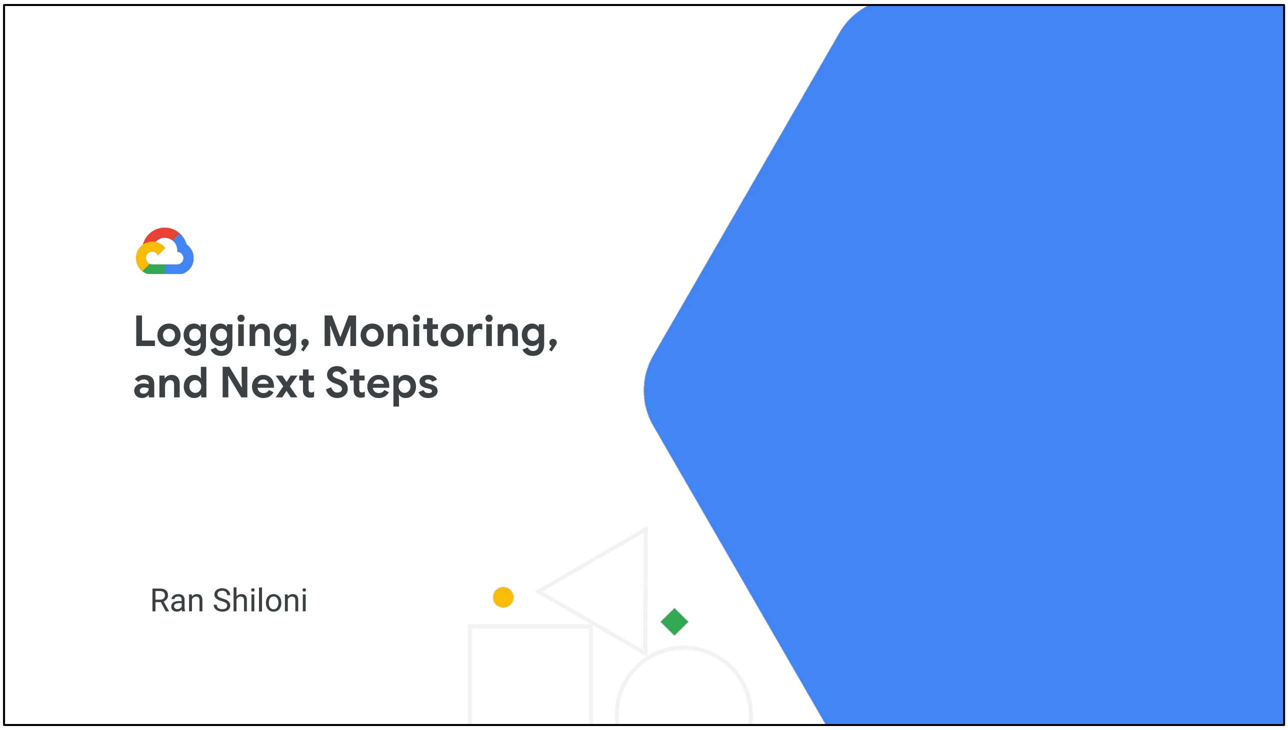
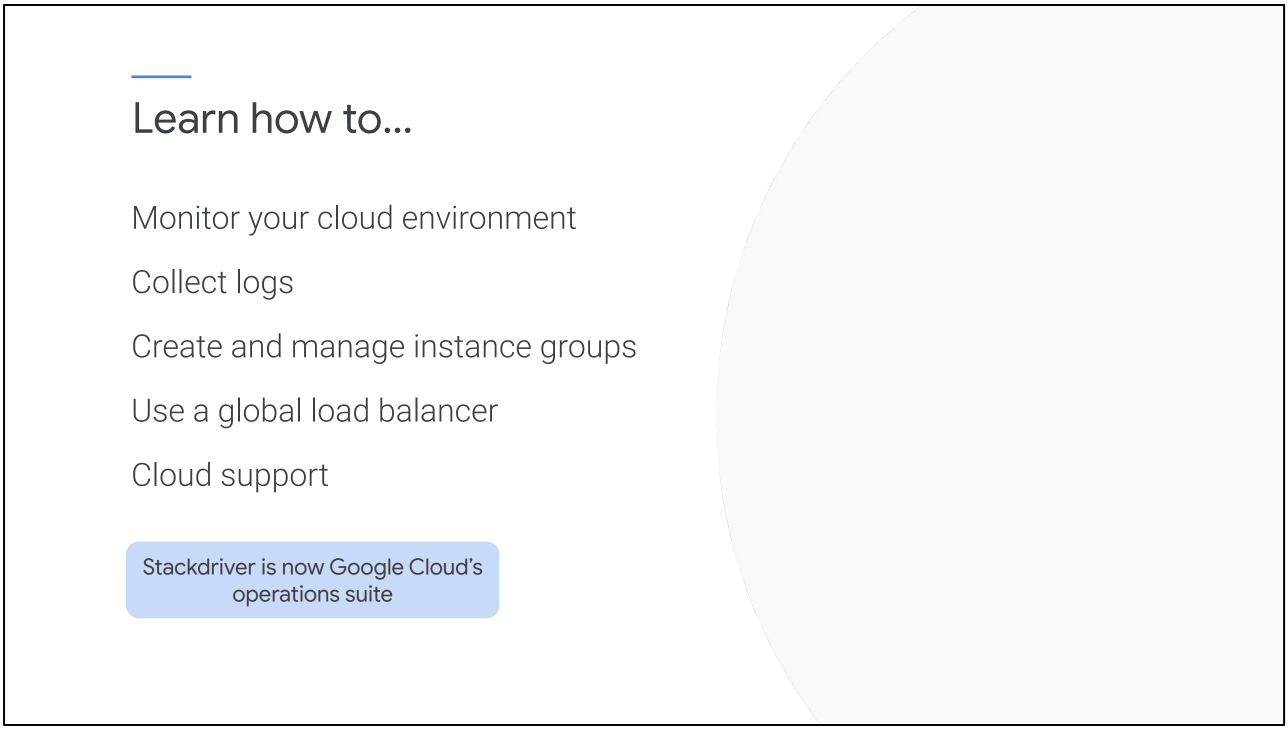


Logging, Monitoring, and Next Steps

Ran Shiloni

Hello, and welcome to the Logging, Monitoring, and Next Steps module.



Learn how to...

Monitor your cloud environment

Collect logs

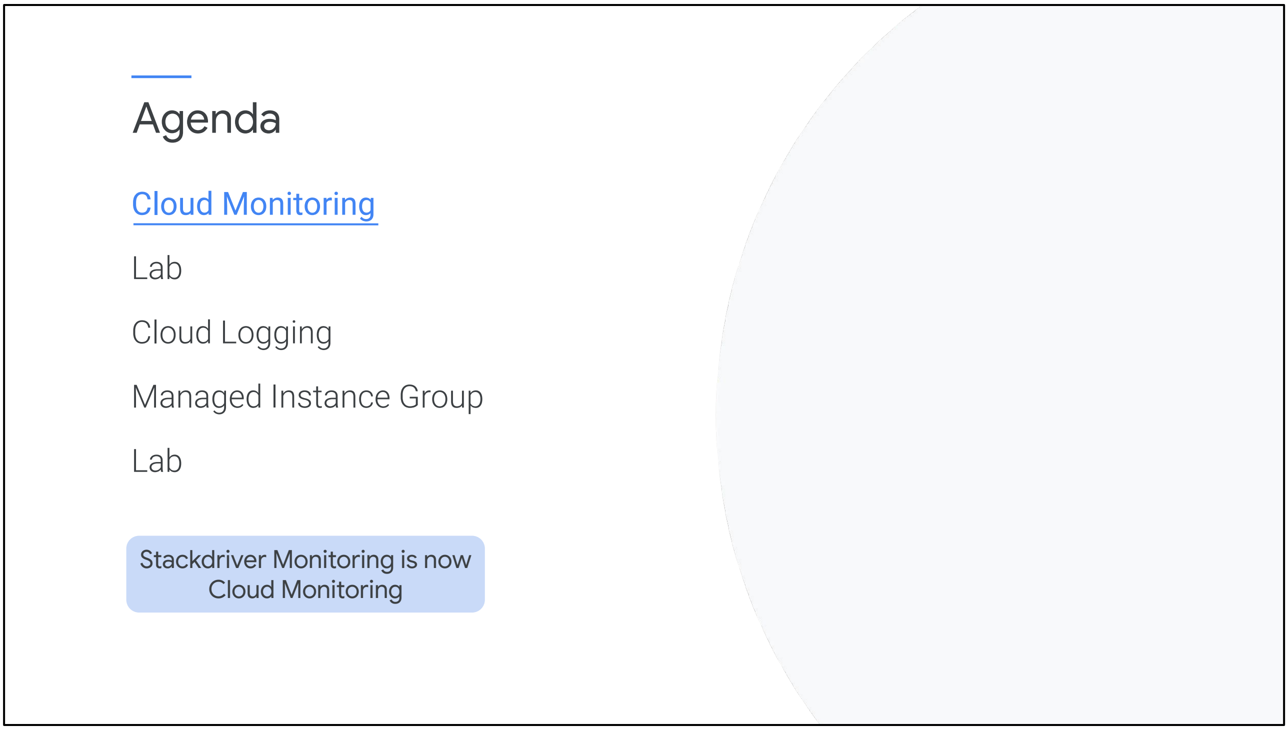
Create and manage instance groups

Use a global load balancer

Cloud support

Stackdriver is now Google Cloud's operations suite

In this module, you will learn how to use Cloud Monitoring and Cloud Logging capabilities to enhance your observability of your cloud environment. You then learn how to use managed instance groups, which are automated groups of virtual machines that can scale up and down based on metrics. You will also learn how to use our global load balancer to frontend the group of virtual machines, and lastly, you will learn about our Cloud Support tiers.



Agenda

Cloud Monitoring

Lab

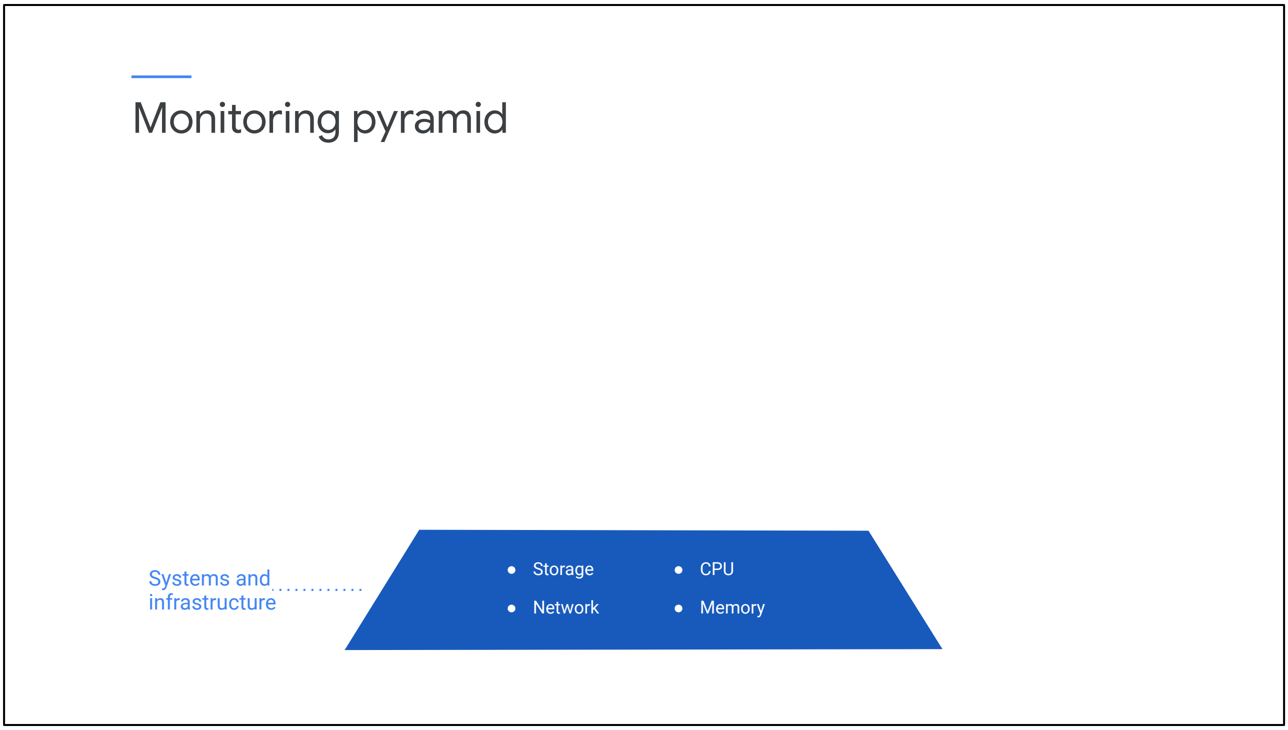
Cloud Logging

Managed Instance Group

Lab

Stackdriver Monitoring is now Cloud Monitoring

In this video, you will learn how to use Cloud Monitoring.

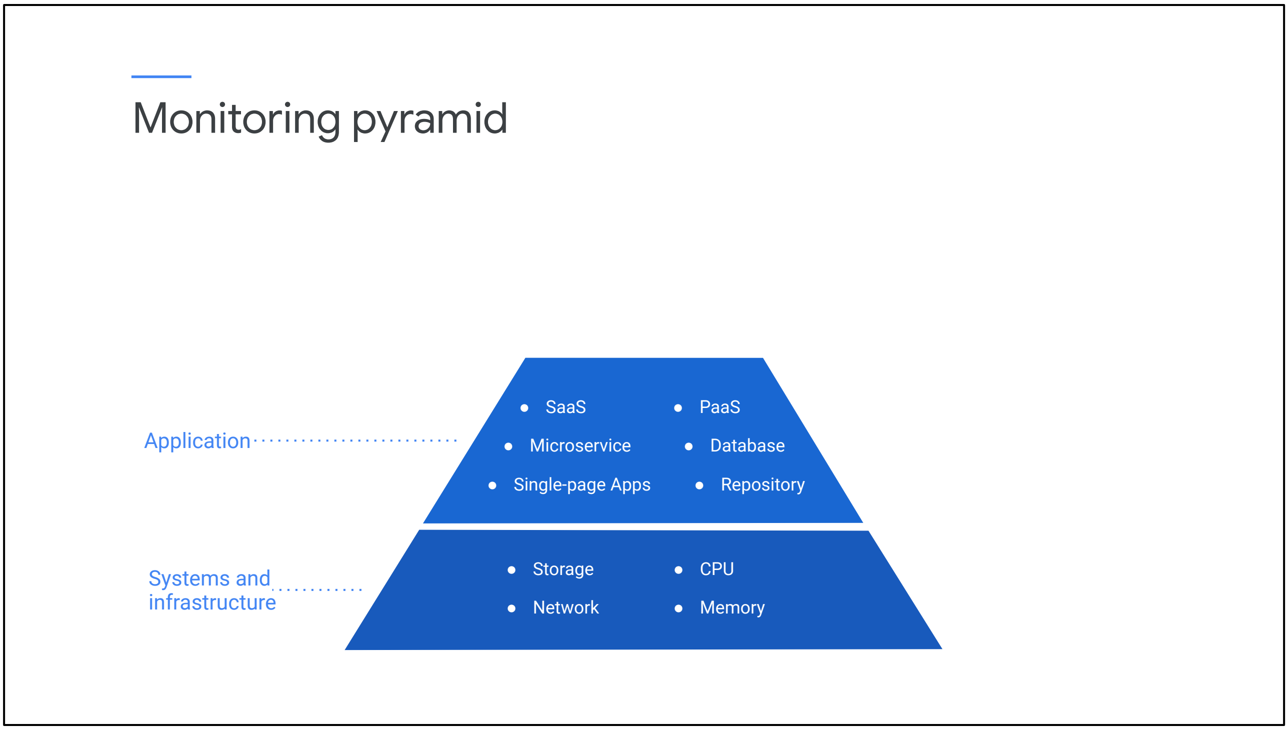


Monitoring pyramid

Systems and infrastructure

- Storage
- CPU
- Network
- Memory

Measuring the systems and infrastructure provides indications about the state of the components that run your infrastructure, like the storage input output per second (IOPS), vCPU utilization, memory allocation, and network performance. These are your lower level indications, at the base of the pyramid.



Monitoring pyramid

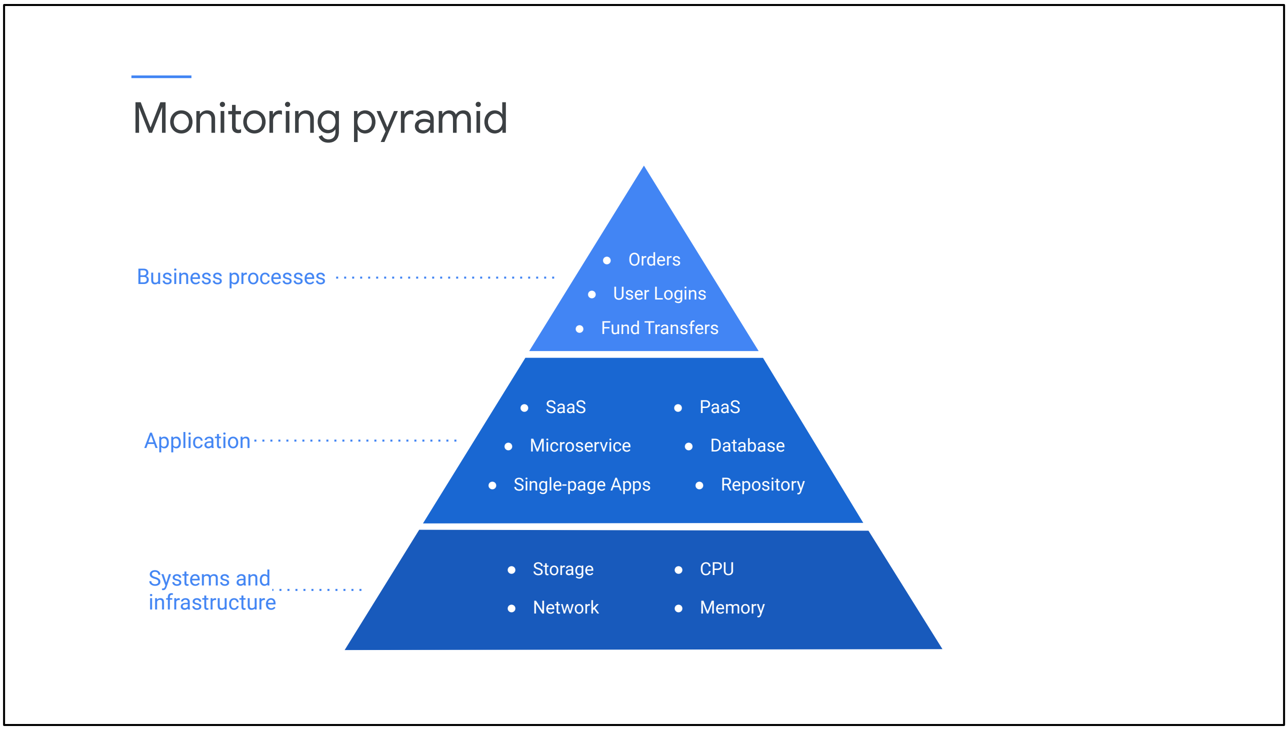
Application

- SaaS
- PaaS
- Microservice
- Database
- Single-page Apps
- Repository

Systems and infrastructure

- Storage
- CPU
- Network
- Memory

The systems and Infrastructure layer function is to run the applications that sit on top of them. Monitoring the application layer provides a holistic picture of the health of your workload. For instance, measuring the vCPU utilization and memory allocation of a database can only provide a partial indication on the health of the service. Measuring the application's metrics, like query latency, provides more granular and specific information.



Monitoring pyramid

Business processes

- Orders
- User Logins
- Fund Transfers

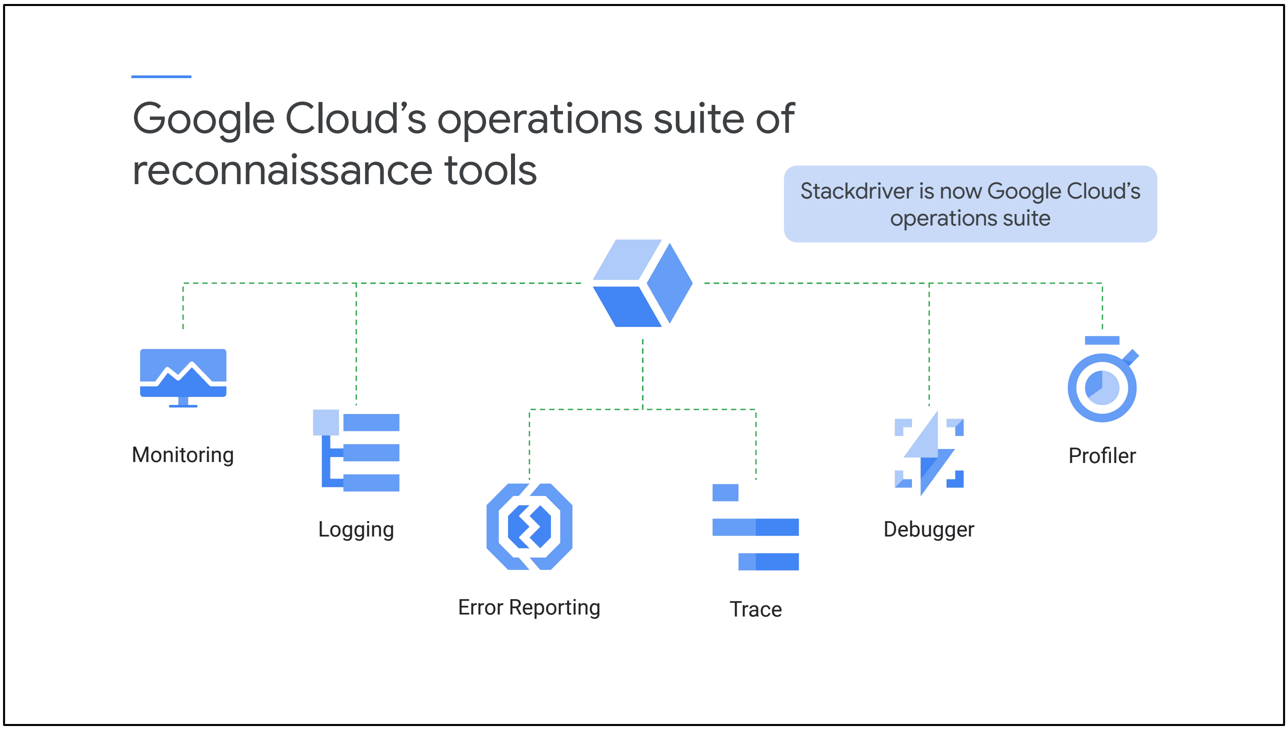
Application

- SaaS
- PaaS
- Microservice
- Database
- Single-page Apps
- Repository

Systems and infrastructure

- Storage
- CPU
- Network
- Memory

Lastly, at the top of the pyramid are the business processes. Applications and platforms are eventually there to answer business needs--for example, daily active users, orders, or refunds--which in turn should be monitored appropriately,.



Google Cloud’s operations suite of reconnaissance tools

Stackdriver is now Google Cloud’s operations suite

Monitoring

Logging

Error Reporting

Trace

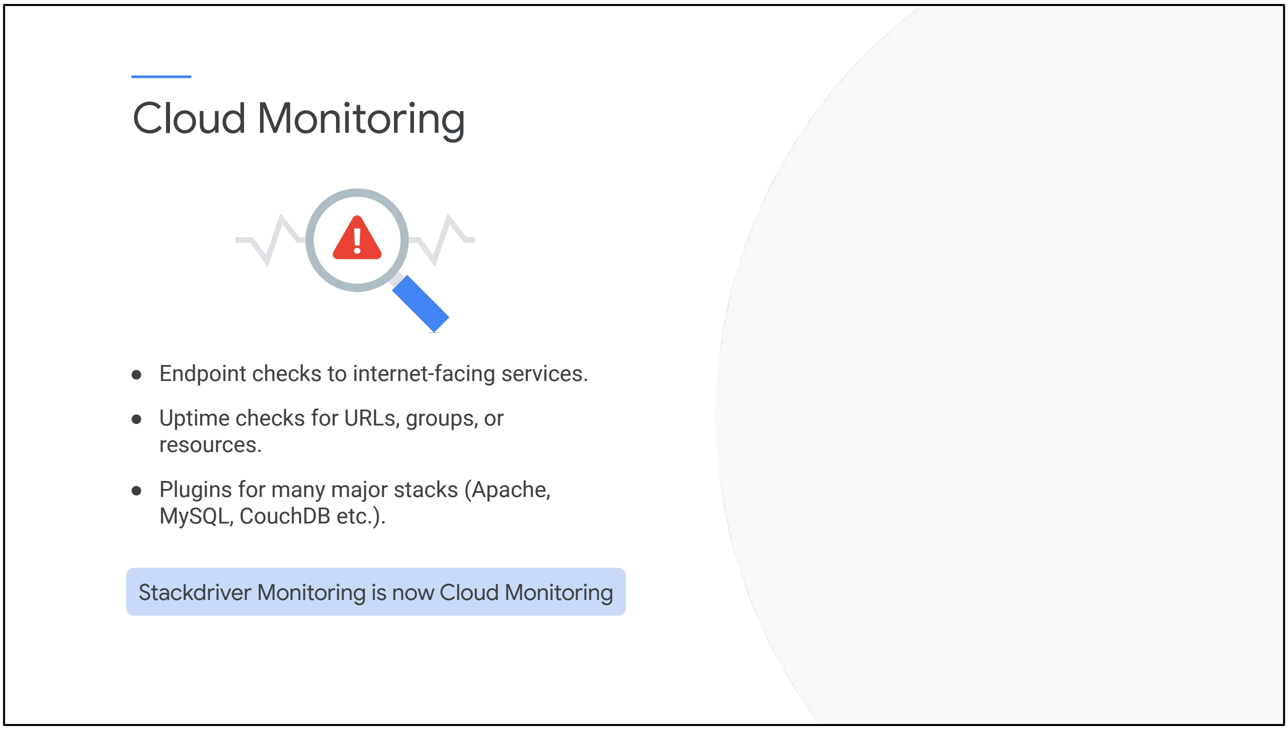
Debugger

Profiler

Google Cloud’s operations suite is a fully-managed native logging and monitoring tool that provides you with observability of all 3 layers.

It gives you access to logs, metrics, traces, and other signals from your infrastructure platform, virtual machines, containers, middleware, and application tier, so that you can track issues all the way from your end user to your backend services and infrastructure.

The tools aggregate metrics, logs, and events from the infrastructure, giving developers and operators a rich set of observable signals that speed root-cause analysis and reduce mean time to resolution. They don’t require extensive integration or multiple “panes of glass,” and won’t lock developers into using a particular cloud provider.



Cloud Monitoring

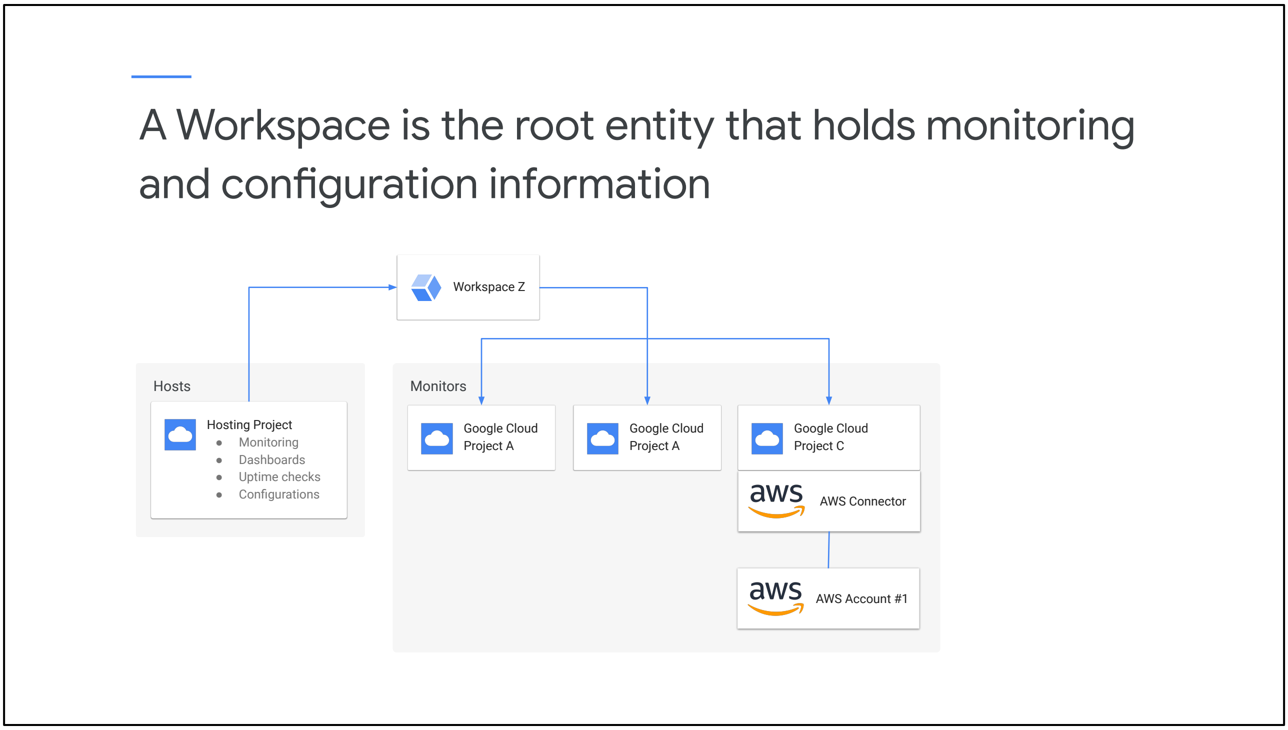
- Endpoint checks to internet-facing services.
- Uptime checks for URLs, groups, or resources.
- Plugins for many major stacks (Apache, MySQL, CouchDB etc.).

Stackdriver Monitoring is now Cloud Monitoring

Cloud Monitoring dynamically configures monitoring after resources are deployed and has intelligent defaults that allow you to easily create charts for monitoring activities.

Rich visualization and advanced alerting help you identify issues quickly, even hard-to-diagnose issues like host contention, cloud provider throttling, and degraded hardware.

This allows you to monitor your platform, system, and application metrics by ingesting data, such as metrics, events, and metadata. You can then generate insights from this data through rich visualization and advanced alerting.

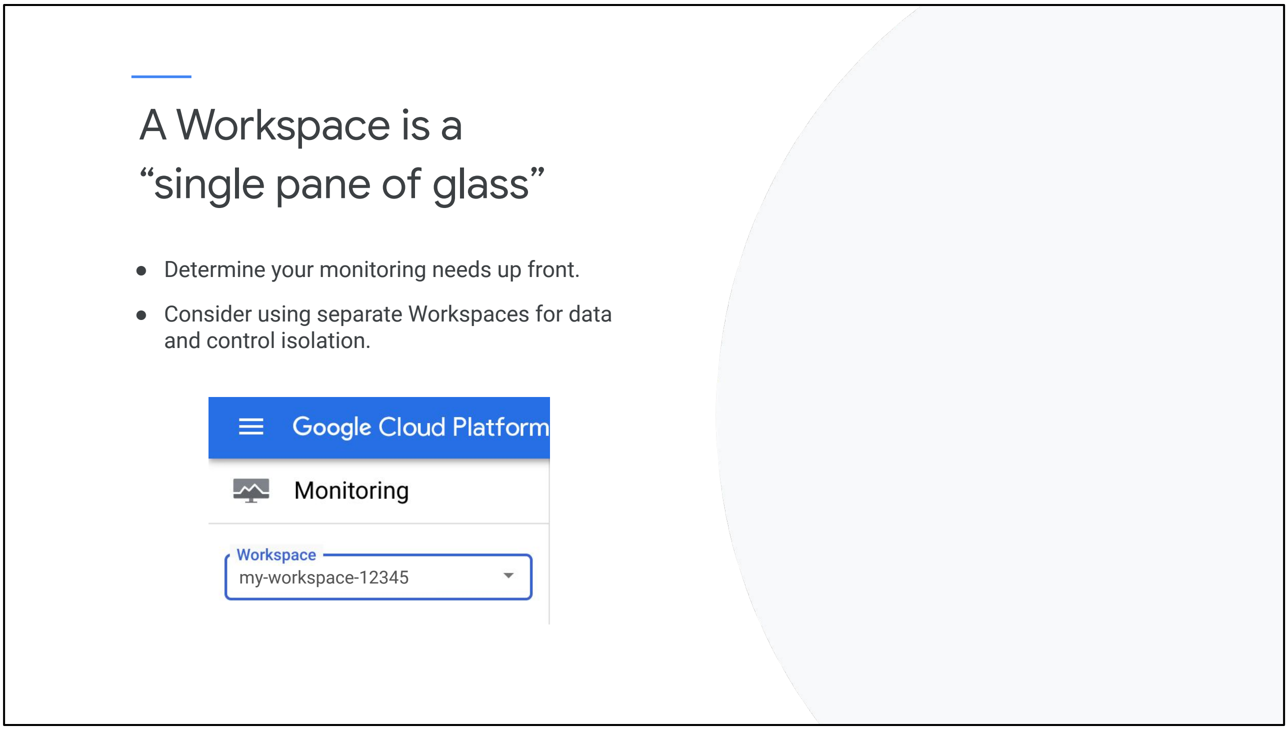


A Workspace is the root entity that holds monitoring and configuration information

A Workspace is the root entity that holds monitoring and configuration information in Cloud Monitoring. Each Workspace can have between 1 and 100 monitored projects. You can have as many Workspaces as you want, but Google Cloud projects and AWS accounts can't be monitored by more than one Workspace.

A Workspace contains the custom dashboards, alerting policies, uptime checks, notification channels, and group definitions that you use with your monitored projects. A Workspace can access metric data from its monitored projects, but the metric data and log entries remain in the individual projects.

The first monitored Google Cloud project in a Workspace is called the hosting project, and it must be specified when you create the Workspace. The name of that project becomes the name of your Workspace. To access an AWS account, you must configure a project in Google Cloud to hold the AWS Connector.



A Workspace is a “single pane of glass”

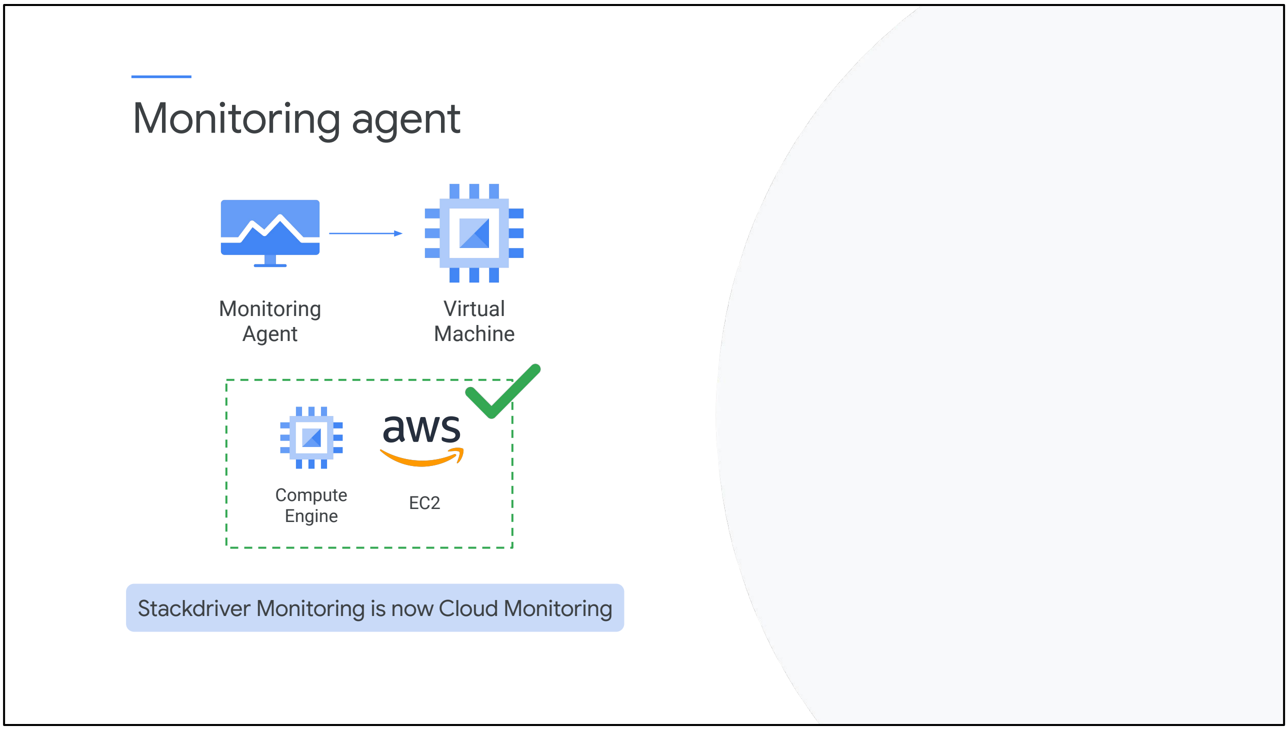
- Determine your monitoring needs up front.
- Consider using separate Workspaces for data and control isolation.

Workspace
my-workspace-12345

Because Workspaces can monitor all your Google Cloud projects in a single place, a Workspace is a “single pane of glass” through which you can view resources from multiple Google Cloud projects and AWS accounts. All users of Google Cloud’s operation suite with access to that Workspace have access to all data by default.

This means that a role assigned to one person on one project applies equally to all projects monitored by that Workspace.

In order to give people different roles per-project and to control visibility to data, consider placing the monitoring of those projects in separate Workspaces.



Monitoring agent

Monitoring Agent

Virtual Machine

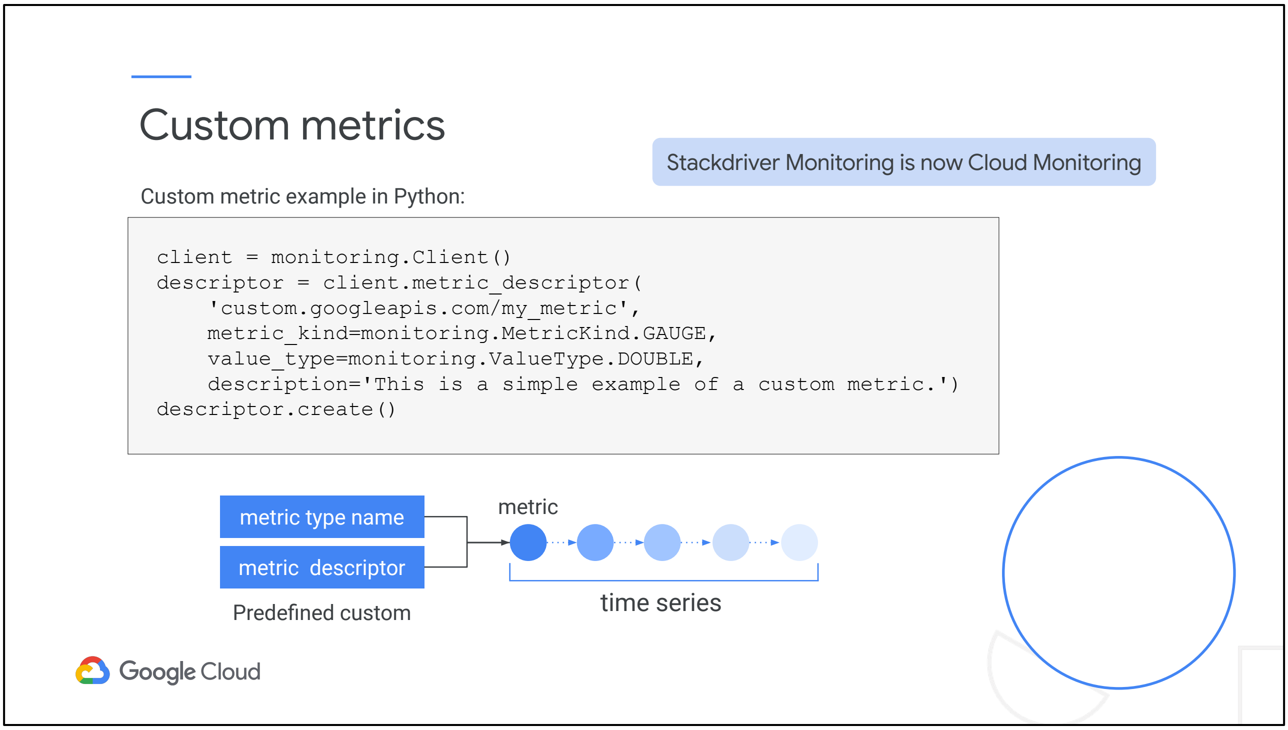
Compute Engine

EC2

Stackdriver Monitoring is now Cloud Monitoring

Cloud Monitoring can access lower level metrics without the Monitoring agent, including CPU utilization, some disk traffic metrics, network traffic, and uptime information. However, to access additional system resources and application services, you should install the Monitoring agent.

The Monitoring agent is supported for Compute Engine and AWS EC2 instances.



Custom metrics

Stackdriver Monitoring is now Cloud Monitoring

Custom metric example in Python:

```
client = monitoring.Client()
descriptor = client.metric_descriptor(
    'custom.googleapis.com/my_metric',
    metric_kind=monitoring.MetricKind.GAUGE,
    value_type=monitoring.ValueType.DOUBLE,
    description='This is a simple example of a custom metric.')
descriptor.create()
```

metric type name

metric descriptor

Predefined custom

metric

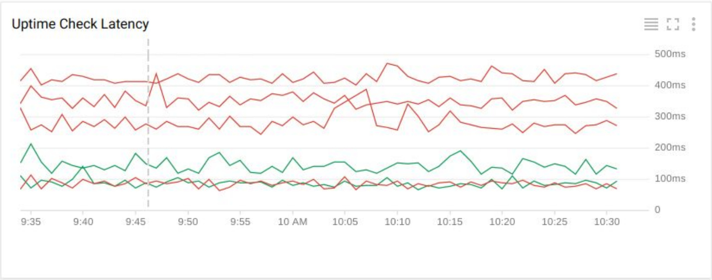
time series

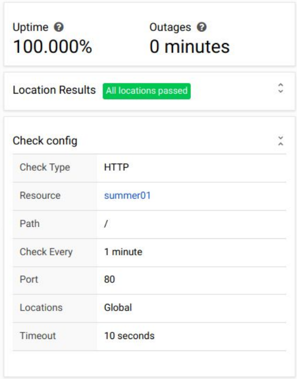
Google Cloud

If the standard metrics provided by Cloud Monitoring do not fit your needs, you can create custom metrics.

For example, imagine a game server that has a capacity of 50 users. What metric indicator might you use to trigger scaling events? From an infrastructure perspective, you might consider using CPU load or perhaps network traffic load as values that are somewhat correlated with the number of users. But with a Custom Metric, you could actually pass the current number of users directly from your application into Cloud Monitoring, and scale your infrastructure based on that custom metric.

Uptime check example





Uptime 100.000%

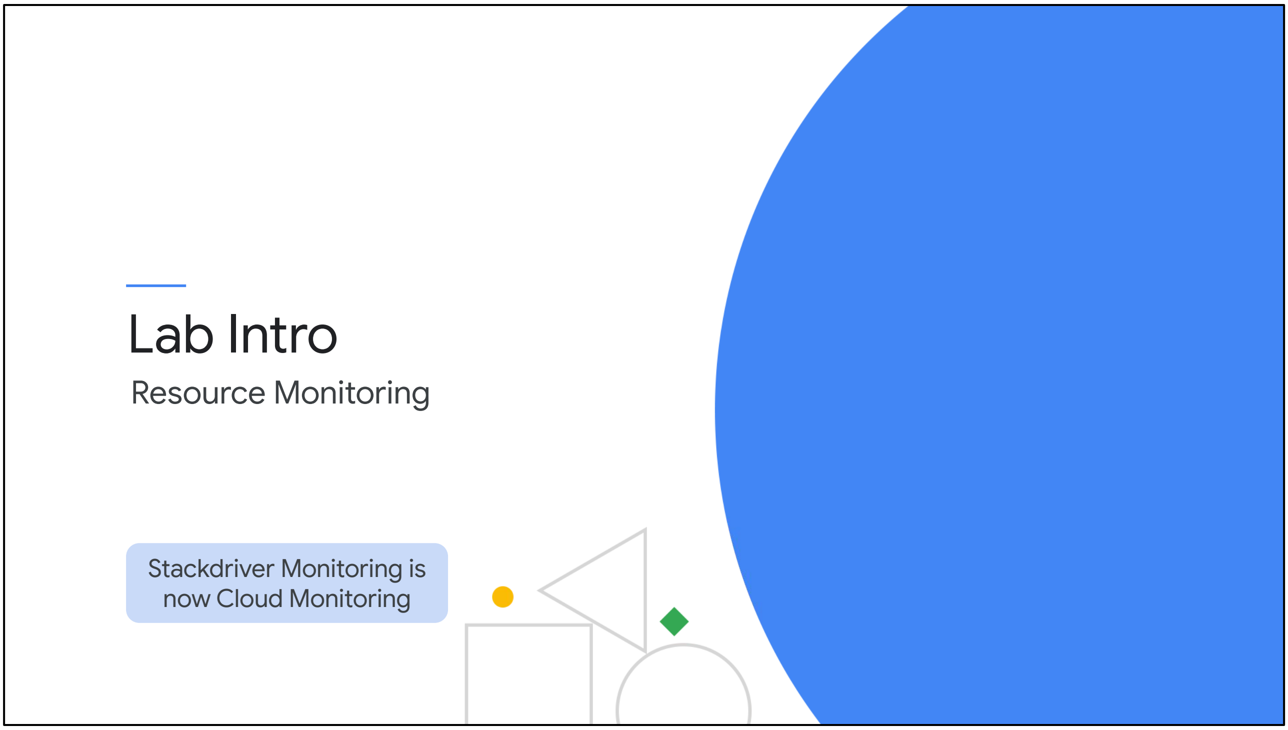
Outages 0 minutes

Location Results All locations passed

Check config

Check Type	HTTP
Resource	summer01
Path	/
Check Every	1 minute
Port	80
Locations	Global
Timeout	10 seconds

Here is an example of an HTTP uptime check. The resource is checked every minute from 6 different locations across the globe. Each uptime check is configured with a timeout period, and checks that do not get a response within a timeout period are considered failures and affect your uptime percentage.

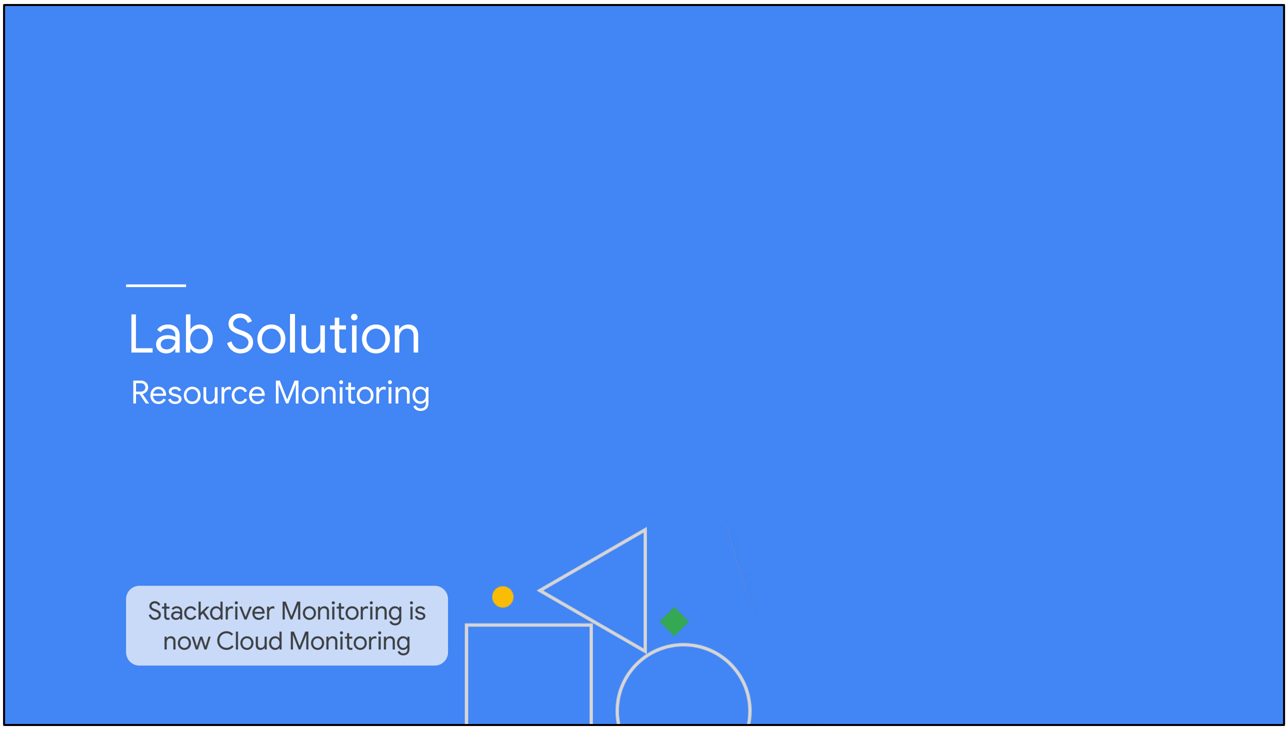


Lab Intro

Resource Monitoring

Stackdriver Monitoring is now Cloud Monitoring

Let's take some of the monitoring concepts that we just discussed and apply them in the lab. In this lab you will learn how to use Cloud Monitoring to gain insight into applications that run on Google Cloud. Specifically, you will enable Cloud Monitoring, add charts to dashboards and create alerts, resource groups and uptime checks.



Lab Solution

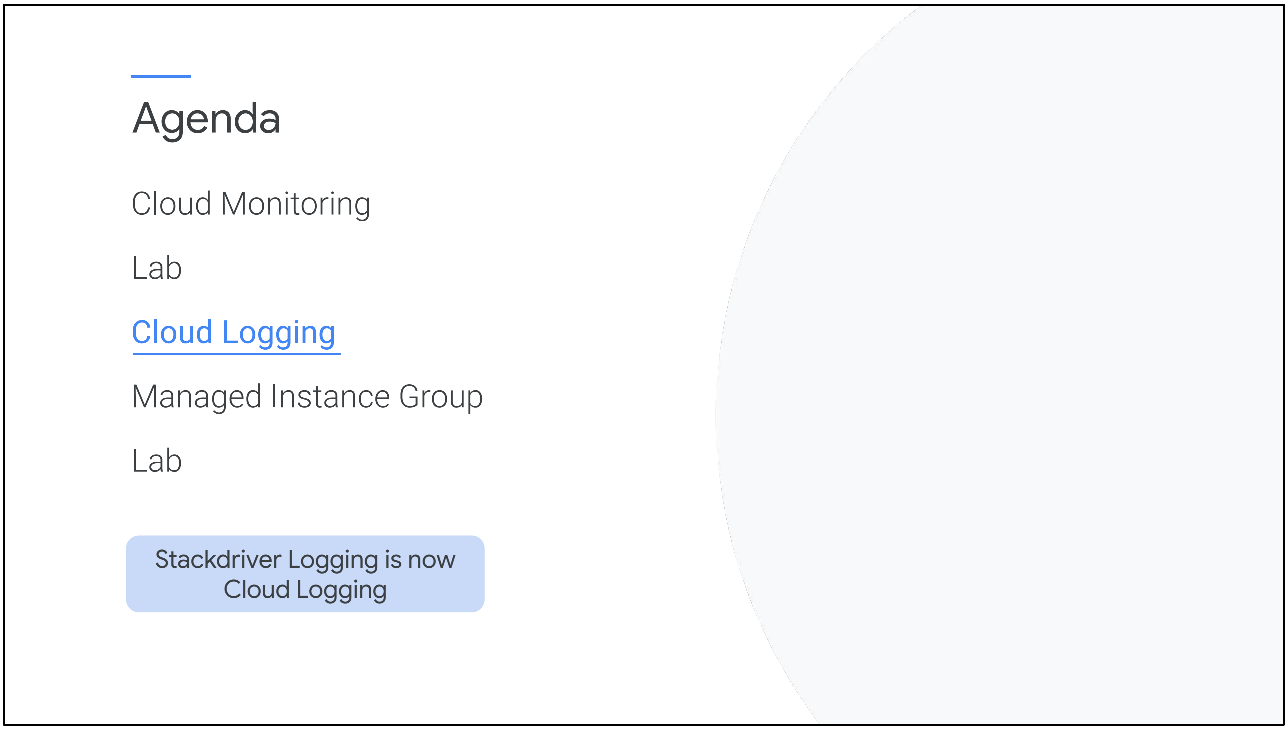
Resource Monitoring

Stackdriver Monitoring is now Cloud Monitoring

In this lab, you got an overview of Cloud Monitoring. You learned how to monitor your project, create alerts with multiple conditions, add charts to dashboards, create resource groups, and create uptime checks for your services.

Monitoring is critical to your application’s health, and Cloud Monitoring provides a rich set of features for monitoring your infrastructure, visualizing the monitoring data, and triggering alerts and events for you.

You can stay for a lab walkthrough, but remember that Google Cloud's user interface can change, so your environment might look slightly different.



Agenda

Cloud Monitoring

Lab

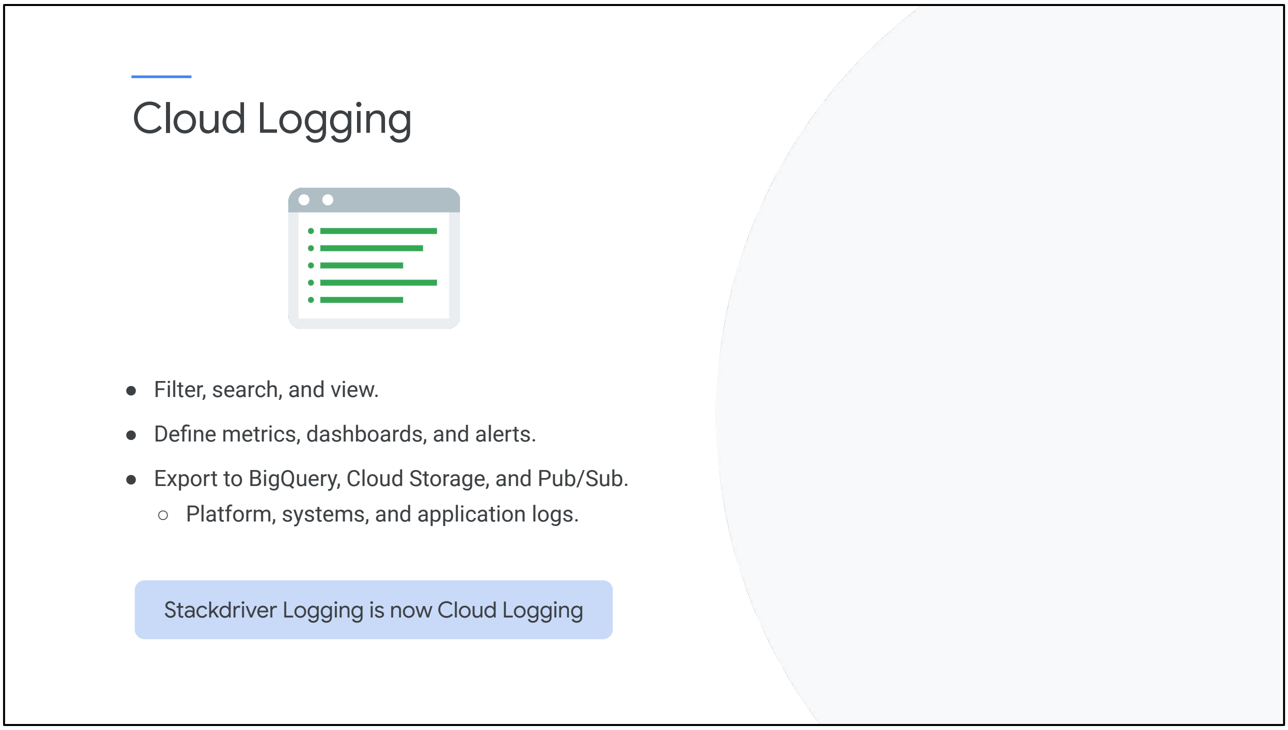
Cloud Logging

Managed Instance Group

Lab

Stackdriver Logging is now Cloud Logging

In this video, you will learn how to use Cloud Logging.



Cloud Logging

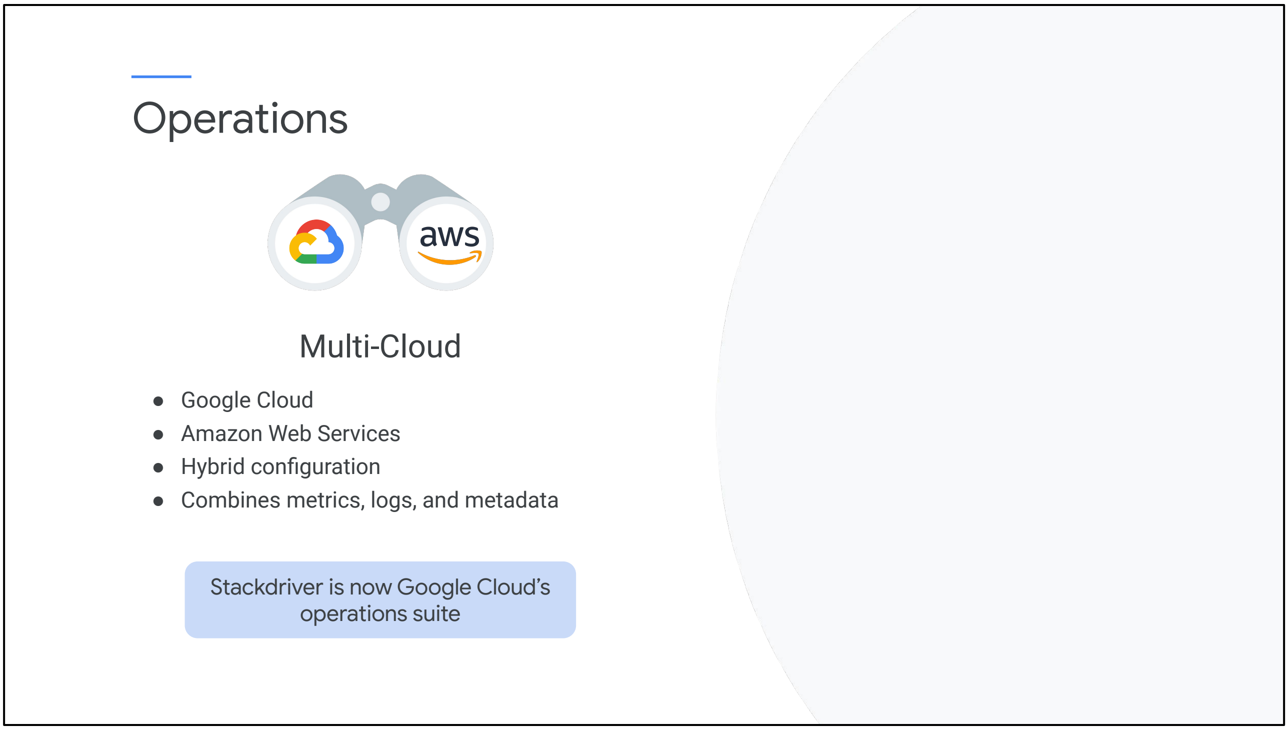
- Filter, search, and view.
- Define metrics, dashboards, and alerts.
- Export to BigQuery, Cloud Storage, and Pub/Sub.
  - Platform, systems, and application logs.

Stackdriver Logging is now Cloud Logging

Cloud Logging allows you to store, search, analyze, monitor, and alert on log data and events from Google Cloud and AWS. It is a fully managed service that performs at scale and can ingest application and system log data from thousands of VMs.

Logging includes storage for logs, a user interface called Logs Explorer, and an API to manage logs programmatically. The service lets you read and write log entries, search and filter your logs, and create log-based metrics.

Logs are only retained for 30 days, but you can export your logs to Cloud Storage buckets, BigQuery datasets, and Pub/Sub topics.



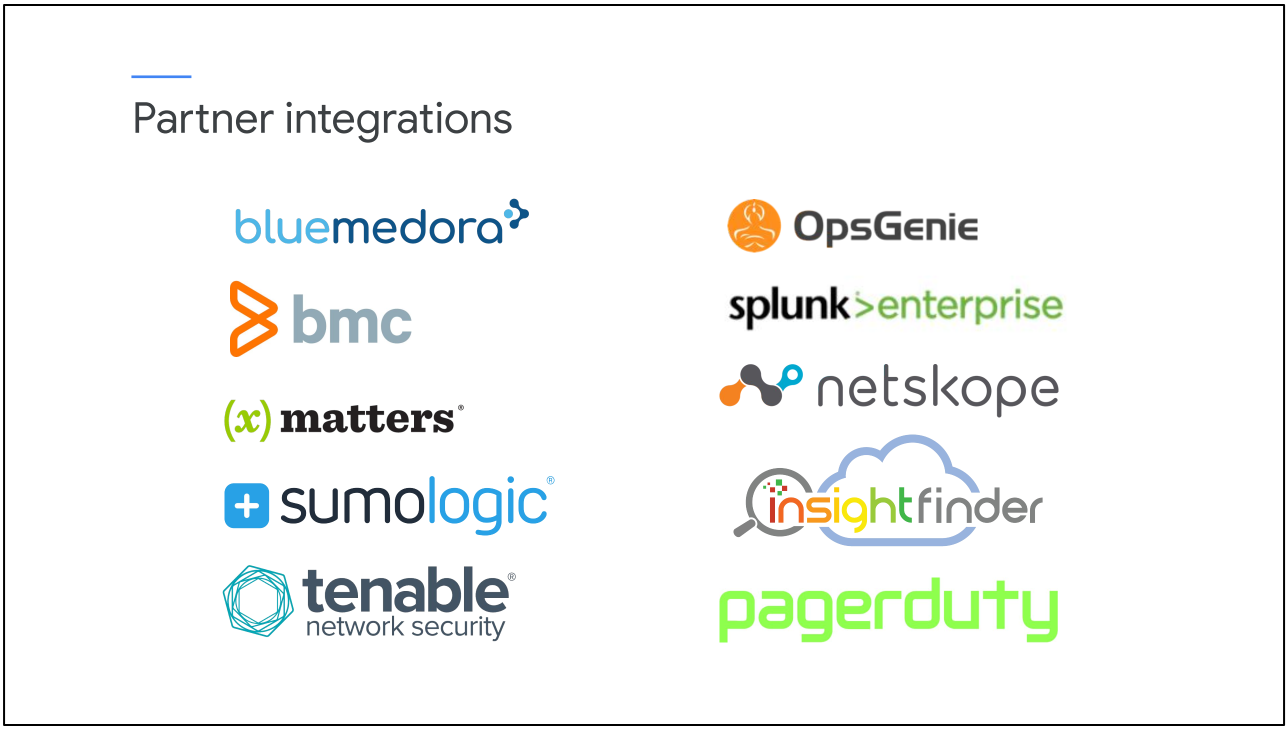
Operations

Multi-Cloud

- Google Cloud
- Amazon Web Services
- Hybrid configuration
- Combines metrics, logs, and metadata

Stackdriver is now Google Cloud’s operations suite

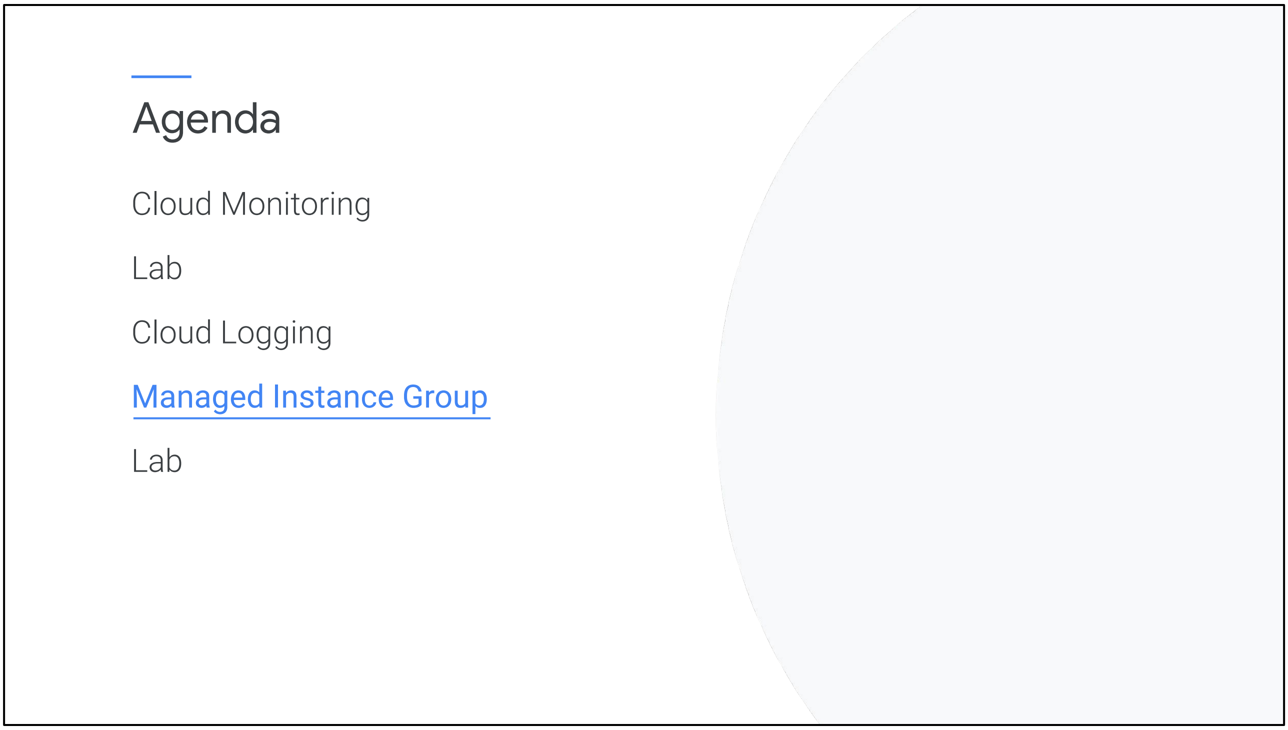
Whether you’re running on Google Cloud, Amazon Web Services, on-premises infrastructure, or with hybrid clouds, Google Cloud’s operation suite combines metrics, logs, and metadata from all of your cloud accounts and projects into a single comprehensive view of your environment, so you can quickly understand service behavior and take action.



Partner integrations

bluemedora
bmc
(x) matters
sumologic
tenable network security
OpsGenie
splunk>enterprise
netskope
insightfinder
pagerduty

Lastly, Google Cloud's operations suite supports a rich and growing ecosystem of technology integrations to expand the IT ops, security, and compliance capabilities available to Google Cloud customers.



Agenda

Cloud Monitoring

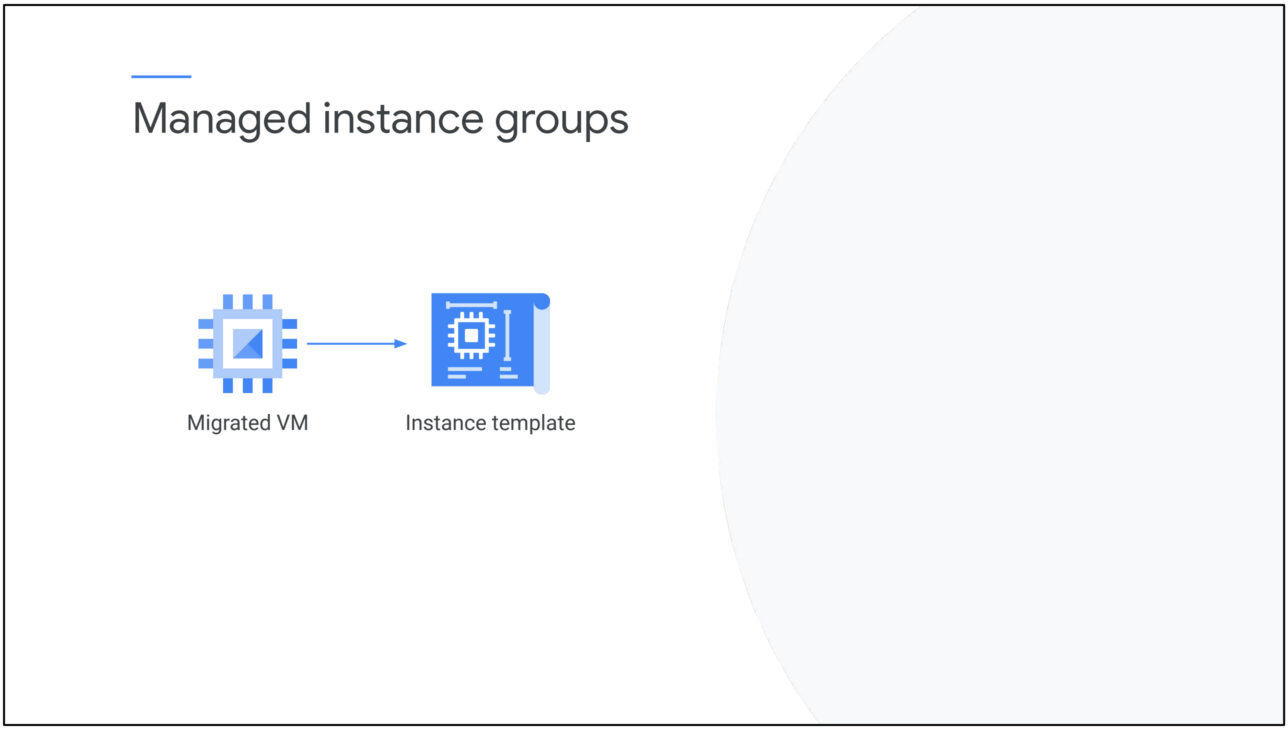
Lab

Cloud Logging

Managed Instance Group

Lab

In this video, you will learn how to use a Managed Instance Group to automate the lifecycle and capacity of your workloads in the cloud, in addition to front ending a group of machines with a global load balancer.

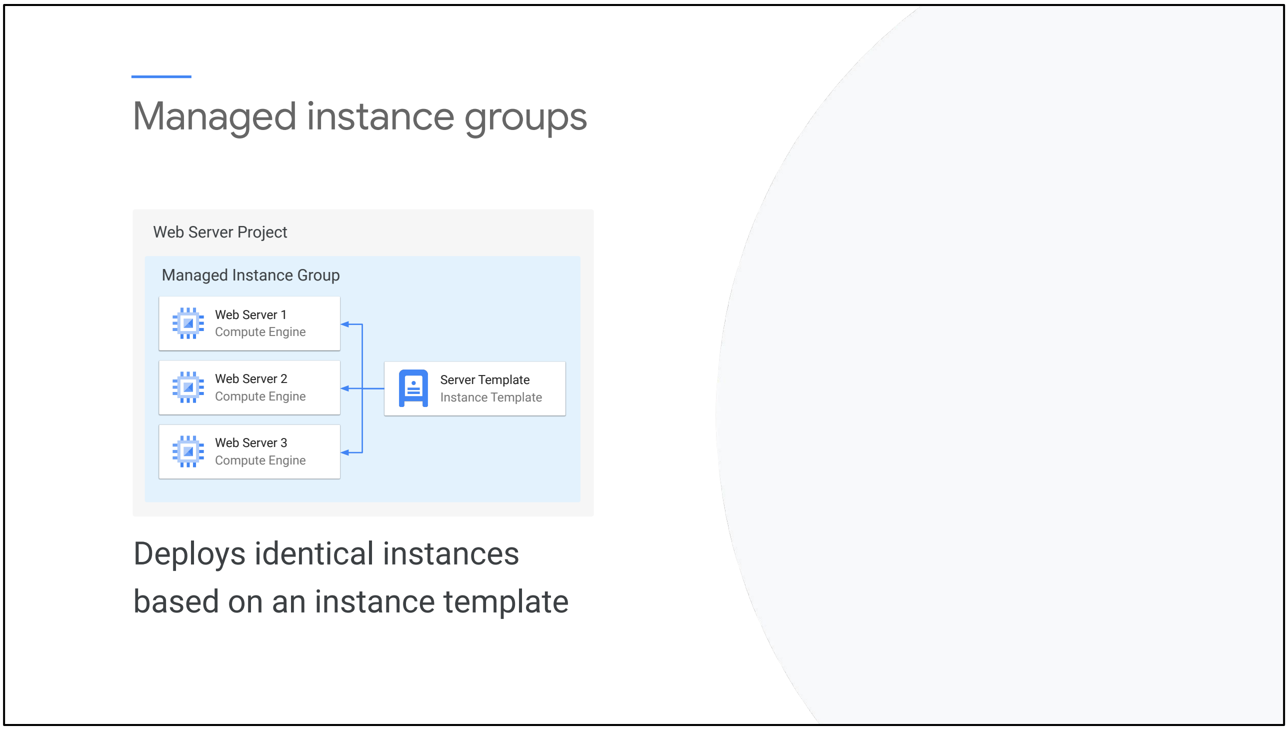


Managed instance groups

After you migrated your virtual machines to the cloud, it’s time to leverage the elasticity, automation and high availability built in to the platform.

Some of your workload can be horizontally scalable, for example, web servers or stateless application. As demand fluctuates, you ideally want to have the right amount of servers operating to serve your users.

The first step is to create an instance template, which is a blueprint of a machine with a pre-existing configuration like machine type, boot disk image, startup script, and other instance properties. For example, your boot disk can be an image of your web server preconfigured, and your startup script can start the service as soon as the machine starts.



Managed instance groups

Web Server Project

Managed Instance Group

Web Server 1
Compute Engine

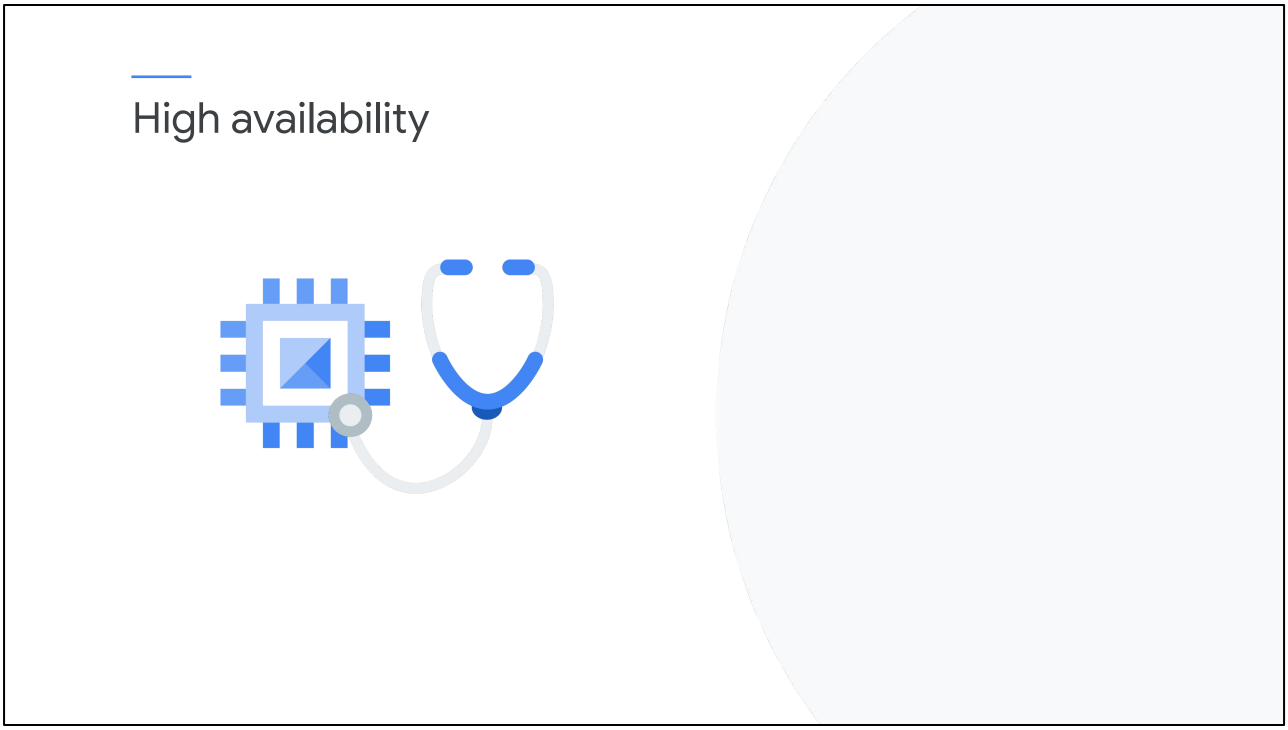
Web Server 2
Compute Engine

Web Server 3
Compute Engine

Server Template
Instance Template

Deploys identical instances based on an instance template

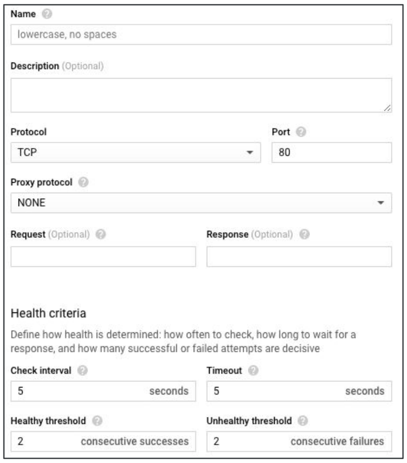
A managed instance group (MIG) is a collection of identical VM instances that you control as a single entity, using an instance template. You can configure it as a static or dynamic number of instances.



High availability

Managed instance groups maintain high availability of your applications by proactively keeping your instances available, which means the instances are all in a RUNNING state. If an instance is not responding, shuts down, or crashes, the managed instance group will replace it with a healthy instance.

High availability



Name

lowercase, no spaces

Description (Optional)

Protocol: TCP

Port: 80

Proxy protocol: NONE

Request (Optional)

Response (Optional)

Health criteria

Define how health is determined: how often to check, how long to wait for a response, and how many successful or failed attempts are decisive

Check interval: 5 seconds

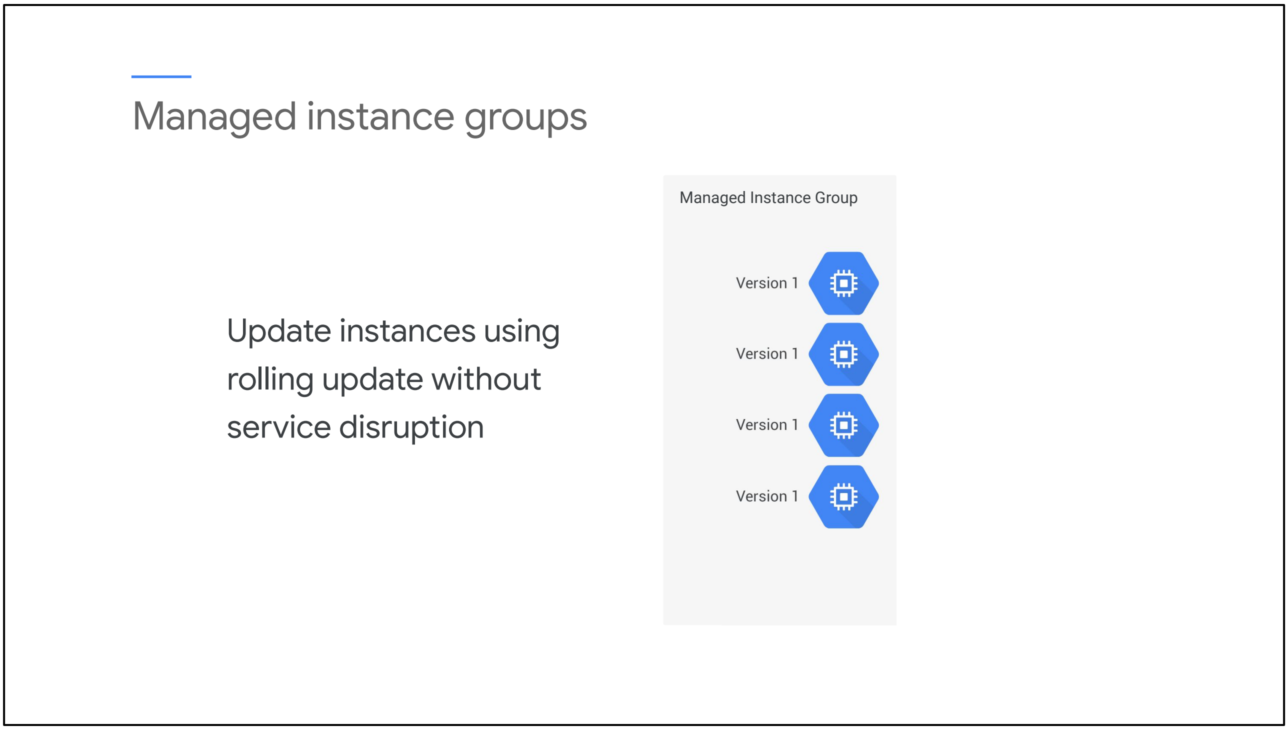
Timeout: 5 seconds

Healthy threshold: 2 consecutive successes

Unhealthy threshold: 2 consecutive failures

However, relying only on instance state may not be sufficient. You may want to recreate instances when an application freezes, crashes, or runs out of memory.

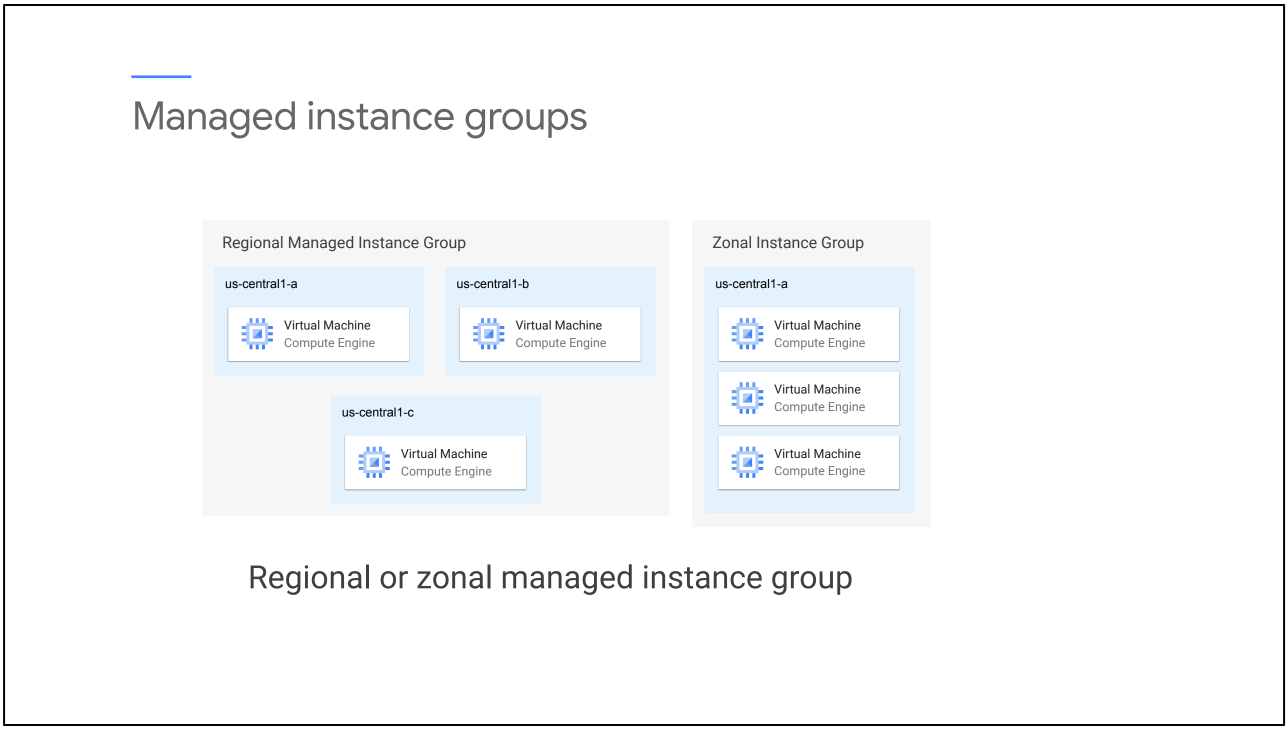
Application-based autohealing improves application availability by relying on a health checking signal that detects application-specific issues such as freezing, crashing, or overloading. If a health check determines that an application has failed on an instance, the group automatically recreates that instance.



Managed instance groups

Update instances using rolling update without service disruption

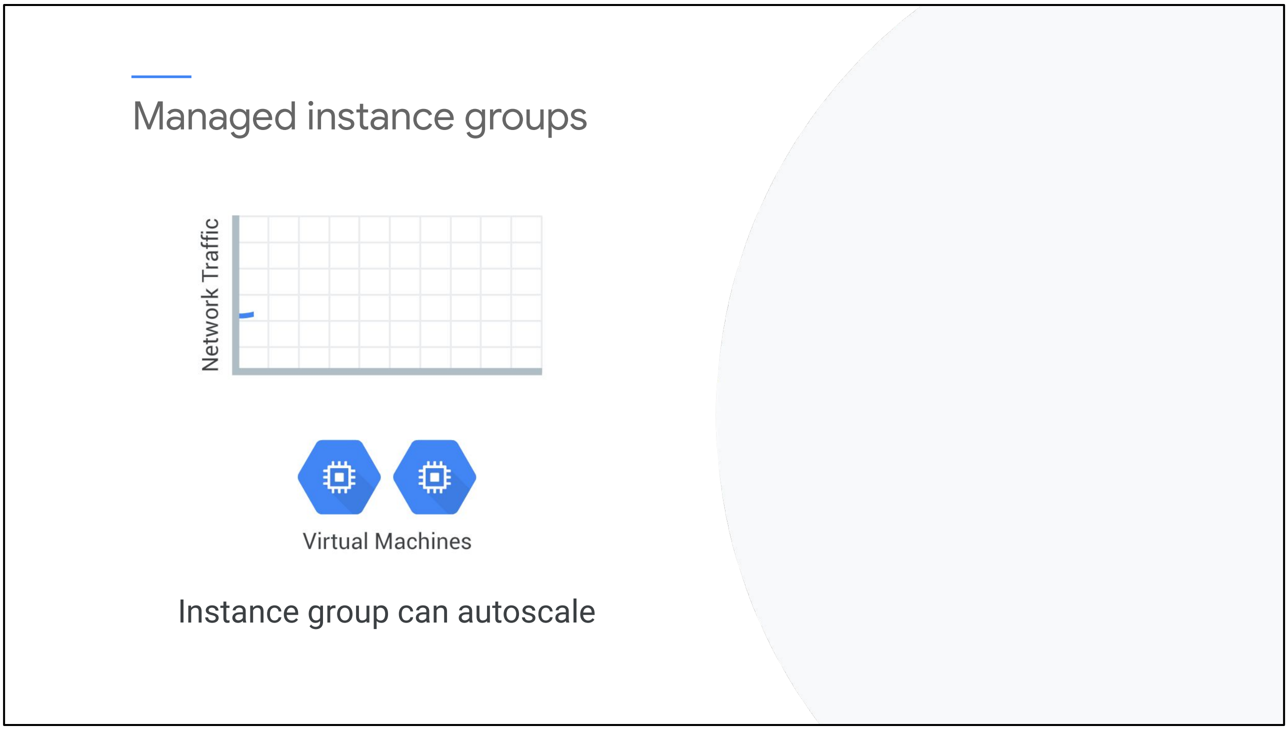
When it’s time to update your application or change the configuration of your machines in the instance group, you can create a new instance template and easily roll out an update with no downtime. A rolling update works by replacing a subset of machines from the group gradually until all machines are created from the new template. This strategy ensures continuous deployment and minimized service interruption. In addition, managed instance groups support a flexible range of other rollout scenarios, such as Canary updates.



Managed instance groups

Regional or zonal managed instance group

Regional managed instance groups are generally recommended over zonal managed instance groups because they allow you to spread your instances across multiple zones, rather than running all your instances in a single zone or having to manage multiple instance groups across different zones. Regional managed instance groups enhance the availability of your instances and protect it from unforeseen scenarios like a single zonal failure.

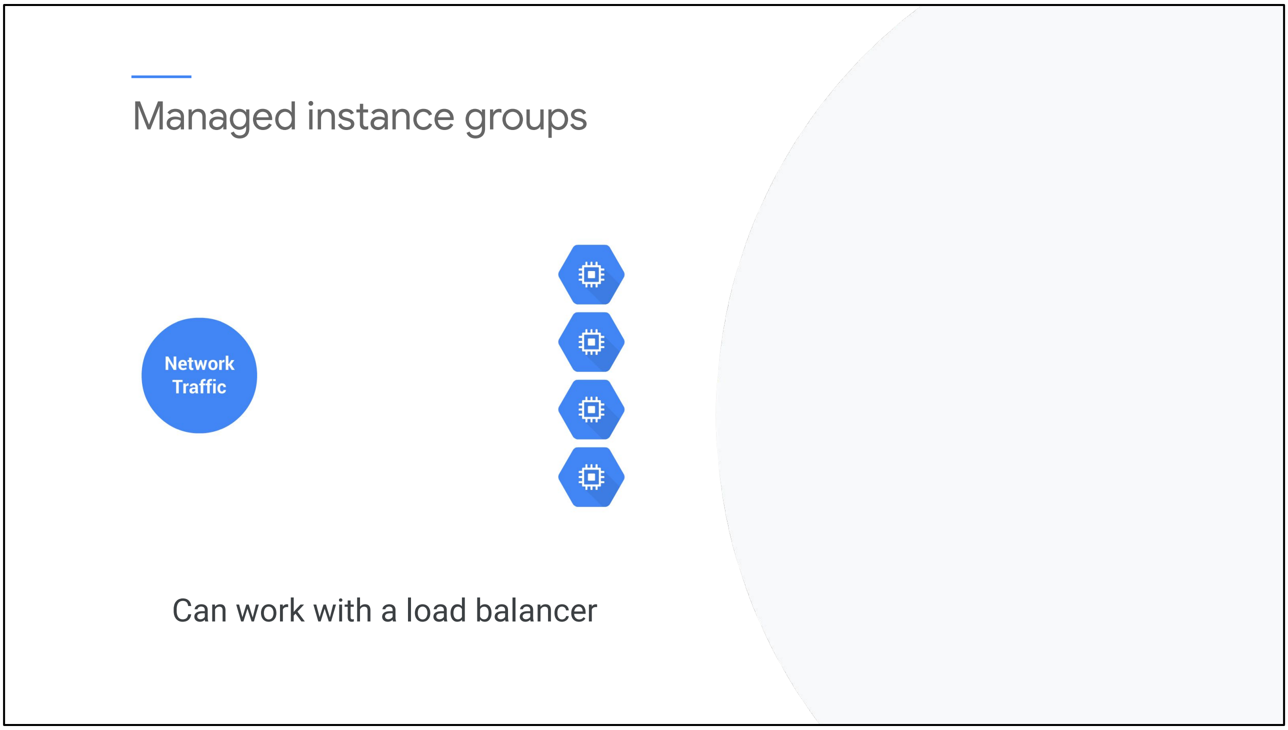


Managed instance groups

Virtual Machines

Instance group can autoscale

Managed instance groups support autoscaling that dynamically adds or removes instances from a managed instance group in response to increases or decreases in load. You turn on autoscaling and configure an autoscaling policy to specify how you want the group to scale. Autoscaling policies include scaling based on CPU utilization, load balancing capacity, or Stackdriver Monitoring metrics.

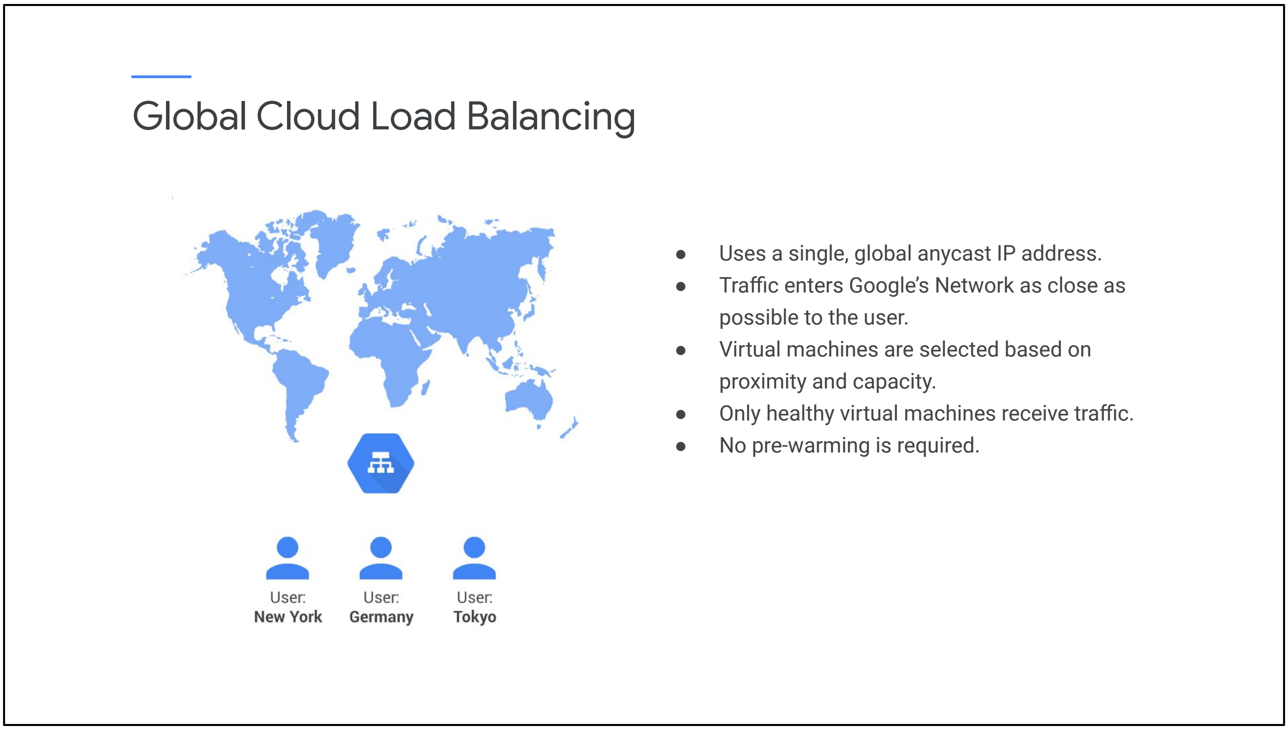


Managed instance groups

Network Traffic

Can work with a load balancer

Managed instance groups can work with load balancing services to distribute network traffic across all of the instances in the group. That way, your instance group can be accessed using a single internal or external IP address.



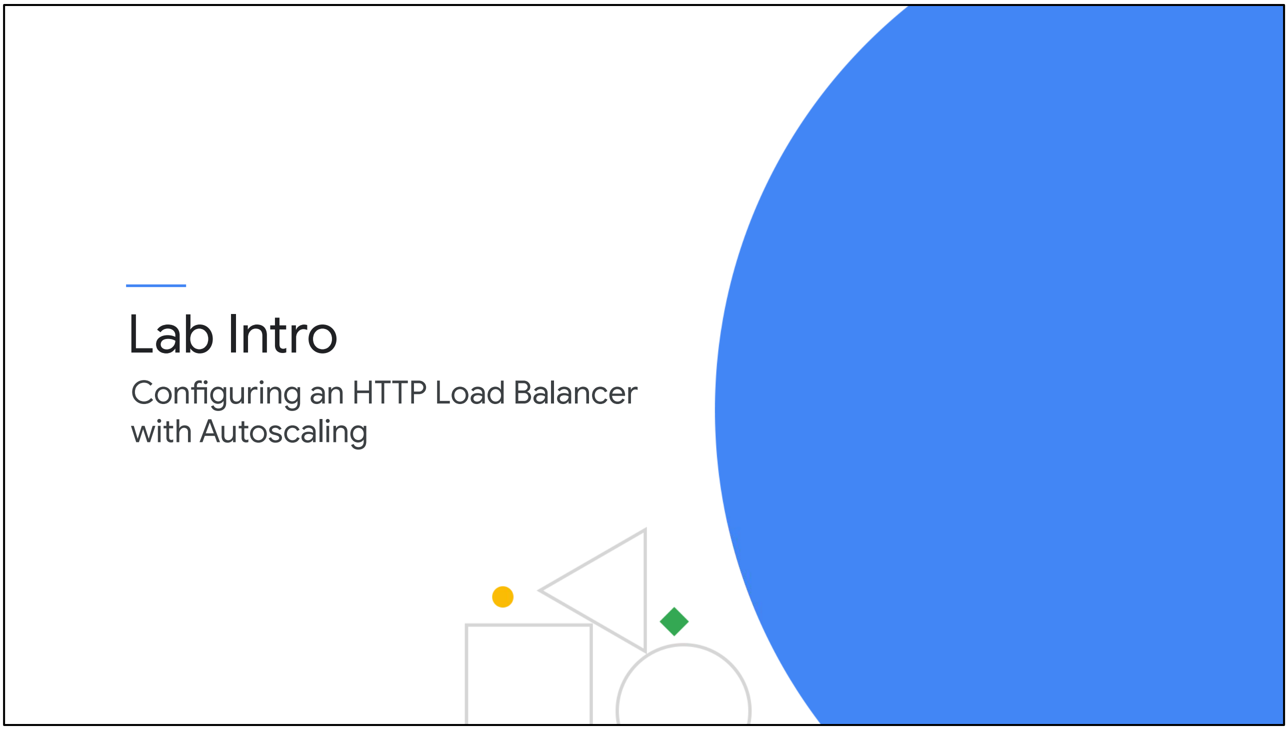
Global Cloud Load Balancing

- Uses a single, global anycast IP address.
- Traffic enters Google's Network as close as possible to the user.
- Virtual machines are selected based on proximity and capacity.
- Only healthy virtual machines receive traffic.
- No pre-warming is required.

Cloud Load Balancing is a fully distributed, software-defined, managed service that directs traffic to your instances. And because the load balancers don't run in VMs you have to manage, you don't have to worry about scaling or managing them. You can put Cloud Load Balancing in front of HTTP(S), TCP and SSL traffic.

With Cloud Load Balancing, a single anycast IP front-ends all your backend instances in regions around the world. It provides cross-region load balancing, including automatic multi-region failover, which elegantly shifts traffic to only healthy instances. Cloud Load Balancing reacts quickly to changes in users, traffic, network, backend health, and other related conditions.

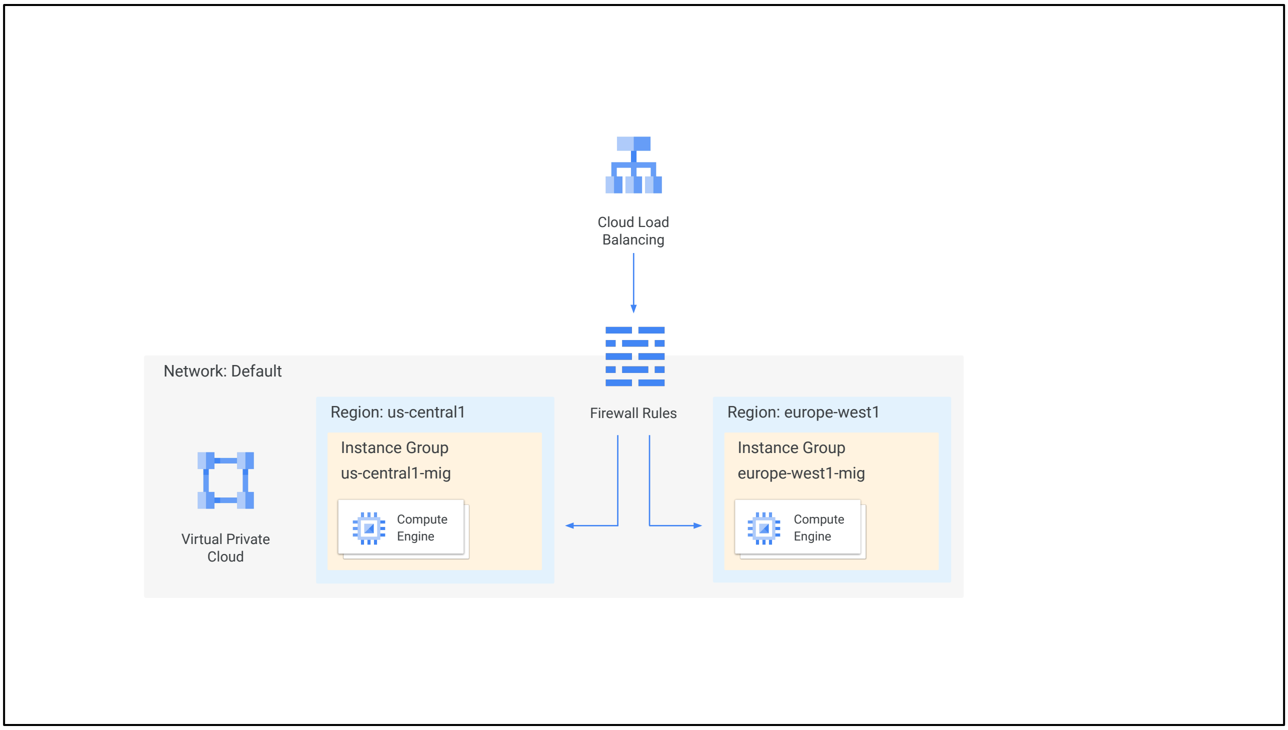
And what if you anticipate a huge spike in demand? Say, your online game is already a hit; do you need to file a support ticket to warn Google of the incoming load? No. No so-called "pre-warming" is required.



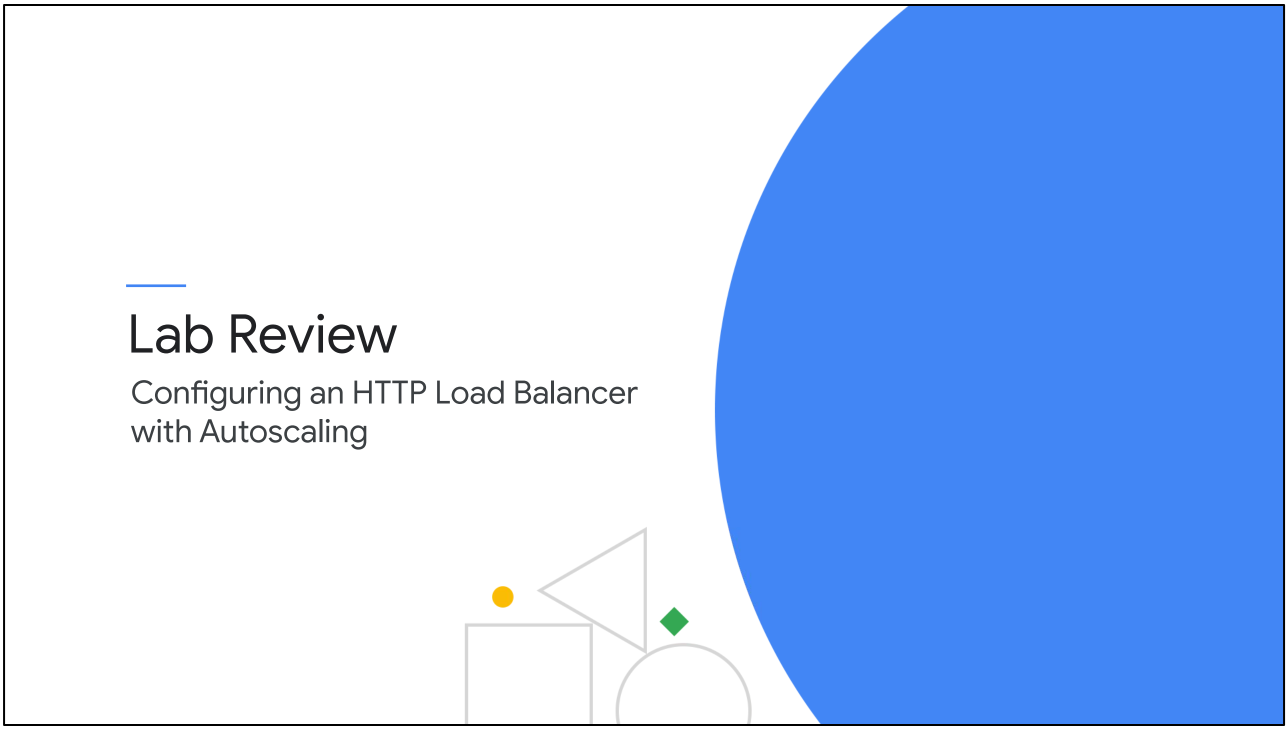
Lab Intro

Configuring an HTTP Load Balancer
with Autoscaling

Let's apply what you just learned. In this lab, you will make an instance template from a migrate virtual machine. You will then create an instance group based on that template and front ended with a global load balancer. You then stress test the instance group which triggers the autoscaler to deploy more machines to handle the load.



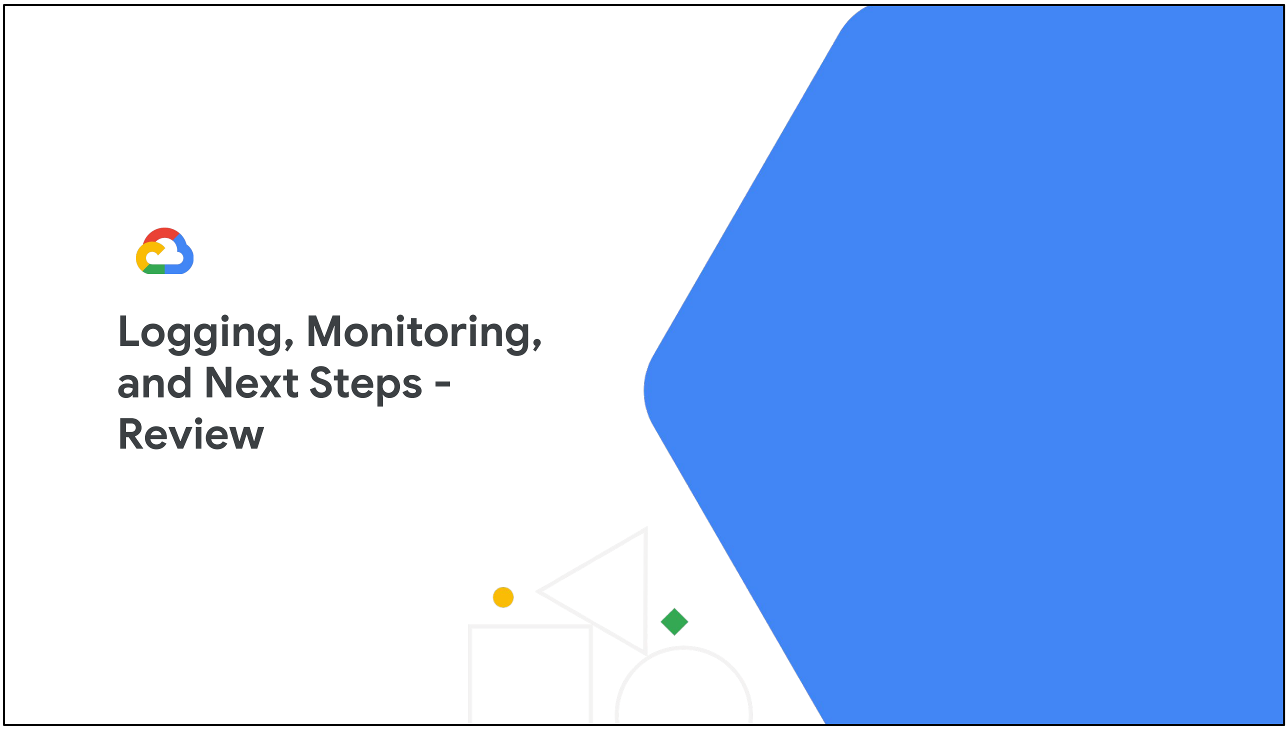
Specifically, you create two managed instance groups that serve as backends in us-central1 and europe-west1. Then, you create and stress test a load balancer to demonstrate global load balancing and autoscaling.



Lab Review

Configuring an HTTP Load Balancer with Autoscaling

In this lab, you created an instance group from the migrated virtual machine, configured a global load balancer and triggered and auto-scaling, which demonstrate the elasticity and flexibility of the Cloud.



Logging, Monitoring, and Next Steps - Review

In this final module, you have learned about how to use Cloud Monitoring and Cloud Logging capabilities to enhance the observability of your cloud environment. You also learned how to use managed instance groups to automate groups of virtual machines that can scale up and down based on metrics, as well as maintaining high availability of your applications by proactively keeping your instances available. You also learned how to use Google Cloud's HTTP load balancer to provide global load balancing.

Now it's your turn. Use what you learnt in this course to migrate your workloads to Google Cloud. Good luck and see you next time!