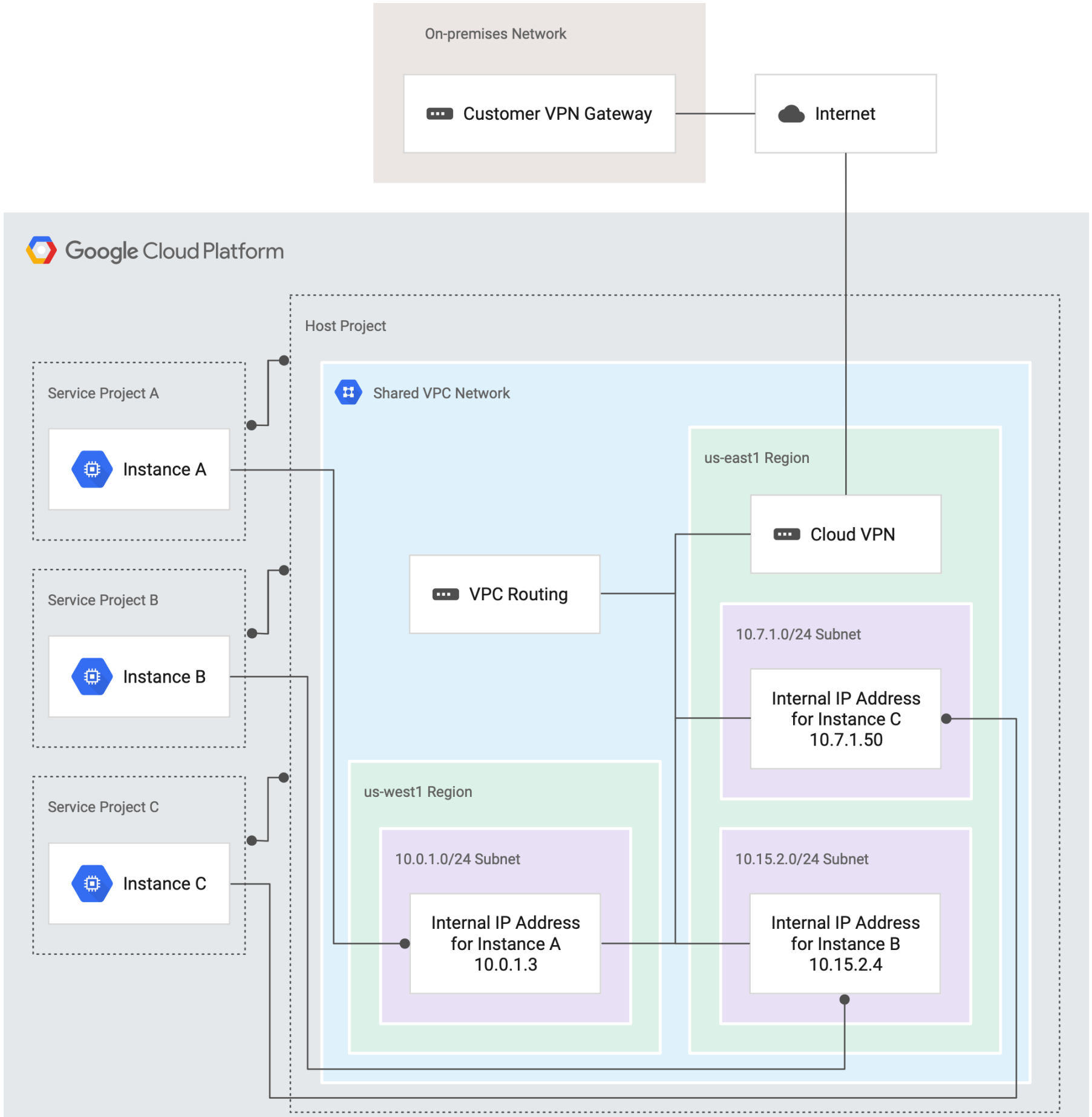
Choose a Lesson
VPC Concepts
Firewalls
Shared VPC

Shared VPC

Previous
Next

Hybrid Cloud

- VPN gateway connection to single shared VPC
- Resources worked on by different teams
- Access restricted by different projects



Return to Table of Contents

Choose a Lesson

- VPC Concepts
- Firewalls
- Shared VPC

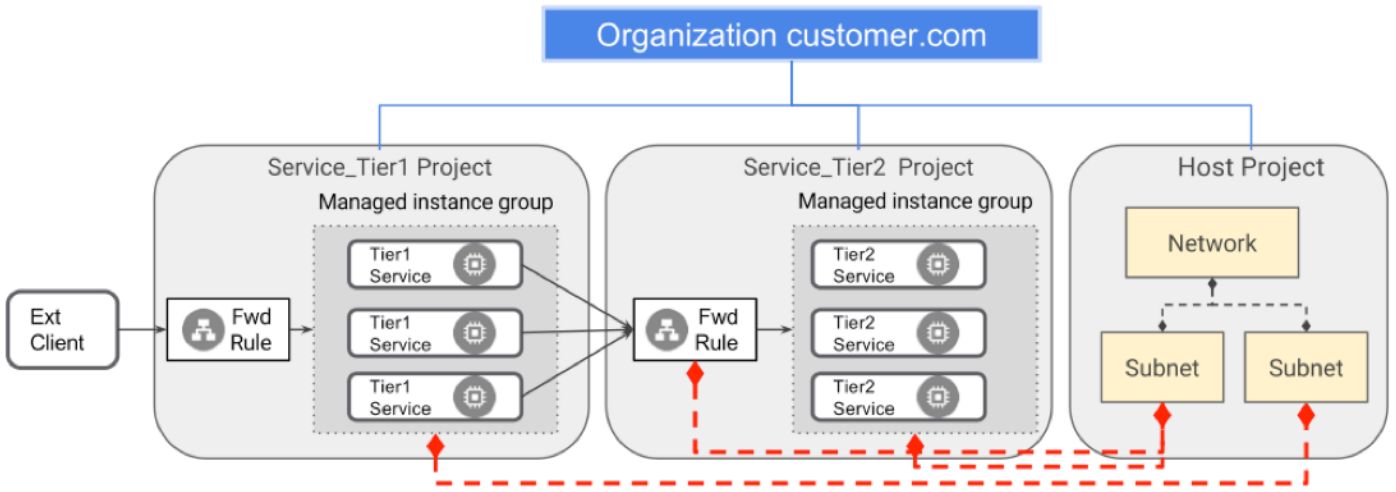
Shared VPC

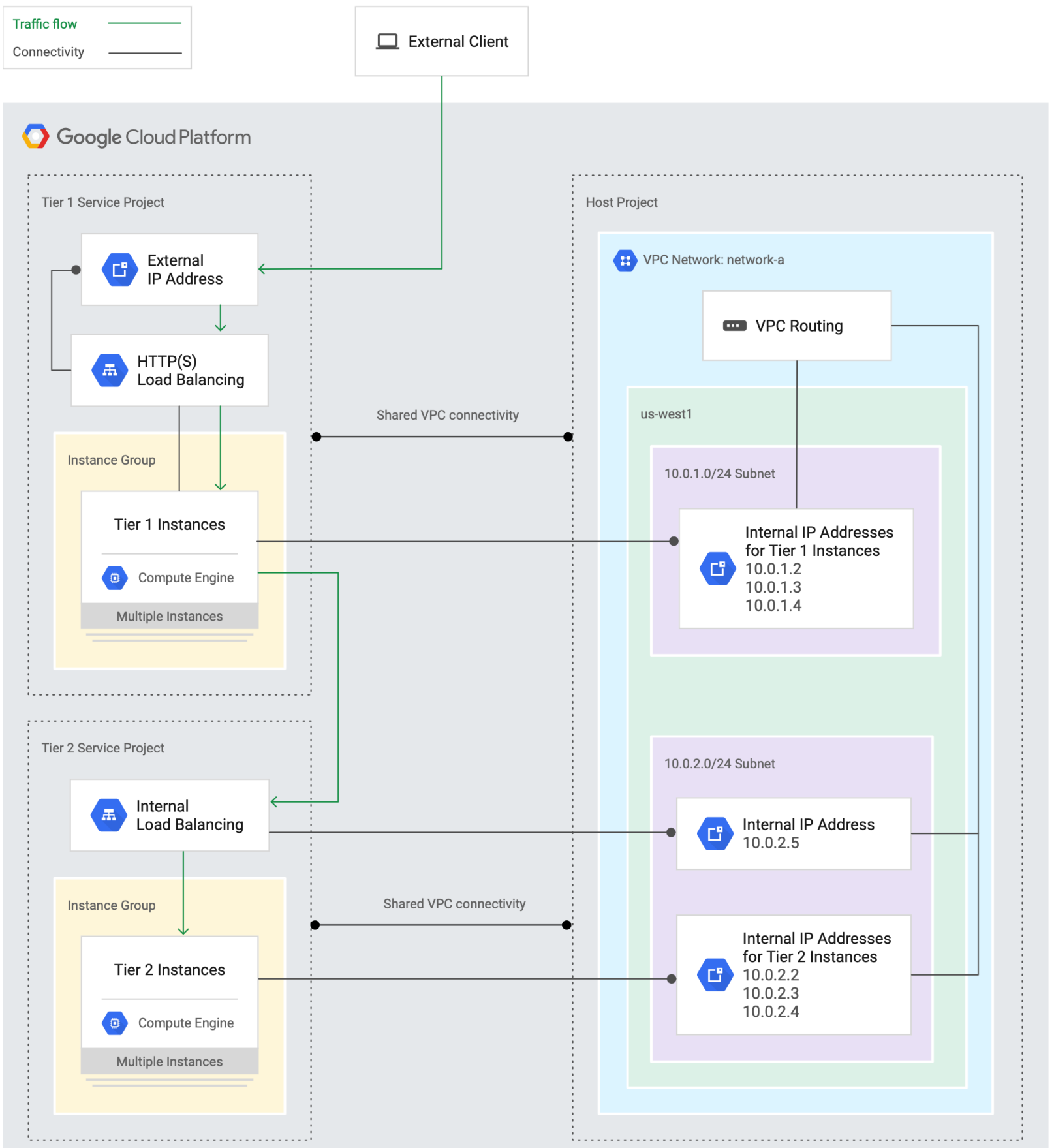
Previous

Next

Two-tier Web Application

- Again, team access to specific tier separated by projects (same for billing)





Return to Table of Contents

Choose a Lesson

VPC Concepts

Firewalls

Shared VPC

Shared VPC

Previous

Next

IAM roles

- Organization Admin
- Shared VPC Admin: compute.xpnAdmin
  - Organization level role
  - Configure shared VPC
  - Associate service projects with host projects
  - Grant Network User role
- Network User: compute.networkUser
  - Project level role
  - Create resources to use shared VPC
  - Discover shared VPC assets
  - Requires project admin role (Project Owner, Editor, Compute Engine Admin)

Return to Table of Contents

Choose a Lesson

- VPC Concepts
- Firewalls
- Shared VPC

Shared VPC

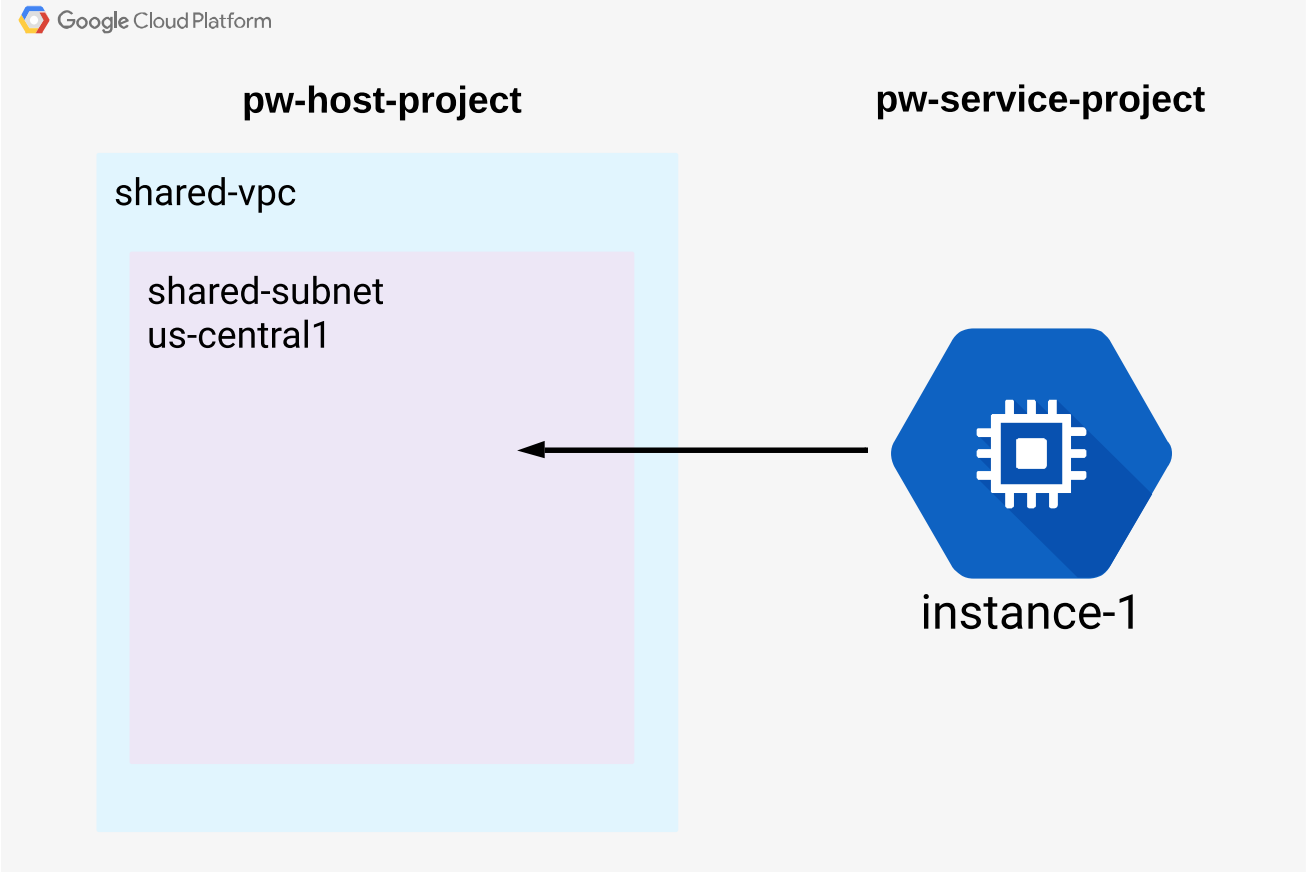
Previous

Hands-On Guideposts

- Create host and service project
- Enable VPC admin account - org level (folder level in beta)
- Enable Compute Engine API in both projects
- Host project - create custom VPC network
  - Enable shared VPC
  - Share subnet(s)
  - Specify project name(s) (attach service project)
  - Select users as service project admins
- Go to service project, view shared VPC settings
- Create compute engine instance, attach to shared VPC

Deprovision shared VPC

- Remove service project resources
- Detach service project (must remove resources first)
- Disable host project
- Delete project



Return to Table of Contents

Choose a Lesson

Return to Table of Contents

Choose a Lesson

The Power of the Network

Connecting Your Network to Google

Google Cloud VPN

Google Cloud DNS

The Power of the Network

Google's network is AWESOME

- Biggest distinction compared to other platforms
- All regions are on one global private network
- **By default**, communication between regions and on-premises (peering/interconnect) never touches public Internet!
  - i.e., "global by default"
- As a result, networking is handled differently

Software-Defined Networking (SDN)

- Traditional network/data center – manage network hardware
  - Switches, routers, load balancers, firewalls, storage devices etc.
  - Detailed device configurations, monitor network software, high management overhead
- Software-Defined Networking – everything is virtualized
  - Removes overhead
  - Rapidly customize and scale services
  - High throughput
  - Global availability
  - Seamless upgrades
- **Note**: Traditional networking concepts still apply, such as subnet, routes, firewall rules, DNS, etc.

How does this affect networking on GCP?

- Single global/cross-region VPCs
  - No managing multiple private networks for global availability
- Global internal DNS/load balancing/firewalls/routes
  - Separate resources with tags
- Global public DNS
- Rapid scaling with global load balancers (Layer 7/HTTP)
- Subnets within VPC group resources by region/zone
- IP ranges between subnets are dynamically expandable

Extending your GCP network to external sources (on-premises)

- Cloud Interconnect
- VPN
- Peering

Return to Table of Contents

Choose a Lesson

The Power of the Network

Connecting Your Network to Google

Google Cloud VPN

Google Cloud DNS

Connecting Your Network to Google

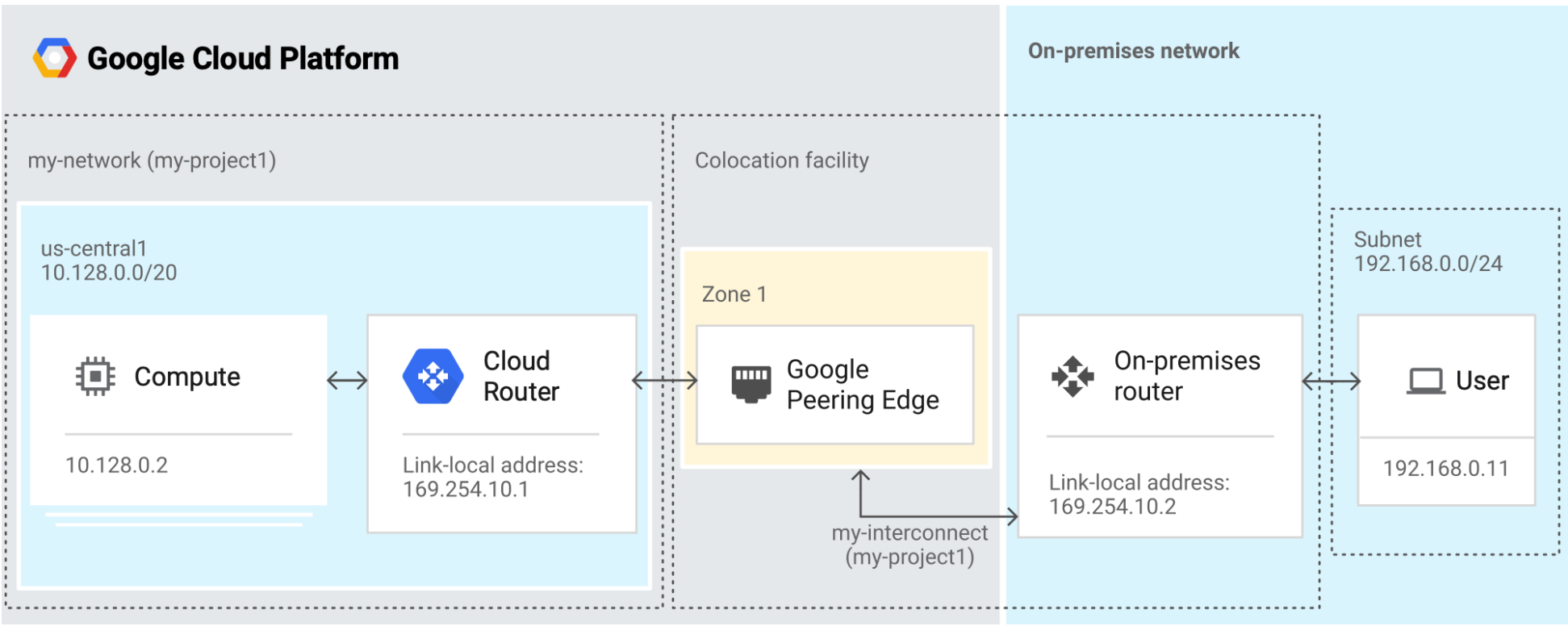
Next

Extend your network to Google's network

- Cloud Interconnect
- Cloud VPN
- Peering
- VPC Network Peering

Cloud Interconnect

- Physically connect on-premises network to GCP VPC network via Google edge location
  - Enterprise grade connection to GCP
- Must be at supported peering location
- Can be direct with Google (direct) or through carrier (partner)
- (Direct) $1700 per 10 Gbps link, up to 80 Gbps total
  - Partner options at variable speeds/costs
- Reduced egress fees



When to use Cloud Interconnect over other options?

- Don't want traffic to touch public Internet
  - Dedicated physical connection either through Google or partner
  - Fewer hops, less points of failure
- Direct integration with internal network—no VPN tunnels to set up
  - Same RFC 1918 IP space
- Need extremely high speed connections—up to 80 Gbps per interconnect
- High volume of egress (outgoing) traffic from GCP—reduced egress fees
- Use Private Google Access for on-premises hosts (i.e., private Cloud Storage access)

When not to use?

- Don't require high speed/low latency
  - Interconnect is the 'heavy-duty' option
- Peering location not available
- Need to access external Internet
- Save costs
- Cloud VPN may be a better choice

Return to Table of Contents

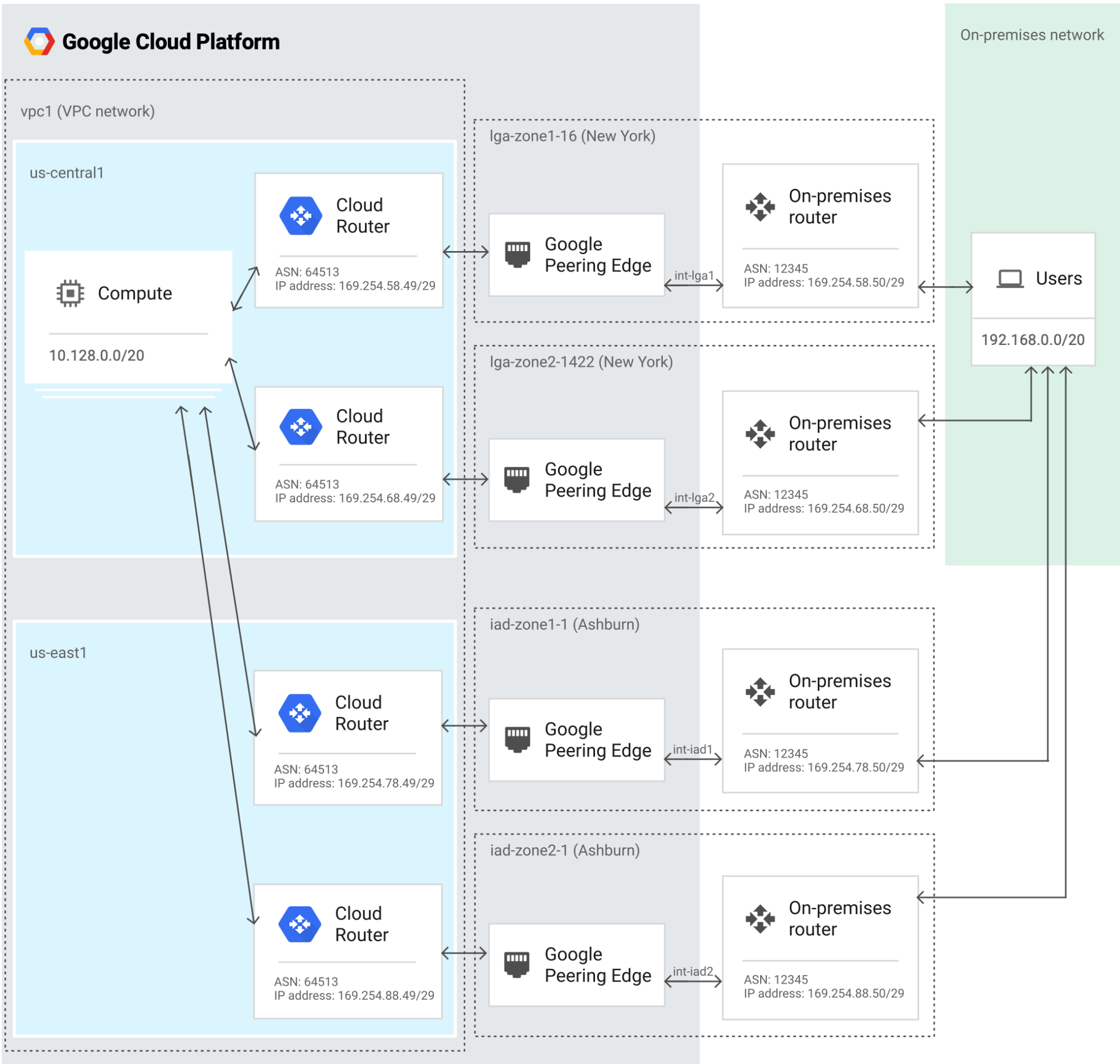
Connecting Your Network to Google

Previous

Next

Choose a Lesson

Multiple Interconnects for high availability



Return to Table of Contents

Choose a Lesson

- The Power of the Network
- Connecting Your Network to Google
- Google Cloud VPN
- Google Cloud DNS

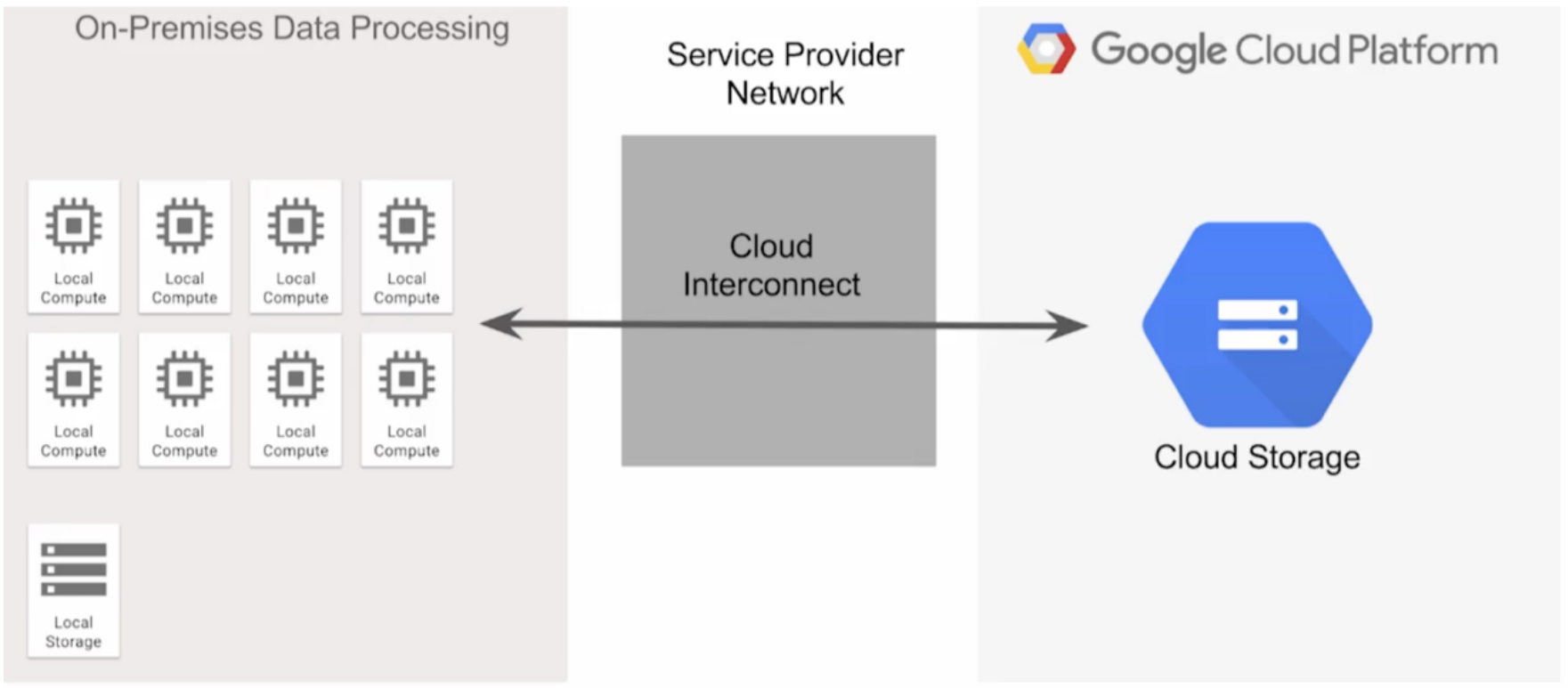
Connecting Your Network to Google

Previous

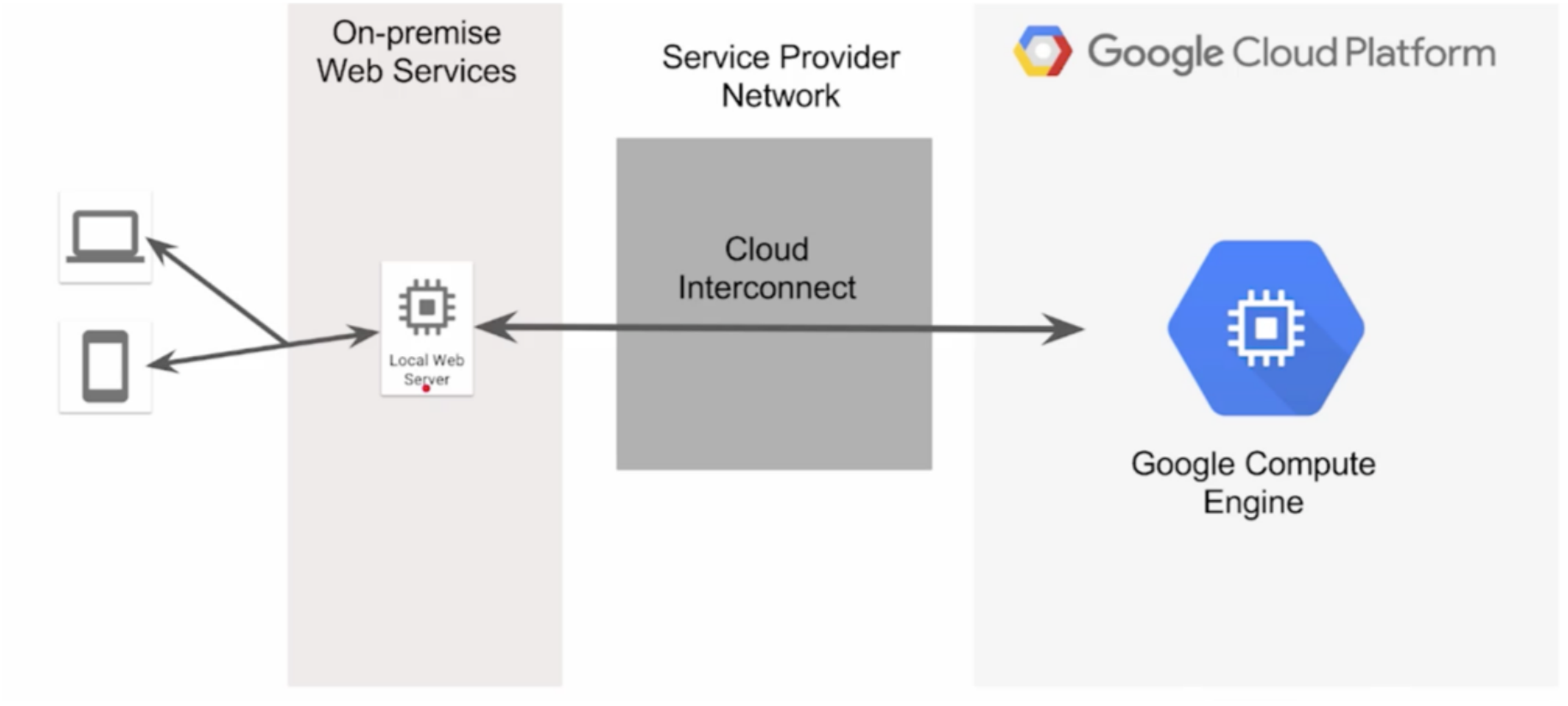
Next

Cloud Interconnect Use Cases

Heavy (volume) Processing



Low-latency/high bandwidth



Return to Table of Contents

Choose a Lesson

- The Power of the Network
- Connecting Your Network to Google
- Google Cloud VPN
- Google Cloud DNS

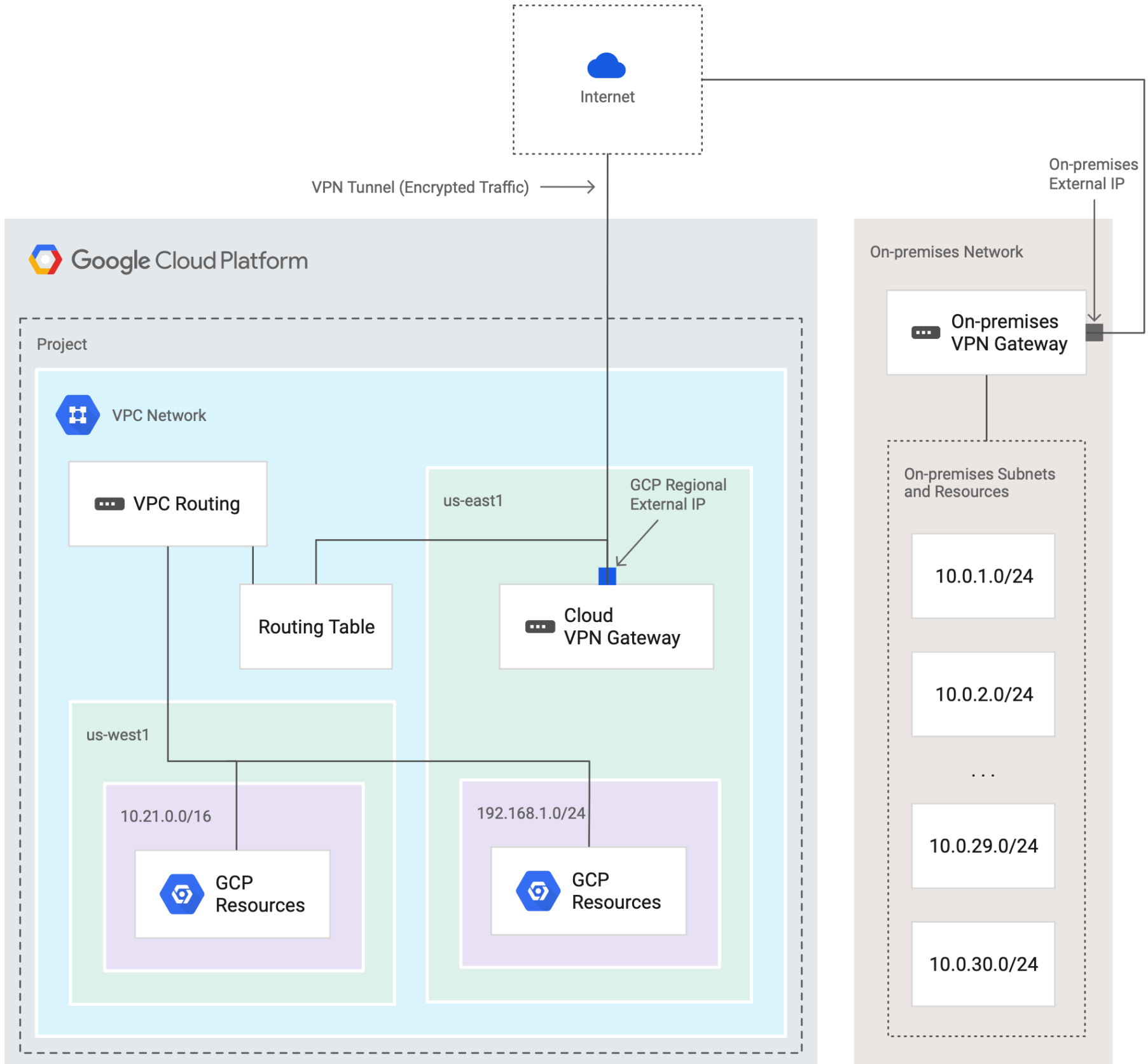
Connecting Your Network to Google

Previous

Next

Cloud VPN

- The option for "most of us"
- Site to site VPN connection over IPSec
- Connect internal network to GCP over encrypted tunnel over public Internet
- Up to 1.5 Gbps per tunnel
- Can use multiple (up to 8) tunnels for increased performance
  - 1.5 Gbps x 8 = 12 Gbps per gateway combined
  - Example: Transfer legacy resources to GCP
- Static and dynamic routes (using Cloud Router)
- Supports IKEv1 and IKEv2 using shared secret
- Connect on-premises to GCP or connect two different VPCs on GCP
- No site to client option available
  - Example: Connecting to GCP VPN via laptop



Return to Table of Contents

Choose a Lesson

The Power of the Network

Connecting Your Network to Google

Google Cloud VPN

Google Cloud DNS

Connecting Your Network to Google

Previous

Next

Peering

- Direct and Carrier Peering
- Connect business directly to Google (not just GCP/VPC)
  - Over public Internet or dedicated connection
- Not GCP specific, but exchanging Internet traffic with Google
- Exchange Border Gateway Patrol (BGP) routes
- Does not connect to external Internet
- Useful for connecting directly to Google (not just GCP)
  - Example: Connect directly to G Suite services
- Also save on egress fees
- 10 Gbps per link (direct), variable for carrier

VPC Network Peering

- Connect two GCP VPC networks
- Private communication (RFC 1918) over GCP's private network
- VPCs can be in different organizations
- Compared to VPN:
  - Lower latency
  - No traffic exposed to public Internet
  - Save on egress costs (due to no external IPs)

Exam perspective

- Interconnect and VPN primary focus
  - When to use Interconnect over VPN
    - Higher bandwidth
    - Massive data transfer per day (measured in TBs)
  - Avoid overlapping subnets with GCP and external network

Return to Table of Contents

Choose a Lesson

- The Power of the Network
- Connecting Your Network to Google
- Google Cloud VPN
- Google Cloud DNS

Connecting Your Network to Google

Previous

Decision Tree

Return to Table of Contents

Choose a Lesson

The Power of the Network

Connecting Your Network to Google

Google Cloud VPN

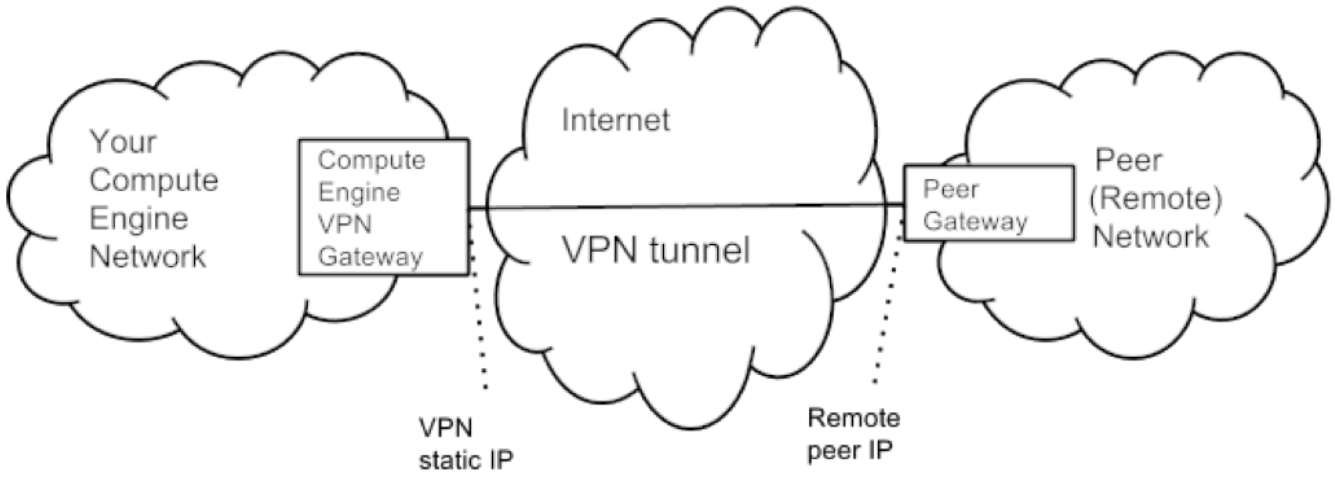
Google Cloud DNS

Google Cloud VPN

Next

What is Cloud VPN?

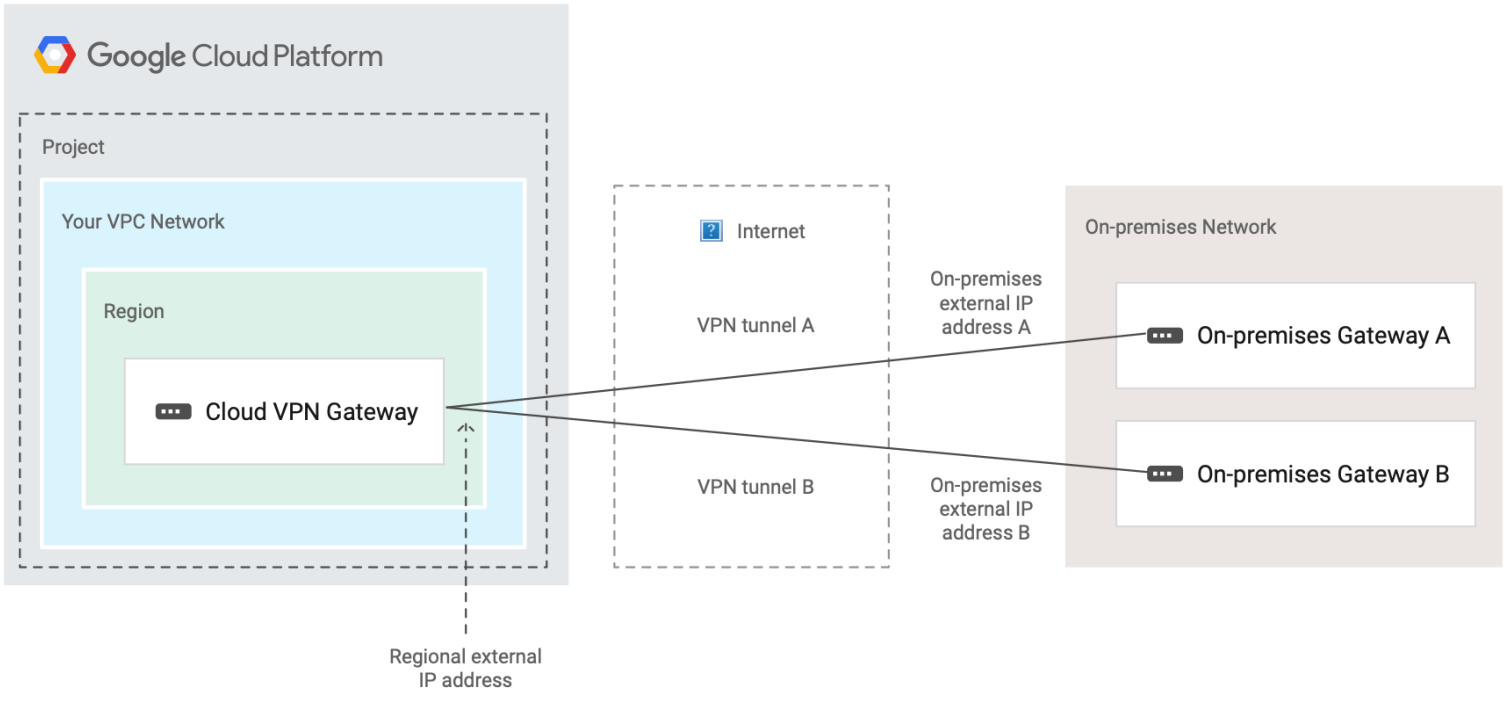
- Connect on-premises network to GCP Virtual Private Cloud (VPC)
- IPSec connection over VPN over public Internet
- Traffic encrypted by one gateway, then decrypted by other gateway



Cloud VPN Traits

- 99.9% SLA
- Site-to-site VPN only, no site-to-client (road warrior)
- Up to 1.5 Gbps per tunnel (3 Gbps per direct peering link), up to 8 tunnels per VPN gateway for increased performance and redundancy
- Static and dynamic routes (with Cloud Router)
- Supports IKEv1 and IKEv2 using shared secret

Multiple Tunnels



Return to Table of Contents

Choose a Lesson

- The Power of the Network
- Connecting Your Network to Google
- Google Cloud VPN
- Google Cloud DNS

Google Cloud VPN

Previous

Next

Requirements

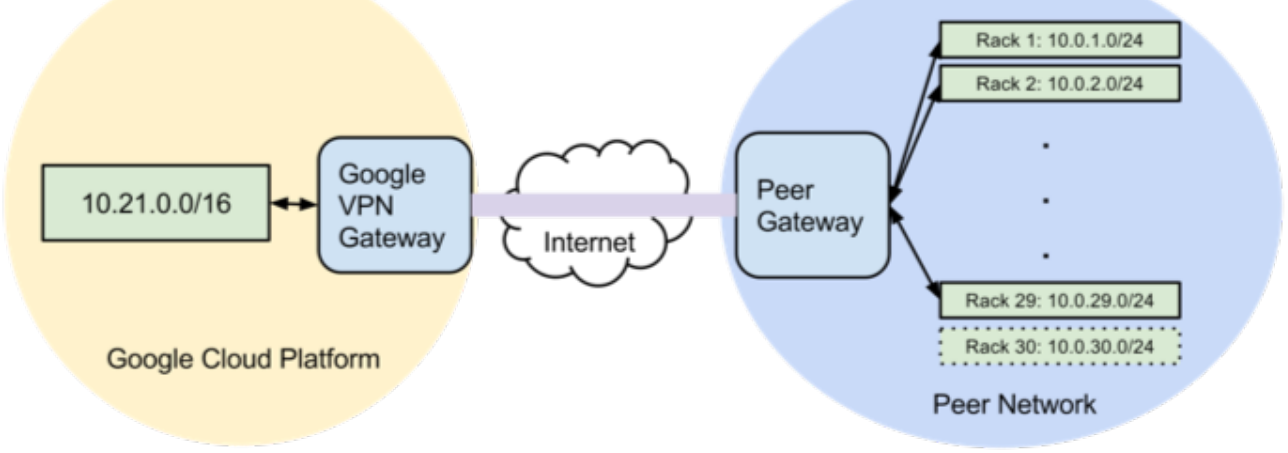
- VPN gateway on both ends (peer)
  - Non-GCP = on-premises VPN server/router
- Peer gateway must have static IP address
  - If behind firewall, configure to pass ESP and IKE traffic
- Non-conflicting CIDR range/subnet with rest of network (Networking 101)

Dynamic routing with Cloud Router

- Not required for VPN router, but makes things much easier
- Static vs. dynamic routing
- Static – create routing table for all existing and new routes. Can't re-route traffic if link fails
- Dynamic – networks automatically discovery topology changes via BGP
  - Can re-route traffic if link fails
- i.e., "the easy way"

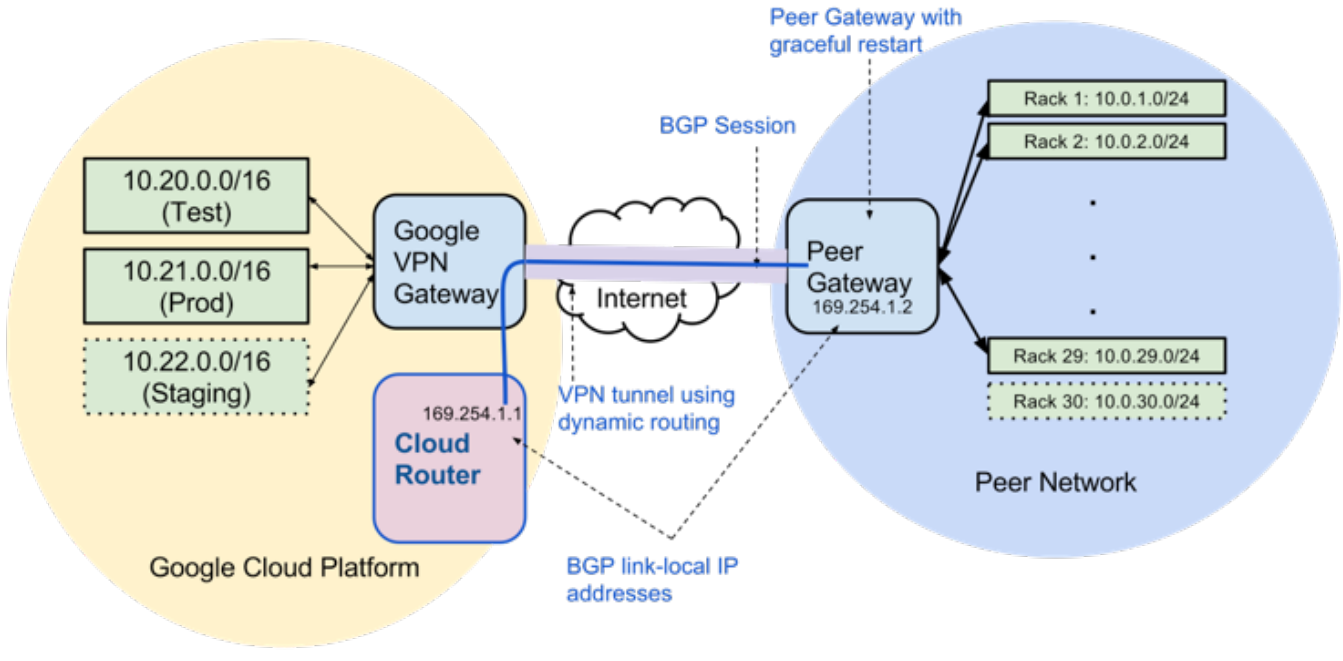
Static Routing

Manually enter every subnet route



Dynamic Routing

Cloud Router automatically discovers new subnet routes



Return to Table of Contents

Choose a Lesson

- The Power of the Network
- Connecting Your Network to Google
- Google Cloud VPN
- Google Cloud DNS

Google Cloud VPN

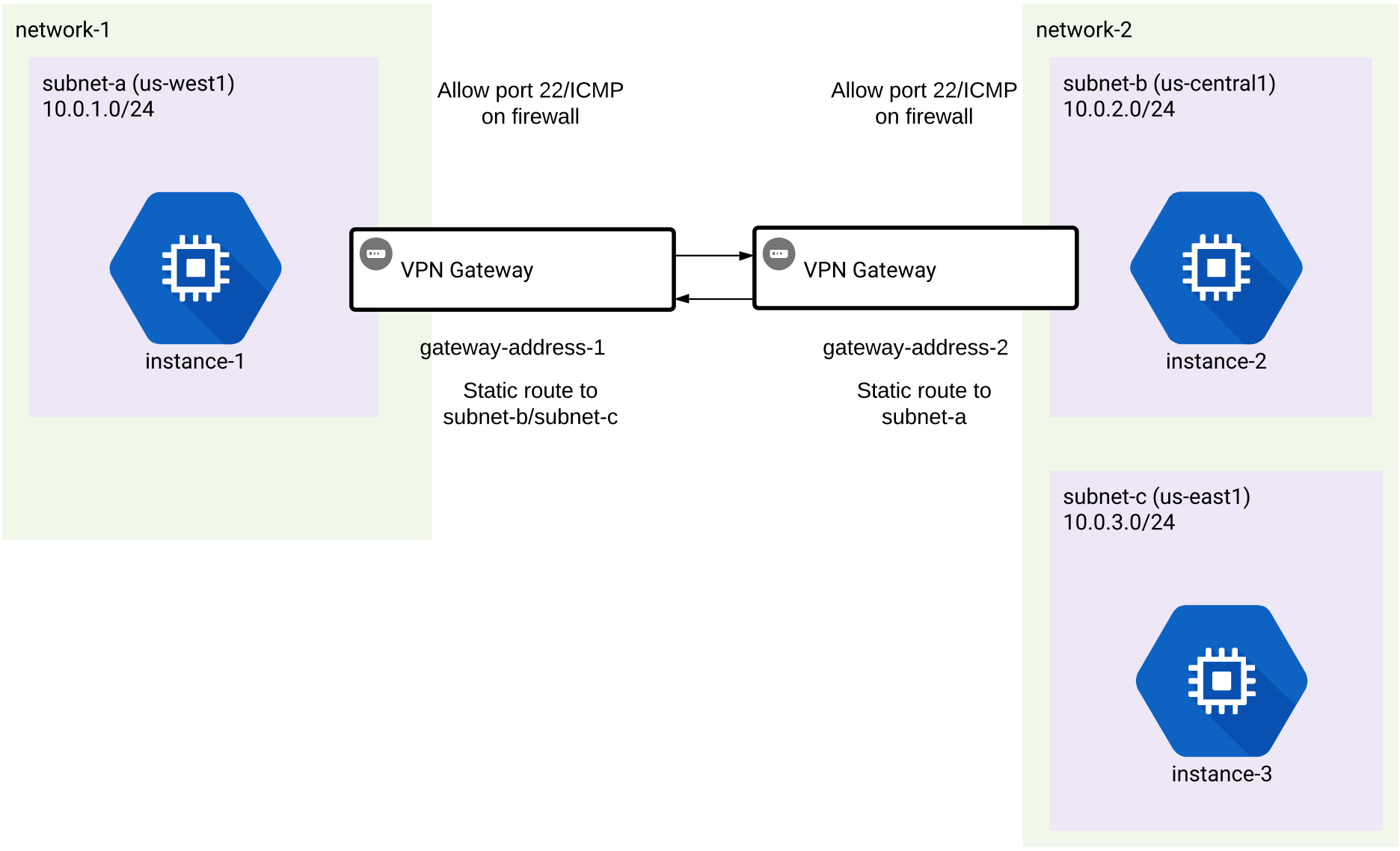
Previous

Next

Hands-On Guideposts

- Create static IP address for both VPN gateways
- Create first VPN gateway
- Create VPN tunnel to second VPN gateway
- Create second VPN gateway
- Create VPN tunnel to first VPN gateway

VPN with static routes



Return to Table of Contents

Choose a Lesson

- The Power of the Network
- Connecting Your Network to Google
- Google Cloud VPN
- Google Cloud DNS

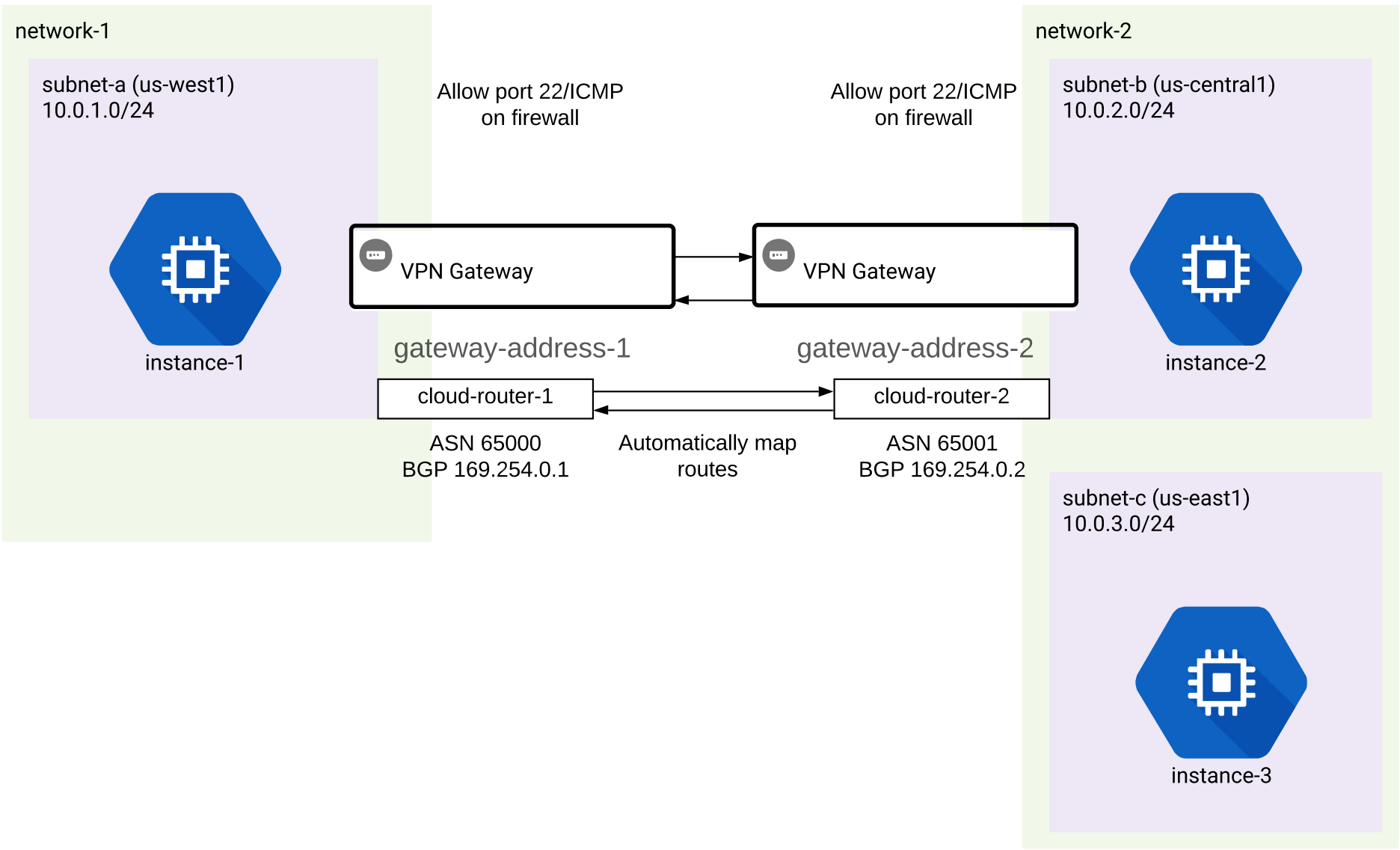
Google Cloud VPN

Previous

Hands-On Guideposts

- Create static IP address for both VPN gateways
- Create Cloud Router for each VPC
- Enable global dynamic routing on each VPC
- Create first VPN gateway
- Create VPN tunnel to second VPN gateway (using Cloud Router)
- Create second VPN gateway
- Create VPN tunnel to first VPN gateway (Using Cloud Router)

VPN with dynamic routes (via Cloud Router)



Return to Table of Contents

Choose a Lesson

- The Power of the Network
- Connecting Your Network to Google
- Google Cloud VPN
- Google Cloud DNS

Google Cloud DNS

What is it?

- Globally available Domain Name System (DNS) service
- DNS: Translating network addresses into human-readable names
  - e.g., https://172.217.1.46 = *google.com*
- AWS equivalent = Route 53
- **Only** GCP service with 100% SLA/availability
- Exam perspective - not prominent on Architect exam, more so on Network Engineer exam

Key Concepts and Terminology

- Hosted in a single project
- Managed zones
  - Public/Private DNS zones
  - Public: DNS records of public domain (e.g., *linuxacademy.com*, *google.com*)
  - Each domain has its own zone
  - Private (Beta): Manage DNS records of GCP resources without public exposure
    - Specify which VPC networks can query private zone data

Return to Table of Contents

Choose a Lesson

Return to Table of Contents

Choose a Lesson

Compute Engine Deep Dive

Disks

Images

Snapshots

Startup and Shutdown Scripts

Preemptible VMs

Compute Engine Deep Dive

Why are we focusing on this so much?

- VM's are the heart and soul of GCP
- GCE, GKE, and GAE all run on VM's
- Substantial portion of exam will be VM focused
- Start with single VM's
- Move up to 'force multipliers' – the true 'magic' of cloud computing
  - Automation
  - Auto-scaling
  - Managed instance groups
  - Load balancers

In this deep dive:

- Learn the main 'building blocks'
- Disk manipulation
- Custom images
- Snapshots
- Startup/shutdown scripts
- Preemptible VM's
- gcloud commands for all of the above
- Prepare you to work with 'force multiplier' concepts

Return to Table of Contents

Choose a Lesson

- Compute Engine Deep Dive
- Disks
- Images
- Snapshots
- Startup and Shutdown Scripts
- Preemptible VMs

Disks

Next

Disk options

- All instances have a single root disk for the OS (persistent disk)
- **Persistent Disk**
  - Most common, default option
  - Not directly attached - network attached storage
  - Standard and SSD variety
- **Local SSD**
  - Directly attached to your VM
- **Cloud Storage buckets**
  - Niche option
  - High collaboration, 'infinite' space
  - Not block storage

Persistent disks

- **Only boot option - the default option**
- **Network-attached, array of multiple disks**
  - Redundancy, reliability, and performance (mirroring + striping)
  - Benefits of RAID - no configuration needed
- Very flexible and powerful - less physical constraints - **Modular!**
  - Independent from the VM instance – not physically attached
  - Can detach/move disks - read from multiple instances at once (read only mode)
  - Preserve data after deleting instances
  - Resize, move, attach additional disks - even when in use!
  - No partitioning necessary - just resize or add new disks
- Performance scales with size
- SSD option available
- Encrypted – either with Google provided keys or bring your own
- Use case: File server - https://cloud.google.com/solutions/filers-on-compute-engine

Local SSD

- Highest performance, but with caveats
  - **Physically attached** to VM
  - **Cannot be boot device** - must be attached disk
  - Must create on instance creation
- **375GB in size – non-configurable** – can attach up to 8
- Trade performance for reliability/flexibility
  - Not automatically replicated
  - All data lost if instance terminated
- Best practice – fast scratch disk, replicate workload across multiple instances
- Encryption – Google supplied, cannot use your own
- Can attach both Local SSD and Persistent Disks to a single instance

Cloud Storage Bucket

- Niche use case
- Uses GCS Fuse application
- Not a boot disk
- Most flexible, scalable, and durable storage option
- Lower performance than other disk options
- Global accessibility vs. zone for other disks
  - Instances in multiple regions/zones can write to same bucket

Return to Table of Contents

Choose a Lesson

- Compute Engine Deep Dive
- Disks
- Images
- Snapshots
- Startup and Shutdown Scripts
- Preemptible VMs

Disks

Previous

Next

Breakdown of different disk/storage options on Compute Engine

	Standard persistent disks	SSD persistent disks	Local SSDs	Cloud Storage buckets
Storage type	Efficient and reliable block storage	Fast and reliable block storage	High-performance local block storage	Affordable object storage
Maximum space per instance	65 TB	65 TB	3 TB 375GB/disk X 8 disks	Almost infinite
Scope of access	Zone Dual-zone in beta	Zone Dual-zone in beta	Instance	Global
Boot disk	Yes	Yes	No	No
Data redundancy	Yes	Yes	No	Yes
Encryption at rest	Yes	Yes	Yes	Yes
Custom encryption keys	Yes	Yes	No	Yes
Machine type support	All machine types	All machine types	Most machine types	All machine types
Storage type	Network attached array Block storage	Network attached array Block storage	Single direct-attached, physical disk Block storage	Object storage Not block storage

Disk Performance

	Standard persistent disks	SSD persistent disks	Local SSD (SCSI)	Local SSD (NVMe)
Maximum sustained IOPS				
Read IOPS per GB	0.75	30	266.7	453.3
Write IOPS per GB	1.5	30	186.7	240
Read IOPS per instance	3,000	15,000 - 40,000*	400,000	680,000
Write IOPS per instance	15,000	15,000 - 30,000*	280,000	360,000
Maximum sustained throughput (MB/s)				
Read throughput per GB	0.12	0.48	1.04	1.77
Write throughput per GB	0.12	0.48	0.73	0.94
Read throughput per instance	180	240 - 800*	1,560	2,650
Write throughput per instance	120	240 - 400*	1,090	1,400

Return to Table of Contents

Choose a Lesson

Compute Engine Deep Dive

Disks

Images

Snapshots

Startup and Shutdown Scripts

Preemptible VMs

Disks

Previous

Next

Pricing

TYPE	PRICE (PER GB / MONTH)
Standard provisioned space	$0.04
SSD provisioned space	$0.17
Local SSD provisioned space (min 375GB disk)	$0.08
Snapshot storage	$0.03

Return to Table of Contents

Choose a Lesson

- Compute Engine Deep Dive
- Disks
- Images
- Snapshots
- Startup and Shutdown Scripts
- Preemptible VMs

Disks

Previous

Hands-On Guideposts

- View new instance disk options
- Create and attach a disk (pay attention to zone!)
- Resize disk
- Attach disk to instances
- Resize existing disk while running
- Use both web console and command line
- Command line:
  - Create disk:
    - `gcloud compute disks create (DISK_NAME) --type=(DISK_TYPE) --size=(SIZE) --zone=(ZONE)`
  - Resize disk
    - `gcloud compute disks resize (disk_name) --size=(size) --zone=(zone)`
  - Attach disk
    - `gcloud compute instances attach-disk (instance) --disk=(disk_name) --zone=(zone)`

Mount disk to Linux instance

- Attach new disk, format, mount
- View available disks
  - `sudo lsblk`
- format
  - `sudo mkfs.ext4 -m 0 -F -E lazy_itable_init=0,lazy_journal_init=0,discard /dev/sdb`
- create mount directory
  - `sudo mkdir -p /mnt/disks/disk2`
- mount disk
  - `sudo mount -o discard,defaults /dev/sdb /mnt/disks/disk2`
- set read/write permissions
  - `sudo chmod a+w /mnt/disks/`

Resize existing Linux Disk

- Identify the disk to resize
  - `sudo lsblk`
- Resize (grow) the partition
  - `sudo growpart /dev/sda 1`
- Extend file system to use added space
  - `sudo resize2fs /dev/sda1`
- Verify file system is resized
  - `df -h /dev/[DEVICE_ID]`

Return to Table of Contents

Choose a Lesson

Compute Engine Deep Dive

Disks

Images

Snapshots

Startup and Shutdown Scripts

Preemptible VMs

Images

Next

Why this matters

- Necessary to create boot disks for GCE instances
- Unlike disks, not limited to zone - much more accessible
- Public and Custom images
- Public - provided and maintained by Google/community/vendors
  - All projects have access
- Custom - Customized boot disks you create
  - Only available to your project (and other sources you control)

Differences between Images and Snapshots

- Images
  - Purpose - create new instances, configure instance templates
  - Recommended to shut down instances before creating new image
  - Access across projects (even different organizations)
- Snapshots
  - Purpose - periodic incremental backup of existing disk/instance
  - Can create while running
  - Cross-project access in 'transition'

Create from several sources

- Persistent disk
- Another image in same project
- Imaged shared from another project
- Compressed image from Cloud Storage

Managing Custom Images

- Image families simplifies image versioning
  - Useful for instance templates and scripts
- Groups related images together
  - Roll forward and back between image versions
  - Family always points to newest non-deprecated version

Deprecating images

- As custom images are continually updated, need to retire older versions
- Transition users away from older unsupported versions in manageable way
- Deprecation states
  - Deprecated – still works but gives warning
  - Obsolete – new users cannot use it - error if attempt to use, existing links still work
  - Deleted – all users cannot use it
  - Active – mark deprecated image as active again (command line only)

Sharing and moving images

- Share across projects
- Requires **Compute Engine Image User** role to host project
  - Example: User in Project A wants to use images from Project B
  - User in Project A must have Compute Engine Image User role granted for project B
  - Role grants access to all images in project
- For managed instance groups, Project A service account must be granted role to Project B

Export image to Cloud Storage

- Ideal for sharing with projects without host project access
- Export image as a tar.gz to Cloud Storage
- Linux only, not available for Windows
- Sharing with Image User Role is preferable

Return to Table of Contents

Choose a Lesson

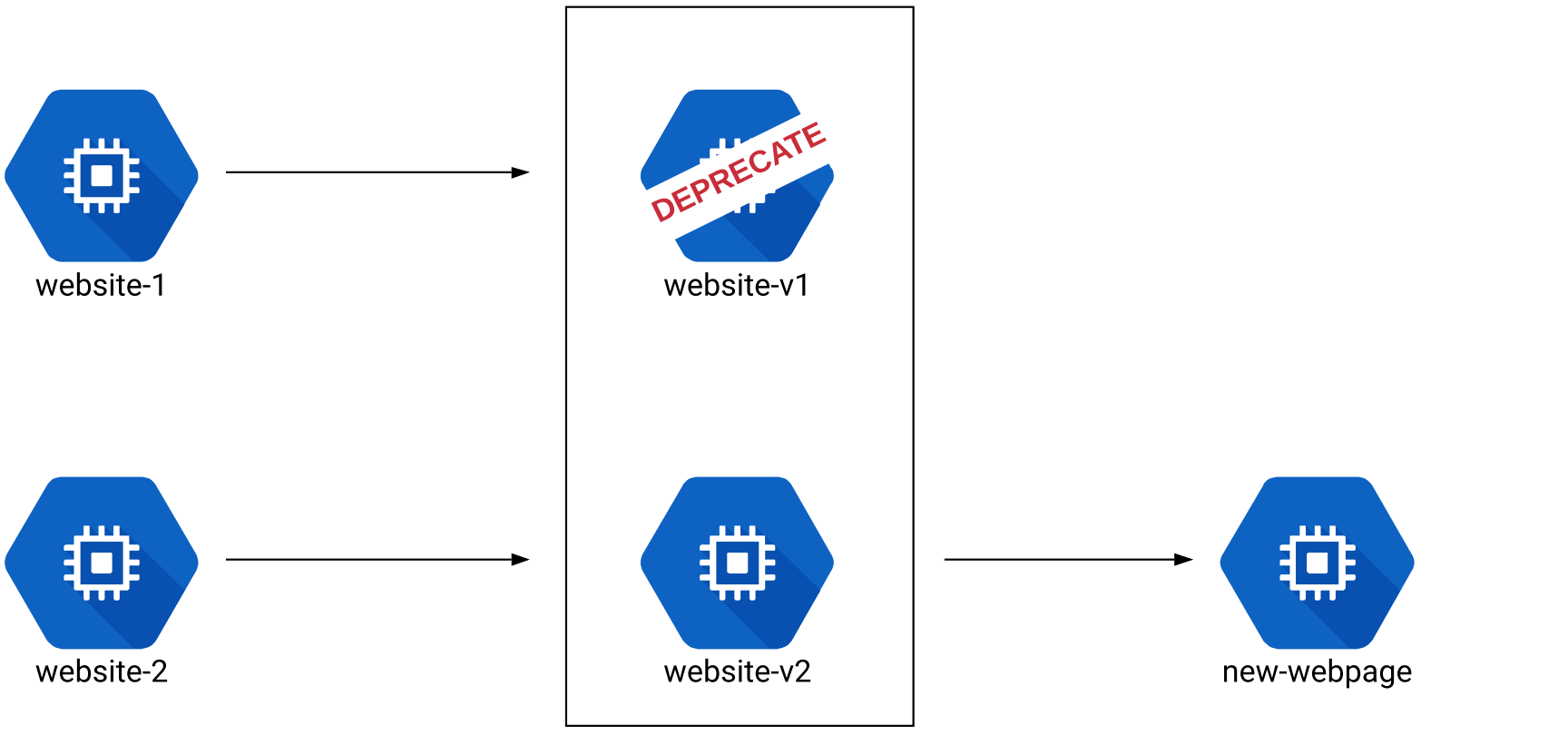
- Compute Engine Deep Dive
- Disks
- Images
- Snapshots
- Startup and Shutdown Scripts
- Preemptible VMs

Images

Previous

Hands-On Guideposts

- Start with two versions of web server
- Create custom image
  - Add to image family
  - View/use command line reference
- Create image via command line
  - `gcloud compute images create (image_name) --source-disk (disk_name) --source-disk-zone (zone) --family (image_family)`
- View image family info
  - `gcloud compute images describe-from-family (image_family)`
- Deprecate/set active image version
  - `gcloud compute images deprecate (image_name) --state (STATE)`
- Delete image
  - `gcloud compute images delete (image_name)`



Return to Table of Contents

Choose a Lesson

Compute Engine Deep Dive

Disks

Images

Snapshots

Startup and Shutdown Scripts

Preemptible VMs

Snapshots

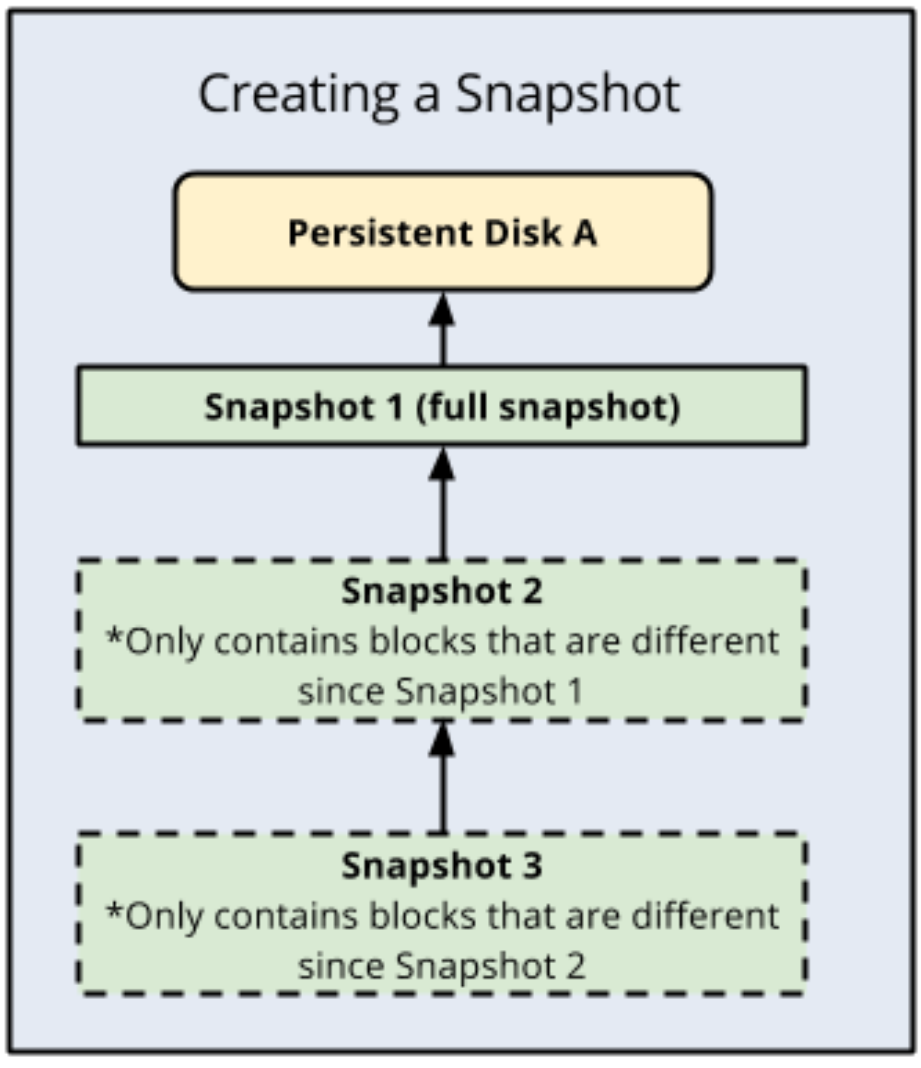
Next

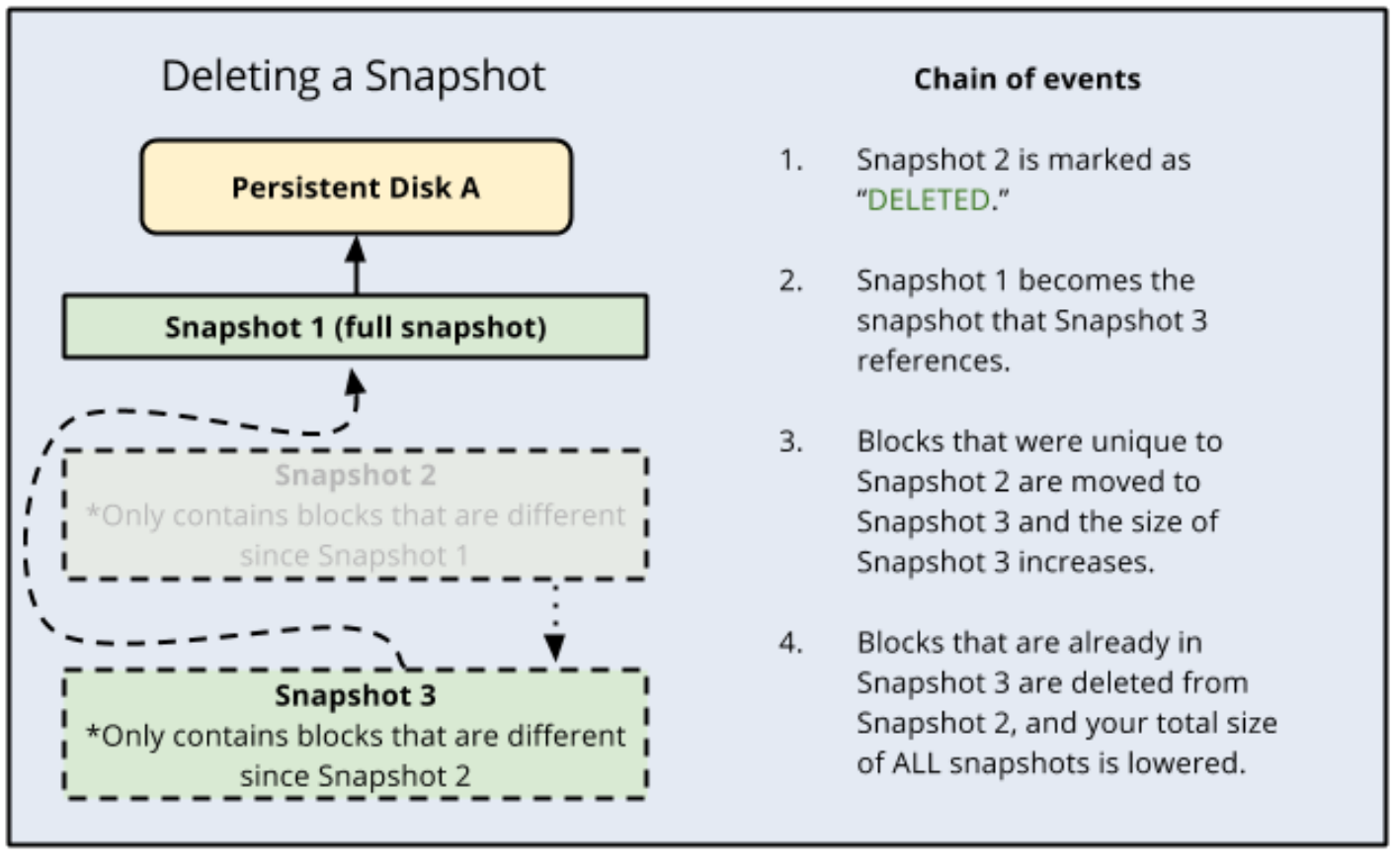
What are snapshots?

- Simply put: **Instance/disk backups**
- Periodic backups via a point in time snapshot of disk
  - Incremental backups
- Can create while instance is running
- Share across projects via command line only (may change in the future)
- Can create instance copies in new zones
- Can create snapshot of boot disk or attached disks

Incremental backups with Snapshots

- **First snapshot full disk copy**
- **Subsequent snapshots are only what's different compared to previous snapshot**
- **Snapshot restore combines previous snapshots to create 'whole disk'**





Return to Table of Contents

Choose a Lesson

Compute Engine Deep Dive

Disks

Images

Snapshots

Startup and Shutdown Scripts

Preemptible VMs

Snapshots

Previous

Snapshot best practices

- All about reducing activity when backing up
- Prepare disk for best consistency
  - Pause applications/processes that write data, then flush disk buffers
  - If possible, unmount disk completely
  - For Windows, use VSS Snapshots
  - Use ext4 for Linux
- Take only one snapshot at a time per disk
- Schedule during off-peak hours
- Use multiple persistent disks for large data volume
- Run fstrim before snapshot (Linux) to clean up space

Return to Table of Contents

Choose a Lesson

- Compute Engine Deep Dive
- Disks
- Images
- Snapshots
- Startup and Shutdown Scripts
- Preemptible VMs

Startup and Shutdown Scripts

Why is this important?

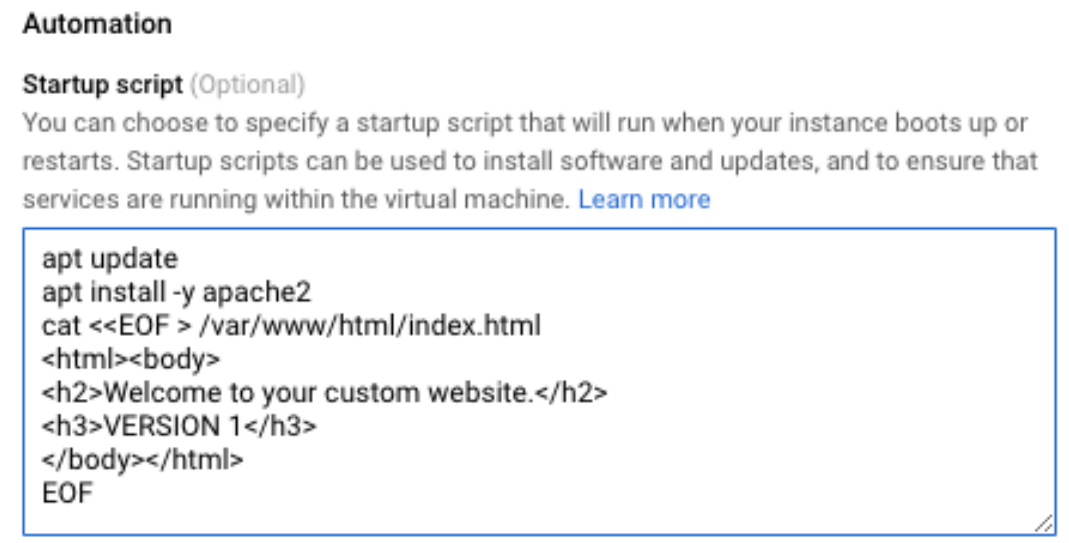
- Automation! Automation! Automation!
- Eases management of large number of VM's
- Automate software installation, updates, services, and much more
- Easily and programmatically customize VM's
- Key component in instance groups and scaling capabilities

Considerations

- Scripts always run as root/administrator
- Can run whatever script types OS recognizes (bash, Python, .bat files)
- Compute Engine will run the script verbatim, regardless of type

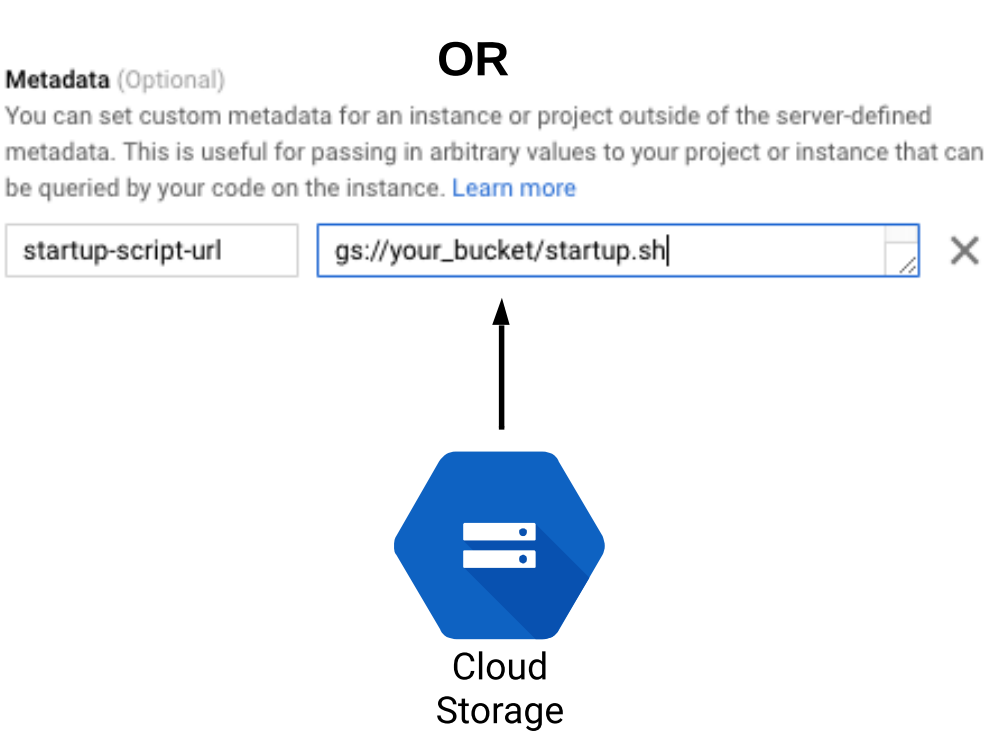
Input Methods

- Direct input
  - Paste in the script field in instance properties
- Link to script in Google Cloud Storage
  - Using metadata server URL
  - Very useful for large scale automation
  - Must have access to bucket/object



Shutdown Scripts

- Best-effort basis
- Run during shutdown period. May shut down before script completes.
- Great for managed instance group/autoscaler
- Example: copy processed data to Cloud Storage, back up logs
- Good to pair with preemptible instances



Return to Table of Contents

Choose a Lesson

- Compute Engine Deep Dive
- Disks
- Images
- Snapshots
- Startup and Shutdown Scripts
- Preemptible VMs

Preemptible VM's

Next

What is it?

- Short-lived, low-cost VM:
  - 24 hours max
  - Can be shut down at any time (30-second warning)
- 'Disposable' – not for critical single VM's
- Ideal for **fault-tolerant**, batch processing workloads:
  - Rendering
  - Media transcoding
  - Big data analytics
  - Hadoop and big data
- Most often used in managed instance groups:
  - 'Swarm' tactics
  - Google's docs: Throw more CPUs at it!
- Fixed pricing, up to 80% off regular instance price
- Compute (CPU/memory) is cheaper, Storage and licensing same cost
- Otherwise, exactly like any other VM

Preemptible VM's let you use MANY machines for batch computing at a vastly lower cost

220,000 cores and counting: MIT math professor breaks record for largest ever Compute Engine job

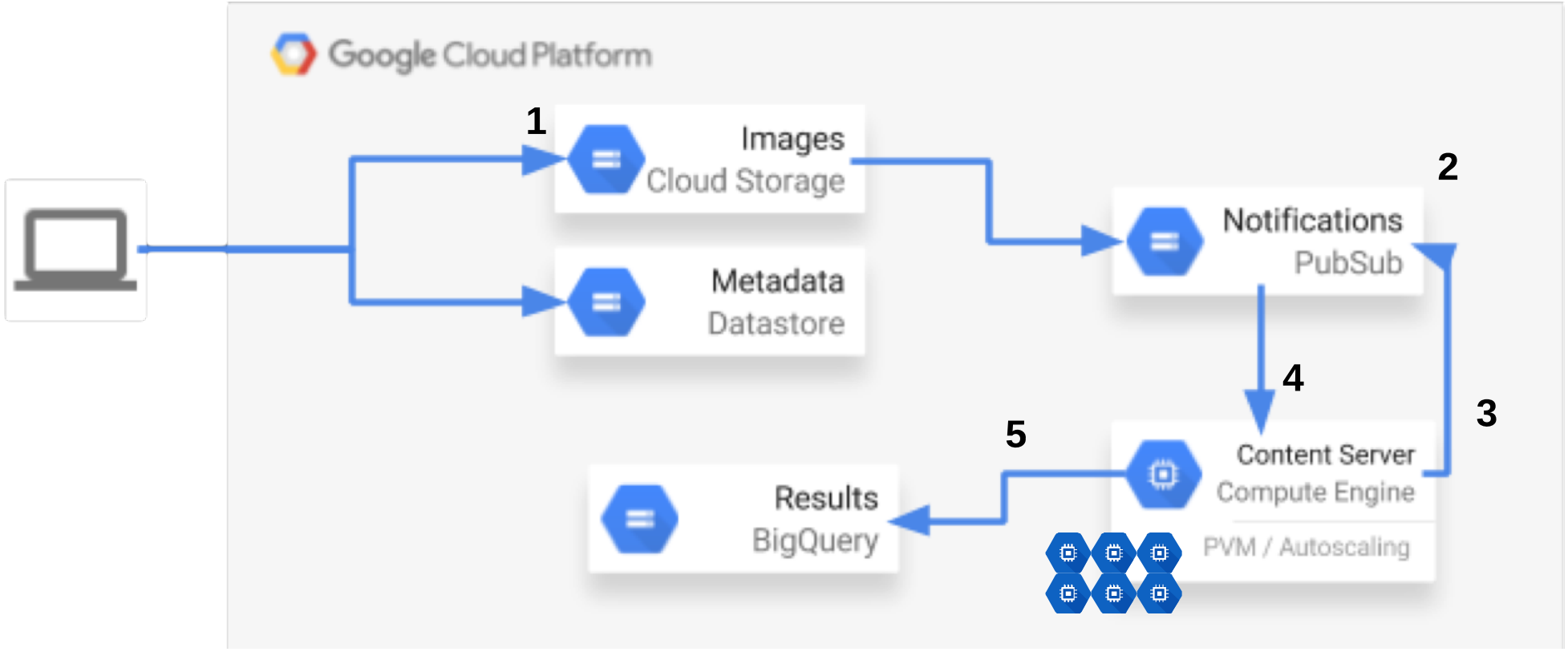
Thursday, April 20, 2017

By Alex Barrett, GCP Blog Editor & Michael Basilyan, Product Manager, Compute Engine

Editor's Note: This post was updated on June 12, 2017.

An MIT math professor recently broke the record for the largest ever Compute Engine cluster, with 220,000 cores on Preemptible VMs, the largest known high-performance computing cluster to ever run in the public cloud.

1. Satellite imagery loaded to Cloud Storage
2. Pub/Sub links images to Managed Instance Group
3. If individual machine finishes processing, notifies Pub/Sub and moves on
4. If machine preempted, Pub/Sub re-pushes load to new machine
5. Each machine outputs results to BigQuery for later analysis



Return to Table of Contents

Choose a Lesson

Compute Engine Deep Dive

Disks

Images

Snapshots

Startup and Shutdown Scripts

Preemptible VMs

Preemptible VM's

Previous

Best Practices (a.k.a "How not to get shut down")

- Use smaller machine types:
  - Many small machines are better than fewer large
  - Less likely to be shut down
- Run jobs during off peak times:
  - Nights and weekends
- Design application for fault/preemption tolerance:
  - Test by manually stopping the instance
  - 'Embarrassingly parallel' operations
- Preserve disk on machine termination (individual instances)
- Use shutdown scripts:
  - Save job progress to pick up where left off

Exam scenarios

- Create and terminate machine to save costs, but preserve disk state.
  - --no-auto-delete --disk example-disk
- Managed instance group with PVM's keep recreating every minute
  - Health check/firewall configuration

Return to Table of Contents

Choose a Lesson

Return to Table of Contents

Choose a Lesson

- Force Multipliers Scalable Computing
- Load Balancers
- Instance Groups and Autoscaling
- Hands-On Guideposts

Force Multipliers Scalable Computing

Now for the fun stuff....

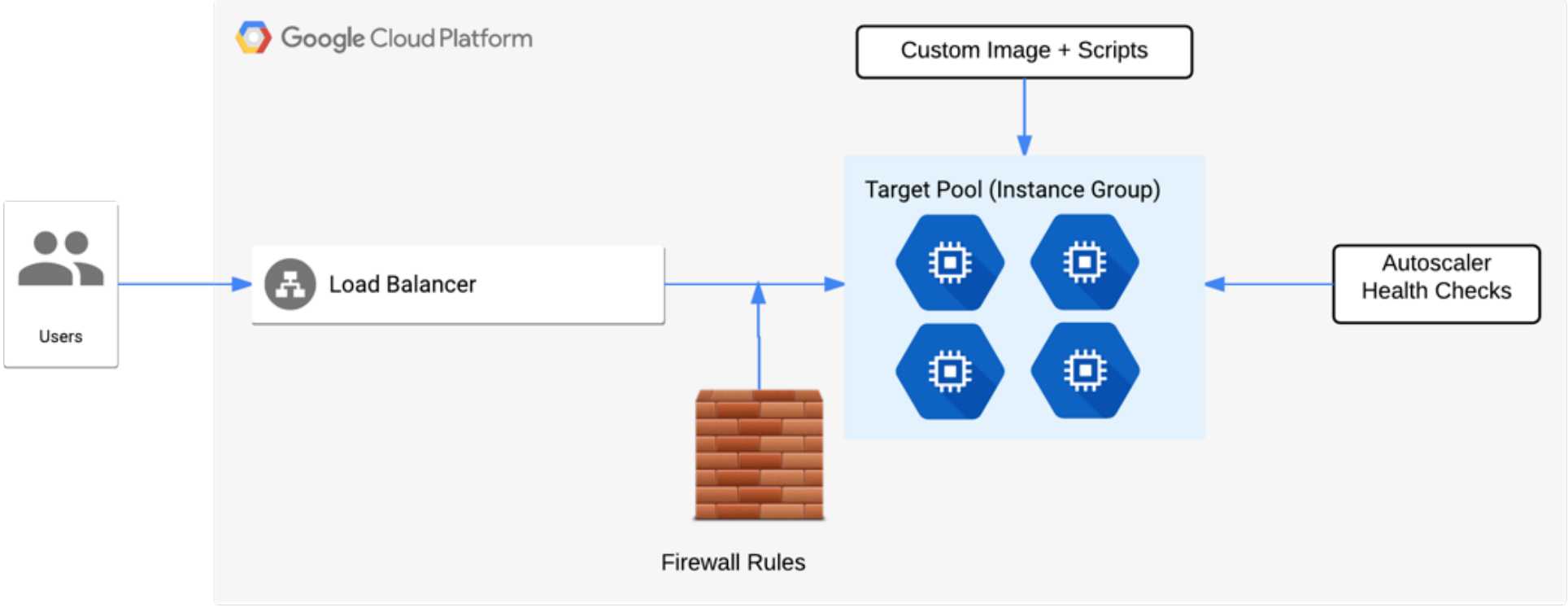
- So far, we've focused on individual VM's
  - Backing up, custom images, etc
- Next: take previous concepts, and apply to managing **many** VM's
- Using force multipliers for massive scalability, full automation

Automate all the things! Scaling and deploying

- Repeatable and documented
  - Easily re-deploy resources if needed
- Scalable
  - Grow and shrink as needed based on demand
- Necessary for large infrastructure
- Reduce complexity
- Fun fact: Google Cloud Labs use automation to create environments

Putting it all together

- Use previous concepts
  - Firewall rules, scripts, custom images, etc
- Combine with scalable components
  - Load Balancers, Instance Groups, Autoscaling



In this section

- Three components – one purpose
- Load Balancer
  - Layer 4 vs. Layer 7
- Instance Groups
- Autoscaling
- Go over concepts on each one individually, then combine them
  - Together, they are incredibly powerful
- Plenty of hands on demos

Return to Table of Contents

Choose a Lesson

- Force Multipliers Scalable Computing
- Load Balancers
- Instance Groups and Autoscaling
- Hands-On Guideposts

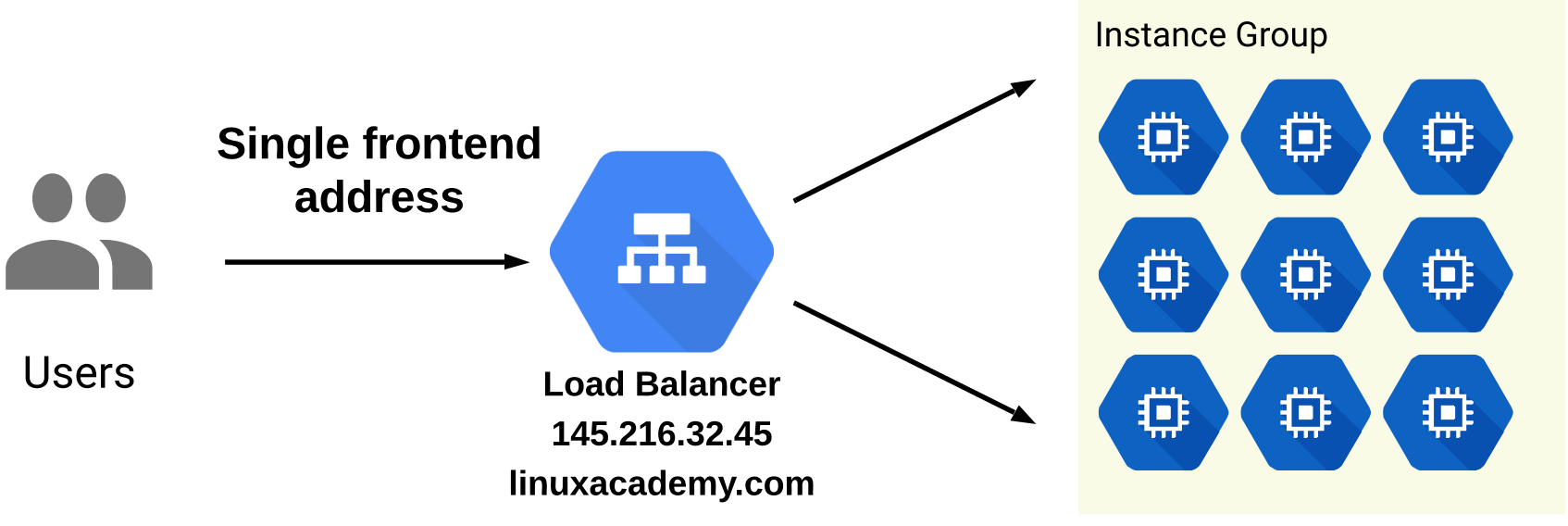
Load Balancers

Next

What is a load balancer?

- Distributes (balances) user network requests among a pool of instances
- Single frontend point of access – multiple backend targets to serve traffic
- Software Defined – not physical
- Necessary for distributed applications
- Global or regional in scope - depends on load balancer type
- Backend traffic subject to firewall rules
- Big picture overview
  - https://cloud.google.com/load-balancing/docs/load-balancing-overview

Simple Example



Load Balancer Types

- HTTP(S) - Layer 7 LB
- SSL Proxy
- TCP Proxy
- Network Load Balancer
- Internal Load Balancer

Load Balancer differences

- Global (multi-regional) vs. Regional
- External vs. Internal
- HTTP vs. TCP/UDP

Multiple backend addresses

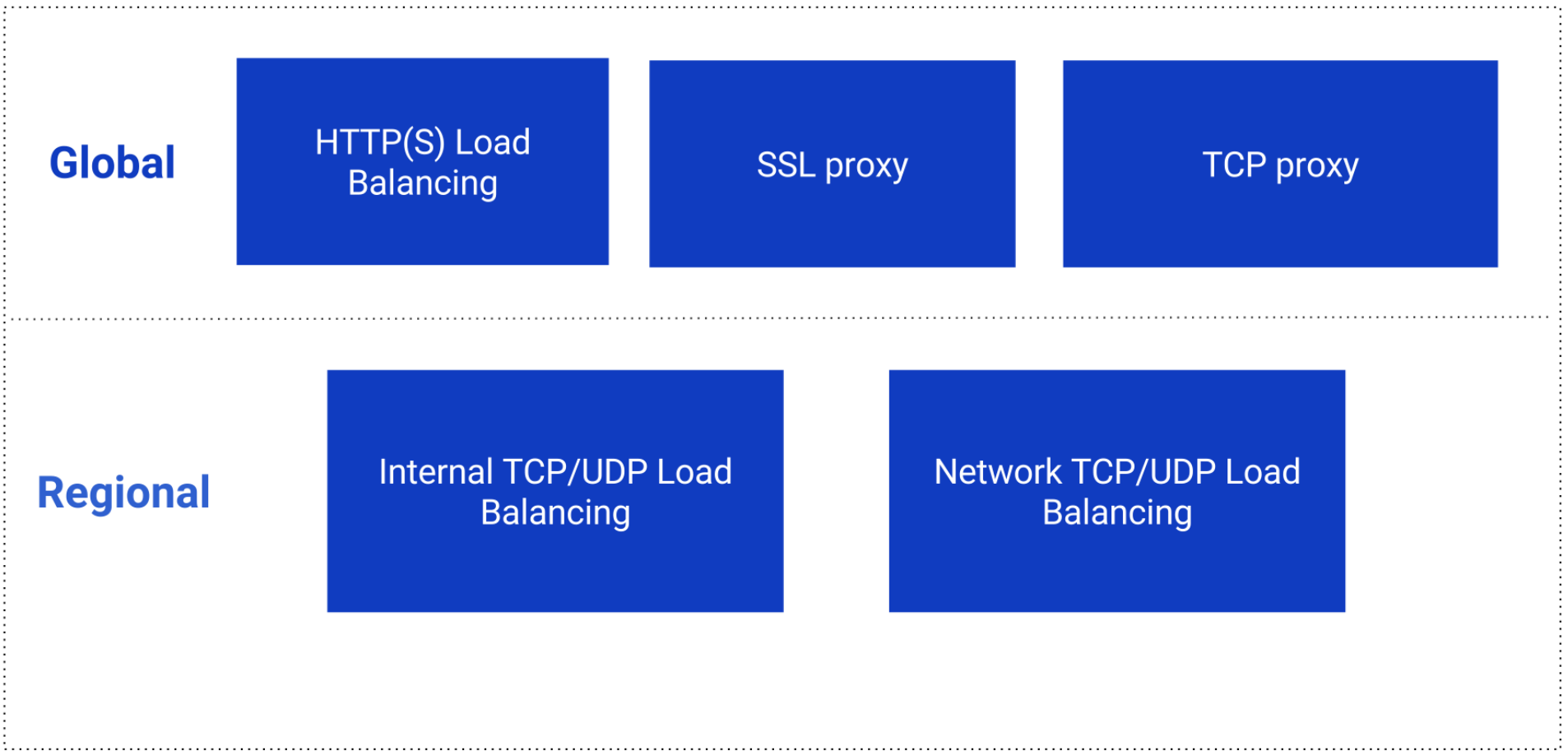
Return to Table of Contents

Choose a Lesson
- Force Multipliers Scalable Computing
- Load Balancers
- Instance Groups and Autoscaling
- Hands-On Guideposts

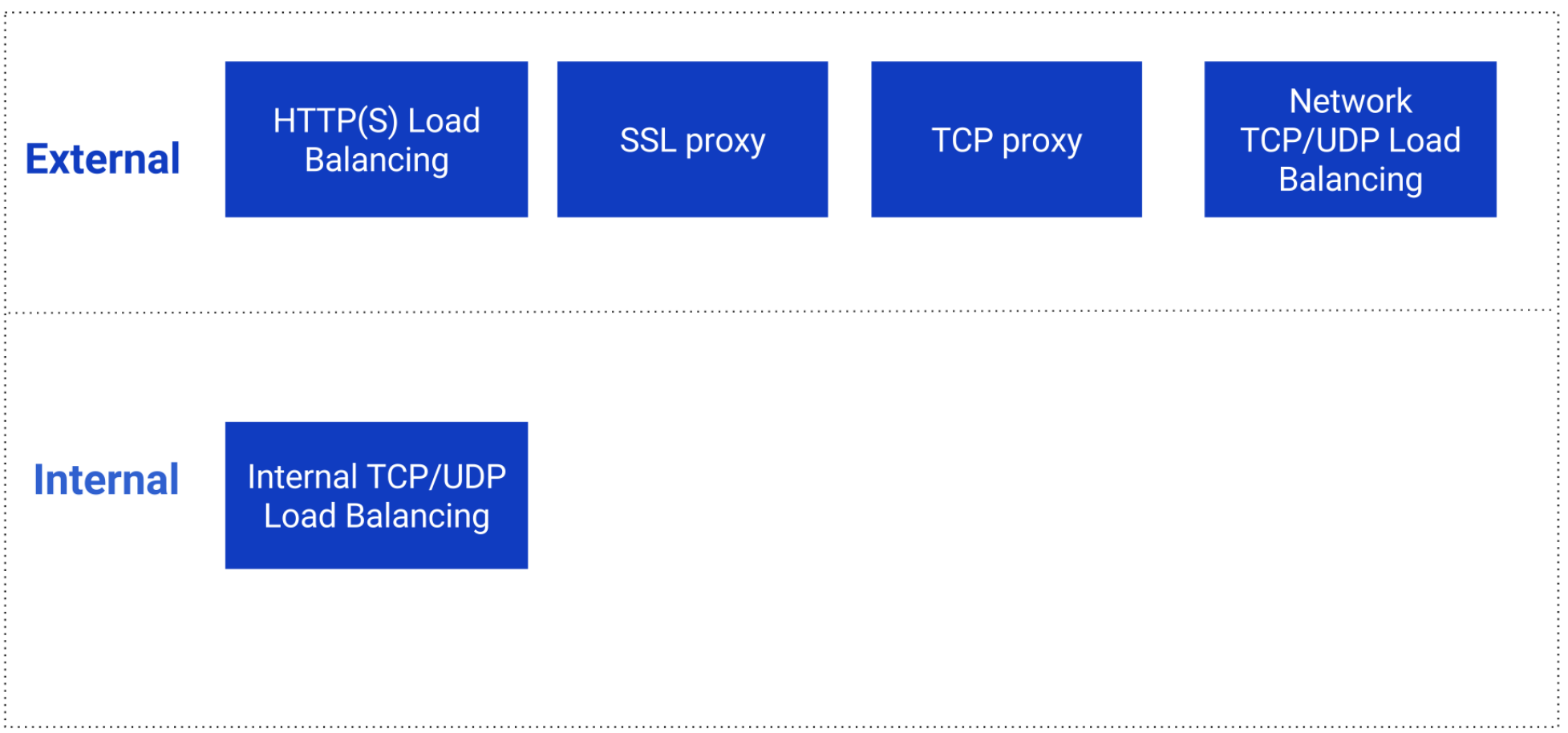
Load Balancers

Previous
Next

Global vs. Regional



External vs. Internal



Return to Table of Contents

Choose a Lesson

- Force Multipliers Scalable Computing
- Load Balancers
- Instance Groups and Autoscaling
- Hands-On Guideposts

Load Balancers

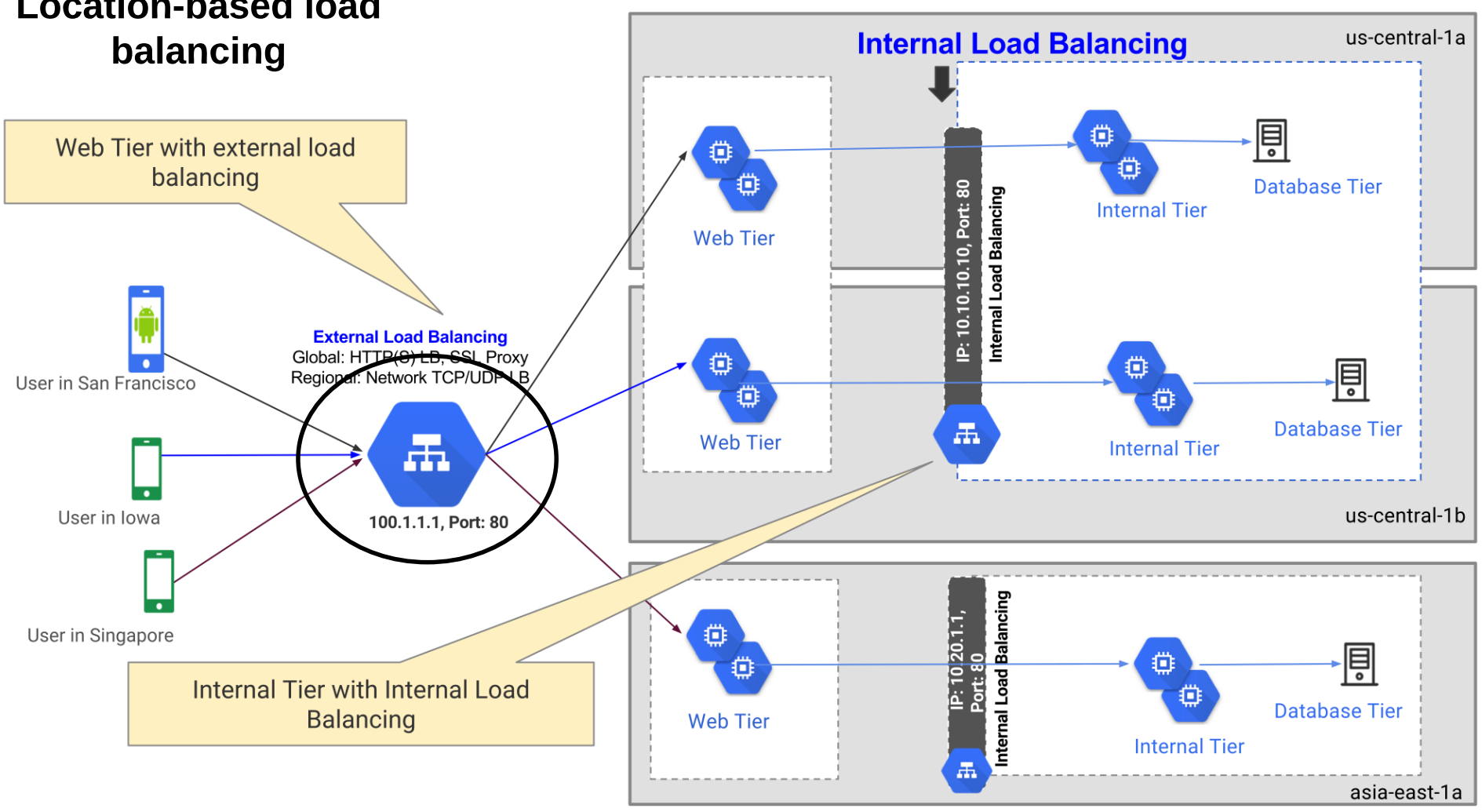
Previous

Next

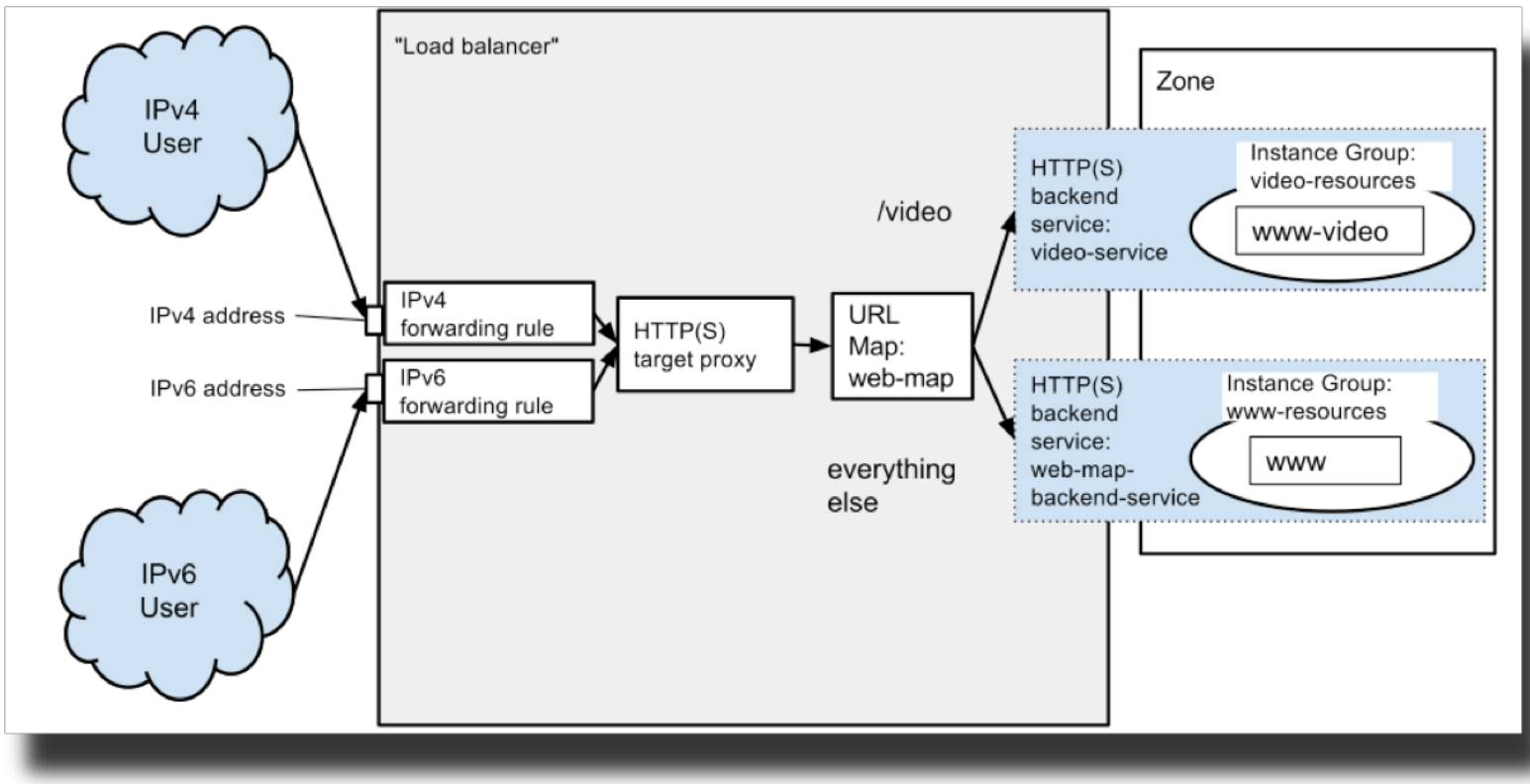
HTTP(S) Load Balancer

- Manages HTTP(S) requests
  - Layer 7 load balancing
- Global scope
  - Distribute traffic to multiple regions
- IPv4 and IPv6 – IPv6 terminated at LB then proxy by IPv4 to backend
- Distribute traffic by location or content requested
  - Forwarding rule – forwards traffic to target pool by matched criteria (location, content)
  - Forwards to target pool
- Paired with instance group for backend
- Native support for websocket protocol

Location-based load balancing



Content-based load balancing



Return to Table of Contents

Choose a Lesson

- Force Multipliers Scalable Computing
- Load Balancers
- Instance Groups and Autoscaling
- Hands-On Guideposts

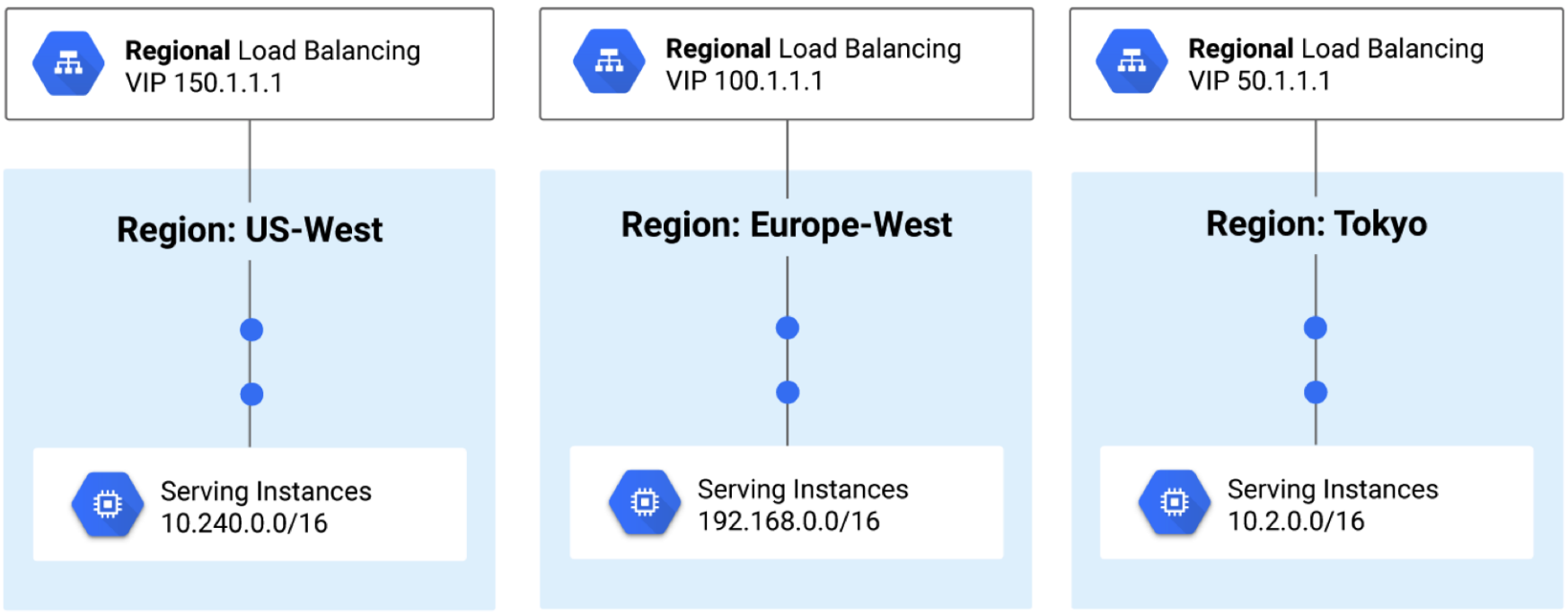
Load Balancers

Previous

Next

Network (External) Load Balancer

- Regional/External
- TCP/UDP traffic (Layer 4 of OSI model)
- "Network" is key differentiating term (vs. HTTPS load balancer)
- Balance requests by IP protocol data (address, port, protocol type)
- How it works:
  - Forwarding Rules – matched criteria = address, protocol, port range
  - Target Pool – group of VM's (usually an instance group)



Network (Internal) Load Balancer

- Regional/Internal
- Private load balancer within same VPC - operate over internal IP addresses
- Same option as Network Load Balancer, with option for internal only
- Often used with multi-tier application – nested LB's
- Affects Cloud Router dynamic routing

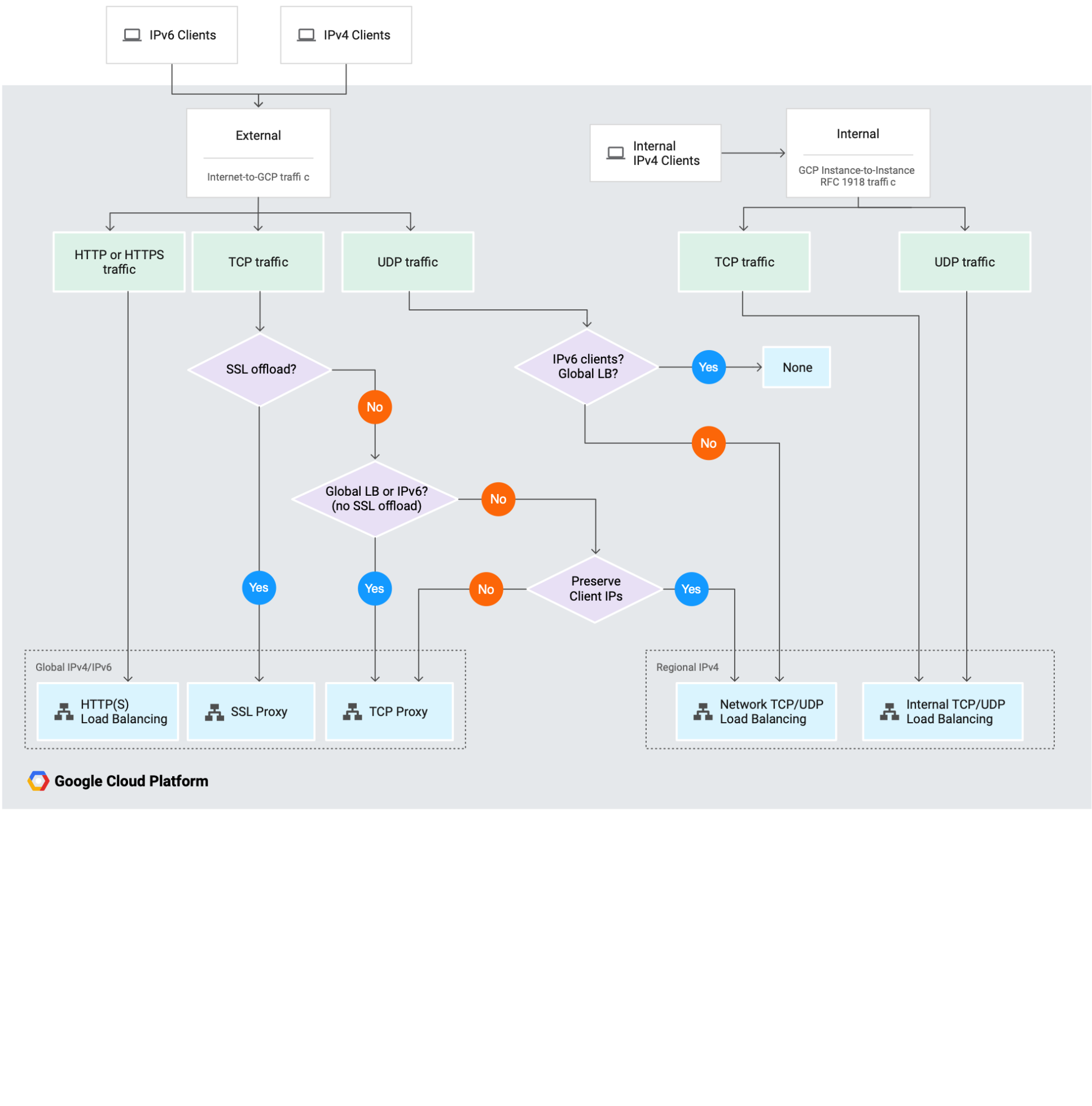
Return to Table of Contents

Load Balancers

Previous

Choose a Lesson

Flowchart Time!



Return to Table of Contents

Choose a Lesson

Force Multipliers Scalable Computing

Load Balancers

Instance Groups and Autoscaling

Hands-On Guideposts

Instance Groups and Autoscaling

Next

What are instance groups?

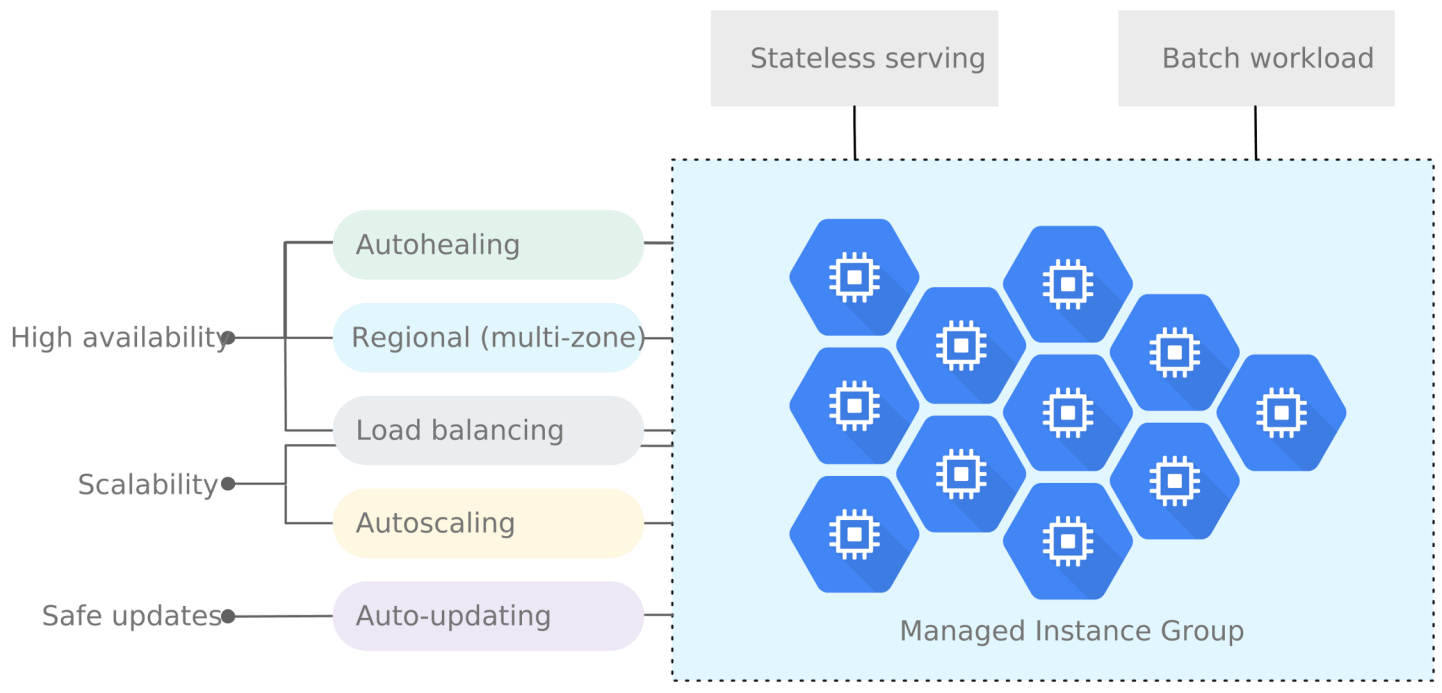
- Group of instances
- Manage as a group, not one at a time
- Managed and unmanaged varieties
  - Unmanaged ideal for migrating existing configurations for load balancing tasks with minimal modification - and that's about it
  - Managed preferred, and what we'll cover

Key Concept: importance of individual servers

- Change in management approach
- Pets vs. Livestock
- Pets - individual, high value servers - lovingly hand crafted
  - Indispensable, can never have downtime
  - Examples: domain controller, mail server, database server
  - If single server goes down, it's a very bad thing
  - Server is individually backed up - backups are very important (snapshots)
  - Limited scalability (vertical scaling)
- Livestock
  - Group (or "herd") of servers, built en masse using automated tools
  - Individual servers not as important - disposable
  - Designed with failure in mind - can be replaced
  - Individual backups not important - don't use snapshots on instance group VM's
  - Much greater scalability - horizontal scalability
    - Highly available (multiple regions) and scalable (increase as load increases)
  - Examples: managed instance groups, Kubernetes, App Engine, Hadoop/MapReduce
  - **Stateless** workloads - don't need to save state on machine

Features of managed instance groups

- Automatically scale
- Work with load balancers
- Health checks – auto-healing groups



Return to Table of Contents

Choose a Lesson

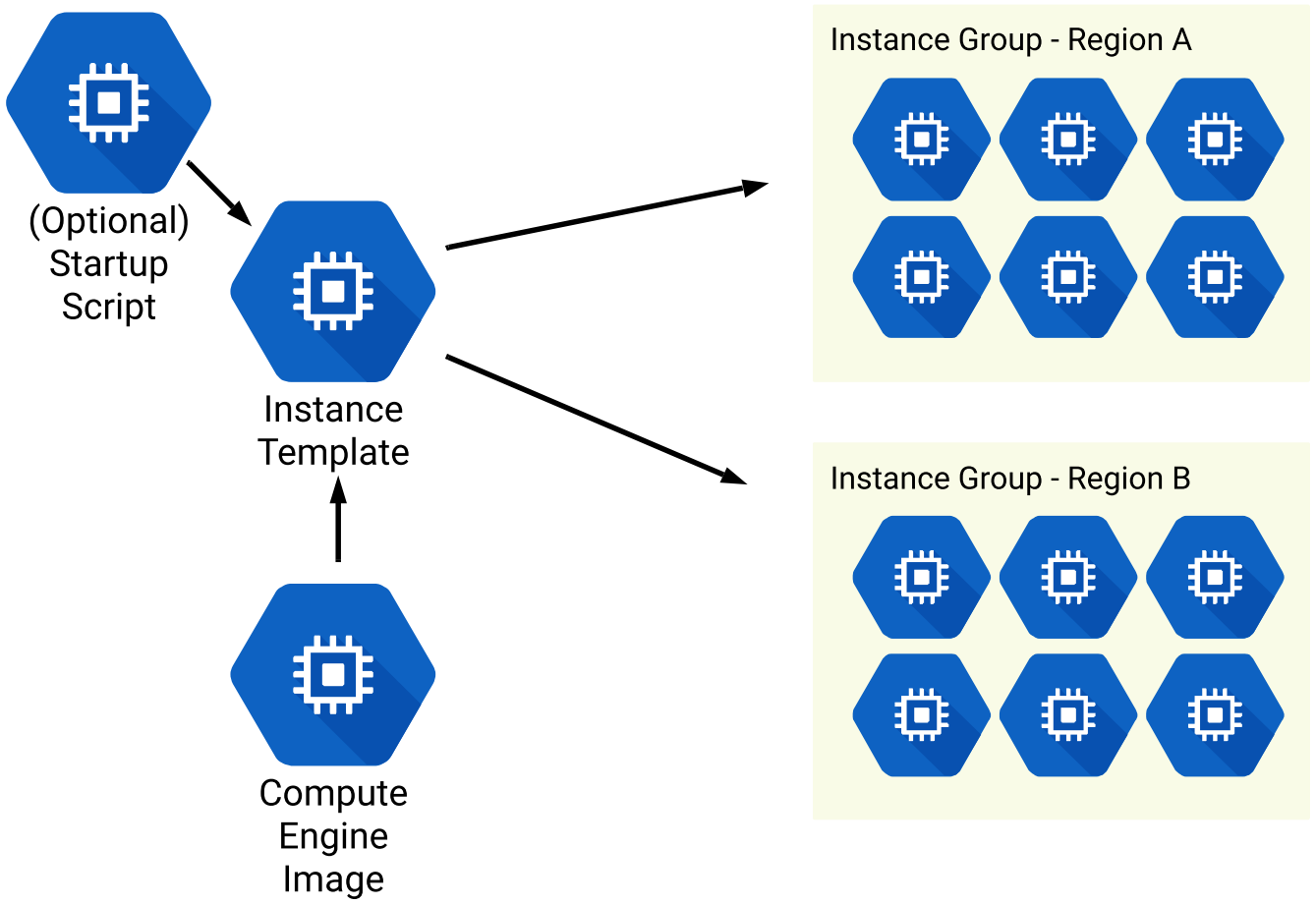
- Force Multipliers Scalable Computing
- Load Balancers
- Instance Groups and Autoscaling
- Hands-On Guideposts

Instance Groups and Autoscaling

Previous
Next

How it works

- Short version (managed instance groups):
  a. Create instance template
  b. Create instance group from instance template
- Instance Template
  - Defines group configuration
  - Machine type, zone, image, scripts
  - Re-usable for multiple group configurations
  - Instance template = Global – not region bound
  - Can specify zonal resources (i.e. read only disk), which effectively binds it
- From template – create managed instance group
  - Instance group = Regional, can use more than one zone in single group



Use startup script or custom image (or both)?

- Startup script = easy to create quick changes
  - However, wait longer until ready
- Images = more involved setup process
  - However, VM is ready to go right away
- Use case for both, though exam tests one or the other

Return to Table of Contents

Choose a Lesson

- Force Multipliers Scalable Computing
- Load Balancers
- Instance Groups and Autoscaling
- Hands-On Guideposts

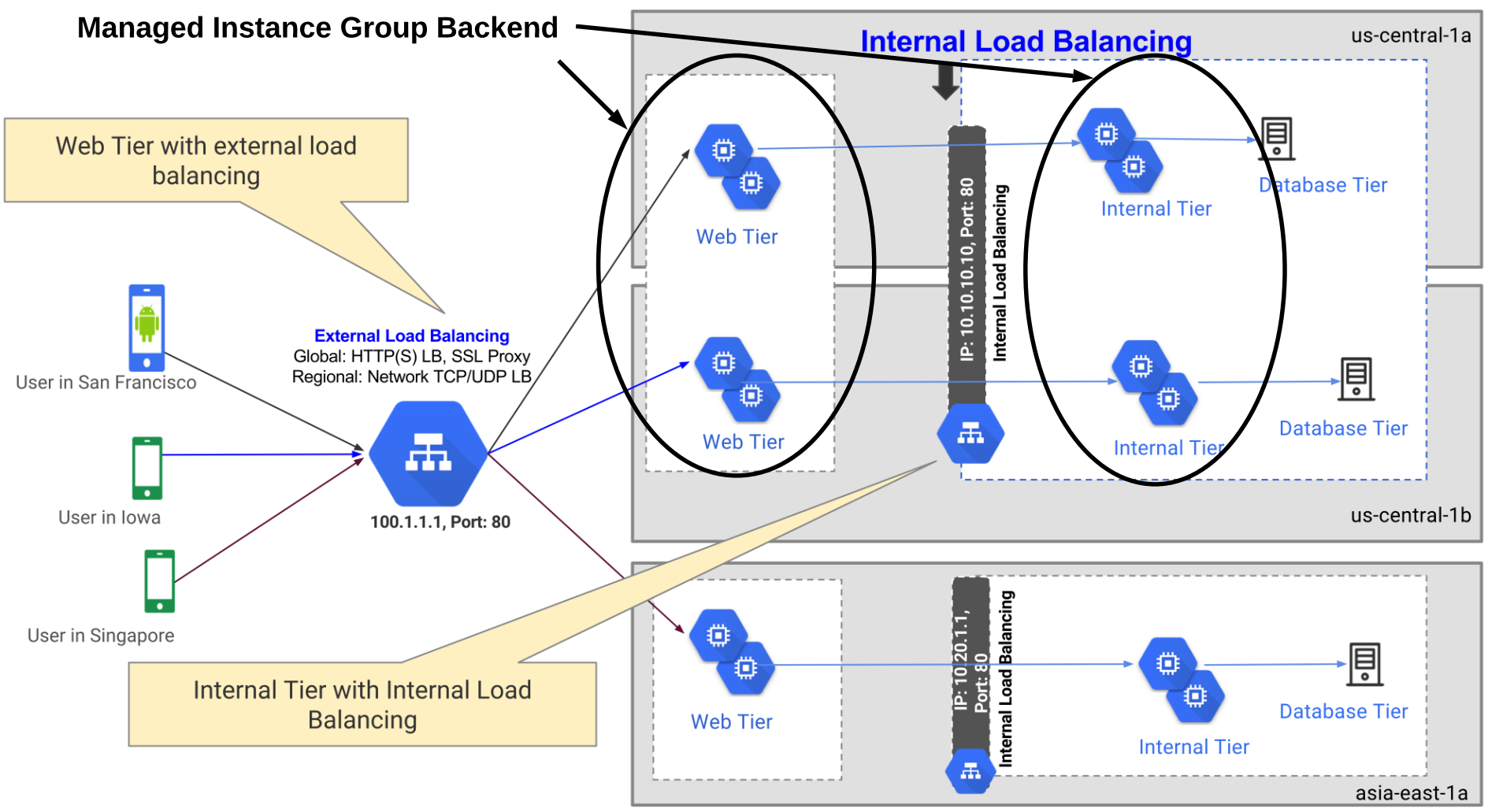
Instance Groups and Autoscaling

Previous

Next

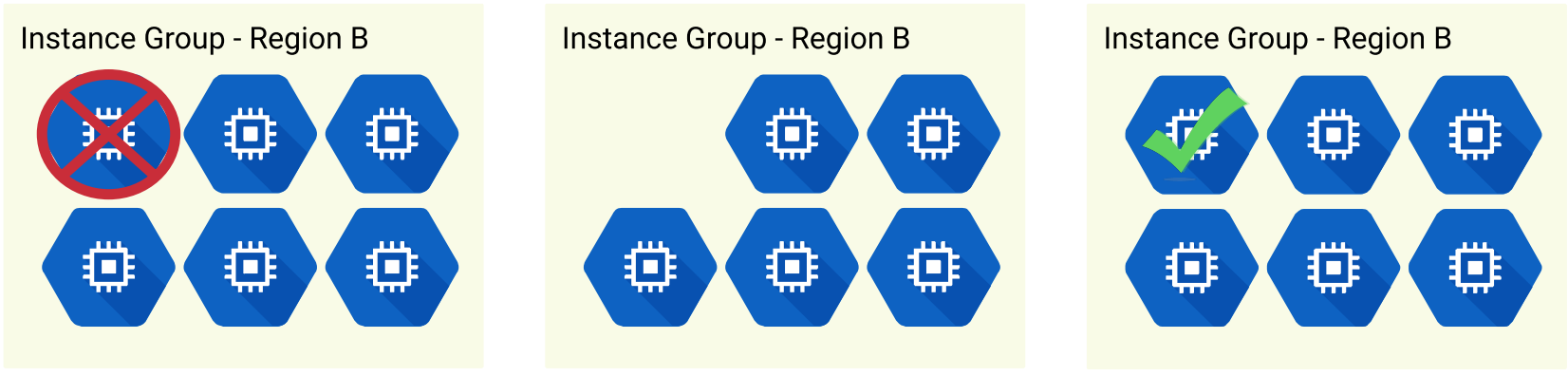
Networking

- Paired with load balancers - distribute traffic across instances in group (or even multiple groups)
  - Load balancers must be assigned to a backend – target pool or instance group
  - HTTP Load balancers must use instance group
  - Load balancer contains one or backend service
  - Backend service links to one or more backends
  - Backend links to one instance group
  - Backend service knows which backends to use – directs traffic
- Subject to firewall rules for allowed traffic
  - Rules applied to instances, not load balancer



Health Checks

- Auto-healing
- If an instance fails or service fails – delete and recreate identical instance
- Managed instance groups only



Return to Table of Contents

Choose a Lesson

- Force Multipliers Scalable Computing
- Load Balancers
- Instance Groups and Autoscaling
- Hands-On Guideposts

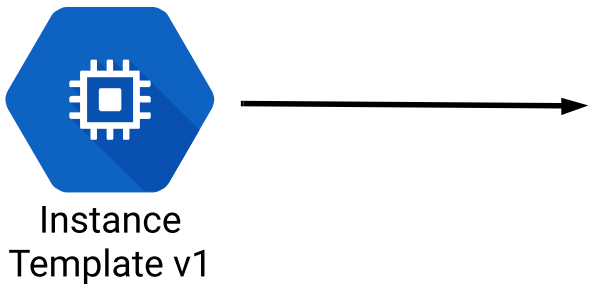
Instance Groups and Autoscaling

Previous

Updating managed instance groups

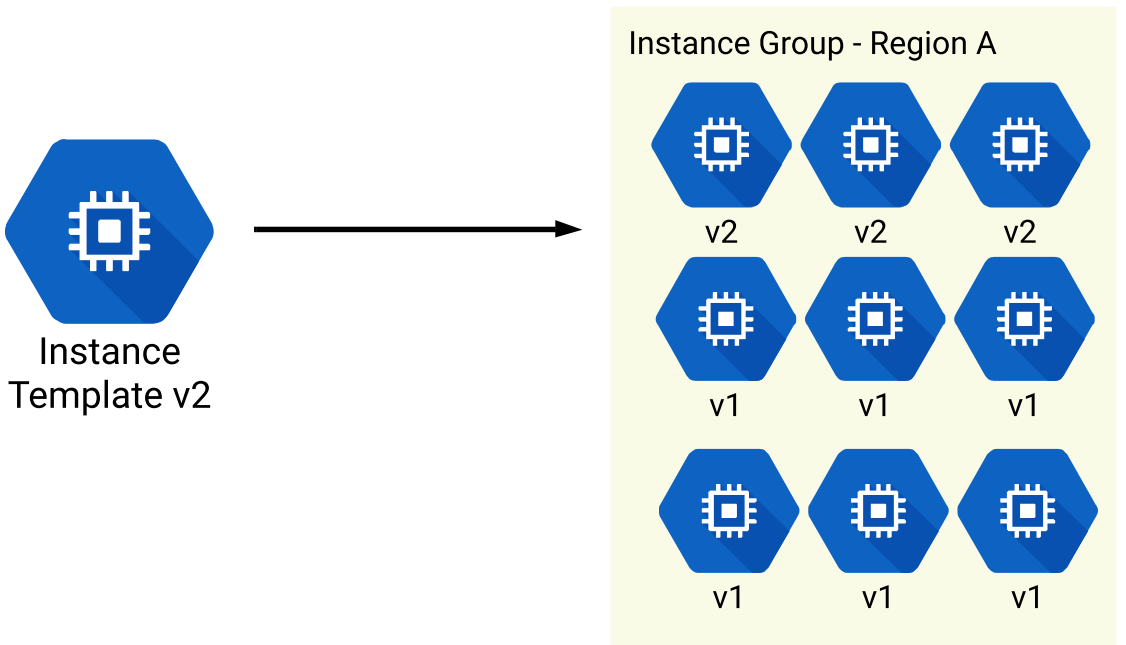
- Short version: replace one machine with a different machine
- Managed Instance Group Updater
  - Update entire group – not just individual machines
  - Deploy new versions of software
  - Control pace of update rollout
    - Rollout happens automatically
    - Can do partial rollouts for canary testing
  - Deploy inside existing managed instance group

Instance Template v1 deploys first version of instance group

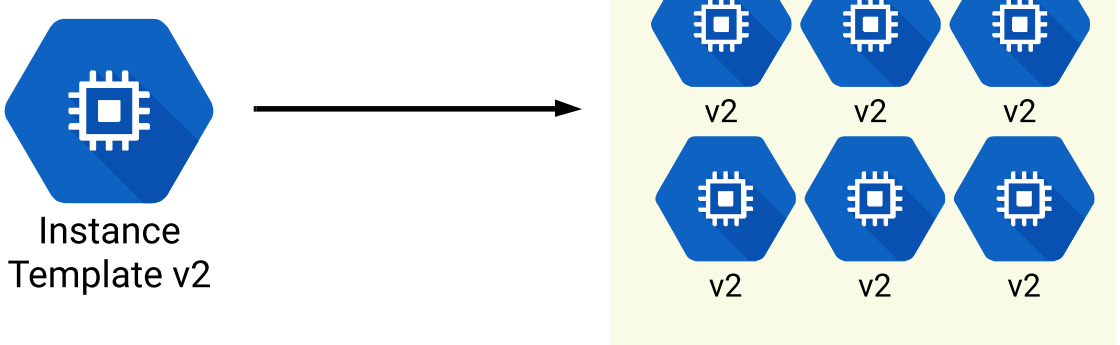


NEW Instance Template v2 rolls out 'v2' machines to replace v1.

Staged replacement to avoid downtime



Instance Template v2 eventually replaces all machines in group



Return to Table of Contents

Choose a Lesson

Force Multipliers Scalable Computing

Load Balancers

Instance Groups and Autoscaling

Hands-On Guideposts

Hands-On Guideposts

What are we doing?

- Start with two single-VM versions of website + stress tester
- Create custom images
- Create instance template using version 1 image
- Create instance group
  - Setup autoscaling and health checks
- Create load balancer
  - Add instance group as backend
- Update instance group with 'version 2' using new instance template
- Force autoscaling via stress test

Return to Table of Contents

Choose a Lesson

Return to Table of Contents

Choose a Lesson

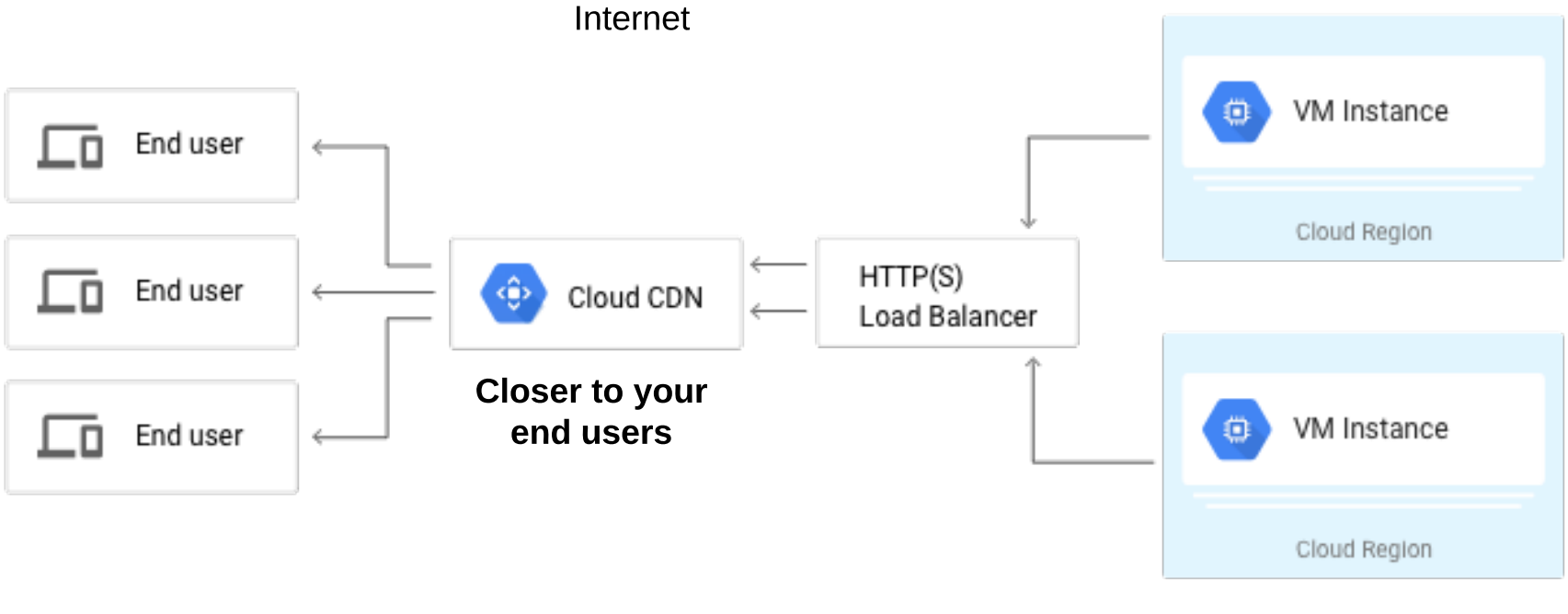
Cloud CDN Concepts

Cloud CDN Concepts

Next

What is it?

- Short for **Content Delivery Network**
- One sentence answer: CDN caches website and application content closer to your users for better performance
- Low cost, low latency content delivery using Google's global network
  - Remember the power of the network? Same thing here
- The closer your data is to the end user, the better the performance and user experience
- CDN caches website/application data closer to user via Cloud CDN locations
  - Often referred to as points of presence (POPs)
  - Currently over 90 locations worldwide
  - Hop off point between Google's private network and the public Internet



Return to Table of Contents

Choose a Lesson

Cloud CDN Concepts

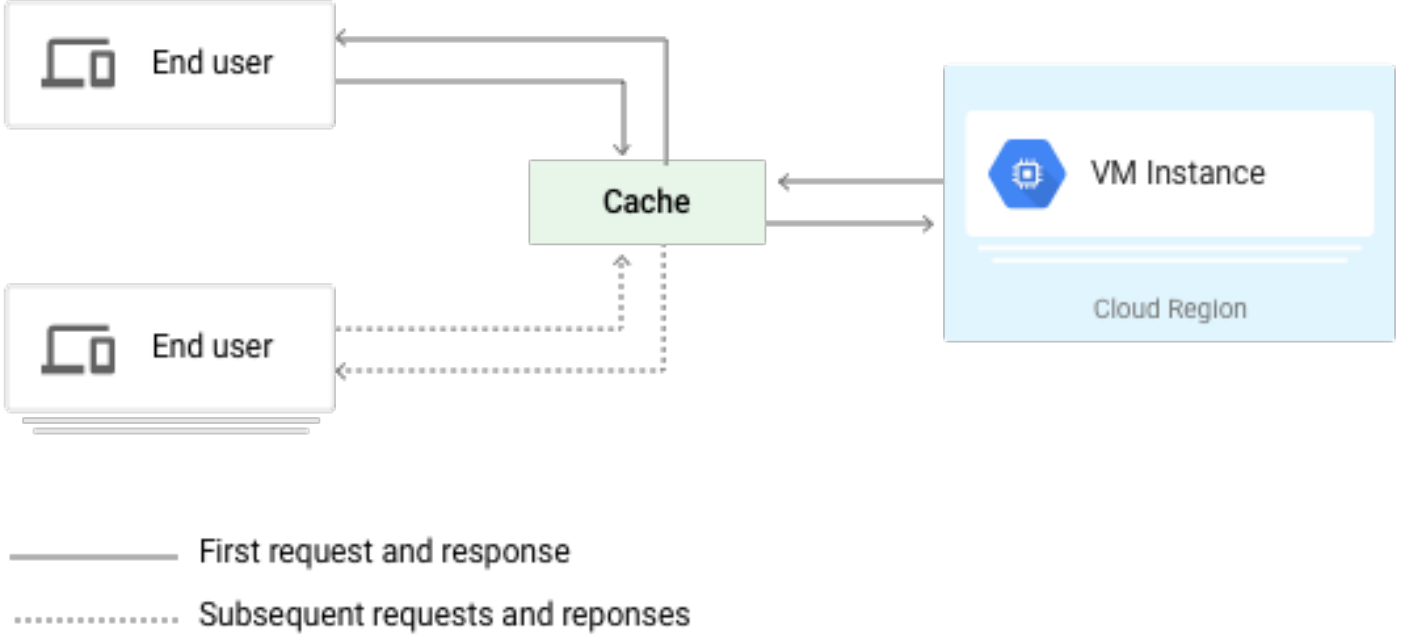
Cloud CDN Concepts

Previous

Next

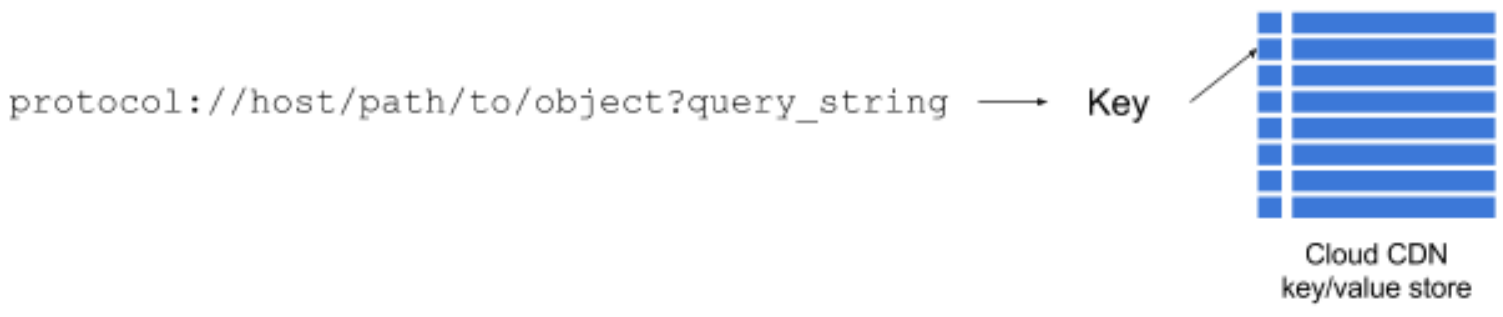
How it works

- Option on HTTP(S) load balancers
- Caches contents from backends (instance groups, Cloud Storage buckets)
- The first time content is requested, it is pulled from backend and cached at a CDN location (edge location)
  - Usually MUCH closer than the datacenter where backend resides
  - Later requests from location will serve from CDN location, not backend



Terminology

- Cache miss - First time content request not at cache location
- Cache fill - after requested for first time, cache loaded by either backend or another nearby cache
  - Subsequent requests will pull from cache location
- Cache key - Identifier for cached content
  - By default, in form of complete URL (https://linuxacademy.com/images/logo.jpg)
  - Must have exact match



Return to Table of Contents

Choose a Lesson

Cloud CDN Concepts

Cloud CDN Concepts

Previous

Improving Cache Key hit ratio

- By default, uses entire URL
- Optimize **cache hit ratio** for better performance/scalability
  - Example- same company logo on different domains
- How to customize
  - Remove unneeded URL aspects
    - Company domain
    - Protocol (HTTP/HTTPS)
    - Part of query string

Return to Table of Contents

Choose a Lesson

Return to Table of Contents

Choose a Lesson

Cloud Deployment Manager Concepts

Hands On Deployment Manager

Cloud Deployment Manager Concepts

Next

What is Cloud Deployment Manager?

- Infrastructure deployment service
  - Automates creation/management of GCP resources
  - Create and manage resources with configuration files and templates

Why is it important?

- As infrastructure grows in size and complexity, so does the chance of human error
- Standardized and repeatable
  - Create resources over and over with repeatable results
  - Highly structured templates and configuration
  - Document infrastructure in easy to understand format
- Used by GCP Marketplace to create easy, one-click deployments

How it works

- Deploy with command line only
- Infrastructure as code
  - Calls on API resources
- Configuration file – YAML format
  - Lists each resource to create and its properties
  - Contains resources section followed by list of resources
  - Resource components
    - Name – user-defined string to identify (my-deployment-project)
    - Type – type of resource to deploy (compute.v1.instance, compute.v1.disk)
    - Properties – resource parameters (zone: us-central1, boot: true)

Resource Type
appengine.v1.version
appengine.v1beta4.version
appengine.v1beta5.version
bigquery.v2.dataset
bigquery.v2.table
bigtableadmin.v2.instance
bigtableadmin.v2.instance.table
cloudfunctions.v1beta2.function
cloudresourcemanager.v1.project
clouduseraccounts.beta.group
clouduseraccounts.beta.user
compute.beta.address
compute.beta.autoscaler
compute.beta.backendBucket
compute.beta.backendService
compute.beta.disk

Return to Table of Contents

Choose a Lesson

Cloud Deployment Manager Concepts

Hands On Deployment Manager

Cloud Deployment Manager Concepts

Previous

Templates

- Configuration file can contain templates
- Separate configurations into smaller chunks
  - Update and re-use
- Python or Jinja2 format
- Advantages:
  - Easier to manage and maintain
  - Reusable
  - Keep consistent definitions in one place

Manifest

- Read only output of final configuration
- Includes configuration YAML, imported templates, expanded resource list
- When troubleshooting, consult the manifest

Exam Perspective

- Cloud Engineer exam goes into much more depth
- Architect exam is more conceptual, know that it is an infrastructure as code product for automating resource deployment
- Same type of service as Terraform, Ansible, etc

Return to Table of Contents

Hands-On - Deployment Manager

Choose a Lesson

Cloud Deployment Manager Concepts

Hands On Deployment Manager

Hands-on Guideposts

- Review configuration file
- Create a deployment
  - `gcloud deployment-manager deployments create (deployment_name) --config (config_file.yaml)`
- View manifest
- Delete deployment (web console or command line)
  - `gcloud deployment-manager deployments delete (deployment_name)`
- Preview a configuration without actually deploying it
  - See what it creates without actually deploying it
  - Same as regular created deployment, but add --preview as an option
- Deploying a previewed deployment
  - `gcloud deployment-manager deployments update (deployment_name)`
- View and deploy template
  - Review template deployment manifest

Return to Table of Contents

Choose a Lesson

Return to Table of Contents

Choose a Lesson

Where Should I Run My Code?

Where Should I Run My Code?

Next

Big Picture Compute Options

- Method for hosting applications/code on GCP
- Where should I run my code? It depends

Compute Options

- Google Compute Engine
- Google Kubernetes Engine
- Google App Engine
- Google Cloud Functions

Each option can take advantage of the rest of GCP services, such as:

- Storage
- Networking
- Big Data
- Security

Why do we care?

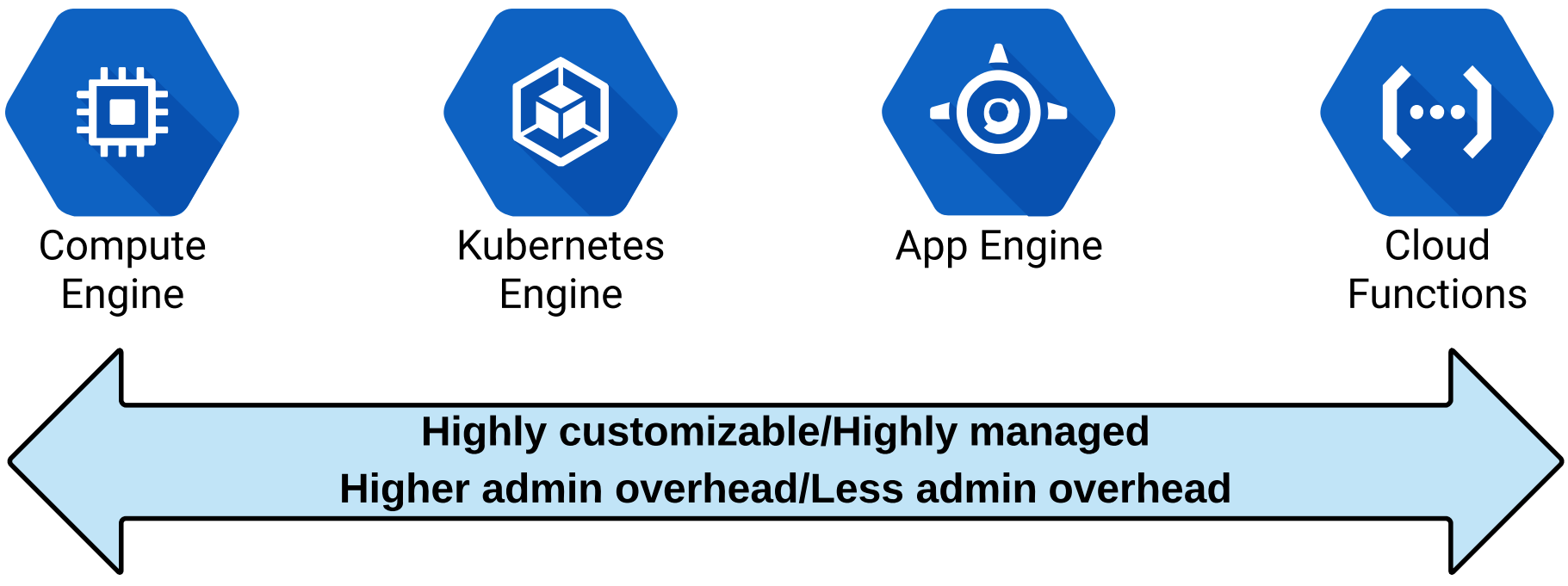
- Role of the Google Cloud Architect:
  - Given a set of business and technical requirements, choose and implement the right compute option tool for the task

Considerations

- What you want
- What we need to think about
- What Google handles
- Constraints
- Use cases

Big Picture Sliding scale between full control and highly managed

The higher the flexibility, the more 'stuff' that you have to think about



Return to Table of Contents

Choose a Lesson

Where Should I Run My Code?

Previous

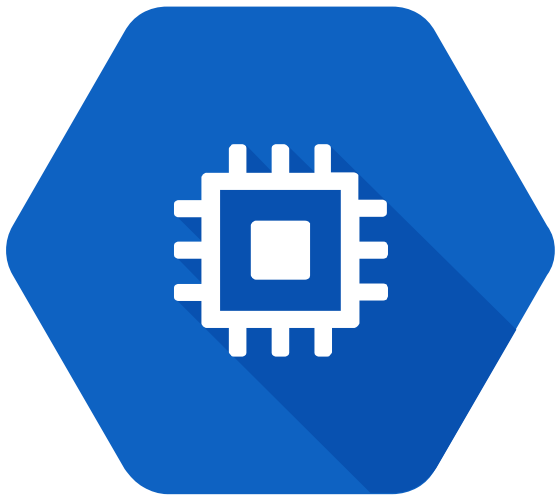
Next

Where Should I Run My Code?

Compute Engine - Infrastructure as a Service (IaaS)

What you want

- Virtual version of physical computer
- Full control/flexibility for customized purpose
- 'Default' option of other managed services don't fit



Compute Engine

What you need to think about:

- CPU/GPU
- Memory
- Disk space
- OS - including patching and updates
- Networking controls - firewalls, load balancers, VPN

Constraints:

- None, but comes with most admin overhead

What does Google handle?

- Hardware management
- Virtualization layer

Short version: Offers complete control and most flexibility, but also comes with most operational overhead.

- Use case: Moving existing servers into the cloud - most flexibility
- Applications with specific OS requirements

Return to Table of Contents

Choose a Lesson

Where Should I Run My Code?

Previous

Next

Where Should I Run My Code?

Kubernetes Engine

What you want:

- Kubernetes/container management
- Manage applications, not machines
- Portability of containers
- No dependencies on OS version
- Network protocols beyond HTTP/S
- CI/CD pipelines

What you think about:

- Applications over computers
- How are programs connected
- How is state stored (for stateful applications)
- Cluster configuration (machine type, GPU's etc)

What does Google handle?

- Managing the cluster - Kubernetes the easy way
  - Node management, software updates, autoscaling

Constraints:

- Must use containers
- Some architecture constraints - inter-app communication

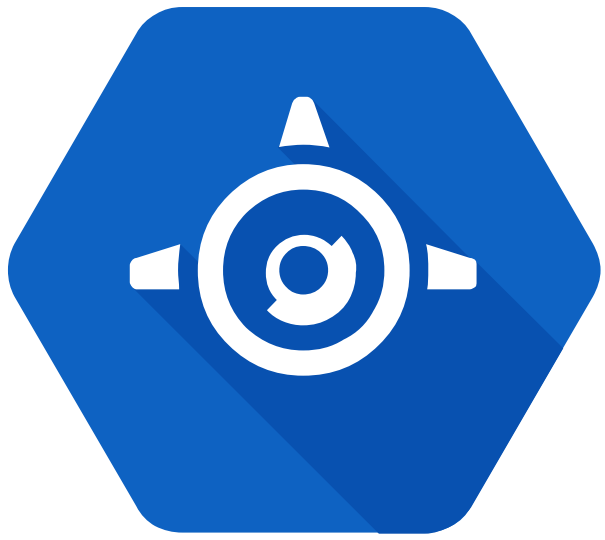
Use cases:

- Containerized workloads
- Cloud-native micro-services architectures

App Engine - Platform as a Service (PaaS)

What you want

- Focus on code - and that's it
  - Don't want to deal with having to manage infrastructure, even GKE infrastructure
- App Engine often referred to as 'serverless before serverless was cool'
- Handle variable load, from zero to massive (Standard)



App Engine

What you think about

- Your code
- HTTP requests
- Versioning

What does Google handle?

- Everything else
- All infrastructure, updates, scaling, networking, etc
- Supports very rapid scaling (Standard environment)

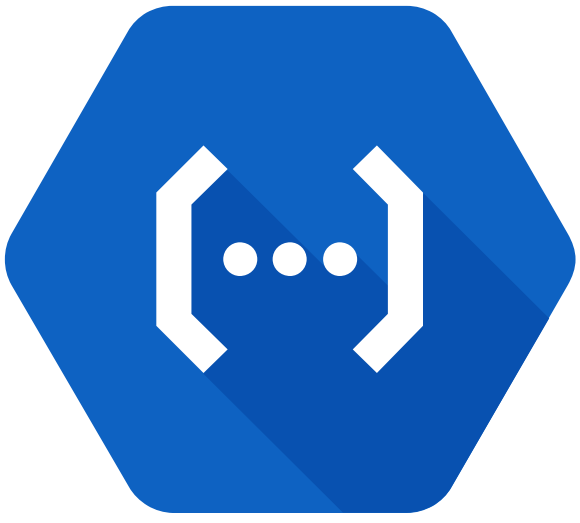
Constraints:

- Standard - limited runtimes (Python, Java, PHP, Go, NodeJS)
- Flexible - no scale to zero, slower scaling up/down compared to Standard

Use cases

- Web sites
- mobile app backends
- IoT apps

Cloud Functions



Cloud Functions

What you want

- True serverless computing
- Ability to respond to events -Event driven compute
- Function executes as a 'trigger' in response to a cloud based event
- Simple, single purpose functions
- Scale down to zero - pay/autoscale with usage

What you think about

- Events and defining your functions
- Google handles everything else

What Google handles

- Everything else

Constraints

- Limited runtimes - NodeJS, Python, Go
- Event driven only

Use cases

- Response to Pub/Sub and Cloud Storage events
- Example: A file is uploaded to Cloud Storage (event), function executes in response to event (trigger)

Exam perspective: Very minor - know what role Cloud Functions serves and the ability to scale down to zero

Return to Table of Contents

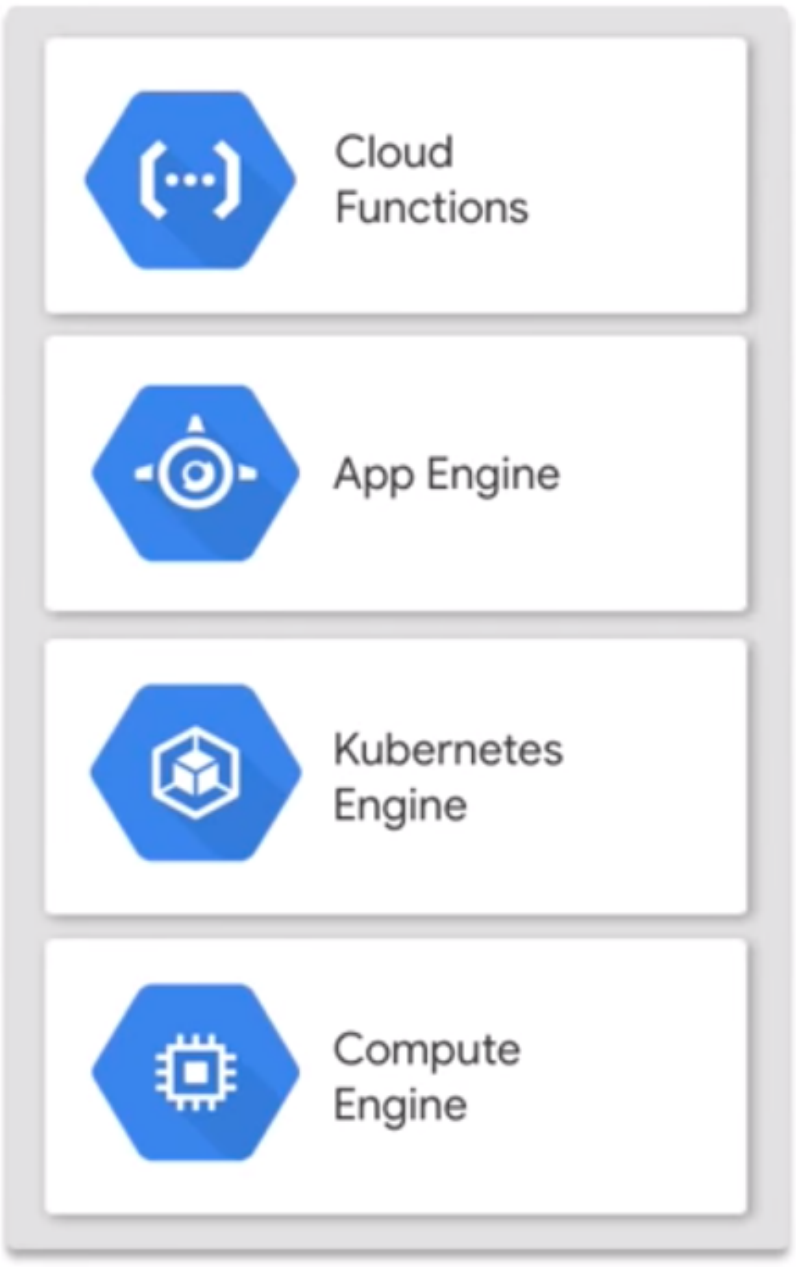
Choose a Lesson

Where Should I Run My Code?

Previous

Where Should I Run My Code?

Summing it all up



Return to Table of Contents

Choose a Lesson

App Engine Management Concepts

Return to Table of Contents

Choose a Lesson

App Engine Management Concepts

App Engine Management Concepts

Next

Course scope

- You should already be familiar with working with App Engine
- **App Engine Deep Dive** course if a refresher needed
- More concerned with how to manage App Engine then creating applications

Standard vs. Flexible Environments

Standard

- More constraints
- Python, Java, PHP, Go, Node.js
- Faster scale-up time (seconds)
- Intend to run for free or very low cost
  Only pay for what you need when you need it (if no traffic, no instances in use)

Flexible

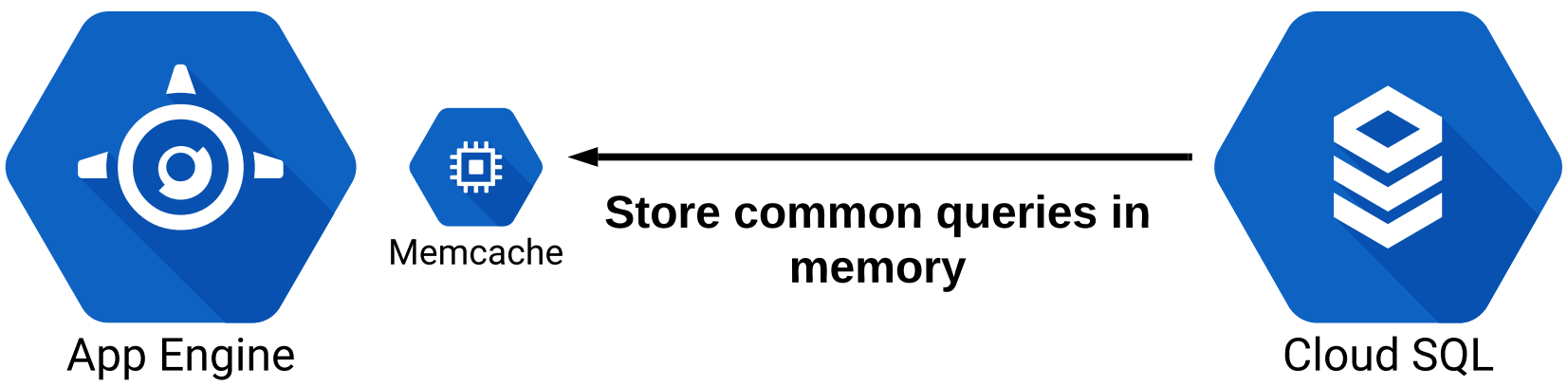
- Less constraints, more customization, but more management
- Python, Java, Node.js, PHP, Go, Ruby, or .NET
  - Or any other in own Docker container
- Slower scale-up time (minutes)
  - Consistent traffic, gradual scale up/down
- Load OS-dependent packages
- VPC access (including VPN connections)

Return to Table of Contents

App Engine Management Concepts

Choose a Lesson
App Engine Management Concepts
Previous
Next

Improving Performance with Memcache



- In-memory storage of commonly cached queries
- Two service levels: shared and dedicated
- Shared: Free, default option
  - Best effort basis on caching query operations
- Dedicated: Fixed cache capacity assigned for only your application
  - Cost assigned
- Scenario: You need to dedicate memcache capacity to improve query times to Cloud SQL backend to improve app performance

Return to Table of Contents

Choose a Lesson

App Engine Management Concepts

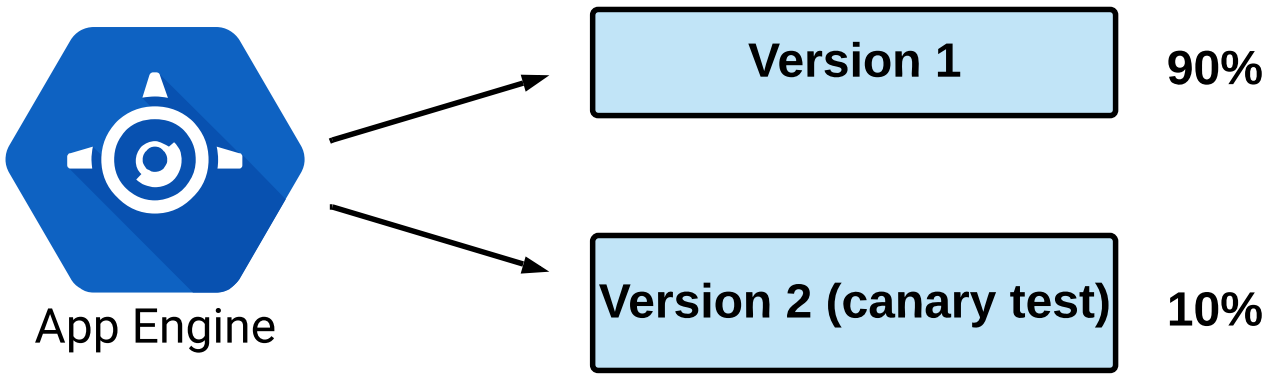
App Engine Management Concepts

Previous

Next

App Engine Version Management

- Very easy to manage App Engine versions — web interface or command line
- Split traffic by percentage between versions of app
- Scenario: You need to deploy a risky update to production — deploy app using **--no-promote** flag to deploy to prevent traffic going to new version
  - Slowly direct a little traffic to 'risky' update as canary test



Return to Table of Contents

Choose a Lesson

App Engine Management Concepts

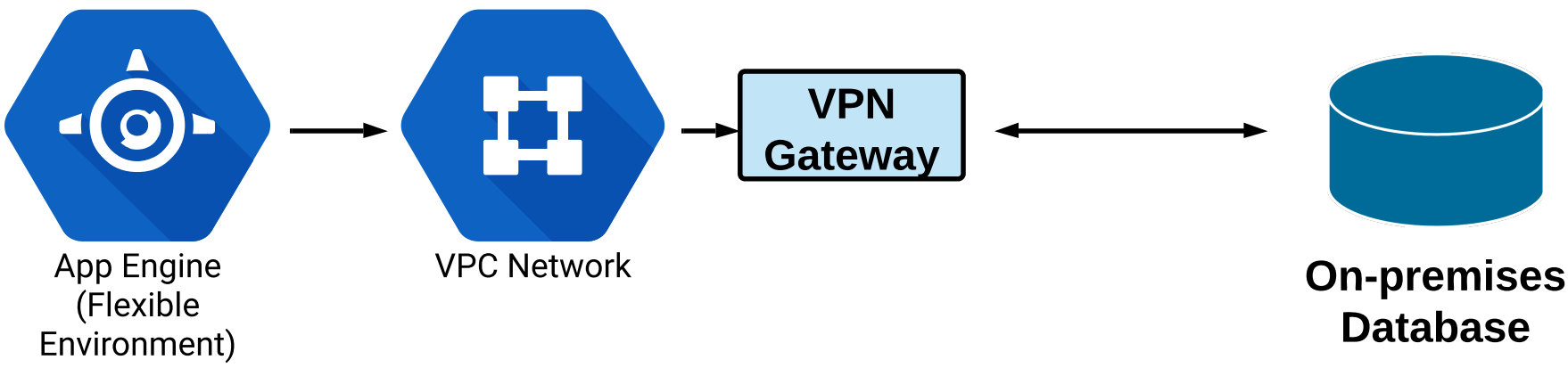
App Engine Management Concepts

Previous

Next

Connecting to external resources

- Scenario: Need to connect App Engine application to external/on-premises database
- Two options:
  - Make on-premises database publicly accessible, and use on-premises firewall to restrict traffic to only your app
  - Use VPN to connect App Engine app to on-premises database
    - Must use Flexible Environment to deploy to VPC — required for VPN solution



Return to Table of Contents

Choose a Lesson

App Engine Management Concepts

App Engine Management Concepts

Previous

Hands-On

- Deploy example bookshelf application
  - `git clone  https://github.com/GoogleCloudPlatformTraining/cp100-bookshelf`
  - `cd ~/cp100-bookshelf/app-engine`
  - `pip install -r requirements.txt -t lib`
  - `gcloud app deploy`
- Adjust memcache settings
- Deploy 'risky' update and manipulate versions
  - Use **--no-promote** flag to deploy new version with no traffic
  - `gcloud app deploy --no-promote`

Return to Table of Contents

Choose a Lesson

Return to Table of Contents

Choose a Lesson

Container Resources

GKE Administration Concepts

GKE Commands and Hands-On

Container Resources

Why this matters

- Exam will test high level understanding of GCP container deployment process
- Tools are part of CI/CD pipeline — discussed in detail later
- Big picture: Deployment Process
  - Build your container
  - Container Registry – “store it”
  - Google Kubernetes Engine – “deploy and run it”
- Other tools/variations (Jenkins/Spinnaker/etc.) will be discussed later

"Build it"

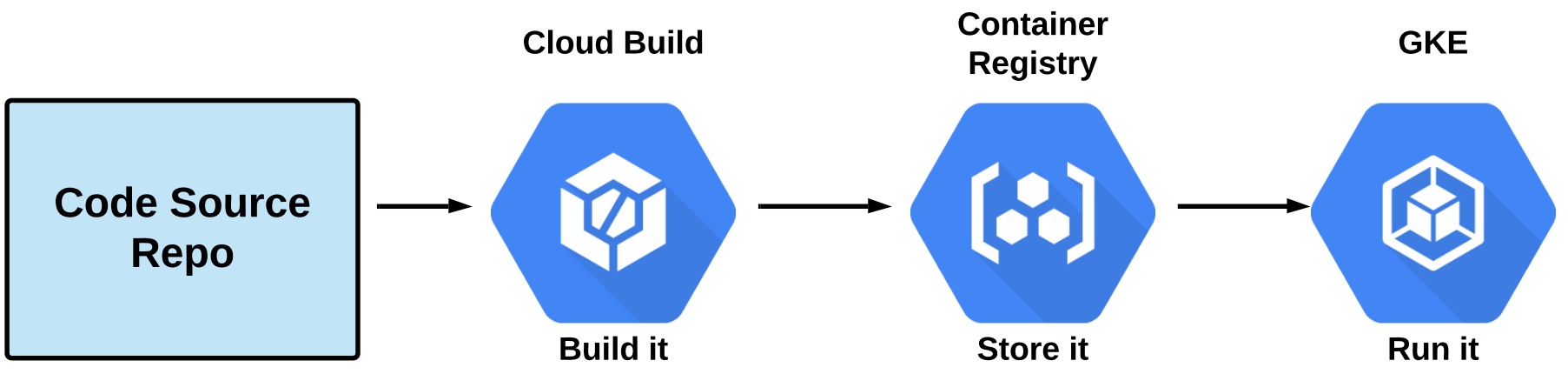
- Create Docker container image from source code
- Push created images stored in **Container Registry**
- **Cloud Build** service both builds and pushes for you
  - Pull code from multiple locations:
    - Google Cloud Storage
    - Google Cloud Source Repositories
    - GitHub
    - BitBucket

Container Registry "store it"

- Private Docker repository
- Integrate with GCP and external container services
- Supports CI/CD model
- Push images to the registry
- Pull images from registry
- Can deploy to GKE, GCE, or GAE (Flexible)
  - Or any other service that runs Docker containers

Kubernetes Engine "deploy and run it"

- Managed Kubernetes orchestration service
- “Kubernetes the easy way”
- Run Kubernetes in mixed environments



Note on newly announced services (Cloud Run, Anthos, etc.)

- Newly announced services (April 2019) will add new capabilities to container management/administration
- Not on the current version of the Architect exam

Return to Table of Contents

Choose a Lesson

Container Resources

GKE Administration Concepts

GKE Commands and Hands-On

GKE Administration Concepts

Next

GKE Course scope

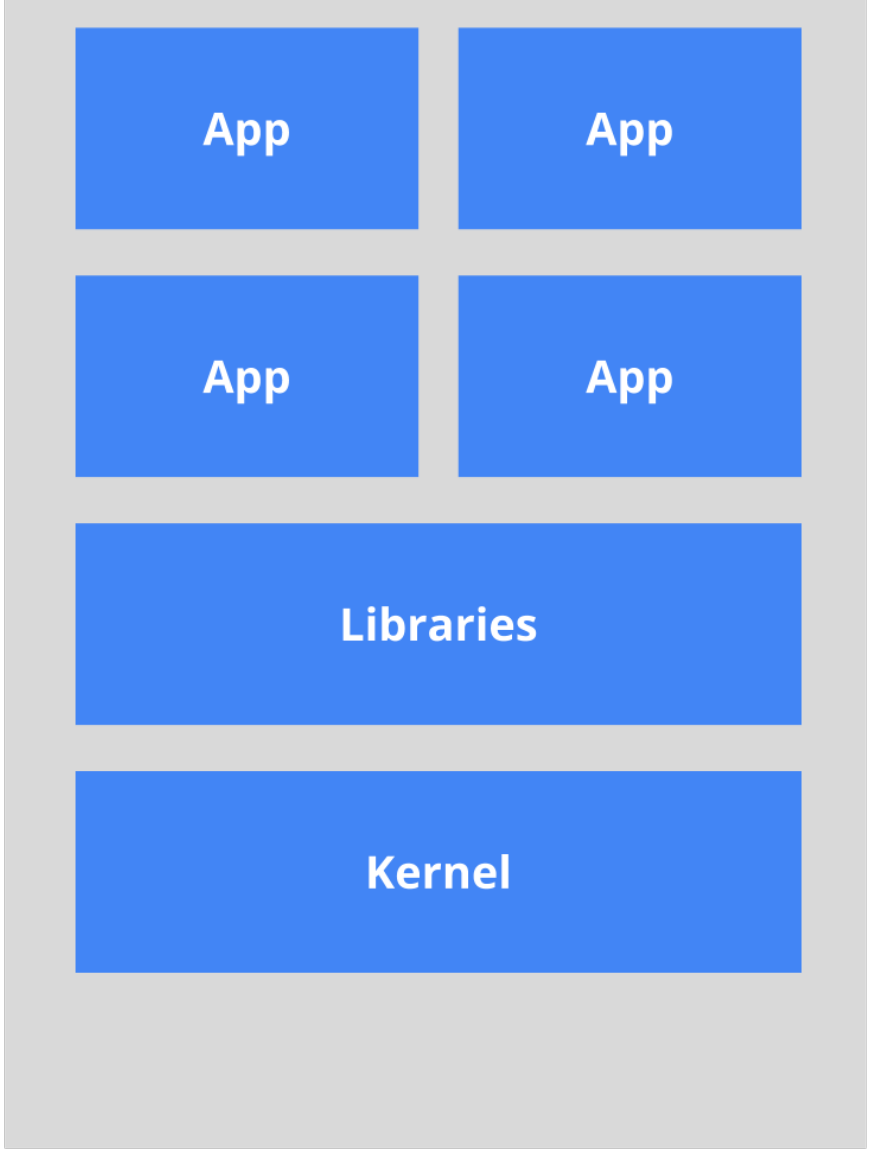
- Cover management concepts and administrative tools
- Hands-on with command line focus
- Prior GKE/Kubernetes experience assumed
- Further study resources
  - **Google Kubernetes Engine Deep Dive course**
  - **Beginner Kubernetes Learning Path**

Crash Course - Containers? Kubernetes? GKE?

What are containers?

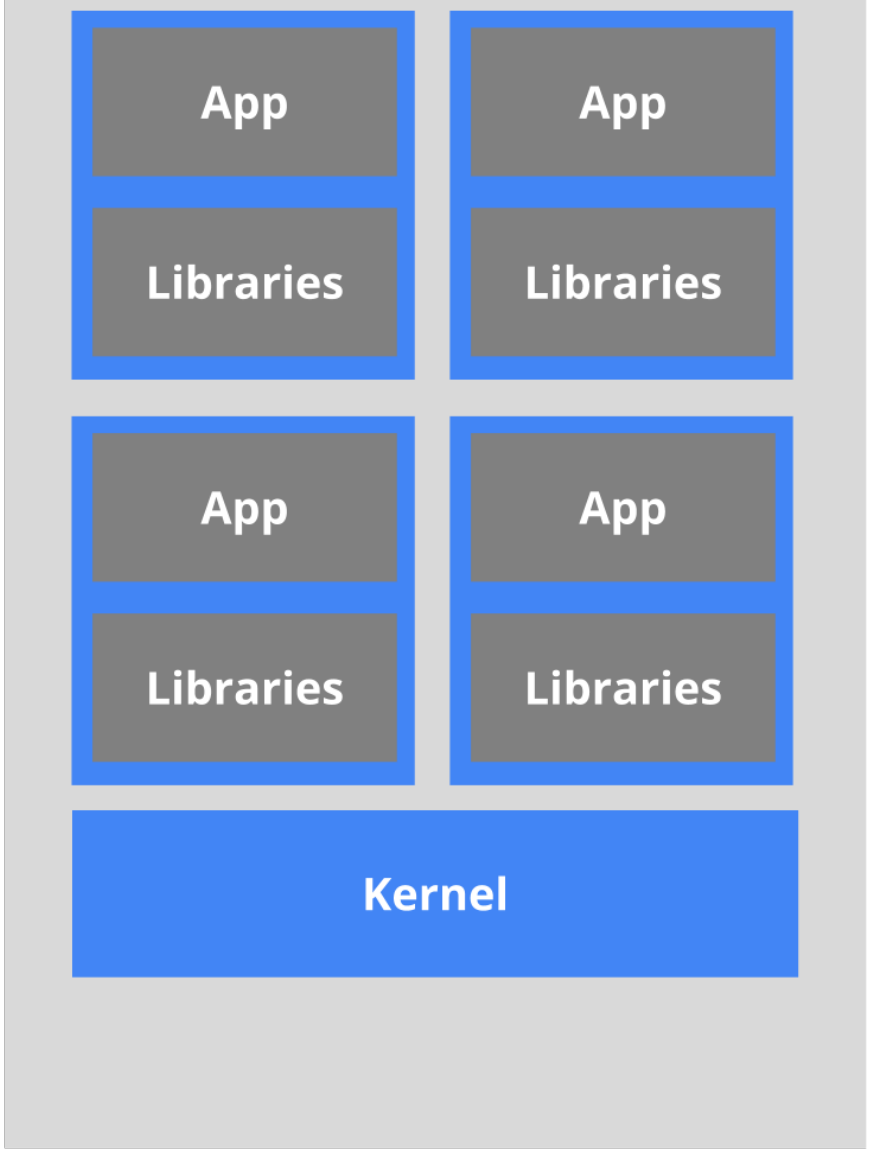
- Answer to the problem: How to make software more portable?
  - "It works in my environment, but not this other one."
  - Non-identical environments (different dependencies) = compatibility issues
- Containers = bundle entire runtime environment in one package
  - Runtime environment = dependencies, libraries, config files
  - By bundling everything in one neat package, OS and infrastructure concerns are abstracted away
  - Result: containers are more lightweight and use fewer resources than an entire VM
  - Often intertwined with **microservices** - smaller, modular app components
- Summary: Containers are lightweight, portable, self-contained packages that can be run virtually anywhere

The old way: Applications on host



Heavyweight, non-portable
Relies on OS package manager

The new way: Deploy containers



Small and fast, portable
Uses OS-level virtualization

Return to Table of Contents

Choose a Lesson

- Container Resources
- GKE Administration Concepts
- GKE Commands and Hands-On

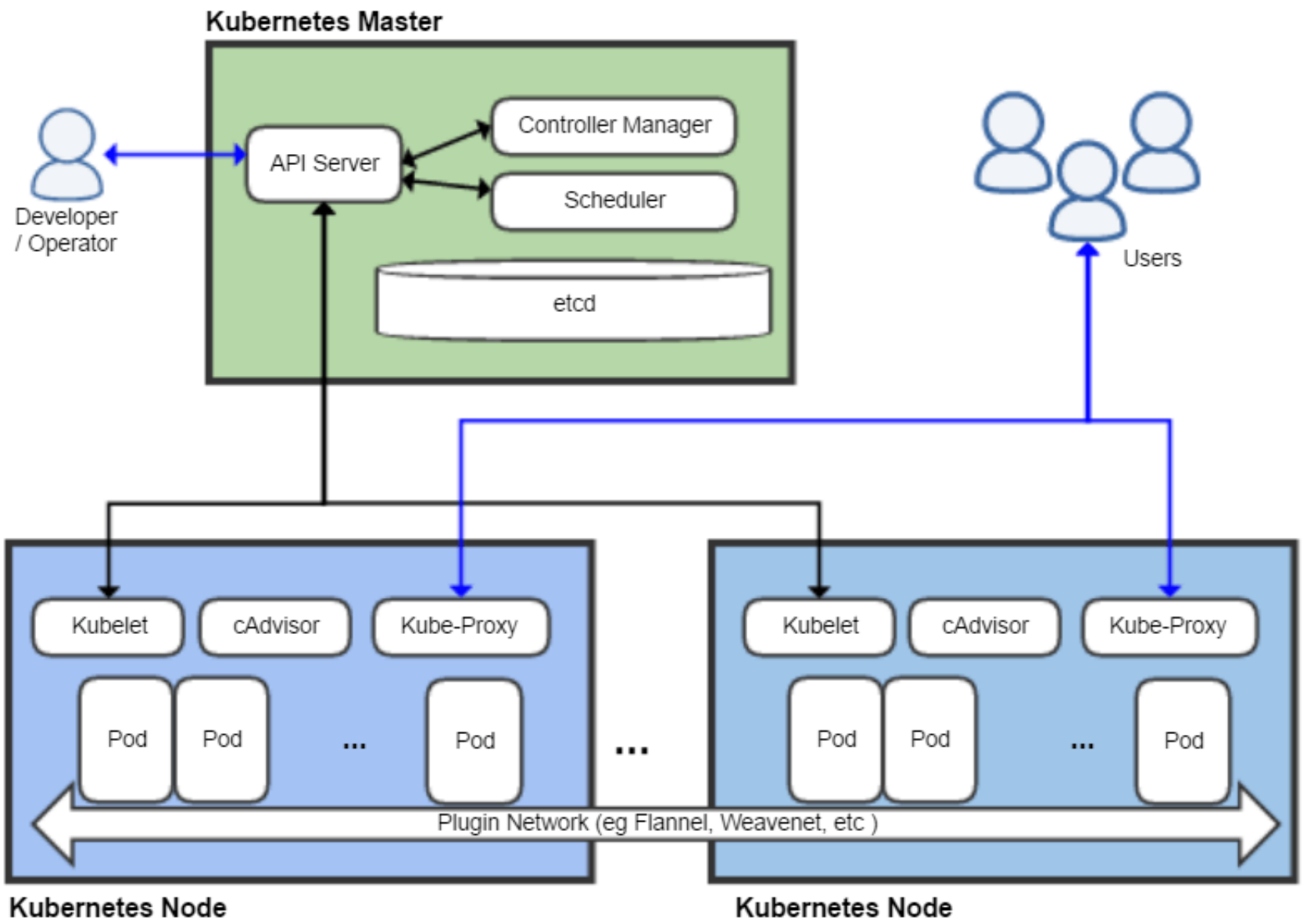
GKE Administration Concepts

Previous

Next

What is Kubernetes?

- Platform for managing container workloads
  - Create containers, manage them with Kubernetes
  - Kubernetes handles:
    - Deployment
    - Scaling
    - Updates (new versions of app)
    - Load Balancing
    - And much more!
  - Invented and used by Google internally, then open sourced in 2014



Return to Table of Contents

Choose a Lesson

Container Resources

GKE Administration Concepts

GKE Commands and Hands-On

GKE Administration Concepts

Previous

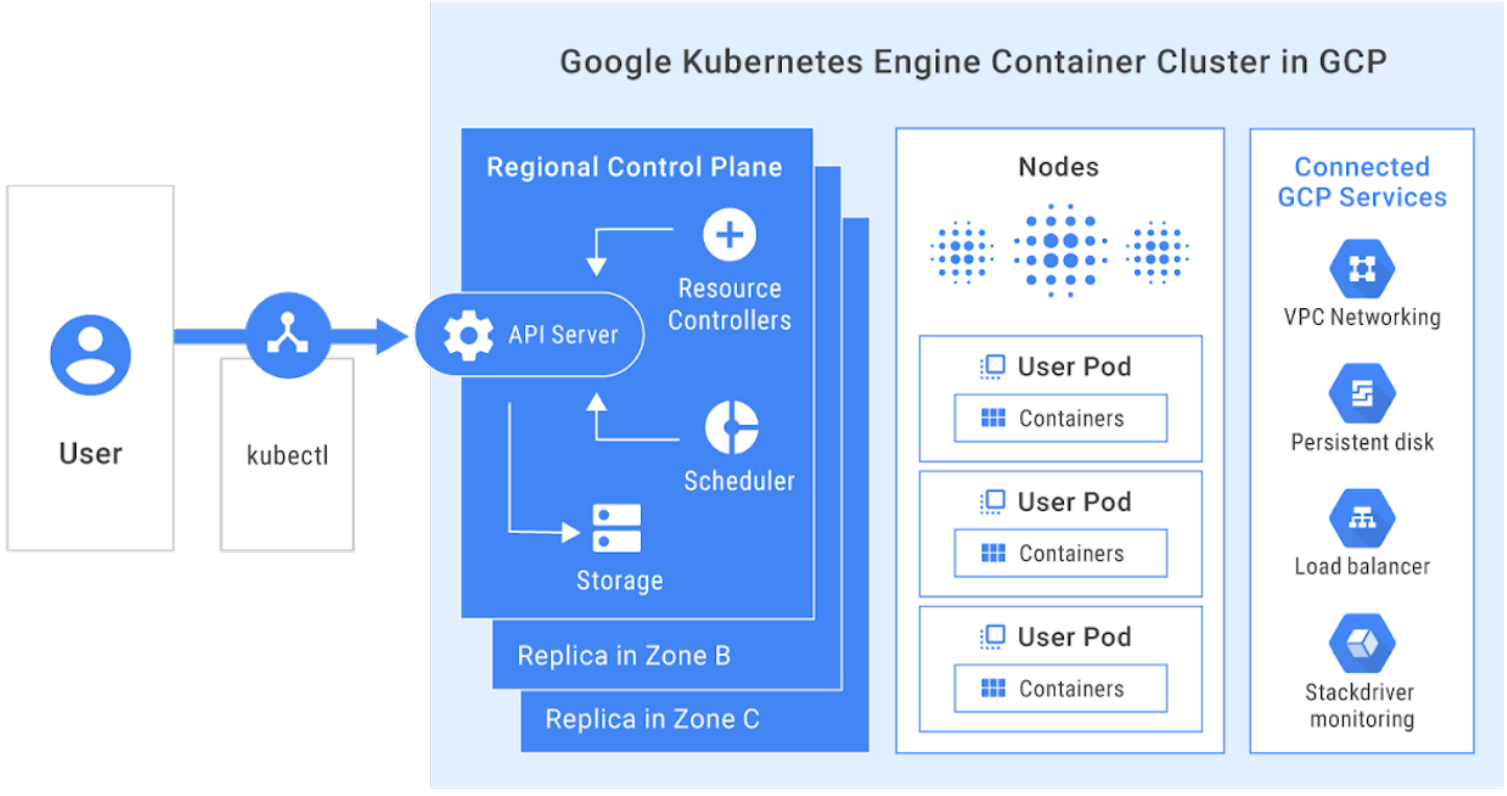
Next

Kubernetes manages containers, but you need to manage:

- Cloud Infrastructure Provisioning
- Setting up a CA and TLS Cert Generation
- Setting up TLS Client Bootstrap and RBAC Authentication
- Bootstrapping a H/A etcd cluster
- Bootstrapping a H/A Kubernetes Control Plane
- Bootstrapping Kubernetes Workers
- Configuring the Kubernetes Client - Remote Access
- Managing the Container Network Routes
- Deploying the Cluster DNS Add-on
- …. or you can just have all that done for you in a few clicks with Kubernetes Engine

What is Kubernetes Engine (GKE)?

- Managed Kubernetes service for container workloads
- Google handles OS management, Master node, scaling, health checks, replication controller, services, and all the above "stuff"
    - You just focus on your containers



Key Benefits

- Rich, powerful UI
- SRE monitoring
- Automated repair of apps
- Resource optimized app deployments
- Load balancing & auto-scaling of resources
- Global Virtual Private Cloud
- SLA of 99.95%

Return to Table of Contents

Choose a Lesson

Container Resources

GKE Administration Concepts

GKE Commands and Hands-On

GKE Administration Concepts

Previous

Next

Essential Terminology

Nodes

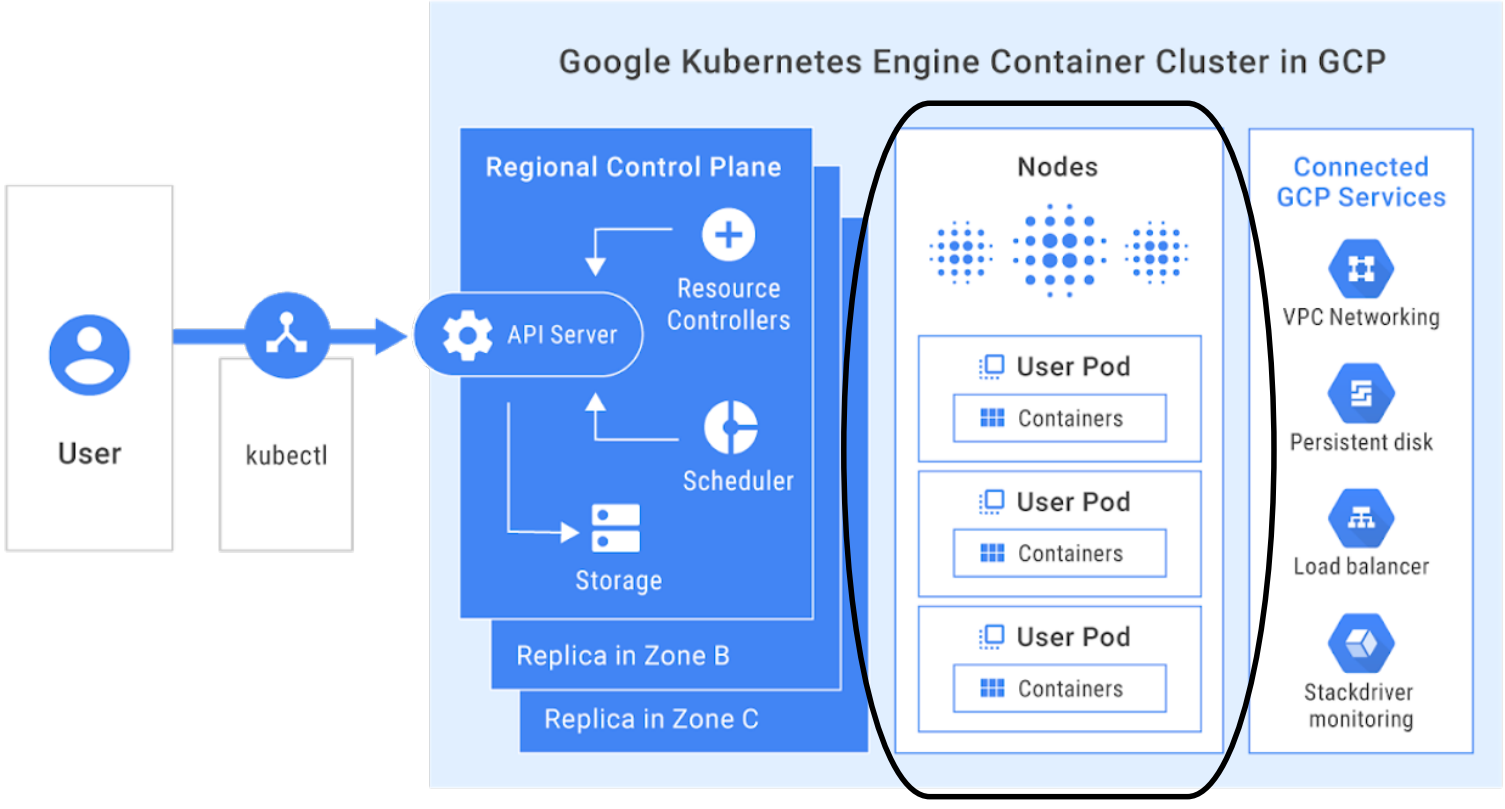
- **Nodes**: VM's containers are hosted on
  - GCE instances
  - Can have multiple containers/pods on single node
- **Node pool**: Group of nodes
  - Form of managed instance group
- **Cluster**: One or more node pools
- **Node image**: Node-level OS image
  - Container Optimized OS/Ubuntu
  - This is different than container image

Pods

- Smallest deployable unit - deployed to Nodes
- Pods are one or more containers bundled together
- Deploy one more more pods to nodes

Container Image

- Base image used in container (not node)
  - Pro tip: Use Alpine Linux for super slim container image



Return to Table of Contents

Choose a Lesson

- Container Resources
- GKE Administration Concepts
- GKE Commands and Hands-On

GKE Administration Concepts

Previous

Next

Interacting with GKE applications

- Direct commands or invoke config files in YAML/JSON format
- Use both gcloud and kubectl commands

kubectl vs. gcloud

- gcloud: Interact with GCP resources (GKE cluster/nodes, disks, API's)
- kubectl: Interact with application on nodes (pods)
  - Deploy, scale, update pods
- See this in action in the next lesson

Management Concepts

Stateful vs. stateless pods

- Stateless applications (pods) do not save state
- Stateful applications (pods) save data to persistent disk storage
  - Client-entered data is saved to server (persistent disk)
  - Kind of like pets vs. livestock
  - Stateful: Pets, Stateless: Livestock (expendable)
  - Updating and scaling/autoscaling stateful sets is much more deliberate

Container Image Best Practices

- **Deployment optimization**
  - Use slim base image like Alpine Linux - smaller deployments and less security surface
  - In Dockerfile, order matters; don't copy source files before installing dependencies, causes unnecessary reinstalls
- **Reliable container deployments**
  - Tag different versions of container deployments with version number, not just "latest"
  - Set pull policy to "IfNotPresent", not "Always"
    - Skip pulling an image if it already exists vs. forcing a pull regardless

GOOD

```
FROM alpine

RUN apt-get update && apt-get install -y
python python-pip

COPY . /src
```

NOT SO GOOD

```
FROM ubuntu: 16.04

COPY . /src

RUN apt-get update && apt-get install -y
python python-pip
```

Return to Table of Contents

Choose a Lesson

Container Resources

GKE Administration Concepts

GKE Commands and Hands-On

GKE Administration Concepts

Previous

Time for some hands on!

- Package a Docker container
- Push it to Container Registry
- Scale our node pool/cluster
- Scale our application (this is different!)
- Update images on pods (new version of app)
- Look at logs
- Cover other commands in addition to the hands on

Return to Table of Contents

Choose a Lesson

- Container Resources
- GKE Administration Concepts
- GKE Commands and Hands-On

GKE Hands-On and Commands

Next

In this hands-on demo:

- Heavy command line focus
- Both gcloud and kubectl commands
- Commands may be on exam
- First thing first: Go to Cloud Shell

Set up

- Set default region/zone
  - gcloud config set compute/region us-central1
  - gcloud config set compute/zone us-central1-a
- Enable API's
  - gcloud services enable container.googleapis.com
  - gcloud services enable containerregistry.googleapis.com

Clone GitHub for sample application and browse to directory

- git clone https://github.com/linuxacademy/content-gc-essentials
- cd content-gc-essentials/gke-lab-01

Package code into Docker container - tag as version 1 ("build it")

- docker build -t gcr.io/$DEVSHELL_PROJECT_ID/hello-la:v1 .
  - Note: the period at the end is required to build it at this location
  - $DEVSHELL_PROJECT_ID resolves to your current project ID
- Check status of images to ensure success
  - docker images

Push container into Container Registry ("push and store it")

- Authenticate gcloud as a Docker credential helper
  - gcloud auth configure-docker
- Push Docker container into Container Registry
  - docker push gcr.io/$DEVSHELL_PROJECT_ID/hello-la:v1

Create Kubernetes Engine Cluster (Nodes) named 'hello-cluster'

- gcloud container clusters create hello-cluster --num-nodes=2
- Authenticate kubectl to point to the cluster we just made (already done for last created cluster)
  - gcloud container clusters get-credentials hello-cluster

Deploy your app ("deploy and run it") - listen on port 80

- kubectl run hello-la --image=gcr.io/$DEVSHELL_PROJECT_ID/hello-la:v1 --port 80
- Check out our pods on the nodes
  - kubectl get pods

Create load balancer and expose application to the Internet on port 80

- kubectl expose deployment hello-la --type=LoadBalancer --port 80 --target-port 80

Find our load balancer frontend IP address

- kubectl get service

Return to Table of Contents

Choose a Lesson

Container Resources

GKE Administration Concepts

GKE Commands and Hands-On

GKE Hands-On and Commands

Previous

Next

Scale up deployment - add static number of replicas (pods)

- kubectl scale deployment hello-la --replicas=3

On second thought, let's just autoscale our application instead....

- Horizontal Pod Autoscaler
- kubectl autoscale deployment hello-la --max 6 --min 4 --cpu-percent 50

Maybe we should statically resize the node pool/cluster as well?

- gcloud container clusters resize hello-cluster --size 3
- If more than one pool per cluster, specify pool with --node-pool (pool_name)

On second thought (again), let's also enable autoscaling for our cluster

- gcloud container clusters update hello-cluster --enable-autoscaling --min-nodes 2 --max-nodes 8

Let's update our website!

Make changes to source code, then build as Docker file as VERSION 2

- docker build -t gcr.io/$DEVSHELL_PROJECT_ID/hello-la:v2 .

Second verse, same as the first - Push to Container Registry, also as version 2

- docker push gcr.io/$DEVSHELL_PROJECT_ID/hello-la:v2

Update our website - Apply rolling update to deployment with image update

- kubectl set image deployment/hello-la hello-la=gcr.io/$DEVSHELL_PROJECT_ID/hello-la:v2

Get log info

Logs are written to pods, by default also written to Stackdriver Logging

View log on pod

- kubectl logs (POD_ID)

Return to Table of Contents

Choose a Lesson

Container Resources

GKE Administration Concepts

GKE Commands and Hands-On

GKE Hands-On and Commands

Previous

Other commands/scenarios

- Upgrade version of Kubernetes on cluster
  - gcloud containers clusters upgrade (cluster_name)
- Change machine type on cluster
  - Trick question: You must create a new cluster/pool and migrate workload to new one

Return to Table of Contents

Choose a Lesson

Return to Table of Contents

Choose a Lesson

Big Data and Machine Learning Services

Data Lifecycle

Big Data and Machine Learning Services

Big Data services and Data Lifecycle

Next

- Exam perspective: "Little bit of everything", "Mile wide, inch deep"
- Conceptual understanding of major big data services
  - Exam questions give following services as correct and incorrect answers
  - 'Big picture' understanding of how data is handled - 'data lifecycle'
  - How the different pieces of the puzzle fit together
  - Translate technical and business requirements into a multi-step data solution

Big Data and Machine Learning services

- Think of this lesson as learning about key puzzle pieces
- Next lesson - how those puzzle pieces fit together
- Lesson objective: Introduction to big data and machine learning services not yet covered
- Thought process:
  - What does this service do?
  - What problem does this solve?
  - In Data Lifecycle lesson: Where does this fit in the big picture puzzle?

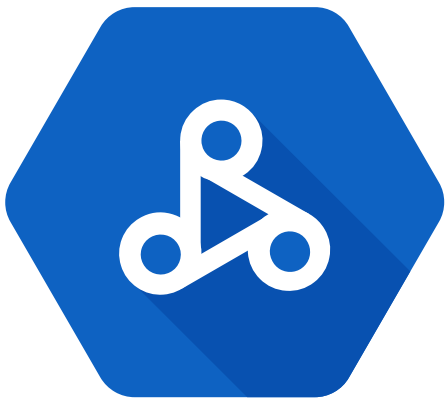
Return to Table of Contents

Choose a Lesson
Big Data and Machine Learning Services
Data Lifecycle

Big Data and Machine Learning Services

Previous
Next

Cloud Dataproc

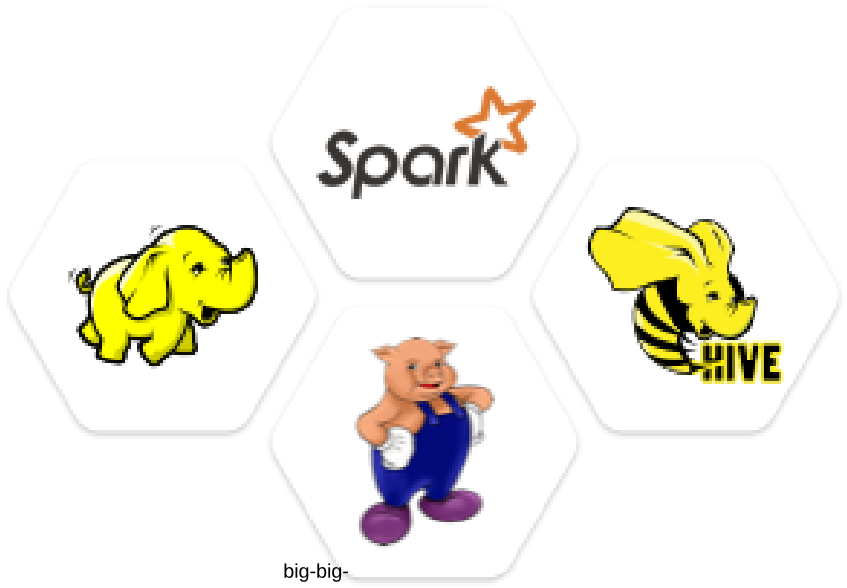


What does this do?

- Managed Hadoop/Spark clusters/management
- What is Hadoop ecosystem?
  - Very popular ecosystem of big data products
  - Hadoop, Spark, Pig, Hive
  - Dataproc manages infrastructure to allow focus on Hadoop/Spark workflows
  - Spark is a Hadoop ecosystem machine learning product - different from GCP specific services

What problem does this solve?

- Currently running Hadoop/Spark on-premises
- Do not want to change workflows when migrating to cloud
- Dataproc = Continue your Hadoop/Spark workflows/jobs with the power of GCP



Role in the 'big picture' Data Lifecycle puzzle?

- Data Processing and Analysis

big-big-

Return to Table of Contents

Choose a Lesson
Big Data and Machine Learning Services
Data Lifecycle

Big Data and Machine Learning Services

Previous
Next

Cloud Dataflow

What does it do?

- Batch and streaming data processing
- Change and transform data from one format into another
- Can process both streaming and batch data in the same pipeline
- Built on Apache Beam

Streaming/batch data?

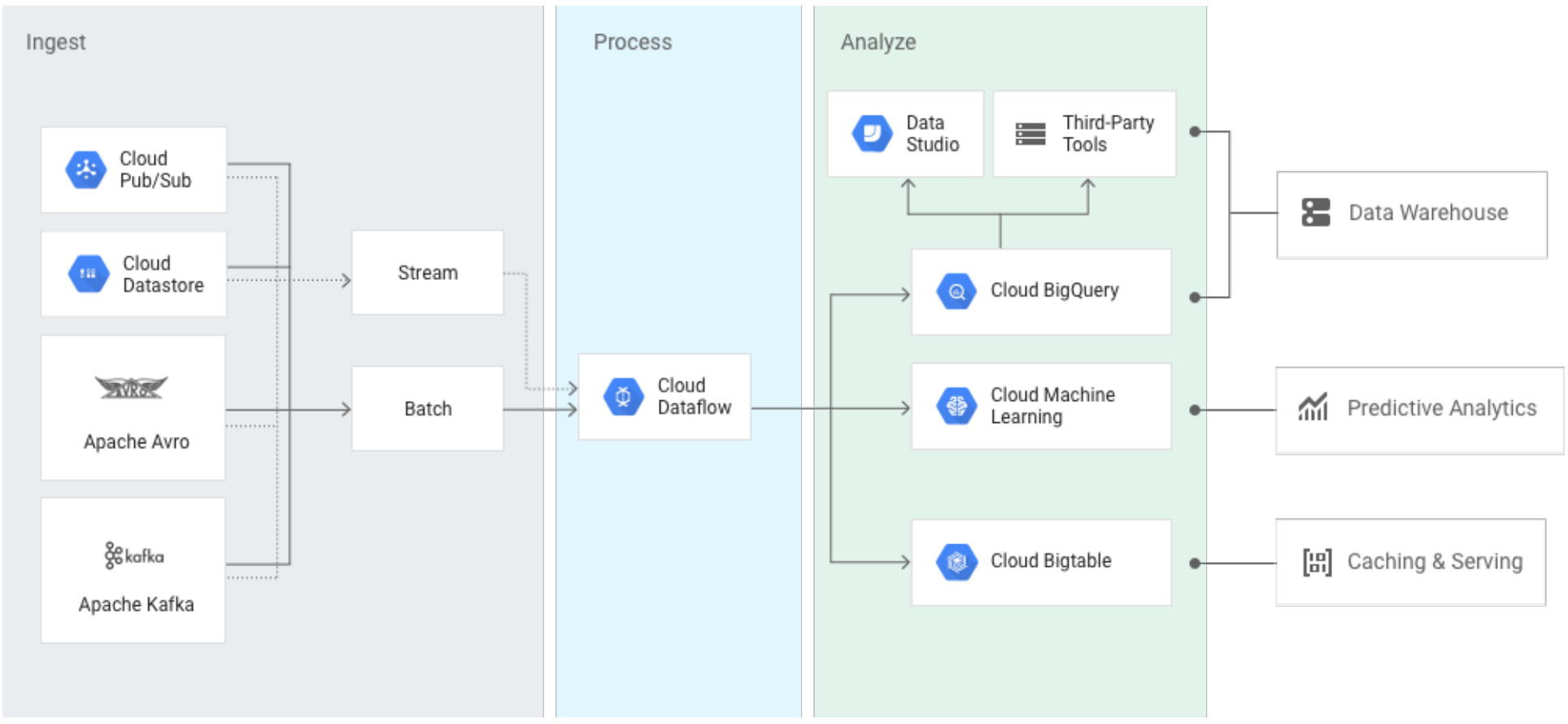
- Stream = continuous stream, asynchronous
  - Small bits
  - Many sources
  - Continuous flow
  - Examples: Sensor data, user events, IoT devices
  - Think Pub/Sub!
- Batch = large amounts of stored data
  - Transferred in bulk
  - Large 'chunks' of data
  - Few sources
  - Once in a while
  - Think Cloud Storage

What problem does this solve?

- Need to transform data before storing/using in another service
- Have multiple streaming and batch data sources to keep together

Role in the 'big picture' Data Lifecycle puzzle?

- Data Processing



Return to Table of Contents

Choose a Lesson

Big Data and Machine Learning Services

Data Lifecycle

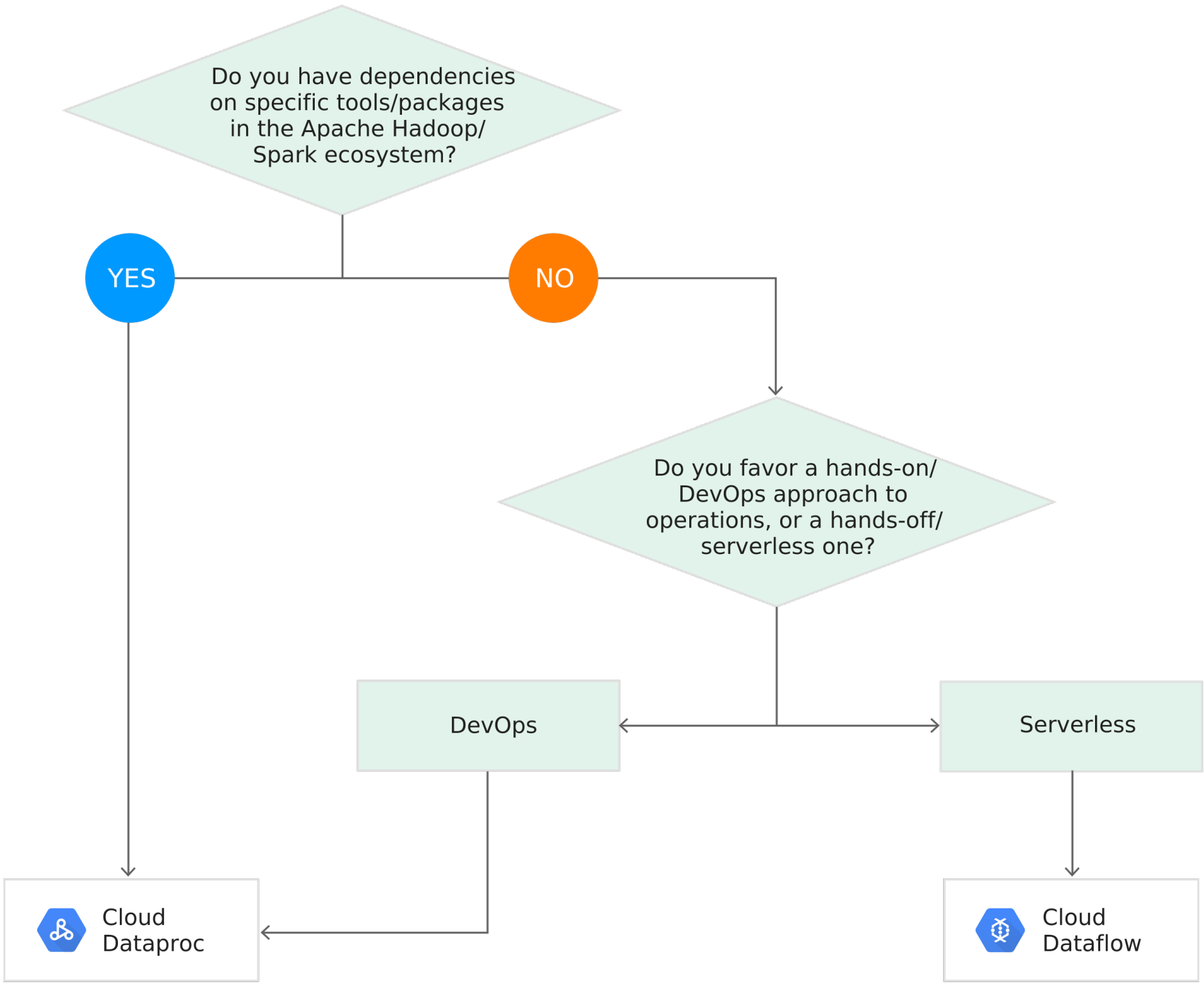
Big Data and Machine Learning Services

Previous

Next

Dataflow (continued)

- Dataproc vs. Dataflow?
- Lots of overlap
- Google's recommendation - use Dataproc if you're tied to Hadoop ecosystem or want to go more 'hands on' with a DevOps approach, otherwise use Dataflow



Return to Table of Contents

Choose a Lesson

Big Data and Machine Learning Services

Data Lifecycle

Big Data and Machine Learning Services

Previous

Next

Cloud Dataprep

What does it do?

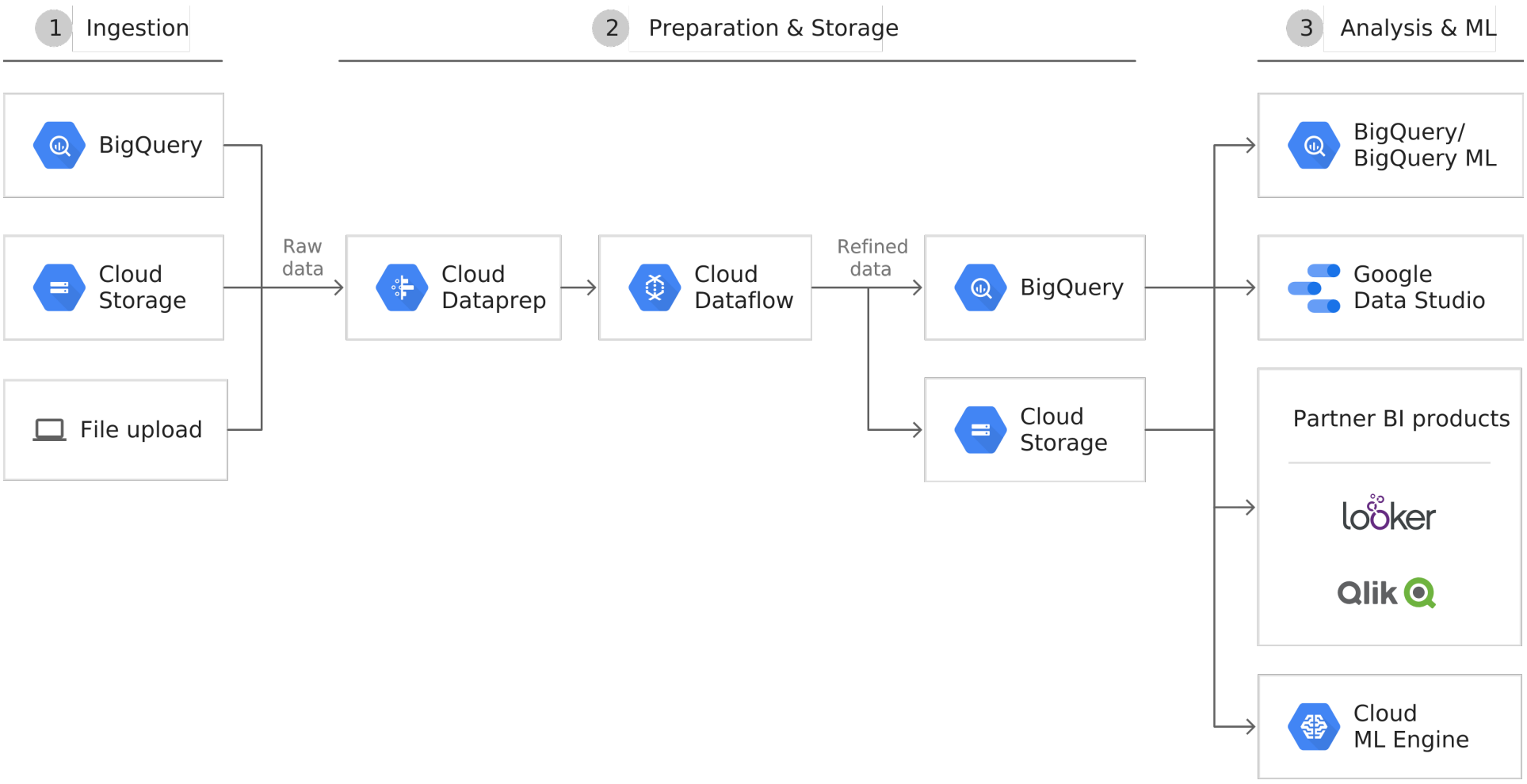
- Data cleaner-upper
  - Dataprep = "**Prep**are your **data**"
- Clean and transform data via web UI/click-and-point interface
- Runs managed Cloud Dataflow jobs via web UI
- Can schedule regular transformation jobs from Cloud Storage/BigQuery

What problem does this solve?

- Data for analysis is in wrong format/needs cleaning
- Need automatic 'cleaning jobs' performed on regularly added data - schedule jobs

Role in the 'big picture' Data Lifecycle puzzle?

- Data processing - uses Dataflow in easy point/click format



Return to Table of Contents

Choose a Lesson

Big Data and Machine Learning Services

Data Lifecycle

Big Data and Machine Learning Services

Previous

Next

Cloud Pub/Sub

What is it?

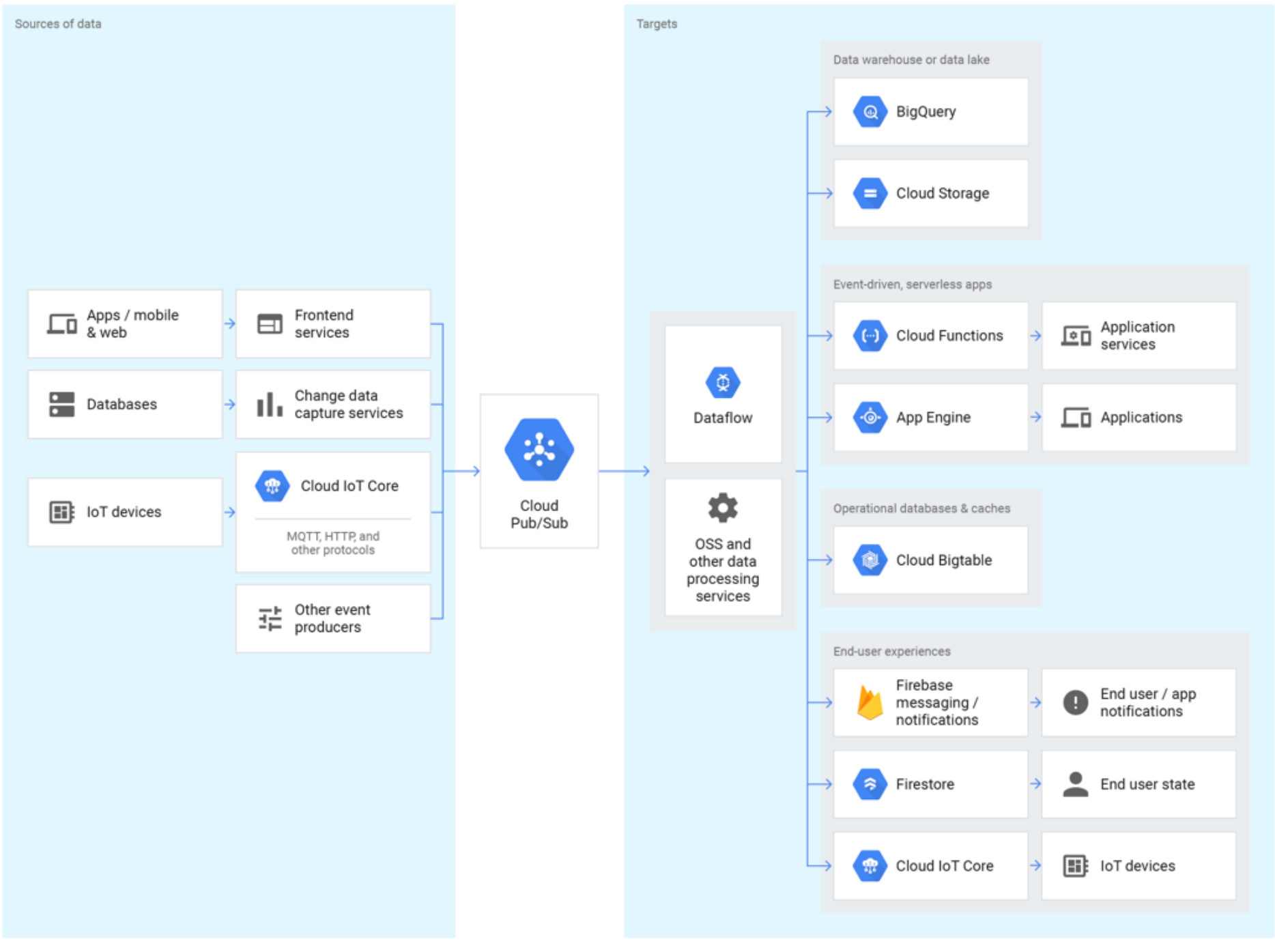
- Asynchronous messaging - 'many to many'
- Decouples senders and receivers = greater flexibility
- Sources publish messages, and other services subscribe to the published messages'
- Pub/Sub = Publish/Subscribe
- Global, infinite capacity data ingestion
- Similar to Apache Kafka

What problem does this solve?

- Ingest streaming data from anywhere in the world without worrying about capacity

Role in the 'big picture' Data Lifecycle puzzle?

- Data ingest
- Often paired with Dataflow for processing after ingest



Return to Table of Contents

Choose a Lesson

Big Data and Machine Learning Services

Data Lifecycle

Big Data and Machine Learning Services

Previous

Next

Machine Learning Services

What is Machine Learning?

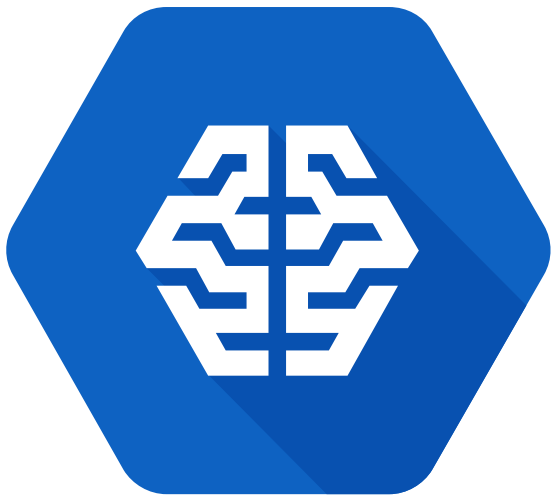
- Machine Learning: Teaching machines to understand data without explicitly programming it
- Examples: Image recognition, predictive analysis
- One of fastest growing fields and rapidly expanding GCP services
- Multiple services: Managed model training (Cloud ML Engine), pre-trained models

What problem does this solve?

- Depends on service:
  - Cloud ML Engine: Provide managed resources for training own custom ML model
  - Pre-trained API's: Google-trained ML models to 'plug in'
  - New solutions in rapid development (BigQuery ML)

Role in the 'big picture' Data Lifecycle puzzle?

- Analysis



Cloud Machine Learning Services

Return to Table of Contents

Choose a Lesson

Big Data and Machine Learning Services

Data Lifecycle

Big Data and Machine Learning Services

Previous

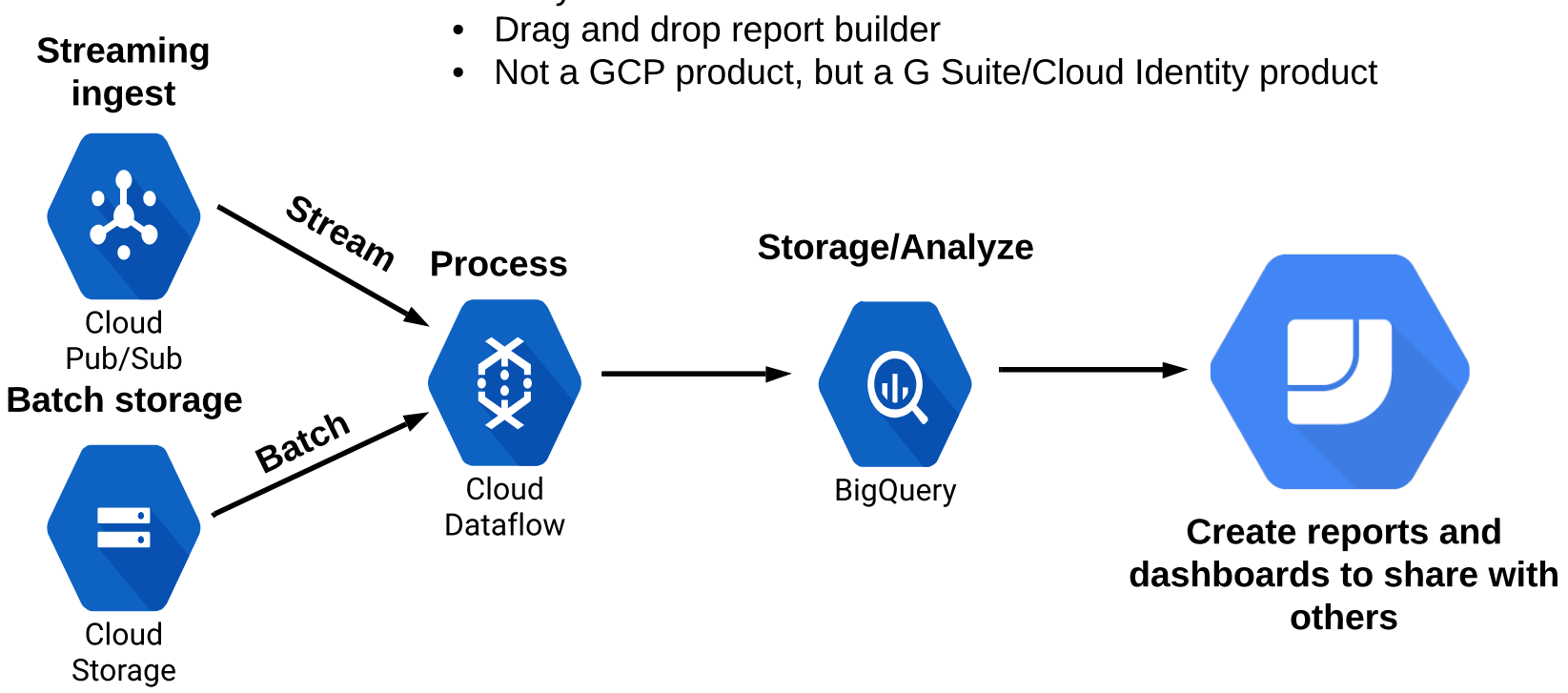
Data Visualization Services

Datalab

- Data Exploration
- Based on Jupyter notebooks
- Visual analysis of data in BigQuery, Cloud ML Engine, and more

Data Studio

- Easy to use data visualization and dashboards
- Drag and drop report builder
- Not a GCP product, but a G Suite/Cloud Identity product



Return to Table of Contents

Choose a Lesson

Big Data and Machine Learning Services

Data Lifecycle

Data Lifecycle

Next

Why is this important?

- Think of data as a tangible object
  - Collected, stored, processed, and used
- Lifecycle from initial collection to final visualization
- Different services (puzzle pieces) connect together to move data along the process
- Data Lifecycle steps:
  - Ingest: Collect raw data
  - Store: Hold data
  - Process and analyze: Transform data from raw format to actionable information
  - Explore and visualize: Get use out of it
    - The final stage is to convert the results of the analysis into a format that is easy to draw insights from and to share with colleagues and peers

Ingest	Store	Process & Analyze	Explore & Visualize
App Engine	Cloud Storage	Cloud Dataflow	Cloud Datalab
Compute Engine	Cloud SQL	Cloud Dataproc	Google Data Studio
Kubernetes Engine	Cloud Datastore	BigQuery	Google Sheets
Cloud Pub/Sub	Cloud Bigtable	Cloud ML	
Stackdriver Logging	BigQuery	Cloud Vision API	
Cloud Transfer Service	Cloud Storage for Firebase	Cloud Speech API	
Transfer Appliance	Cloud Firestore	Translate API	
	Cloud Spanner	Cloud Natural Language API	
		Cloud Dataprep	
		Cloud Video Intelligence API	

Return to Table of Contents

Choose a Lesson

Big Data and Machine Learning Services

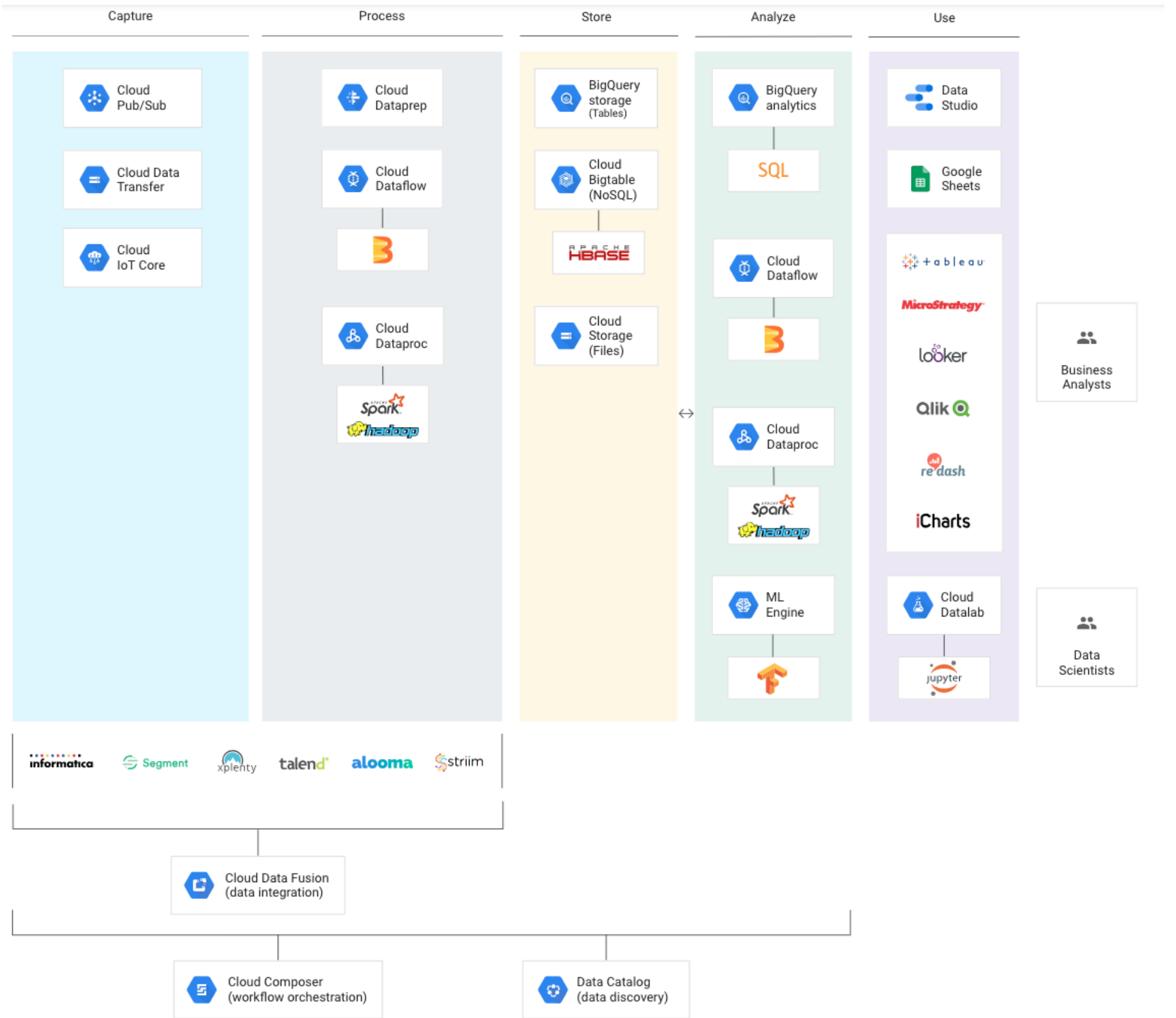
Data Lifecycle

Data Lifecycle

Previous

Next

Exact steps are flexible



Return to Table of Contents

Choose a Lesson

Big Data and Machine Learning Services

Data Lifecycle

Data Lifecycle

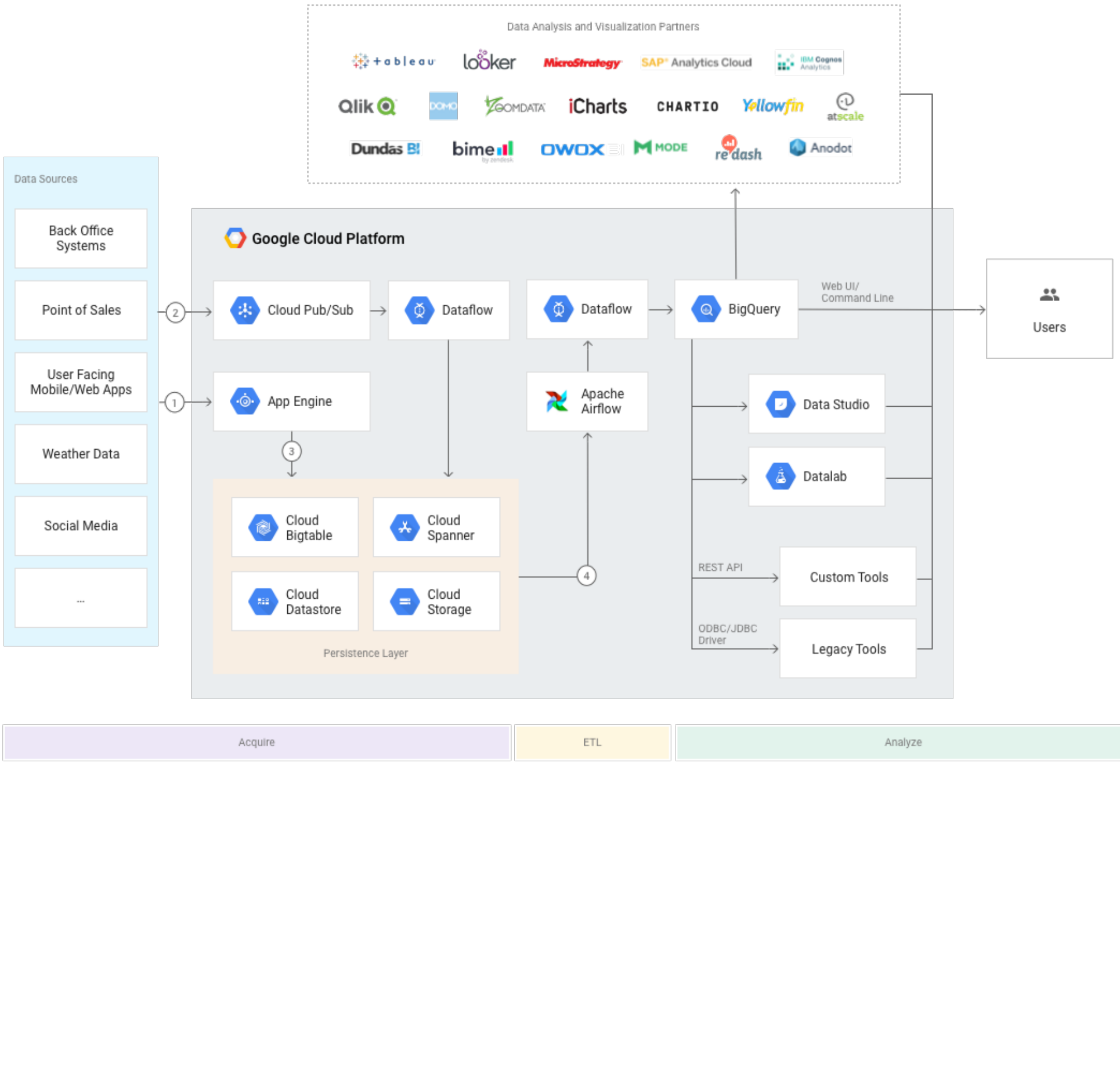
Previous

Think about:

- End-to-end process of collecting, storing, processing, and using data
- What services in the above process
- Examples in upcoming Case Study lessons

Additional references

- Google Cloud Data Lifecycle
- Google Cloud Big Data products at a glance
- Google Solutions Architecture Reference



Return to Table of Contents

Choose a Lesson

Return to Table of Contents

Choose a Lesson

Case Studies Overview

Mountkirk Games

Dress4Win

TerramEarth

Case Studies Overview

Case Studies are a BIG part of the exam

- Three available case studies
  - Expect to see all three
- Same case studies available from Google in advance (and covered here)
- Roughly 30% of exam questions
- Exam will be side-by-side format

Questions to ask for each case study (some overlap with others):

- What does this company do? (Company Overview)
- What are their pain points?
- What are their goals? (Solution Concept)
- What do they care about (big picture)?
- What are their business/technical requirements?
- If applicable, what is their current environment, and how would it translate to GCP given their above requirements?
- If applicable, what does infrastructure/data lifecycle look like?

Return to Table of Contents

Choose a Lesson

- Case Studies Overview
- Mountkirk Games
- Dress4Win
- TerramEarth

Mountkirk Games

Next

Link to case study

- Mountkirk Games Case Study

What does this company do? (Company Overview)

- Mobile games maker (think Pokemon Go, Candy Crush, etc.)

What are their pain points?

- Current mobile game not scaling well
- New game (hosted on GCP) needs to scale well

What are their goals? (Solution Concept)

- Building new game, no need to migrate existing environment
- Two key environments:
  - Create game backend on GCE
    - Hardened Linux distro
    - NoSQL transactional database - **Datastore (Firestore)**
  - Separate analytics setup
  - Different storage service for each

What do they care about (big picture)?

- Scaling!
- Measuring performance - increase efficiency of solutions - Stackdriver Monitoring/Logging
- Reliable experience for users - no downtime
- Managed services
- Analytics - usage patterns
- Global footprint

Mapping Business/Technical Requirements
Business Requirements

- Increase to a global footprint
  - Multiple regional **instance group** backends
    - Served by single global **HTTP load balancer**
  - Multi-regional ingest/storage options
    - Pub/Sub, Datastore, BigQuery, Cloud Storage
- Improve uptime - downtime is loss of players
  - Autoscaling instance groups
  - Low latency global load balancer
  - **Pub/Sub** and **Dataflow** accounts for slow/late data
  - **Pub/Sub** buffers data, no scaling cap, no log backlog
- Increase efficiency of the cloud resources we use
  - Scaling infrastructure (all of the above)
  - Monitor with **Stackdriver**, Stackdriver metrics can drive GCE group scaling
- Reduce latency to all customers
  - Multi-regional GCE backends (served by **HTTP load balancer**), Multi-region **Datastore**

Return to Table of Contents

Choose a Lesson

Case Studies Overview

Mountkirk Games

Dress4Win

TerramEarth

Mountkirk Games

Previous

Next

Technical Requirements

Requirements for Game Backend Platform

- Dynamically scale up or down based on game activity.
  - Autoscaling **managed instance groups**
  - **Stackdriver** to measure performance/efficiency
    - Also drive autoscaling
- Connect to a transactional database service to manage user profiles and game state.
  - **Cloud Datastore** (Firestore) - NoSQL transactional database - perfect for game user profiles and game states
- Store game activity in a timeseries database service for future analysis.
  - Game activity from servers
  - Store in **BigQuery**
  - BigQuery vs. Bigtable?
    - Bigtable = millisecond response time
    - BigQuery = response measured in seconds, scales more efficiently
    - No requirement for low latency analytics response time
    - BigQuery reading from Bigtable also a valid answer, but not best answer (in my opinion)
- As the system scales, ensure that data is not lost due to processing backlogs.
  - **HTTP Load Balancer** - automatically scales to meet demand
  - **Managed Instance Groups** - also autoscales
  - **Pub/Sub** - buffers late/slow data
- Run hardened Linux distro.
  - **Managed instance groups** - custom images

Return to Table of Contents

Choose a Lesson

- Case Studies Overview
- Mountkirk Games
- Dress4Win
- TerramEarth

Mountkirk Games

Previous

Next

Requirements for Game Analytics Platform

- Dynamically scale up or down based on game activity.
  - Autoscaling services, everything below
- Process incoming (streaming) data on the fly directly from the game servers.
  - Connect services with Pub/Sub, process with Dataflow
- Process data that arrives late because of slow mobile networks.
  - **Pub/Sub** scales and buffers messages
  - **Dataflow** accounts for late/out of order data
- Allow queries to access at least 10 TB of historical data.
  - **BigQuery** - SQL queries against data
- Process files that are regularly uploaded by users' mobile devices.
  - Upload to storage (Cloud Storage)
  - Process via Dataflow

Choosing a storage option: https://cloud.google.com/storage-options/

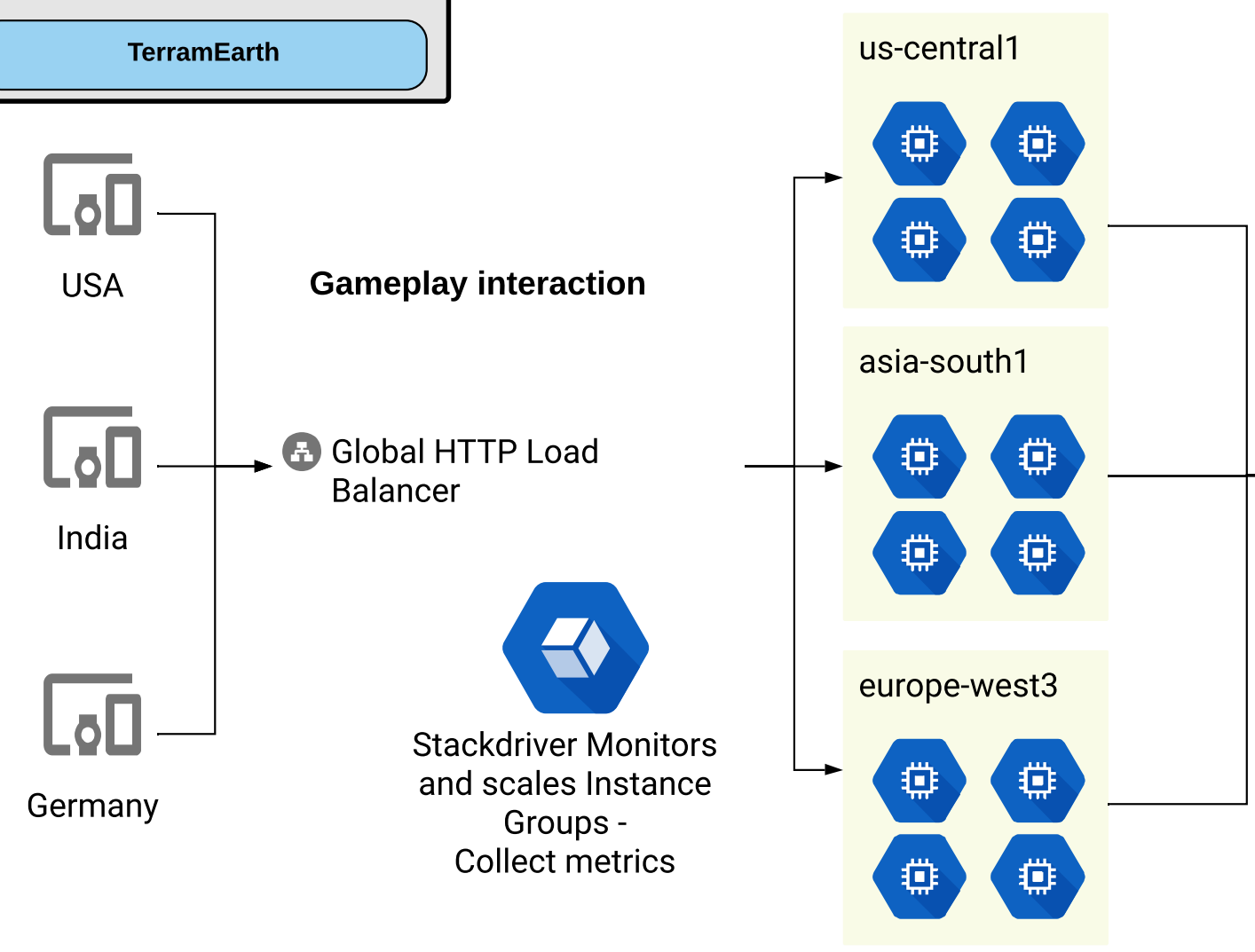
Building a mobile games analytics platform:
https://cloud.google.com/solutions/mobile/mobile-gaming-analysis-telemetry

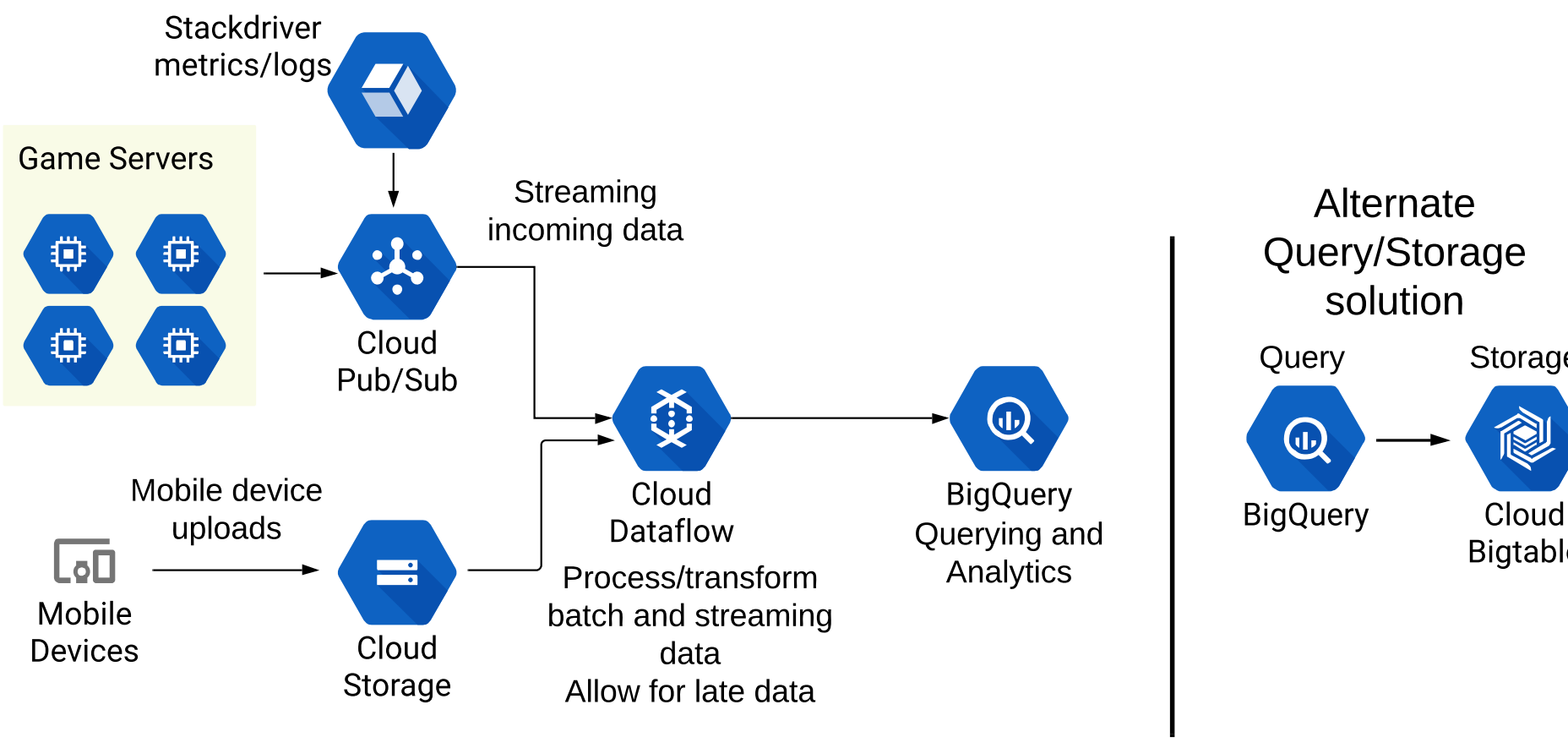
Return to Table of Contents

Choose a Lesson
Case Studies Overview
Mountkirk Games
Dress4Win
TerramEarth

Mountkirk Games

Previous





Return to Table of Contents

Choose a Lesson

- Case Studies Overview
- Mountkirk Games
- Dress4Win
- TerramEarth

Dress4Win

Next

Link to Case Study

Dress4Win

What does this company do? (Company Overview)

- Web-based wardrobe management
- Social networking around fashion
- Monetizes via ads

What are their pain points?

- Current infrastructure can't keep up with company growth
- Need to future-proof themselves
- Innovation stifled due to hardware constraints

What are their goals? (Solution Concept)

- Proof of concept deployment to GCP
  - Migrate dev/test environments
  - Build disaster recovery (backup of current business)
    - Think hybrid networks (GCP connect to on-premises)
- If successful, will work toward full cloud migration
- Need to map current environment to GCP equivalents
- Prefer managed services

What do they care about (big picture)?

- Manage costs
  - Initial cloud investment
  - Scaling down capacity during off-peak times
- Out-innovate competitors
- Business agility
  - Quickly provision resources
  - Have room to innovate without waiting on hardware
- Maintaining secure environment
  - IAM, Customer supplied encryption, firewall rules
- Note: global footprint not a priority - could be a future possibility

Return to Table of Contents

Choose a Lesson

- Case Studies Overview
- Mountkirk Games
- Dress4Win
- TerramEarth

Dress4Win

Previous

Next

Mapping Business/Technical Requirements

Business Requirements

- Build a reliable and reproducible environment with scaled parity of production.
  - As much as possible, create equivalent setup on cloud without having to re-engineer existing applications
- Improve security by defining and adhering to a set of security and Identity and Access Management (IAM) best practices for cloud.
  - Principle of least privilege
  - Separate test/development environments into separate projects
- Improve business agility and speed of innovation through rapid provisioning of new resources.
  - Automate infrastructure creation
    - gcloud/Google Cloud SDK
    - Rapid deployment (deployment manager, etc)
    - Marketplace for easy deployment of existing services (Tomcat, Nginx, Jenkins, etc)
- Analyze and optimize architecture for performance in the cloud.
  - Stackdriver
    - Monitor infrastructure with Stackdriver Monitoring
    - Notified of errors with Stackdriver Logging
    - Troubleshoot errors with Stackdriver Debug/Error Reporting

Return to Table of Contents

Choose a Lesson

- Case Studies Overview
- Mountkirk Games
- Dress4Win
- TerramEarth

Dress4Win

Previous

Next

Technical Requirements

- Easily create non-production environments in the cloud.
  - Best practices for migration
    - Move data first, then applications
    - More detail later in this course
- Implement an automation framework for provisioning resources in cloud.
  - gcloud for automated management (scripts)
    - **Cloud Deployment Manager**
    - Other infrastructure as code products
- Implement a continuous deployment process for deploying applications to the on-premises datacenter or cloud.
  - Discussed further in this course
  - CI/CD pipeline, Jenkins, Spinnaker, **Cloud Build**, etc
- Support failover of the production environment to cloud during an emergency.
  - Replicating environment on Google Cloud
    - MySQL replicating to **Cloud SQL**
    - On-premises/cloud application servers - DNS cutover
- Encrypt data on the wire and at rest.
  - Customer supplied (custom) encryption keys
- Support multiple private connections between the production data center and cloud environment.
  - **Cloud VPN** or **Cloud Interconnect**

Return to Table of Contents

Choose a Lesson

- Case Studies Overview
- Mountkirk Games
- Dress4Win
- TerramEarth

Dress4Win

Previous

Next

Existing Technical Environment
Databases:

- MySQL - **Cloud SQL**
  - Native MySQL support
    - **Lift and shift**
  - 10 TB size limit
  - Single region - no global footprint requirement
  - Migration - create replica server managed by Cloud SQL
    - Once replica is synced:
      - Update applications to point to replica
      - Promote replica to stand-alone instance

Redis 3 server cluster

- Two options:
  - Run Redis server on Compute Engine
  - Use new **Memorystore** managed Redis database

Compute:
40 Web Application servers providing micro-services based APIs and static content.

- Tomcat - Java
- Nginx
- 4 core CPUs
- 32 GB of RAM

Existing environment has lots of idle time

- **Managed instance groups** - autoscaling

Use custom machine types

Alternatively: Re-architect for GKE/GAE for microservices deployments

Return to Table of Contents

Choose a Lesson

Case Studies Overview

Mountkirk Games

Dress4Win

TerramEarth

Dress4Win

Previous

20 Apache Hadoop/Spark servers:
- **Cloud Dataproc**
  - Can connect to **Cloud Storage**

3 RabbitMQ servers for messaging, social notifications, and events:
- **Pub/Sub** likely replacement
  - Can also deploy same environment on Compute Engine instance group (lift and shift)

Miscellaneous servers:
Jenkins, monitoring, bastion hosts, security scanners
  - No managed service equivalents
  - Use GCE instances - custom machine types available

Storage appliances:
- iSCSI for VM hosts/Fiber channel SAN - Backup for MySQL databases
- 1 PB total storage; 400 TB available
  - SAN/iSCSI requires block storage
  - **Persistent disks** working in SAN cluster

- NAS - image storage, logs, backups
  - **Cloud Storage** will be direct replacement
  - Infinite scalability in a single bucket
  - Persistent also an option

Return to Table of Contents

Choose a Lesson

Case Studies Overview

Mountkirk Games

Dress4Win

TerramEarth

TerramEarth

Link to Case Study
[TerramEarth Case Study]

Next

What does this company do?

- Heavy equipment, mining, agriculture
- Bulldozers, tractors, etc.
- 500 dealers all over the world
- Mission = make customers more productive

What are their pain points?

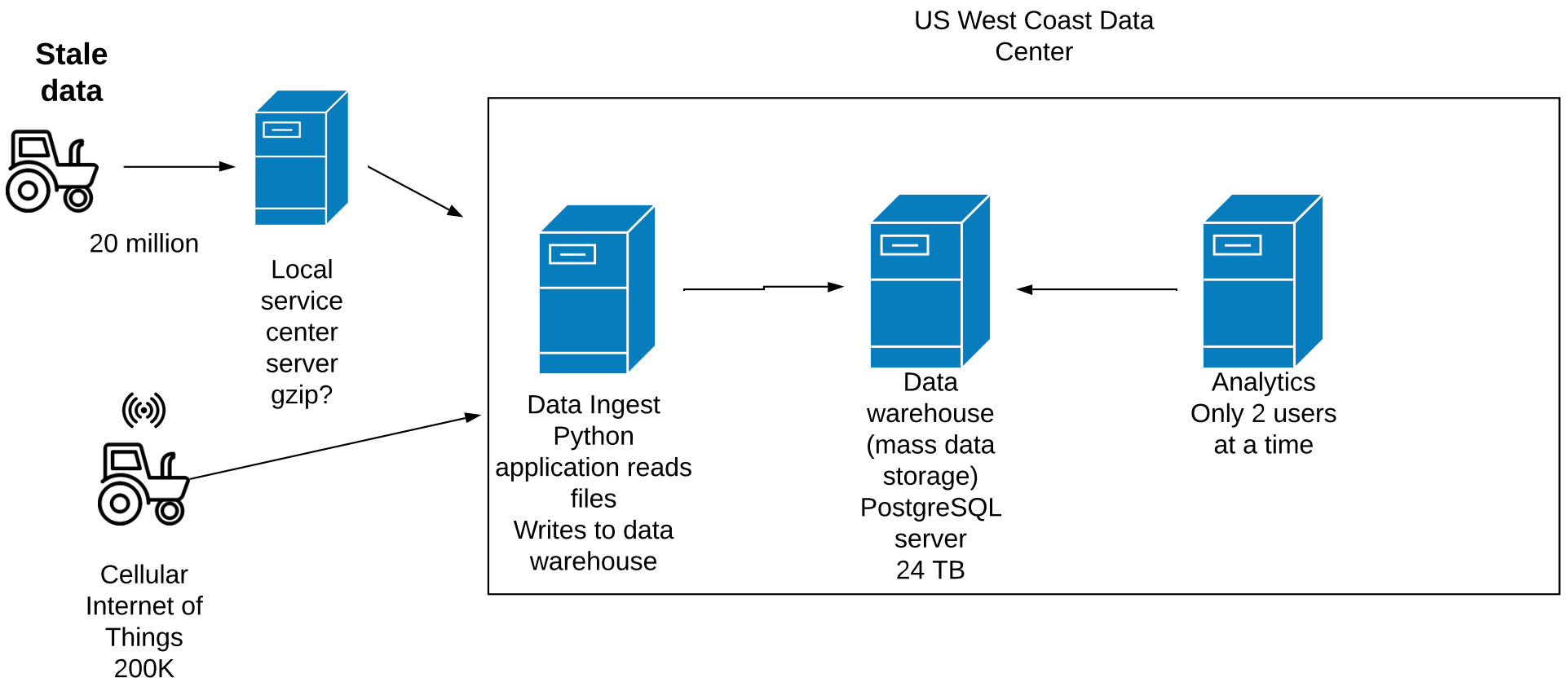
- Current environment allows TerramEarth to preemptively stock replacement parts = reduced downtime
- Not agile enough
  - 4 week turnaround for new parts
  - Results in customer downtime
- Limited access to reporting/analytics system

What are their goals?

- Top priority: Reduce vehicle downtime
  - Get data off the vehicle into analytics FASTER
- Share data with dealer network
- Facilitate partnerships with other companies

Current setup

- Collect analytics on vehicles
  - Increase vehicle efficiency
  - Predict breakdowns and pre-stage replacement parts
- 20 million vehicles - each collect 120 fields per second
  - Data stored locally, then uploaded (batch upload) when at dealer
  - Data sits on the vehicle until in for service - **This needs to change!**
  - Same port adjusts parameters
- 200,000 (10% of above) use cellular connection
  - Always streaming data
  - 9 TB per day total upload



Return to Table of Contents

Choose a Lesson

Case Studies Overview

Mountkirk Games

Dress4Win

TerramEarth

TerramEarth

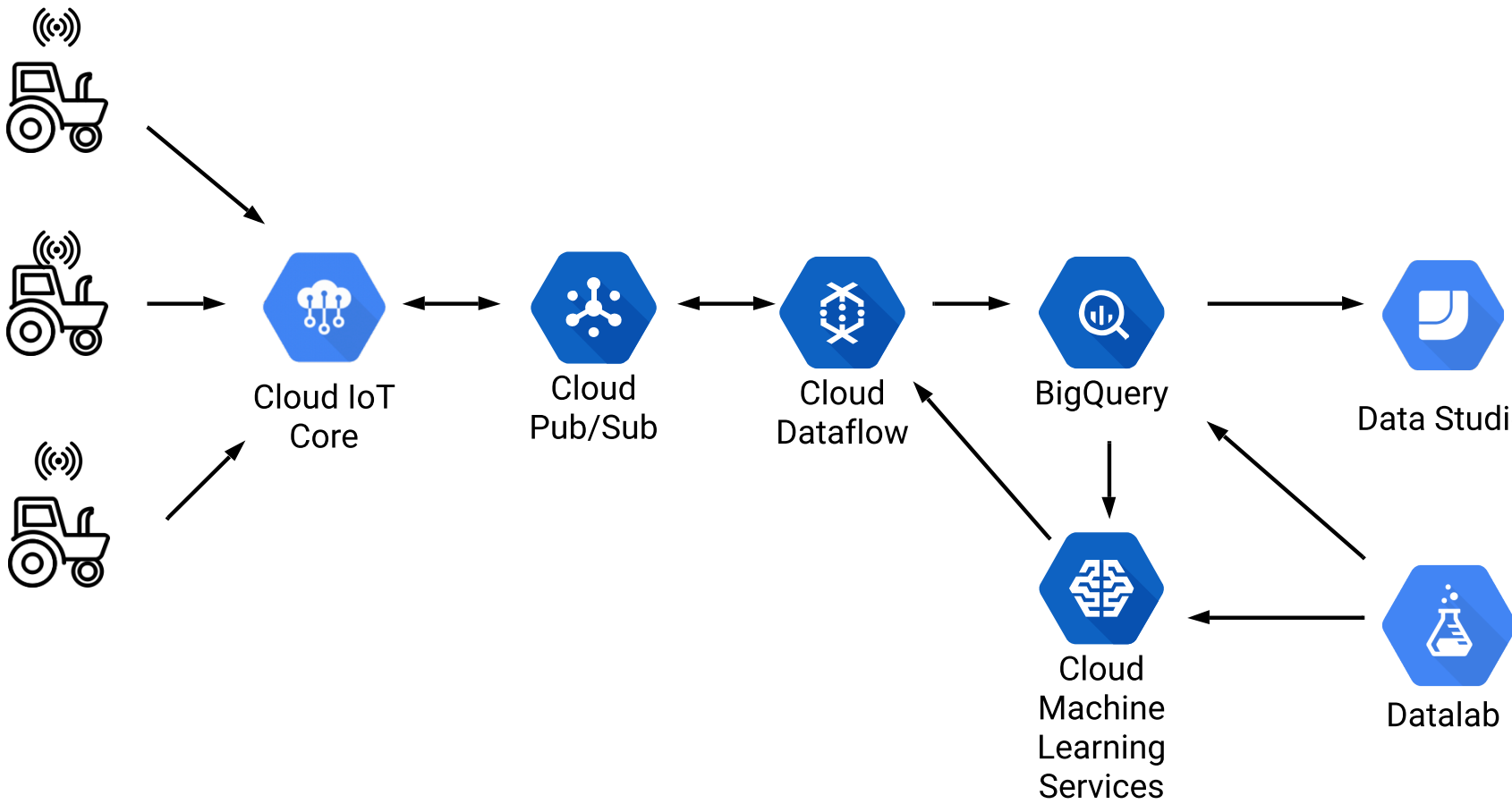
Previous

Next

What do they care about (big picture)?

- Business agility
  - Collect data faster
  - Act on data faster
  - Get results faster
- Maintain global footprint
- Improve dealer relationships
- Improve partnerships

Solution concept - convert to 100% cellular = no stale data



Cellular Solution breakdown:

- Manage cellular devices with **Cloud IoT Core** service
- Use **Cloud Endpoints** service to manage and protect APIs
- **Cloud Pub/Sub** handles ingest from all devices - global footprint
- **Cloud Dataflow** processes data from Pub/Sub, and inserts into **BigQuery** for storage/analytics
- Machine learning services (e.g., **Cloud ML Engine**) use data to predict breakdowns and optimize parameters
- ML services deploy updated parameters back to machines to update config
- Analytics data displayed in **Data Studio**, which can be shared to dealers via dashboards
- **Datalab** provides visual notebooks for working with BigQuery/Cloud ML Engine data for ML/analytics

Return to Table of Contents

Choose a Lesson

- Case Studies Overview
- Mountkirk Games
- Dress4Win
- TerramEarth

TerramEarth

Previous

Business and Technical Requirements

Business Requirements

- Decrease unplanned vehicle downtime to less than 1 week
  - Convert to 100% cellular connectivity
- Support the dealer network with more data on how their customers use their equipment to better position new products and services.
  - Share insights with **Data Studio**
- Have the ability to partner with different companies—especially with seed and fertilizer suppliers in the fast-growing agricultural business—to create compelling joint offerings for their customers.
  - Also share insights with **Data Studio**
  - **BigQuery/ ML** analytics to predict customer needs
  - Tech lead will enable partnerships

Technical Requirements

- Expand beyond a single datacenter to decrease latency to the American midwest and east coast
  - Multi-regional/global services
- Create a backup strategy
  - Regular BigQuery exports to Cloud Storage
- Increase security of data transfer from equipment to the datacenter
  - **Cloud Endpoints** - manage and protect APIs
  - **Cloud IoT Core** also managed security
  - Customer supplied encryption keys an option
- Improve data in the data warehouse
  - **Cloud Dataflow** - transform incoming streaming data to preferred format
  - Alternatively, stage in **Cloud Storage**, clean with **Cloud Dataprep**, and run job (backed by Cloud Dataflow) into **BigQuery**
- Use customer and equipment data to anticipate customer needs
  - Pair **BigQuery** with machine learning services for predictive analytics

Return to Table of Contents

Choose a Lesson

Return to Table of Contents

Choose a Lesson

Making the Case for the Cloud and GCP

Cost Optimization

Architecting Cloud Applications

Making the Case for the Cloud and GCP

Next

Think like a CEO

- Start thinking like upper leadership
- Role of Cloud Architect is to bridge the gap of technical and business requirements
- What do your CEO, CFO, CIO, etc care about?

Questions to ask when planning a cloud transition?

- What does Google Cloud Platform do that we can't do now?
- Why should we migrate our resources to GCP?
- Case studies as a reference.
- Start asking 'why?'
  - Why should your CEO, CFO, or CIO care?

The "why's" of moving to GCP

- Costs
- Future-proof infrastructure
- Scale to meet demand
- Data analytics/big data
- Greater business agility
- Managed services
- Global reach
- Security at scale

Everyone has different "why's"

Return to Table of Contents

Choose a Lesson

Making the Case for the Cloud and GCP

Cost Optimization

Architecting Cloud Applications

Making the Case for the Cloud and GCP

Previous

Next

Cost

- 'Catch-all' for other reasons
- Do more with less cost
- Trade CapEx for OpEx
  - No need to spend big $$$ up front on hardware investments

Future-proof infrastructure

- Hardware does not wear out (end of life).
- Migrating data to new hardware every few years is a pain!
  - Time consuming AND expensive

Scale to meet demand

- Elastic computing (a.k.a. distributed computing)
- Dynamically scale compute up and down as needed
  - Pay for only what you need, at that moment

Greater Business Agility

- ‘Need for speed’
- Create resources faster
- Act on data faster
- Do everything faster
- No waiting on hardware
- Rapid resource provisioning = greater flexibility/experimentation

Return to Table of Contents

Choose a Lesson

Making the Case for the Cloud and GCP

Cost Optimization

Architecting Cloud Applications

Making the Case for the Cloud and GCP

Previous

Managed Services

- Let Google manage infrastructure for you
- Less administrative overhead
- ‘Serverless’

Global Reach

- Easy worldwide presence
- Multi-national resources on same private network (VPC)

Security at Scale

- Economies of scale at work.
- Over 500 security engineers protecting your data.
- They’re really good at this.
- Ease of access management - projects

Going forward: match the "why's" with solutions

Return to Table of Contents

Choose a Lesson

Making the Case for the Cloud and GCP

Cost Optimization

Architecting Cloud Applications

Cost Optimization

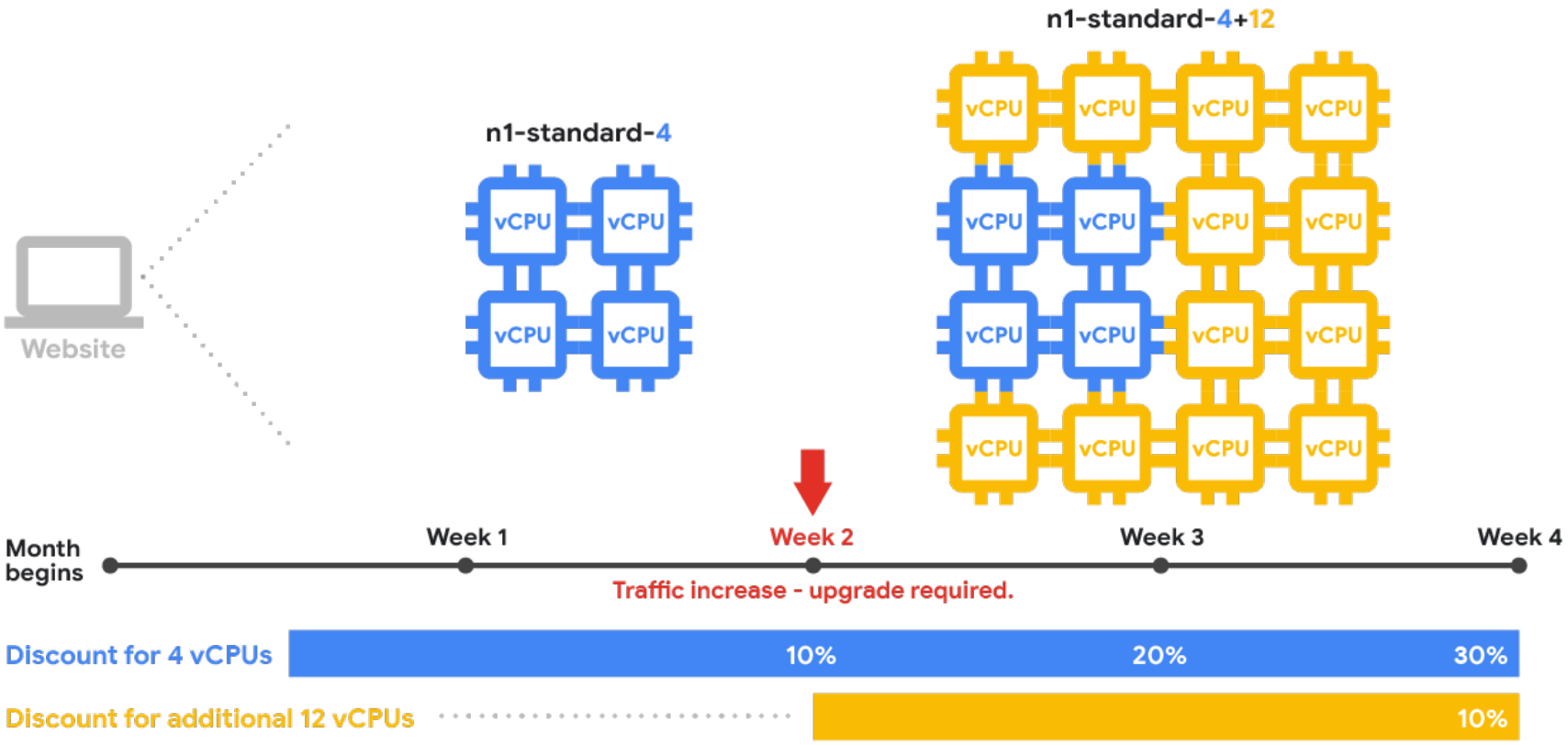
Next

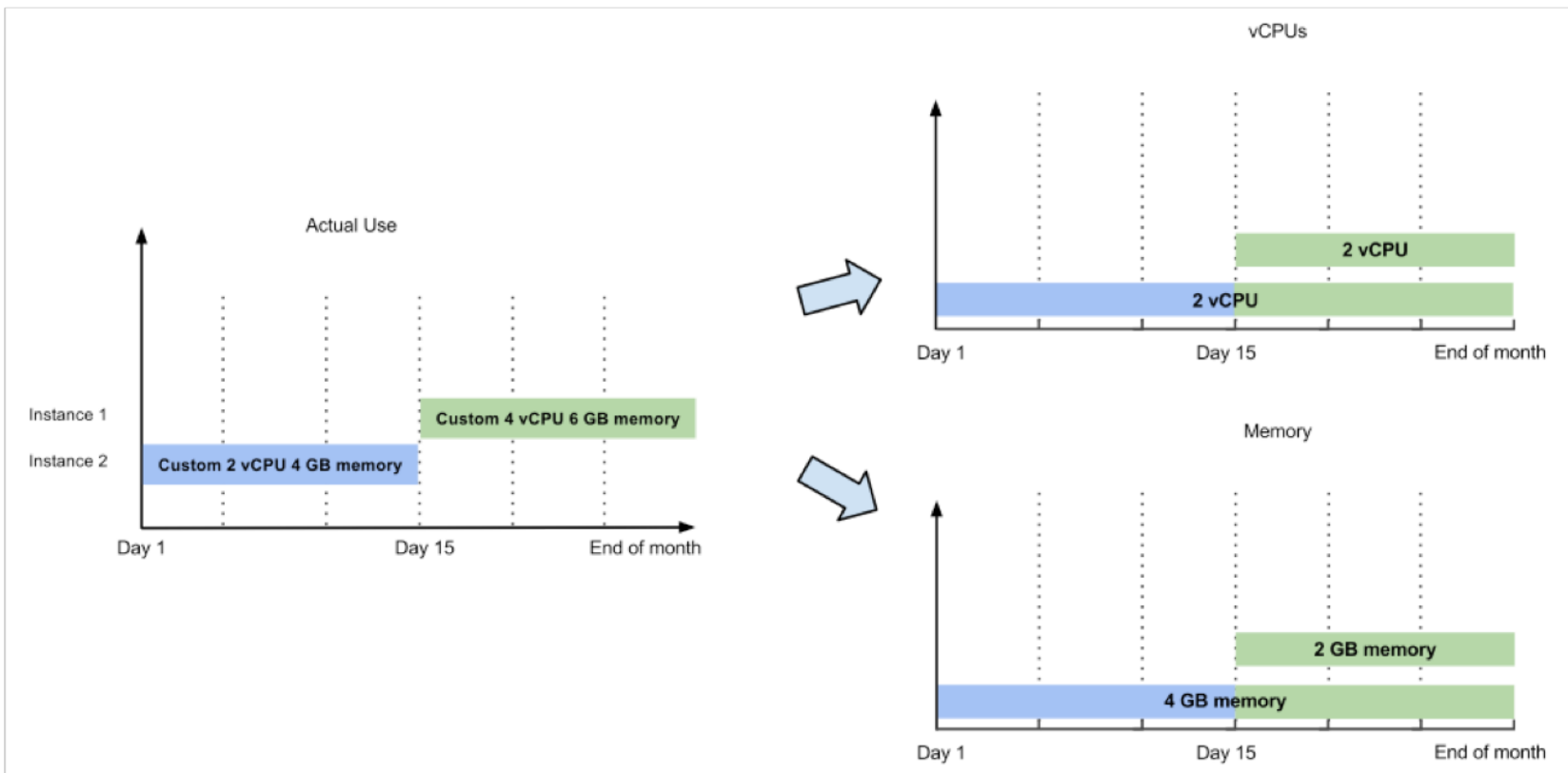
Why is this important?

- All businesses want to do more at less cost.
- 'Bang for your buck'
- GCP has many unique cost saving features to save money for more performance.
- May be testable

Sustained Use Discounts

- Up to 30% discount on Compute Engine and Cloud SQL VM's.
- The longer your compute (CPU/RAM/GPU) is used per month, the greater the discount.
- Discount calculation is inferred across different VM's throughout the month
  - Predefined and custom machine types calculated separately
- High flexibility/agility (both financial and technical)
- No up-front commitment
- Automatic, and simple - Google does the math for you





Return to Table of Contents

Choose a Lesson

Making the Case for the Cloud and GCP

Cost Optimization

Architecting Cloud Applications

Cost Optimization

Previous

Next

Free Tier

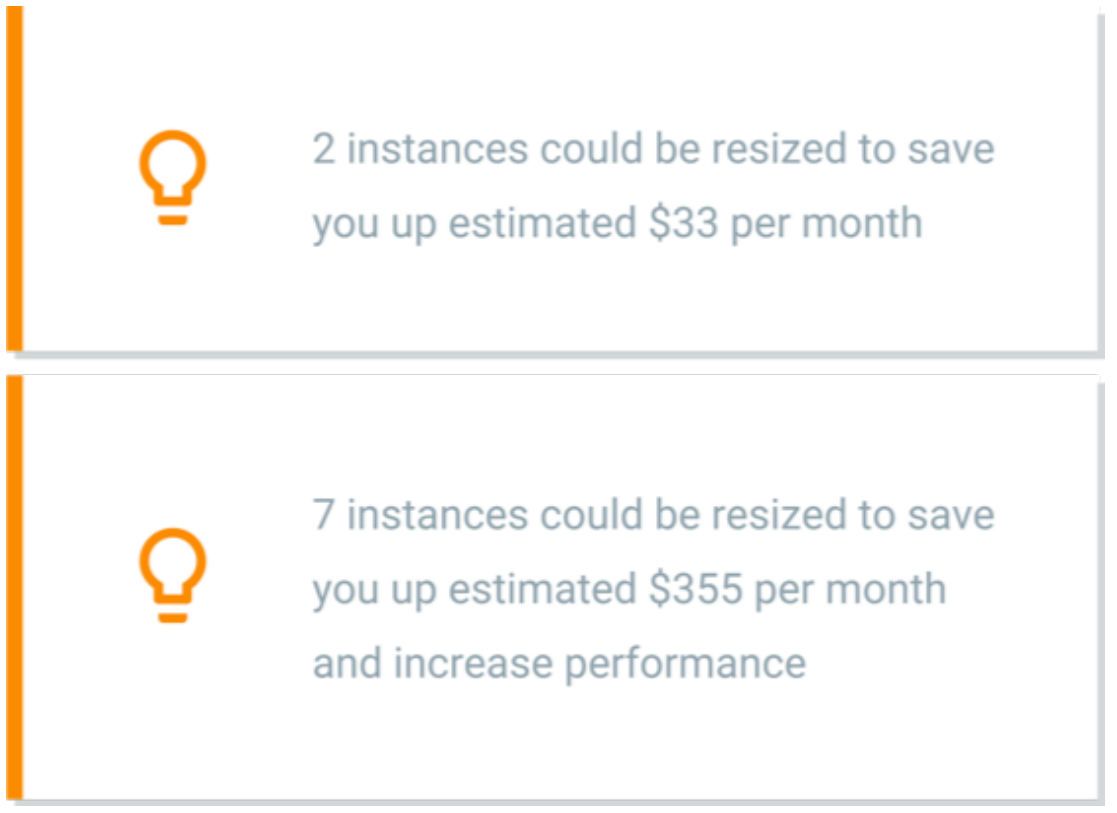
- Separate from free trial
- Low/reduced usage of many (not all) GCP products that are always free, all the time
- Any usage above the free tier usage limits charged at usual rate
- Good for testing out services without worrying about payment
- Automatic
- Use billing budgets and alerts to keep control of costs
- Does not contribute to sustained use discounts
- Learn more: https://cloud.google.com/free/

Custom Machine Types

- Unique to GCP
- Customize # CPU's and RAM amounts
  - Greater customization than predefined types.
- Choose from 0.9 to 6.5 GB RAM per CPU

Rightsizing Recommendations

- Automatically recommend machine type resizing for Compute Engine VM's.
- Takes last 8 days of usage.
- Recommend sizing up or down to increase performance/save costs.
- Recommends custom machine types where applicable



Return to Table of Contents

Choose a Lesson

Making the Case for the Cloud and GCP

Cost Optimization

Architecting Cloud Applications

Cost Optimization

Previous

Preemptible VM's

- Low-cost, short life, interruptible VM's
- Up to 80% discount
- Fixed price, not variable market price = easier to budget for
- Ideal for fault tolerant, batch processing workloads

Nearline and Coldline Cloud Storage

- Low cost, and very low cost cloud storage
- Ideal for archive/disaster recovery data
- Unique to GCP – low cost, but same fast access as premium storage

Committed Use Discounts

- Commit to 1 or 3 year term for set amount of CPU's/RAM ('pool' of CPU/RAM)
- Up to 57% discount
- Billed for CPU/RAM amounts whether or not they're used
- CPU/RAM pool can be used on multiple CE instances
  - Discount automatically applied
- Example: commitment of 10 vCPU's, 30 GB RAM
  - Two VM's with 4 vCPU/10 GB RAM each
  - One VM with 2 vCPU/10 GB RAM
  - Committed use discount applied to all machines
- Does not stack with sustained use discounts

Return to Table of Contents

Choose a Lesson

Making the Case for the Cloud and GCP

Cost Optimization

Architecting Cloud Applications

Architecting Cloud Applications

Next

App Design Requirements

- Five principles of good cloud app design:
  - High availability
  - Scalability
  - Security
  - Disaster Recovery
  - Cost
- Same principles regardless of compute platform (GCE/GKE/GAE)
- Understand how principles work across compute methods

High Availability

- “Can users access the application with minimal latency?”
- Placement of resources key.
- Who are your users?
  - Local company
  - Public users nationwide
  - International audience
- Regional deployment, multi-regional.
- Serve traffic to multiple regions via global load balancing.

Scalability

- More compute when needed, less when not
- Autoscaling
- Best practice - Run load tests
- GCE – Managed instance group with autoscaling
- GKE – Cluster with autoscaling enabled
- GAE – Autoscaler built-in
- GAE considerations
  - Standard – daily spending limit
  - Quota limits on API calls

Security

- Limit access to those who need it
- Principle of least privilege
- Secure administrative access
- IAM roles – limit personnel access
- Firewall rules to restrict traffic

Return to Table of Contents

Choose a Lesson

Making the Case for the Cloud and GCP

Cost Optimization

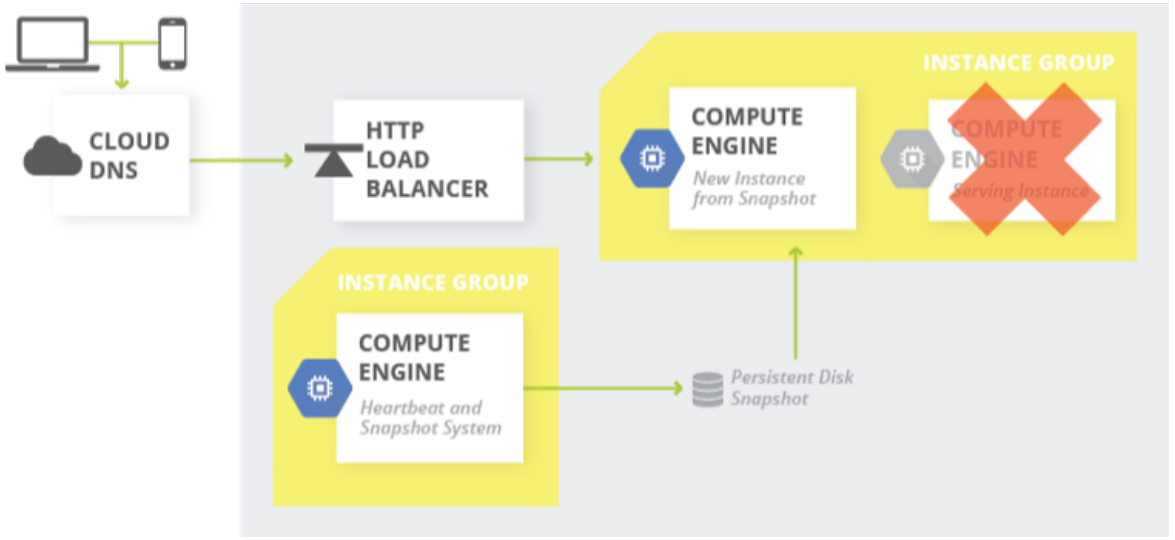
Architecting Cloud Applications

Architecting Cloud Applications

Previous

Disaster Recovery

- What to do when something goes 'boom'?
- Service not available = lost business
- GCE – snapshots for individual instances
- Failover server
- Backup data to cloud storage bucket
  - Database data
- Manage/rollback app versions:
  - GCE – instance group rolling update
  - GKE – rolling updates
  - GAE – traffic splitting/versions
- More detail later in course



Costs

- GAE (Standard)– set daily spend limit
- GAE (Flexible) – set custom machine types
- GCE – managed instance groups w/ autoscaling
- Custom machine types for perfect sized VM
- Preemptible VM's
- Resource Quotas to prevent accidental spikes in usage

Return to Table of Contents

Choose a Lesson

Return to Table of Contents

Choose a Lesson

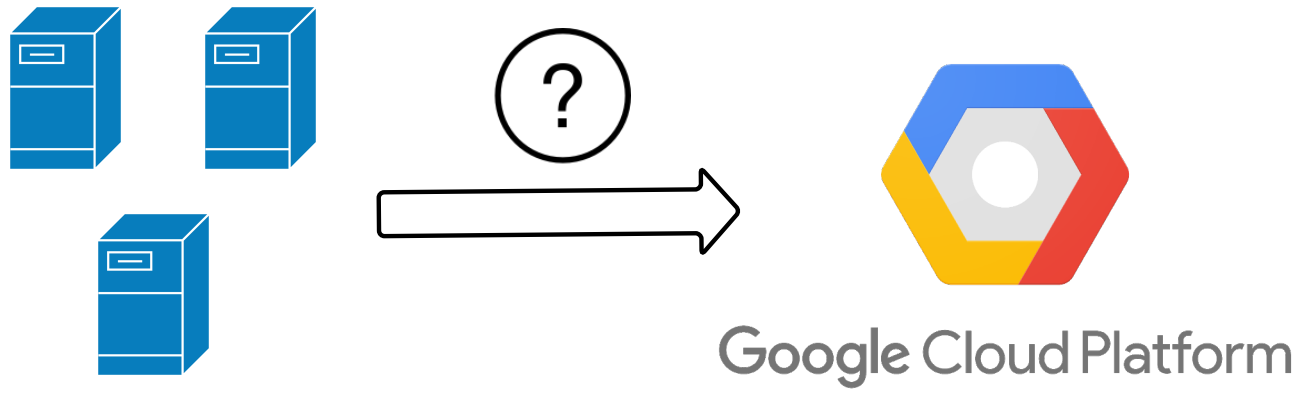
- Planning a Successful Cloud Migration
- Storage Transfer Service
- Data Migration Tips
- Migrating Applications

Planning a Successful Cloud Migration

How do you get your 'stuff' into GCP?

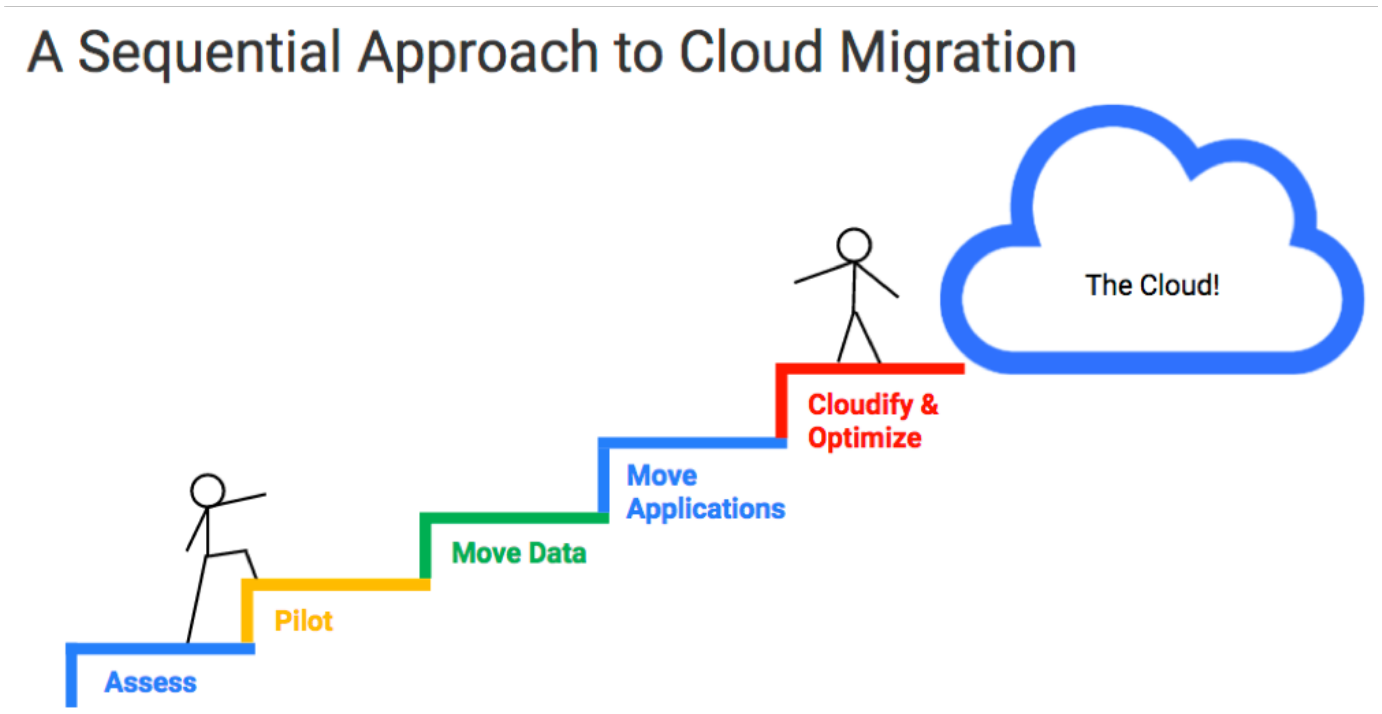
Next

- Stuff: Data (databases, unstructured storage) and applications



Five Phases for a Successful Cloud Migration

- Assess
- Pilot
- Move Data
- Move Applications
- Optimize



Return to Table of Contents

Choose a Lesson

- Planning a Successful Cloud Migration
- Storage Transfer Service
- Data Migration Tips
- Migrating Applications

Planning a Successful Cloud Migration

Previous

Next

Assess ("What should we move?")

- **Easy to move**
  - Newer
  - Fewer Dependencies
  - No/few licensing requirements
  - Tolerant to scaling
  - Great return on investment (ROI)
    - Example: Due for hardware replacement now
    - Example: Expensive to run on-premises
- **Hard to move**
  - More dependencies
  - Less tolerant to scaling
  - Complex licensing requirements
- **Can't move**
  - Require specialized or older hardware
  - Compliance requirements
  - Non-cloud compliant licensing (Oracle has been a common example)

Pilot - "baby steps"

- Proof of concept/test run
- Non-critical or easily duplicated services
- Small steps at first
- Considerations
  - Licensing
  - Roll back plan
  - Process changes
- Start mapping roles/structure
  - Projects
  - Separation of duties
  - Test/Production environments
  - VPCs

Service Type	Data Center	GCP
Compute	Physical hardware, virtualized hardware (VMWare ESXi,	Compute Engine
Storage	SAN, NAS, DAS	Persistent disk, Cloud Storage
Network	MPLS, VPN, hardware load balancing, DNS	Cloud VPN, CDN Interconnect, Cloud Load Balancing, Google Domains, Cloud DNS
Security	Firewalls, NACLs, route tables, encryption, IDS, SSL	Compute Engine firewalls, encryption, IDS, SSL
Identity	Active Directory, LDAP	IAM, GCDS, LDAP
Management	Configuration services, CI/CD tools	Cloud Deployment Manager, configuration services, continuous integration/continuous delivery (CI/CD) tools

Return to Table of Contents

Choose a Lesson

- Planning a Successful Cloud Migration
- Storage Transfer Service
- Data Migration Tips
- Migrating Applications

Planning a Successful Cloud Migration

Previous

Next

Time to REALLY move some stuff....what first?

Data first, then applications!

- Data before applications
- Evaluate storage options
- Transfer methods

Source	Destination	Tool(s)
On-prem data	Google Cloud Storage (GCS)	gsutil, transfer appliance, batch upload, drag and drop
On-cloud data (S3)	GCS	Storage Transfer Service
Database (SQL)	GCS/Cloud SQL/Spanner	Batch import mysqldump
Database (Non-SQL)	GCS	Batch upload to GCS
Database (Non-SQL)	Compute Engine	Backup files to persistent disk Stream to persistent disk

Return to Table of Contents

Choose a Lesson

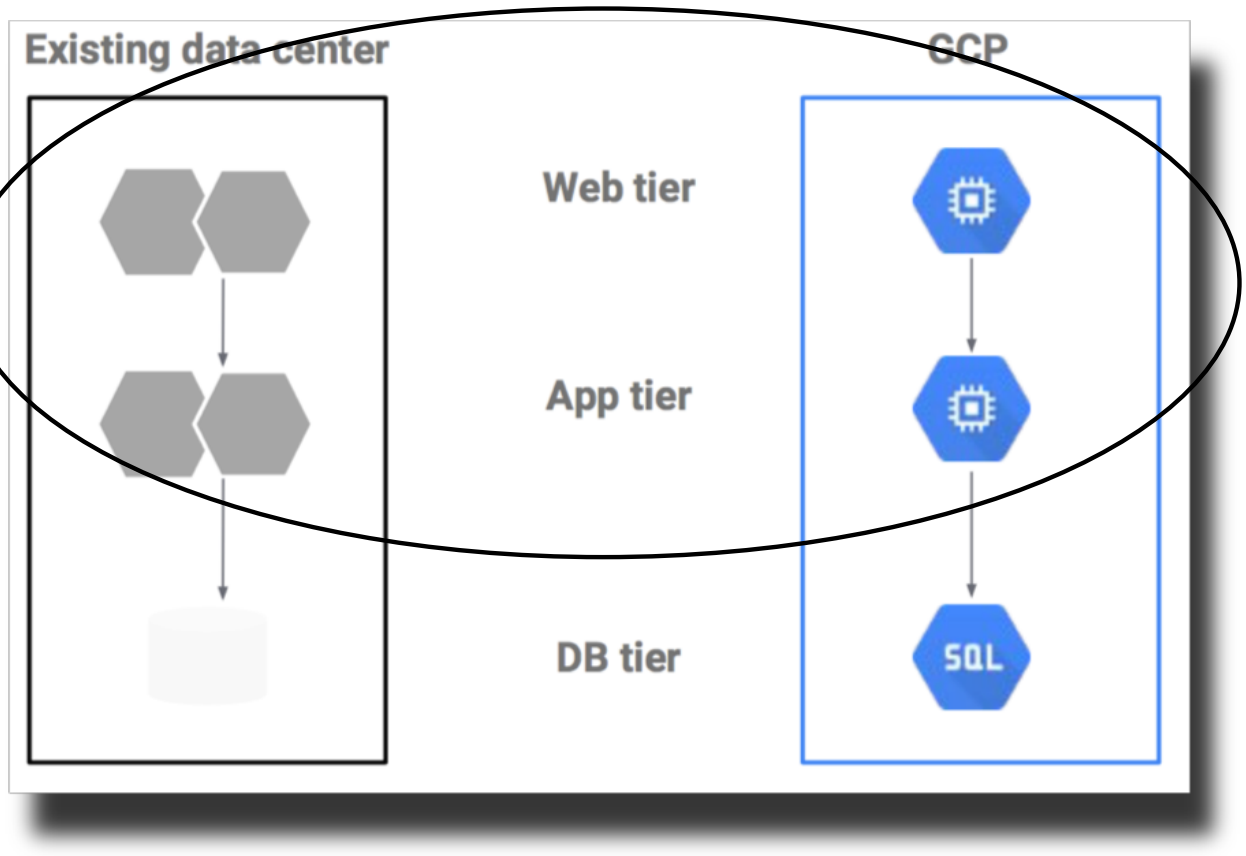
- Planning a Successful Cloud Migration
- Storage Transfer Service
- Data Migration Tips
- Migrating Applications

Planning a Successful Cloud Migration

Previous
Next

Move Applications

- Self service or partner assisted
- Keep it simple (usually) – 'lift and shift' recommended
  - Create duplicate environment of on-prem resources
  - Managed services?
- VM import freely available options via CloudEndure and Velostrata
- Other options:
  - Hybrid – resources in both environments
    - Hint: Dress4Win
  - Backup-as-migration
    - DR makes for a great migration cutover



Return to Table of Contents

Choose a Lesson

- Planning a Successful Cloud Migration
- Storage Transfer Service
- Data Migration Tips
- Migrating Applications

Planning a Successful Cloud Migration

Previous

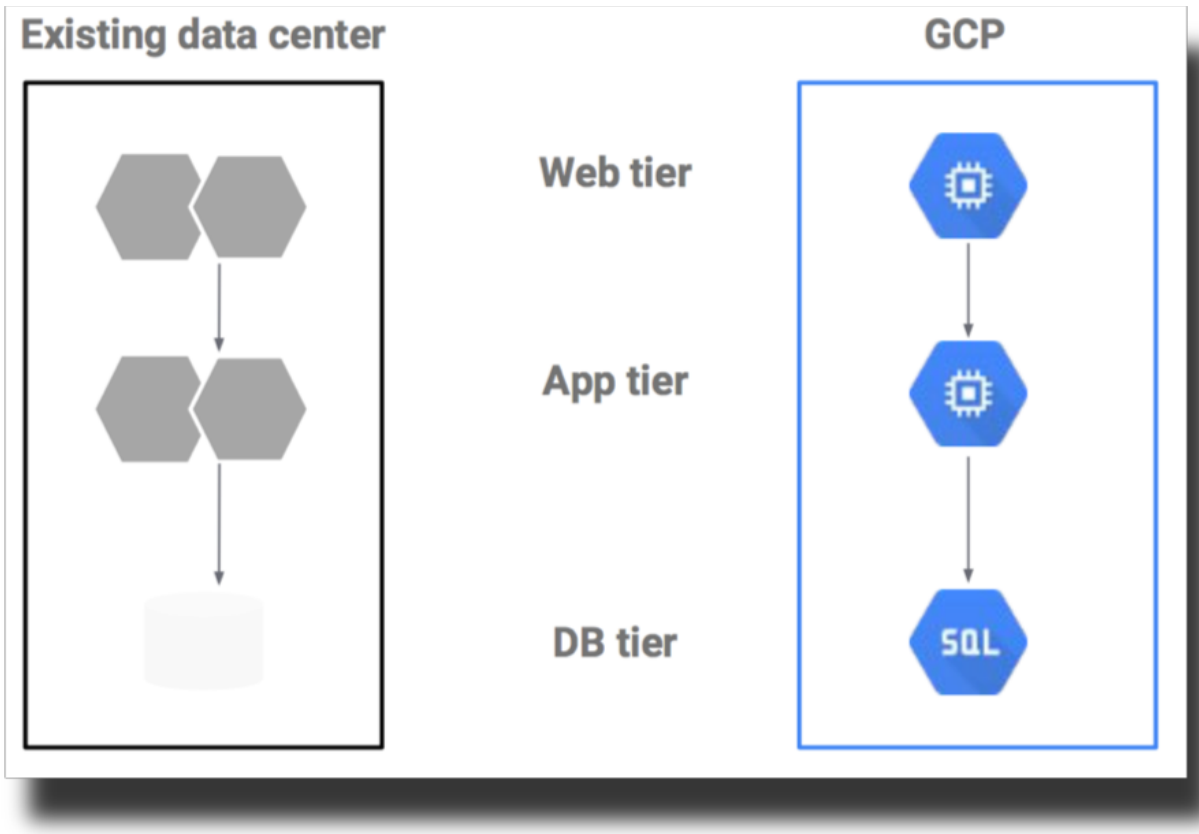
Next

Optimize!

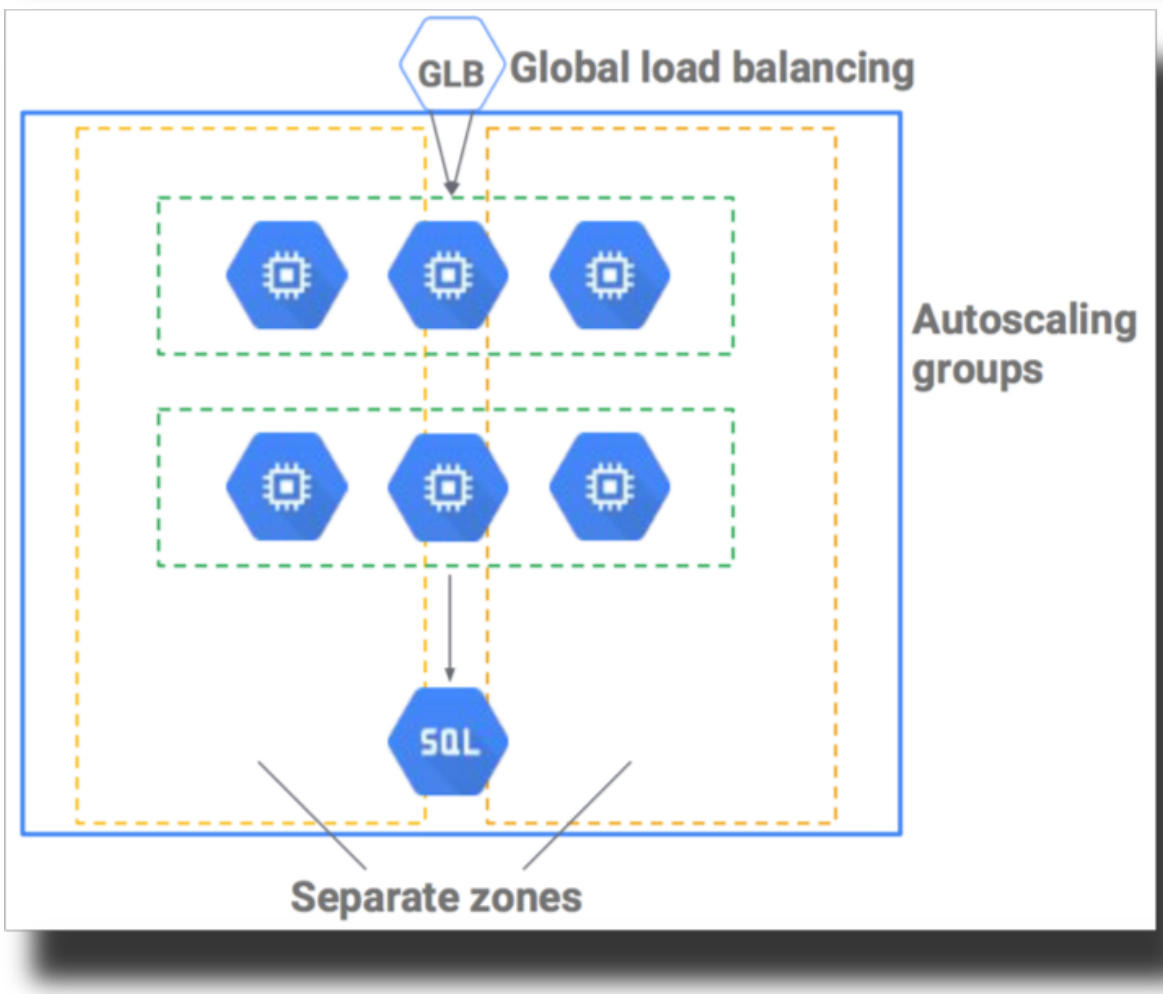
- Time for a cloud makeover!
- Re-tool processes and apps with modern GCP tools
  - Offload static assets to Cloud storage
  - Enable auto scaling
  - Enhance redundancy with different availability zones
  - Enhanced monitoring with Stackdriver
  - Managed services
  - How to launch future resources (with less baggage)
  - Decouple stateful storage from application

Start: Lift and shift into GCP

- Simple server stack (web tier, app tier, SQL database)



Add availability and elasticity



Return to Table of Contents

Choose a Lesson

Planning a Successful Cloud Migration

Storage Transfer Service

Data Migration Tips

Migrating Applications

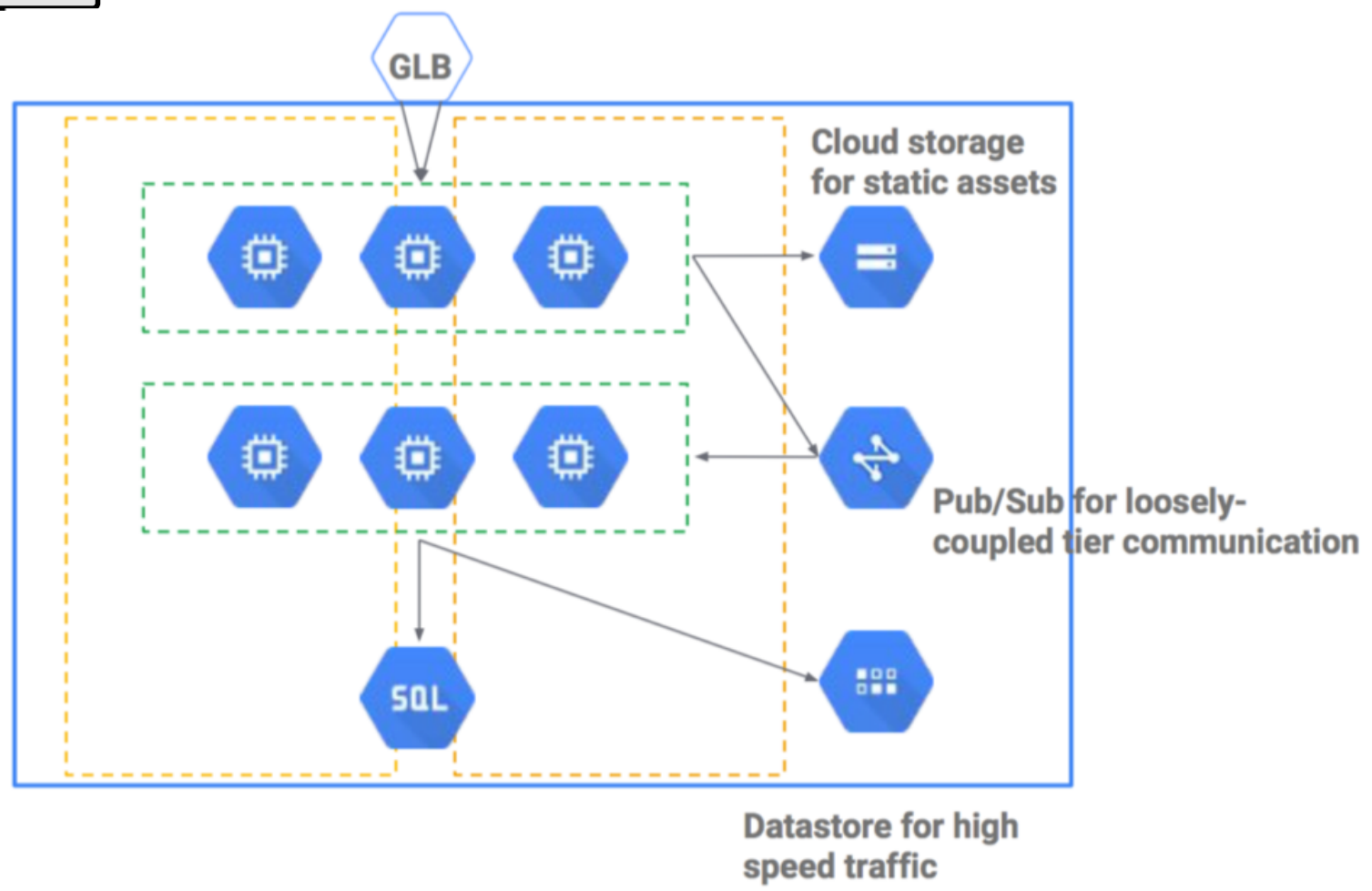
Planning a Successful Cloud Migration

Previous

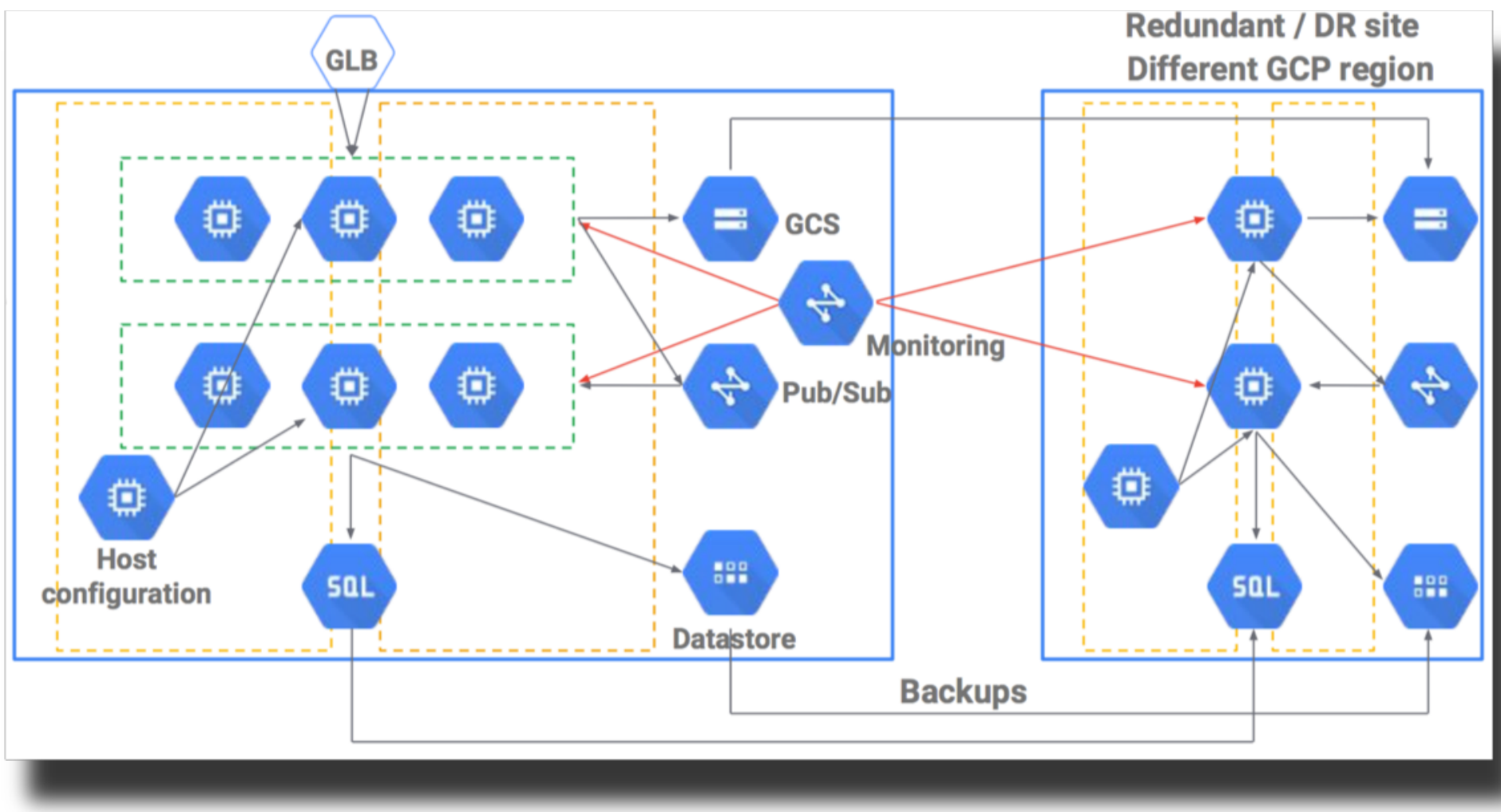
Next

Further optimization

- Consider storage options
- Convert to Datastore for high scale/high availability data
- Pub/Sub for inter-service communication



Full featured cloud solution



Return to Table of Contents

Choose a Lesson

Planning a Successful Cloud Migration

Storage Transfer Service

Data Migration Tips

Migrating Applications

Planning a Successful Cloud Migration

Previous

Additional Study Resources

- Five Phases of Migrating to Google Cloud Platform
- Best Practices for Migrating to Compute Engine

Return to Table of Contents

Choose a Lesson

Planning a Successful Cloud Migration

Storage Transfer Service

Data Migration Tips

Migrating Applications

Storage Transfer Service

What is it?

- Import online data into Cloud Storage:
  - AWS S3 bucket
  - HTTP/HTTPS location
  - Another Google Cloud Storage bucket
- Import from online data source (above) to data sink:
  - Sink = Google Cloud Storage bucket

What can it do?

- Back up data from other storage providers.
- Move data from one GCS bucket to another .
  - Example: Move from Multi-Regional bucket to Nearline bucket to lower costs.
- Configured through transfer job
  - One time or recurring transfer.
  - Delete destination objects if not present in source.
  - Delete source objects after transfer.
  - Periodic sync of data source and data sync.
- Requires owner or editor project IAM role + access to source and sink:
  - Source/sink can be outside of project.
  - Service account accesses source/sink.

gsutil or Storage Transfer Service?

- Cloud storage provider (GCS, AWS, HTTP) - use Storage Transfer Service.
- On-premises location - use gsutil.

Return to Table of Contents

Choose a Lesson

Planning a Successful Cloud Migration

Storage Transfer Service

Data Migration Tips

Migrating Applications

Data Migration Tips

Next

Moving lots of data - how "close" is it?

- From cloud = very close
  - Storage Transfer Service
- From colocation or on-premises datacenter = close
  - Fast bandwidth
  - Copy with gsutil (or third party tools)
- Slower connections = far
  - ‘Mail it in’ - Transfer Appliance

Physical Transfer — Physical / Online Transfer — Online Transfer

	1 Mbps	10 Mbps	100 Mbps	1 Gbps	10 Gbps	100 Gbps
1 GB	3 hours	18 minutes	2 minutes	11 seconds	1 second	0.1 seconds
10 GB	30 hours	3 hours	18 minutes	2 minutes	11 seconds	1 second
100 GB	12 days	30 hours	3 hours	18 minutes	2 minutes	11 seconds
1 TB	124 days	12 days	30 hours	3 hours	18 minutes	2 minutes
10 TB	3 years	124 days	12 days	30 hours	3 hours	18 minutes
100 TB	34 years	3 years	124 days	12 days	30 hours	3 hours
1 PB	340 years	34 years	3 years	124 days	12 days	30 hours
10 PB	3,404 years	340 years	34 years	3 years	124 days	12 days
100 PB	34,048 years	3,404 years	340 years	34 years	3 years	124 days

Return to Table of Contents

Choose a Lesson

Planning a Successful Cloud Migration

Storage Transfer Service

Data Migration Tips

Migrating Applications

Data Migration Tips

Previous

Next

How to speed up data transfer

- Decrease data size
  - Dedupe and compress
  - Both reduces transfer time and storage costs
  - For many small files, compressing and grouping together = faster transfers
- Increase network bandwidth
  - Public Internet connection
  - Cloud Interconnect

gsutil copy command considerations

- Limitations
  - No network throttling
  - Best for one time/manual transfers
  - For ongoing, automated transfers, use cron job
- Tools for faster/better transfers
  - Multi-threaded/processed. Useful when transferring large number of files.
  - Parallel composite uploads. Splits large files, transfers chunks in parallel, and composes at destination
  - Retry. Applies to transient network failures and HTTP/429 and 5xx error codes.
    - Implement exponential backoff method
  - Resumability. Resumes the transfer after an error
  - gsutil cp command will automatically attempt retries using exponential backoff

Return to Table of Contents

Choose a Lesson

Planning a Successful Cloud Migration

Storage Transfer Service

Data Migration Tips

Migrating Applications

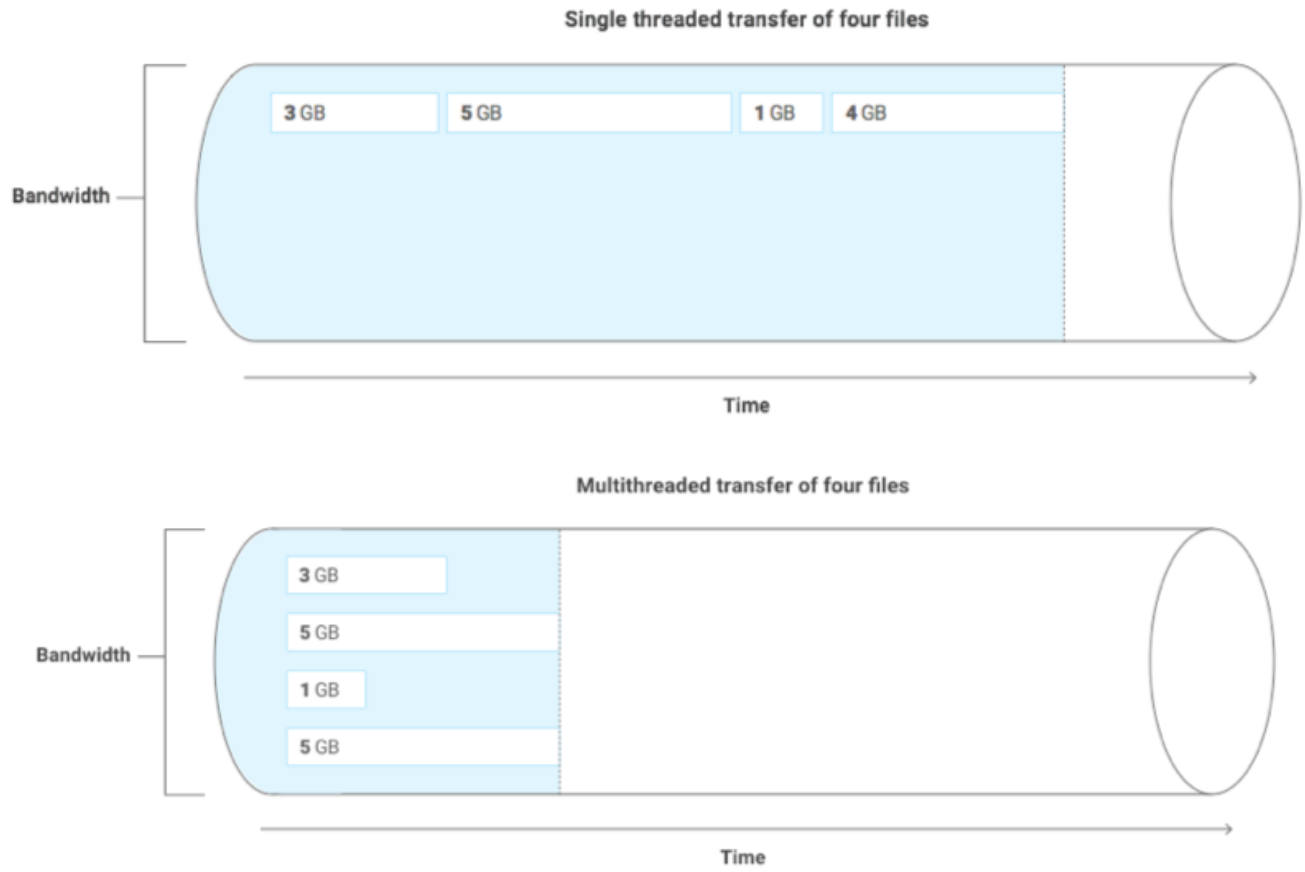
Data Migration Tips

Previous

Next

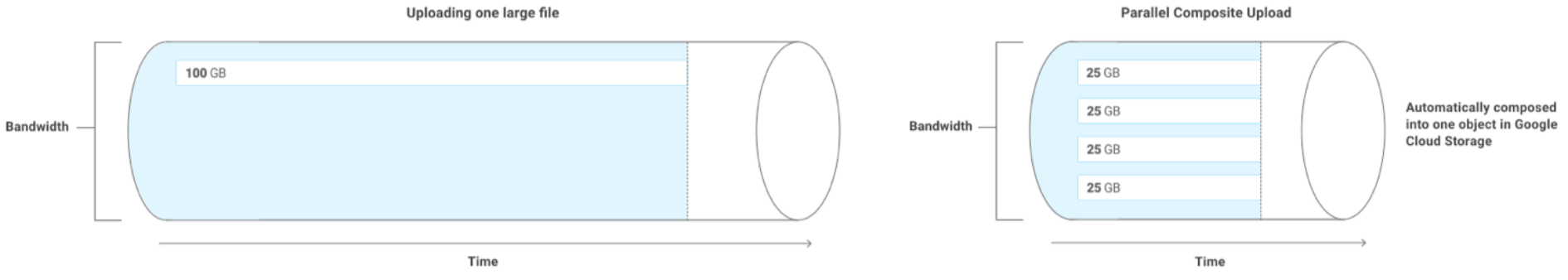
Multi-threading transfer/copy

- -m option
- gsutil -m cp -r [SOURCE_DIRECTORY] gs://[BUCKET_NAME]



Parallel Uploads

- Break single file into chunks for parallel upload
- Don’t use for nearline/coldline buckets – extra charge for ‘modifying’ files on upload
- gsutil -o GSUtil:parallel_composite_upload_threshold=150M cp bigfile gs://your-bucket



Return to Table of Contents

Choose a Lesson

Planning a Successful Cloud Migration

Storage Transfer Service

Data Migration Tips

Migrating Applications

Data Migration Tips

Previous

'Far' option - mail it in

Physical Transfer — Physical / Online Transfer — Online Transfer

	1 Mbps	10 Mbps	100 Mbps	1 Gbps	10 Gbps	100 Gbps
1 GB	3 hours	18 minutes	2 minutes	11 seconds	1 second	0.1 seconds
10 GB	30 hours	3 hours	18 minutes	2 minutes	11 seconds	1 second
100 GB	12 days	30 hours	3 hours	18 minutes	2 minutes	11 seconds
1 TB	124 days	12 days	30 hours	3 hours	18 minutes	2 minutes
10 TB	3 years	124 days	12 days	30 hours	3 hours	18 minutes
100 TB	34 years	3 years	124 days	12 days	30 hours	3 hours
1 PB	340 years	34 years	3 years	124 days	12 days	30 hours
10 PB	3,404 years	340 years	34 years	3 years	124 days	12 days
100 PB	34,048 years	3,404 years	340 years	34 years	3 years	124 days

Transfer Appliance

- Load up and mail your data in
- How it works:
  - Request Transfer Appliance
  - Load data onto physical appliance (data is encrypted)
  - Ship to Google
  - 'Rehydrate' encrypted data and place in Cloud Storage

Return to Table of Contents

Choose a Lesson

Planning a Successful Cloud Migration

Storage Transfer Service

Data Migration Tips

Migrating Applications

Migrating Applications

Server Migration

- Migrating from on-premises = migrating servers
- Application migration = server migration
- First assess what can be moved
- Map to additional GCP services

Before Moving the Server...

- Create a project
- Determine network configuration (VPC)
  - Firewall
  - Regions
  - Subnets
- Determine IAM roles – who needs access to what?

"Lift and Shift" Options

- Recreate server environment on Compute Engine public image
  - Create new GCE instance, install application/import data
- Import direct image
  - Carbon copy snapshot
- Recommended to run on GCE public image
- Reasons for direct image import:
- Require operating system that is not provided as a public image.
  - Already have a set of basic images that you use to create virtual machines in another cloud platform.
  - The work required to migrate application code to one of the public images is greater than the work required to complete the boot disk image import process.

Importing Boot Disk Images

- Manually create disk image file - Cloud Storage - import as custom image
  - Linux only
  - Compress into .tar.gz format (gzip compression)
- Newer option: Use GCE import tool to import virtual disk images
  - Supports VMDK and VHD formats
  - Linux and Windows
- Migrate entire server system with CloudEndure or Velostrata VM migration service
  - Managed service
  - Free service

Return to Table of Contents

Choose a Lesson

Return to Table of Contents

Choose a Lesson

Disaster Recovery Concepts

Backup and Recovery Methods in GCP

Disaster Recovery Concepts

Next

GCP Docs DR Guides

- Disaster Recovery Planning Guide
- Exam-related materials should be review

Always have a plan in case something goes 'boom'

- Network outage
- Application update breaks application
- Natural disaster
- Accidental data deletion
- A well-planned DR plan will ensure that if a catastrophe hits, you can recover in minimal time

Disaster Recovery Metrics - RTO/RPO

- Recovery Time Objective (RTO)
  - Maximum acceptable length of time that your application can be offline
  - Example: Linuxacademy.com website goes down, how long until it's back up and operational
- Recovery Point Objective (RPO)
  - Maximum acceptable length of time during which data might be lost from your application due to a major incident
  - Example: Linuxacademy.com subscription payment data is lost, how long until that data is back to its current state
- **Question to ask:** How quickly after a disaster does your application and application data need to be recovered in its previous state?
- The shorter your RTO/RPO needed, the higher the cost
  - Much greater complexity and overhead required
  - Time to recovery reflected by Service Level Agreements (SLA's)

Return to Table of Contents

Choose a Lesson

Disaster Recovery Concepts

Backup and Recovery Methods in GCP

Disaster Recovery Concepts

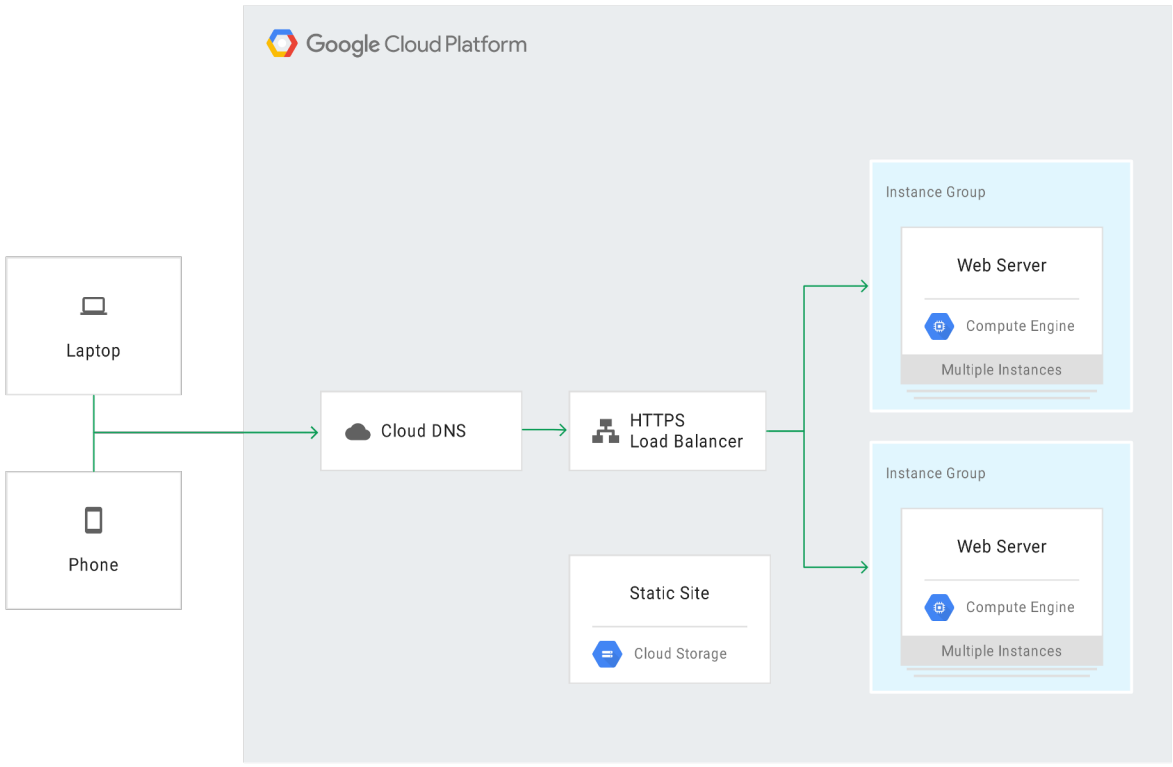
Previous

DR patterns - Cold/Warm/Hot

- How quickly you can recover if something goes wrong
- Specifics are beyond scope of course and exam, but good for foundational understanding
  - Link at lesson start goes into more detail
- Range from manually configure DNS switchover to backup site to 'hot' duplicated environments for instant switchover

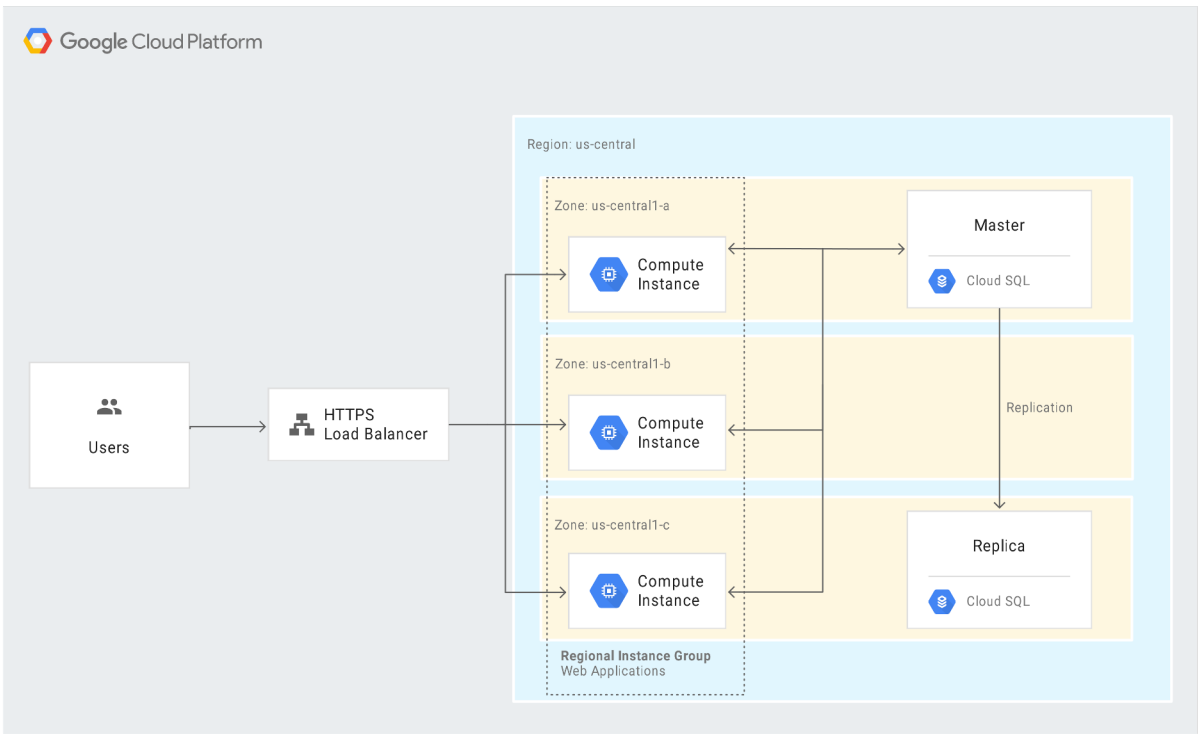
Warm failover

- Manual DNS switchover to static site
- Medium cost, medium recovery time



Hot failover

- Multi-zone instance group, replicated Cloud SQL instance
- Higher cost, shorter recovery time



Return to Table of Contents

Choose a Lesson

Disaster Recovery Concepts

Backup and Recovery Methods in GCP

Backup and Recovery Methods in GCP

Next

Exam Backup/DR topics

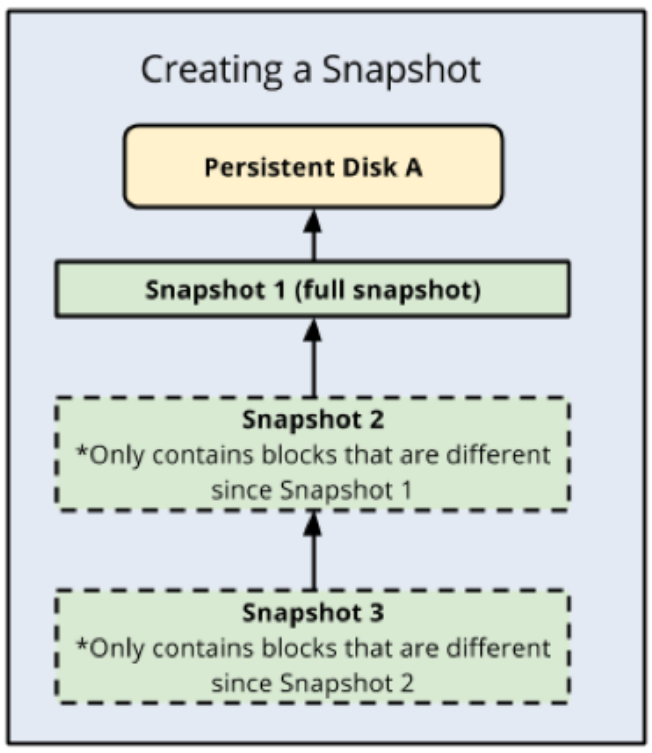
- Mostly review - backup/recovery focus
- Backup individual GCE instances
- Cloud Storage backup/rollback
- Distributed computing application roll back:
  - GCE managed instance group
  - App Engine
- Scheduling automated backups
- Exam scenarios

Backup of individual GCE instances

- Disk Snapshots
- Copy of entire disk
- Incremental:
  - Only backs up what's different from the previous backup

Scheduling Snapshots

- cron job on instance
- Snapshot scheduler



Cloud Storage Backup/Rollback

- Object versioning + lifecycle management:
  - gsutil versioning set on gs://[BUCKET_NAME]
- Delete + rollback protection
- Revert to earlier version

Return to Table of Contents

Choose a Lesson

Disaster Recovery Concepts

Backup and Recovery Methods in GCP

Backup and Recovery Methods in GCP

Previous

Distributed Computing Application Rollback

- GCE managed instance group + Google App Engine
- Individual instances are not backed up
  - No snapshots
- **Rollback to previous app versions**
  - Compute Engine Managed instance group
    - Rolling update - apply previous instance group template
    - Optional: set target % other than 100
  - App Engine
    - Versioning control/split traffic

Exam Scenarios

- Rollback plan for managed instance group serving website – 100's of instances:
  - Object versioning on static data in Cloud Storage
  - Rolling updates
  - NOT snapshots
- App Engine – need to push risky update to live environment:
  - Versioning/traffic splitting
  - Deploy update to small % of traffic as canary update

Return to Table of Contents

Choose a Lesson

Return to Table of Contents

Choose a Lesson

Security Methods in GCP

Network Design for Security and Isolation

Audits and Compliance

Security Methods in GCP

Purpose of Section

- Review of security principles, with some new items
- Exam focus of security scenarios may be tested
- IAM roles play substantial role in most scenarios (but not all)

Separation of Duties

- Don't place all organization resources in one project:
  - Difficult to limit access
- Principle 1: Different scopes of access = separate projects:
  - Example: Separate development and production environments by team
  - Individual user accounts – separate Dev and Prod projects for each team
- Principle 2: Give fewest rights necessary – organization and project levels:
  - Example: Security team needs detailed visibility to all projects in organization
  - Only need to view org-wide = organization viewer/project viewer

Securely Interact with Google Cloud Storage

- Three access control methods:
  - IAM
  - ACL
  - Signed URL
- IAM = Bucket level permissions
- ACL/Signed URL = Object level permissions
- Signed URL does not require GCP account (user uploads)
  - Example: External customer upload PII data to GCS with timed access – does not have GCP account:
    - Signed URL for secure access to only their data
    - Expires after set period of time

Penetration Testing ('pen test')

- Simulated attack on your computer system to find vulnerabilities
- Find holes before the bad guys do
- Exam focus: Choose correct environment to conduct pen test:
  - Pen test should be same avenue as a real attack
  - Example: Publicly available application should test from outside GCP over public Internet

Supporting Encrypted Communications

- Applies to secure communications
- Chat apps, IoT devices
- Use Public/Private Key format
  - Public Key Infrastructure (PKI) is one popular method
  - IoT Core uses Message Queuing Telemetry Transport (MQQT)

Return to Table of Contents

Choose a Lesson

- Security Methods in GCP
- Network Design for Security and Isolation
- Audits and Compliance

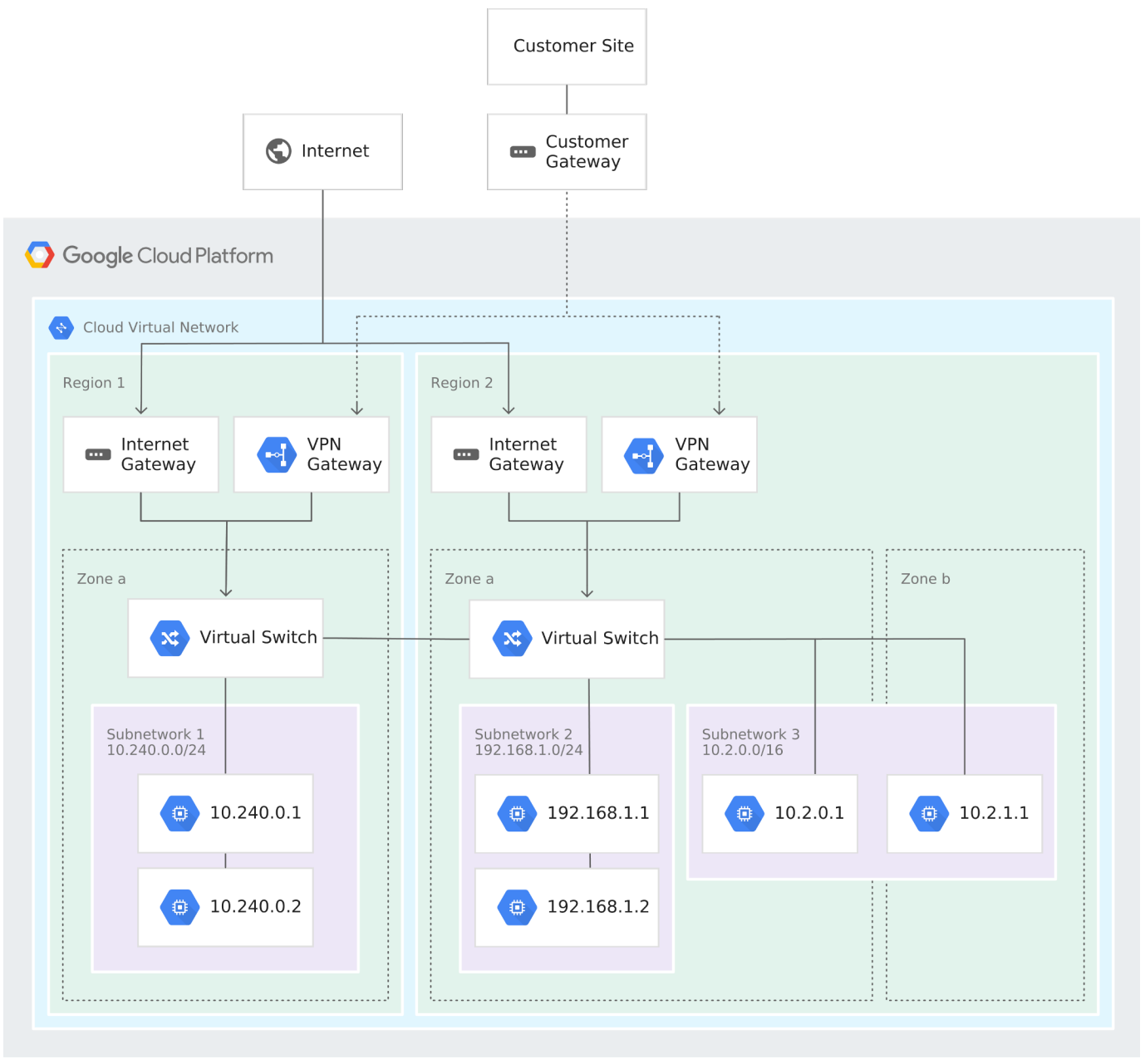
Network Design for Security and Isolation

Next

Enterprise Focus

- Increased layers of complexity
- Exam focus/scenarios

VPCs can get quite complex, but have a few key fundamentals



Return to Table of Contents

Choose a Lesson

- Security Methods in GCP
- Network Design for Security and Isolation
- Audits and Compliance

Network Design for Security and Isolation

Previous

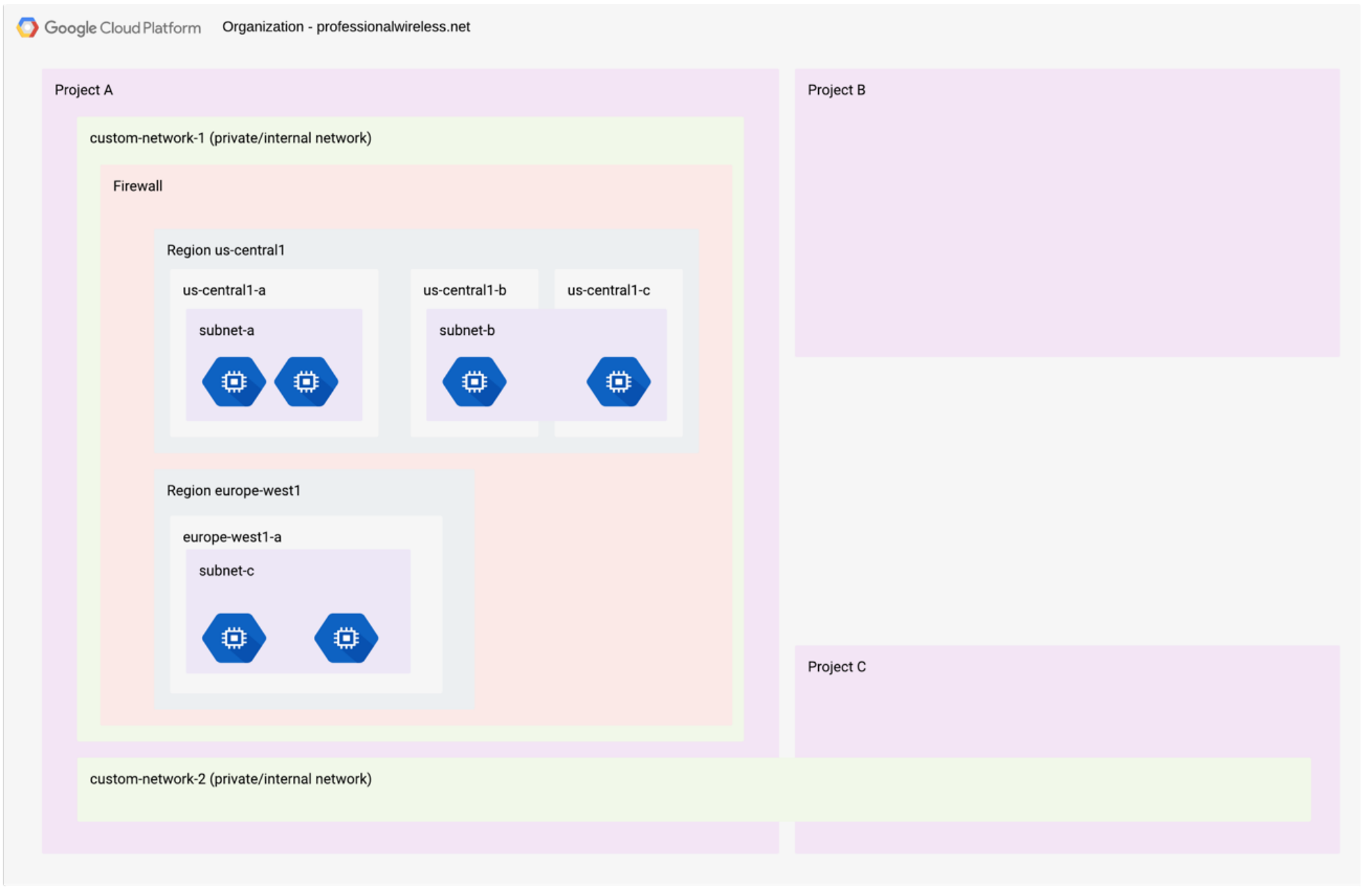
Next

Tools at a Glance

- Methods of isolation/separation, to include:
  - Projects (includes IAM)
  - VPCs
  - Firewall
- Principle of least privilege

Gotta keep them separated!

- Nested structure
- Organization → Projects → VPC → Regions → Subnets
- Organizations have one or more Projects
- Projects have one or more VPCs
- VPCs can span projects via Shared VPC
- VPCs include one or more Regions
- Regions can have one or more subnets



Return to Table of Contents

Choose a Lesson

- Security Methods in GCP
- Network Design for Security and Isolation
- Audits and Compliance

Network Design for Security and Isolation

Previous

Next

Projects

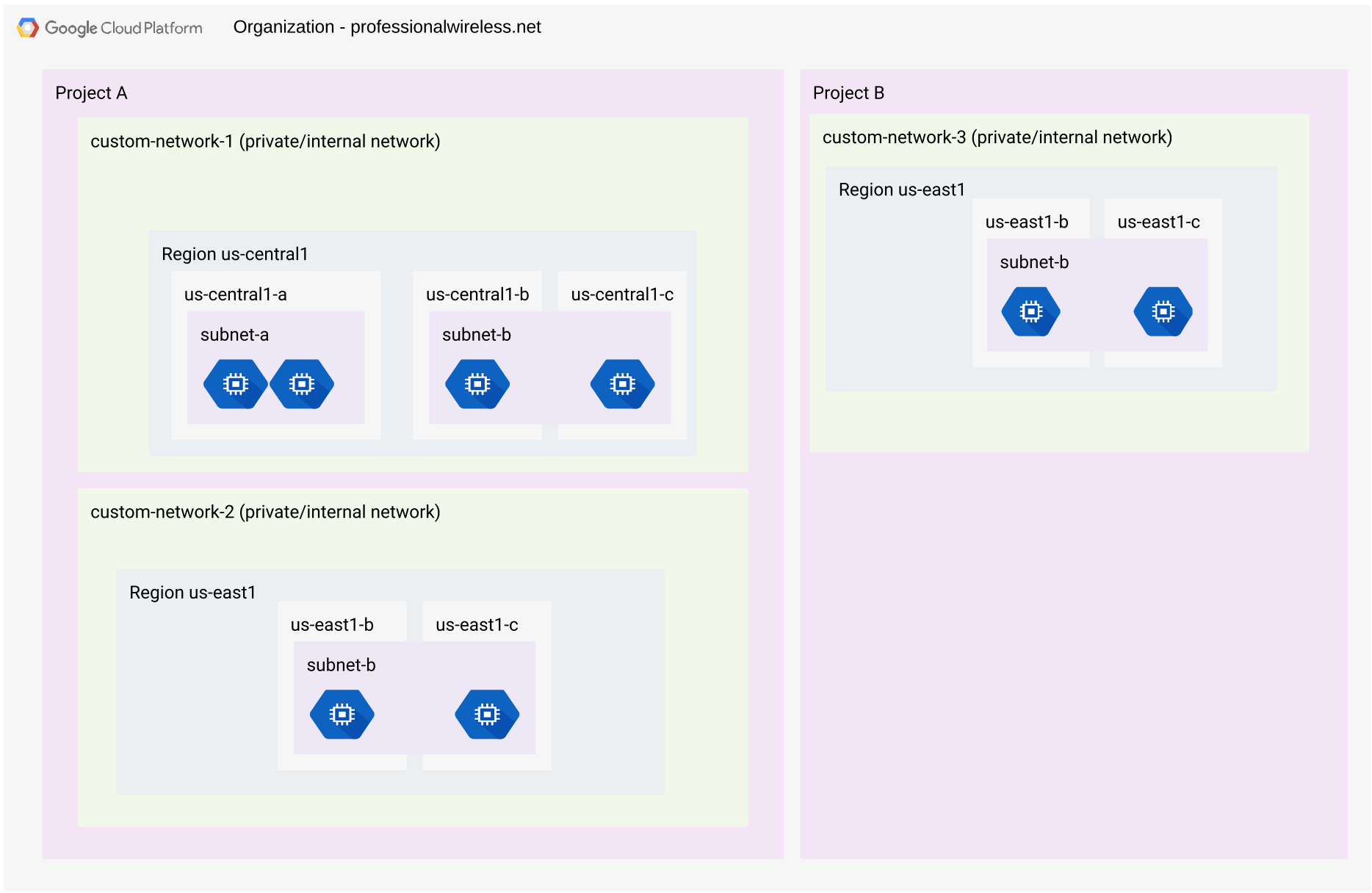
- What do they separate?
  - Separation of people/accounts - IAM to restrict access to project resources:
- Further separation by service, but not by VPC
- Primary method of full isolation between environments
- Projects further divided into VPCs
- Project-wide IAM roles = access to all VPCs within project
  - Example: Project A has 2 VPCs
  - Cannot grant access to one VPC, but not another in same project

User 'Bob' granted Network Admin role to Project A

Bob has rights to both custom-network-1 and custom-network-2 in Project A

Cannot grant access to one and not the other

Bob does not have rights to custom-network-3 in Project B



Return to Table of Contents

Choose a Lesson

- Security Methods in GCP
- Network Design for Security and Isolation
- Audits and Compliance

Network Design for Security and Isolation

Previous

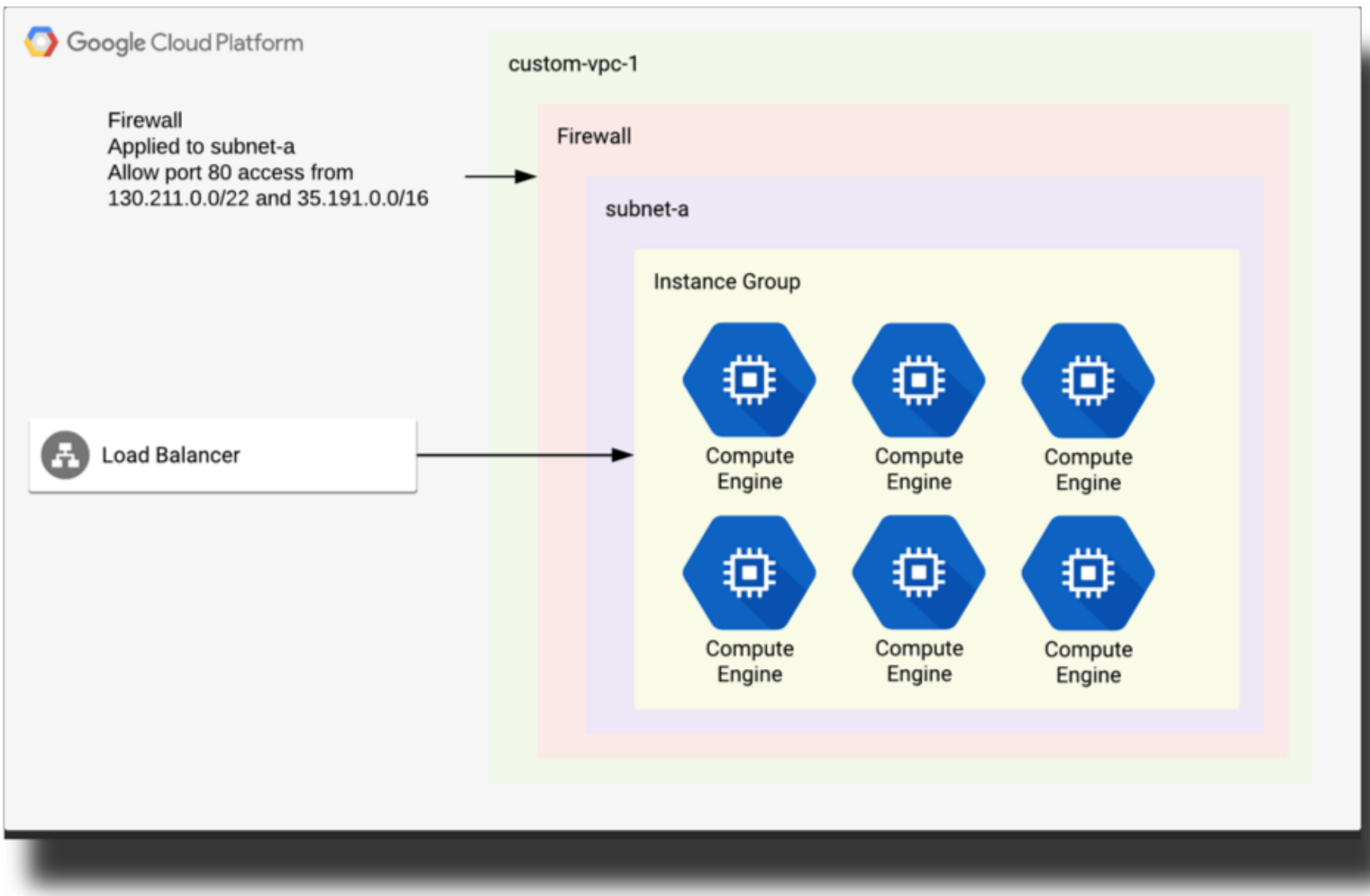
Next

Virtual Private Cloud (VPC)

- What do they separate?
  - Separation of resources (e.g., groups of VM instances)
- Can have multiple VPCs per project
- All VPC resources in same private IP network (RFC 1918 space):
  - Global access to same private network
- Project users have same access to all VPCs:
  - Granular IAM roles divide access by GCP service, not VPC
  - Cannot give Bob access to custom-network-1 in Project A, but restrict access to custom-network-2 in Project A

Firewall

- What do they separate?
  - Separate access by network traffic and location
- Control both ingress and egress traffic
- Limit access by:
  - Port
  - IP address/range (internal only, all external, certain IP ranges)
  - Between subnets
  - Tags
  - Service accounts
- Load balancer/health check interactions:
  - Firewall controls access at instance level, not load balancer
  - Must allow load balancer traffic to connect to backend instances (also allows health check)
  - Network Load Balancer = 209.85.152.0/22, 209.85.204.0/22, and 35.191.0.0/16
  - HTTP(S)/SSL proxy/TCP proxy/internal LB = 130.211.0.0/22 and 35.191.0.0/16



Return to Table of Contents

Choose a Lesson

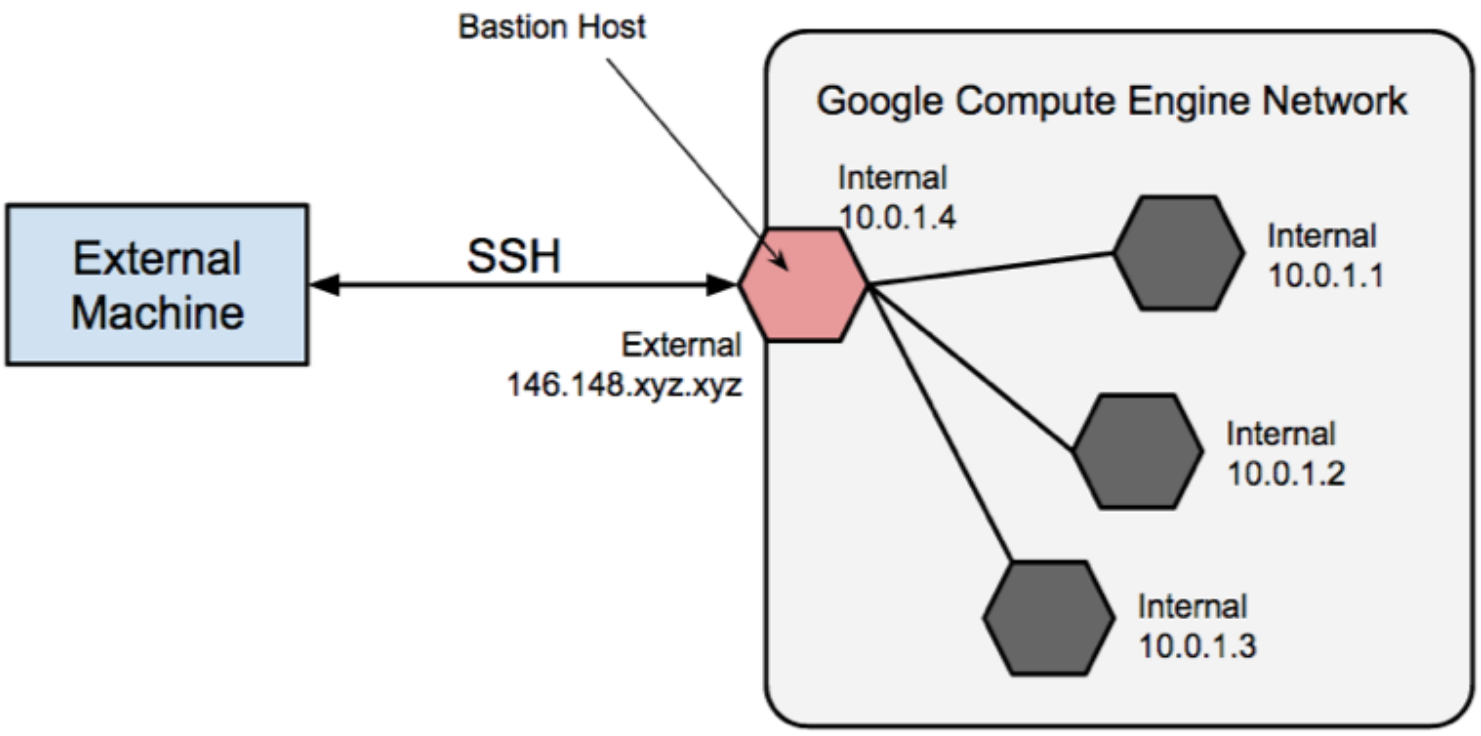
- Security Methods in GCP
- Network Design for Security and Isolation
- Audits and Compliance

Network Design for Security and Isolation

Previous

Further Network Isolation - Disable External IP address

- 'Air gap' resources by removing external IP address
  - Greatly limits exposure
- Problem: Can't administer air gapped resources
- Two primary options:
  - VPN connection – place myself on internal network
  - Bastion host
- Additional new managed option option: Cloud NAT service



Exam Scenarios

- Instance group VMs keep restarting every minute:
  - Failing health check
  - Configure firewall to allow proper access to instance group VMs (subnet, tag) from load balancer IP
- On-premises network access to proper network resources:
  - Restrict ingress firewall access to on-premises network IP range
- Failover from on-premises load balancer hosted application to GCP hosted instance group ('hot standby'):
  - Consider security and compliance
  - Allow firewall access at instance group level (subnet/tag) from outside source

Return to Table of Contents

Choose a Lesson

Security Methods in GCP

Network Design for Security and Isolation

Audits and Compliance

Audits and Compliance

Next

Exam Outline can be Confusing

- Designing for legal compliance. Considerations include:
  - Legislation (e.g., Health Insurance Portability and Accountability Act (HIPAA), Children's Online Privacy Protection Act (COPPA), etc.)
  - Audits (including logs)
  - Certification (e.g., Information Technology Infrastructure Library (ITIL) framework)
- Of the above, **audits** will be a frequently seen topic.

Audits

- If you see the terms audit, auditor, access logs, or compliance, think **Stackdriver Logging**
- Audits = Stackdriver

HOWEVER!

- Similar but different: Billing data is not through Stackdriver:
  - Billing data exported directly to Cloud Storage/BigQuery

Automating/Exporting Logging Data for Audits

- Scenario: Make audit data available for auditors
- Scenario solution:
  - Create sink for needed log data.
  - Export to Cloud Storage
  - Provide auditor access to export location (GCP account/signed URL)

Why Cloud Storage over BigQuery?

- Context: Long term storage of logs with infrequent audits
- Cloud Storage more cost effective, and can be exported or queried by BigQuery if audit need comes up

Return to Table of Contents

Choose a Lesson

Security Methods in GCP

Network Design for Security and Isolation

Audits and Compliance

Audits and Compliance

Previous

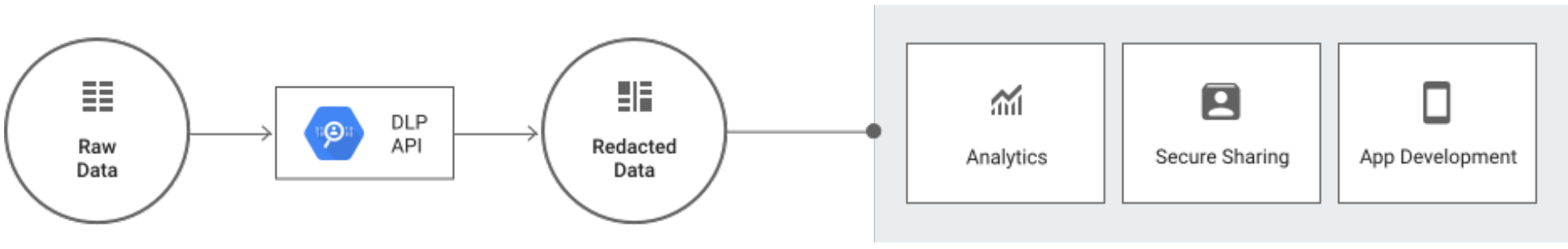
Next

Compliance

- HIPPA, COPPA, GDPR, etc.
- Shared responsibility model
  - Most GCP services meet most compliance standards
  - Customer needs to take steps to meet compliance in their applications
  - Previously covered security measures help ensure compliance
    - IAM, Stackdriver Logging, 2FA authentication
  - Verify GCP compliance certifications at below link:
  - https://cloud.google.com/security/compliance

Data Loss Prevention (DLP) - Sanitize your data

- Problem: Created data has sensitive, personally identifiable information (PII) data that needs to be removed/redacted
  - Internal logs, customer documents, etc.
  - PII data: Credit card numbers, social security numbers, names, phone numbers, etc.
- DLP automatically discovers, classifies, and sanitizes
- Related to meeting compliance requirements
- Scenario: Need to redact customer credit card info/logs with sensitive data



Return to Table of Contents

Choose a Lesson

Security Methods in GCP

Network Design for Security and Isolation

Audits and Compliance

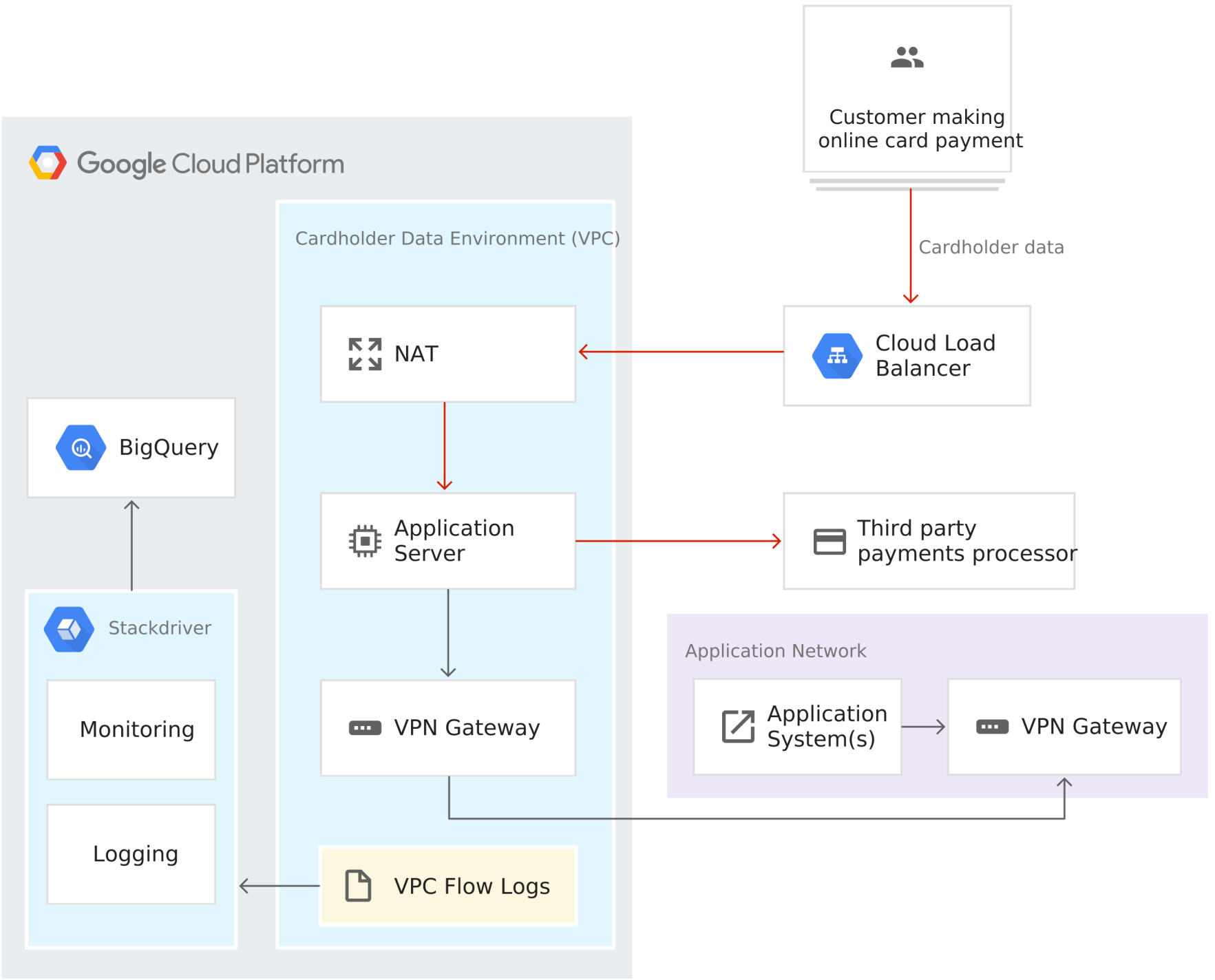
Audits and Compliance

Previous

Next

Analyze PCI Data (Credit Card Data)

- Payment Card Industry Data Security Standard (PCI DSS)
- Securely handle credit card information
- Exam focus: compliant GCP services, stream to BigQuery for analysis
- Compliant services: Almost all services are compliant (refer to GCP's compliance page for verification)



Return to Table of Contents

Choose a Lesson

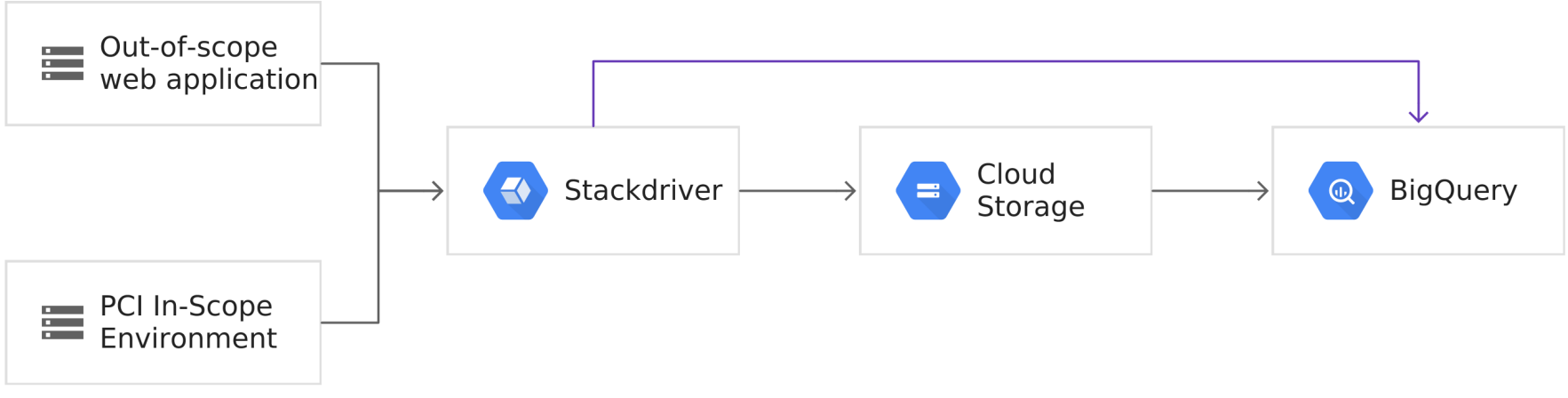
- Security Methods in GCP
- Network Design for Security and Isolation
- Audits and Compliance

Audits and Compliance

Previous

Analyzing Credit Card Swipe data

- CC numbers are tokenized for protection
- Relevant log data is written to Stackdriver Logging
- Export relevant logs (preferably ran through DLP) to BigQuery/Cloud Storage for analysis
- Scoped logs can be shared with auditors as needed



Return to Table of Contents

Choose a Lesson

Return to Table of Contents

Choose a Lesson

Software Development Lifecycle Concepts

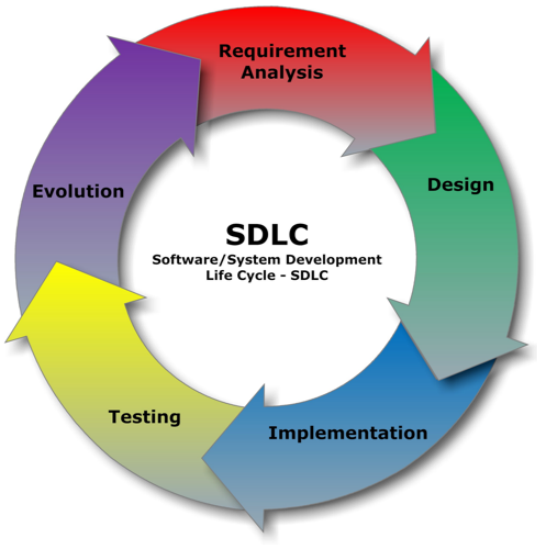
Testing Your Application for Resiliency

Software Development Lifecycle Concepts

Next

How to Approach Development Topics

- Different backgrounds
- Not a deep dive
- Match exam's depth



What We will Cover

- Software Development Lifecycle (SDLC)
- Continuous Integration/Continuous Deployment (CI/CD)
- Best practices for testing and resiliency

What is Software Development Life Cycle?

- Produces software with the highest quality and lowest cost in the shortest time
- Plan to develop, alter, maintain, and replace a software system
- Stages of SDLC – not a set list

Why Does this Matter?

- Exam questions assume using SDLC for application development
- Keep environments separate, with different access for different teams
  - Environments in separate projects, with separate levels of access
  - Separate development/staging/production projects
- Same concepts regardless of compute platform (GCE/GKE/GAE)

Return to Table of Contents

Choose a Lesson

Software Development Lifecycle Concepts

Testing Your Application for Resiliency

Software Development Lifecycle Concepts

Previous

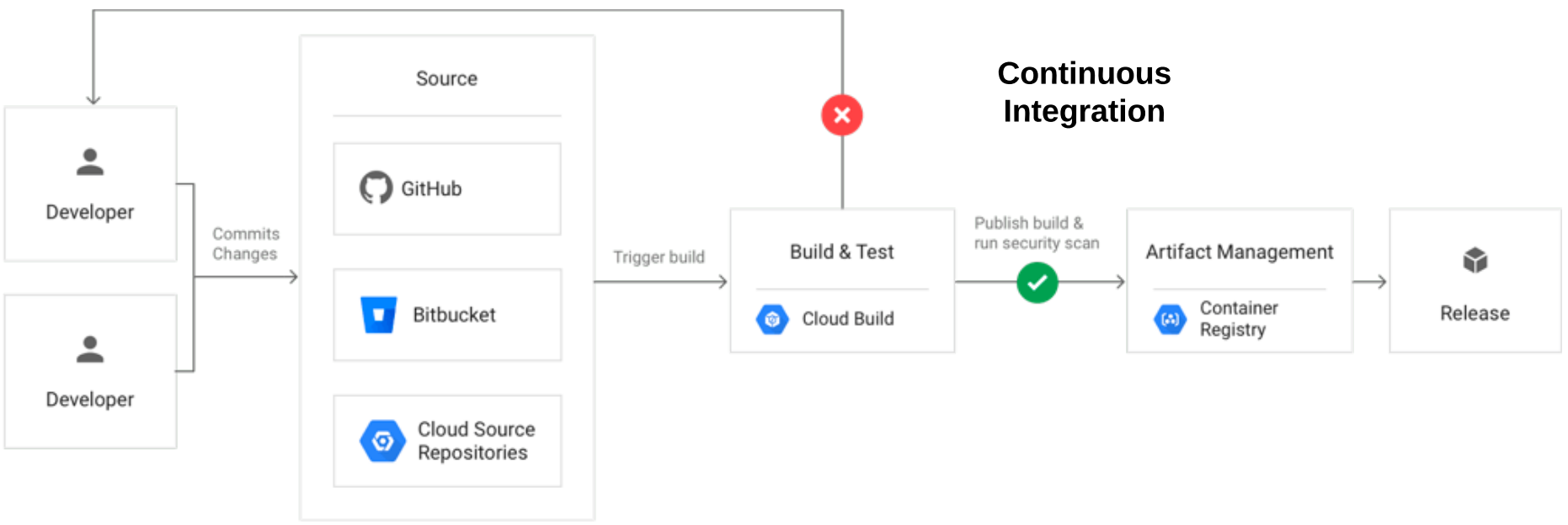
Next

Continuous Integration/Continuous Deployment (CI/CD)

- Continuous Integration
  - Integrate their code into the main branch of a shared repository early and often
  - Minimize the cost of integration
- Continuous Deployment (Delivery)
  - Focus on automating the software delivery process
  - Automatically deploys each build that passes the full test cycle.
  - No waiting for a human gatekeeper
  - Delivery vs. Deployment
    - Delivery: Require manual deployment after verification
- Result: Software deployment isn't an infrequent, scary, massive event, but a continuous process of small improvements

Key Components of CI/CD

- Cloud Source Repositories, Cloud Build, Container Registry, Jenkins, Spinnaker
- Cloud Source Repositories: Private git repository
- Cloud Build: Managed service; automate container packaging and push to Container Registry
- Container Registry: Store container images
  - Scan for vulnerabilities
- Jenkins/Spinnaker: Automation tools for:
  - Building deployment pipelines
  - Blue/green deployments
  - Canary releases
  - Jenkins: Build staging and master branches, verify code changes



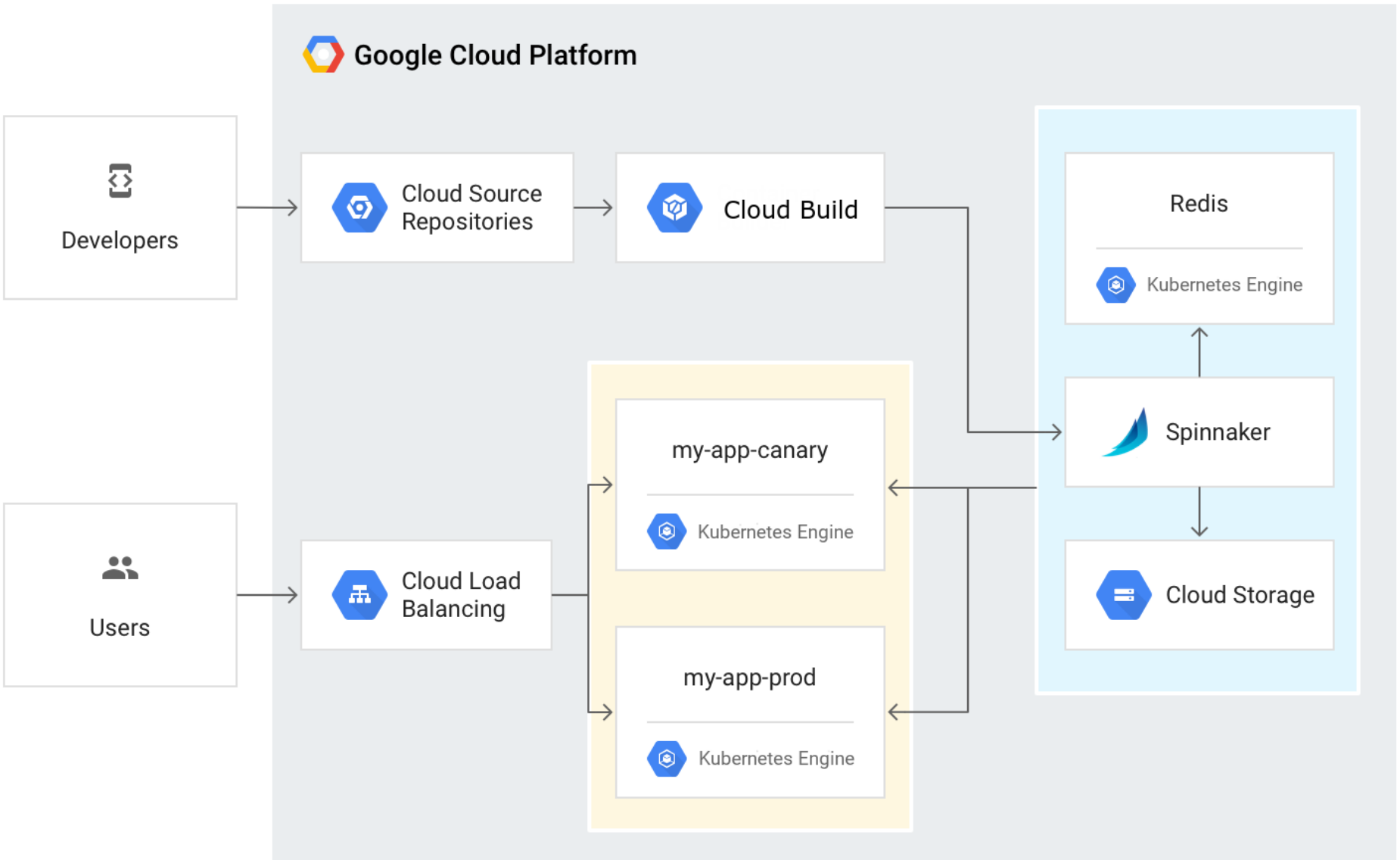
Return to Table of Contents

Software Development Lifecycle Concepts

Previous
Next

Choose a Lesson

Continuous Deployment/Delivery



Return to Table of Contents

Choose a Lesson

Software Development Lifecycle Concepts

Testing Your Application for Resiliency

Software Development Lifecycle Concepts

Previous

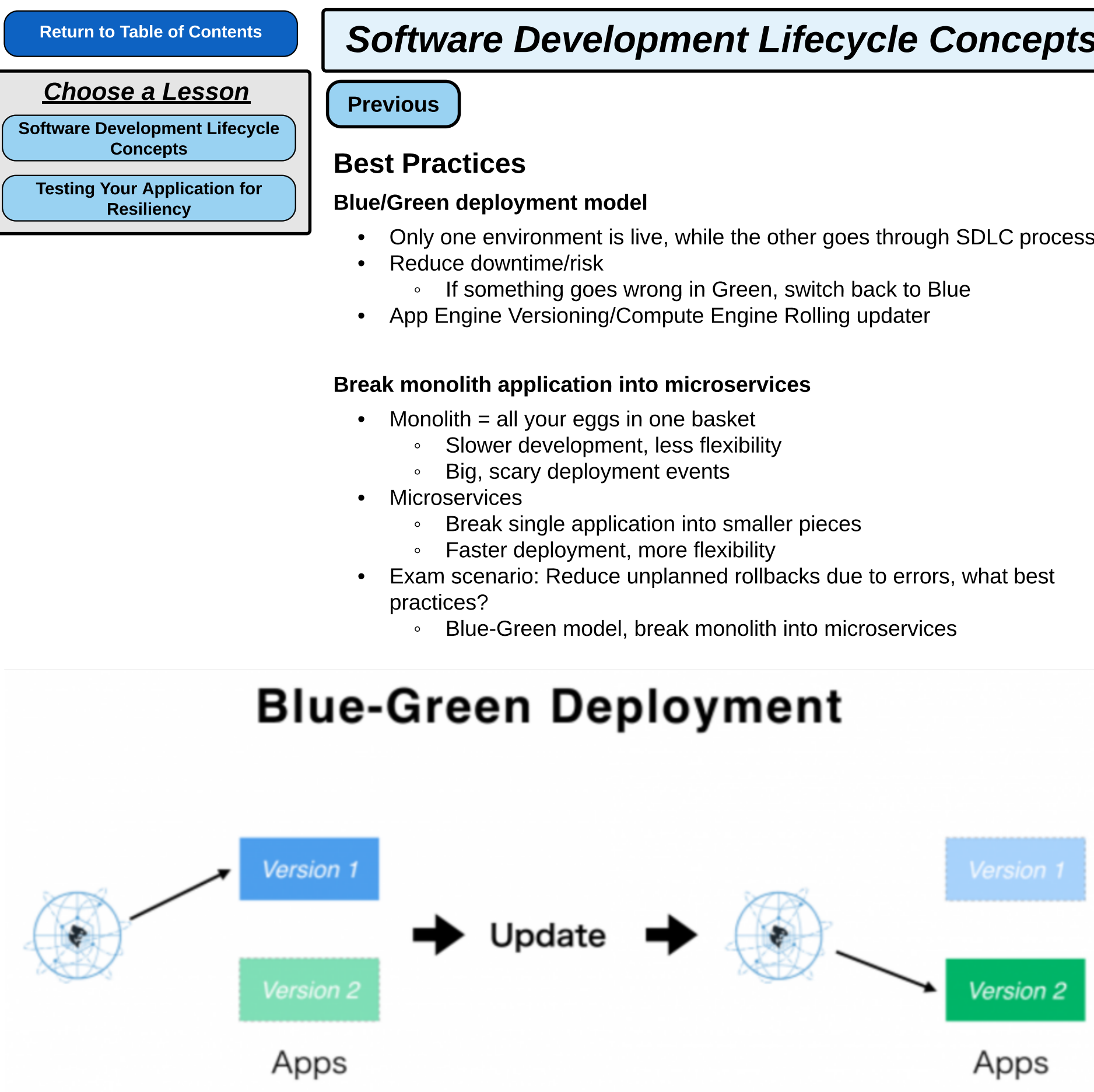
Best Practices

Blue/Green deployment model

- Only one environment is live, while the other goes through SDLC process
- Reduce downtime/risk
  - If something goes wrong in Green, switch back to Blue
- App Engine Versioning/Compute Engine Rolling updater

Break monolith application into microservices

- Monolith = all your eggs in one basket
  - Slower development, less flexibility
  - Big, scary deployment events
- Microservices
  - Break single application into smaller pieces
  - Faster deployment, more flexibility
- Exam scenario: Reduce unplanned rollbacks due to errors, what best practices?
  - Blue-Green model, break monolith into microservices



Return to Table of Contents

Choose a Lesson

Software Development Lifecycle Concepts

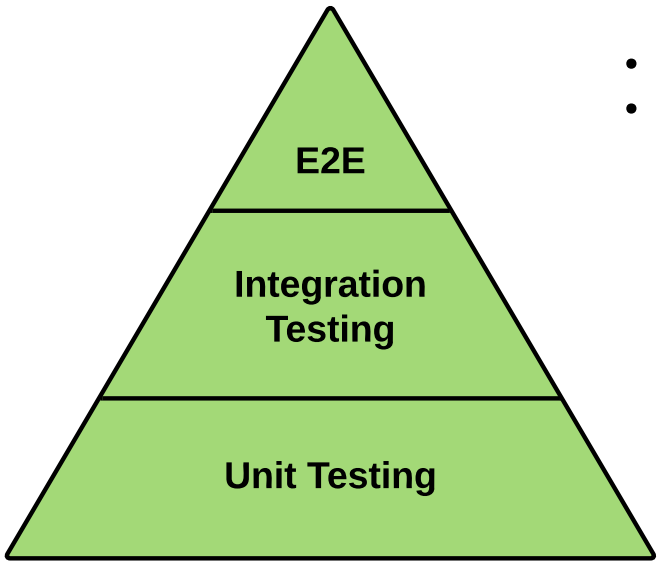
Testing Your Application for Resiliency

Testing Your Application for Resiliency

Next

Testing your application

- We are rolling out a new application, how do we test it?
- Test types
  - Unit tests: Focus on individual components of application
  - Integration Tests: Testing groups of integrated components
  - End to end tests: Simulate real user; test from start of user interaction to end
- How to approach testing: Start small, then go up.
- Focus on unit testing for the majority of your tests. Why?
  - Faster
  - Reliable
  - Isolate failures more quickly
  - Then move up to more involved testing



Testing for resiliency

- Application should properly scale with increased load while also introducing 'chaos'
- Highly available/redundant locations
  - Multi regional/multi-zone
- Load test application - make sure it scales
  - Cloud considerations - Will be able to scale much more than previous testing methods
- Introduce 'Chaos'
  - Shut down machines in different zones
  - Shut down an entire zone
  - Shut down stateful storage - trigger failover
  - Variety of applications to do this - Chaos Monkey, Simian Army

Return to Table of Contents

Choose a Lesson

Software Development Lifecycle Concepts

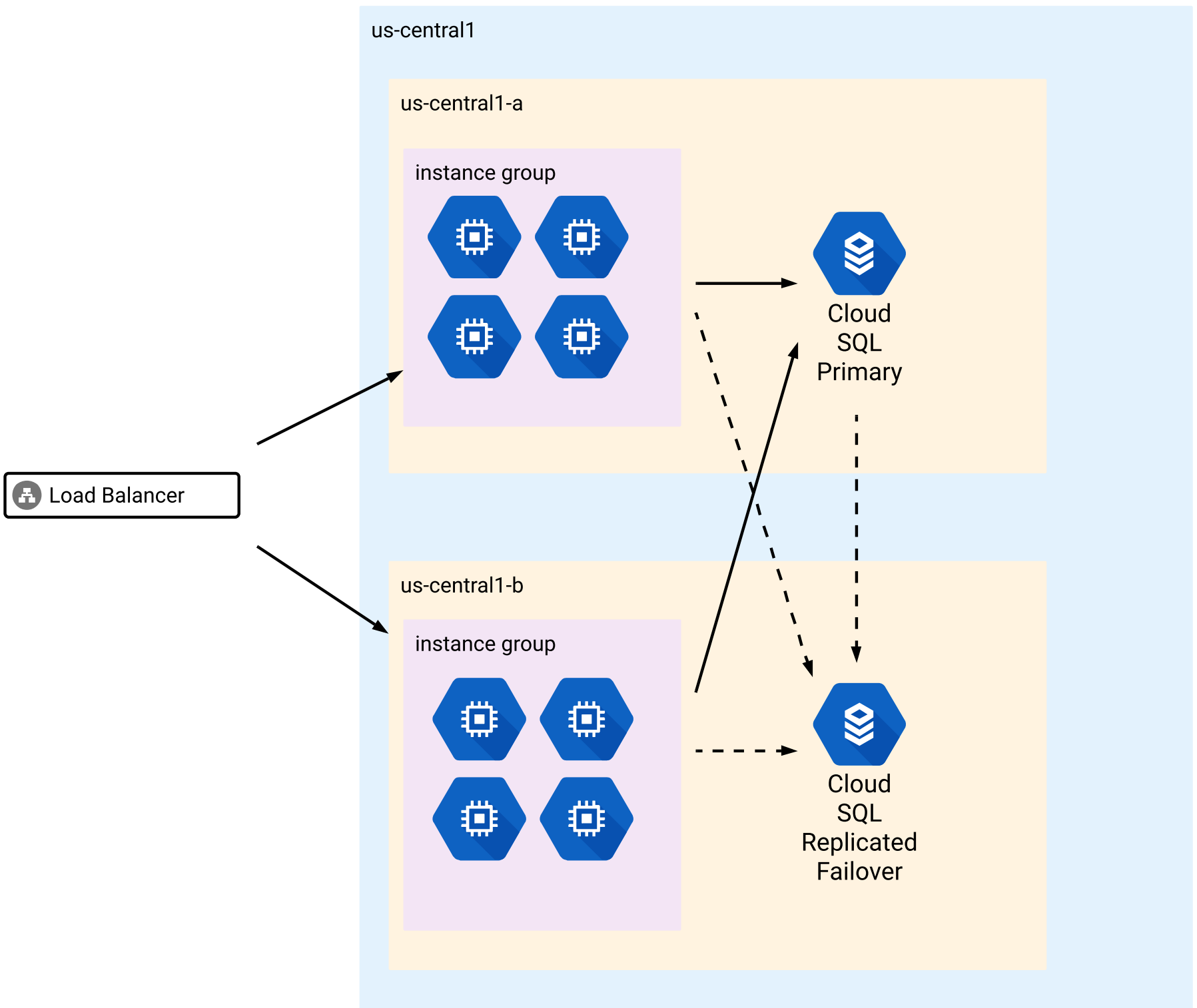
Testing Your Application for Resiliency

Testing Your Application for Resiliency

Previous

Testing example web application

- Separate testing environment/project
- Multi-zone managed instance group
  - Autoscaling enabled
- Cloud SQL with failover instance
- HTTP Load balancer
- Disable VMs, zones, primary databases to test resiliency



Return to Table of Contents

Choose a Lesson

Software Development Lifecycle Concepts

Testing Your Application for Resiliency

Return to Table of Contents

Choose a Lesson

Return to Table of Contents

Choose a Lesson

Return to Table of Contents

Choose a Lesson

Return to Table of Contents

Choose a Lesson

Additional Study Resources

Return to Table of Contents

Choose a Lesson

Additional Study Resources

Additional Study Resources

Codelabs

- https://codelabs.developers.google.com/
- Filter by 'Cloud'
- Use GCP Playgrounds on Linux Academy for practice

Google Cloud Solutions Architecture Reference

- https://gcp.solutions/
- Solutions architecture reference diagrams

GCP in 4 words or less

- https://github.com/gregsramblings/google-cloud-4-words
- Updated reference (with documentation links) of all GCP services

Official Google Cloud Blog

- https://cloud.google.com/blog/
- Many informational articles