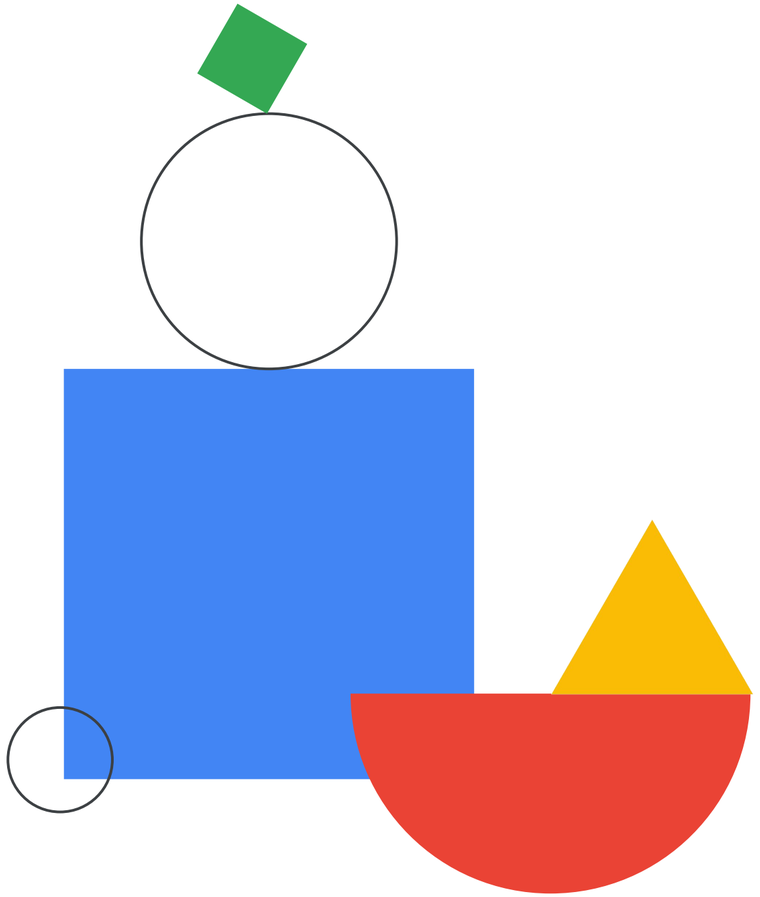
Google Cloud

Load Balancing and Autoscaling



Global and regional load balancers

Global

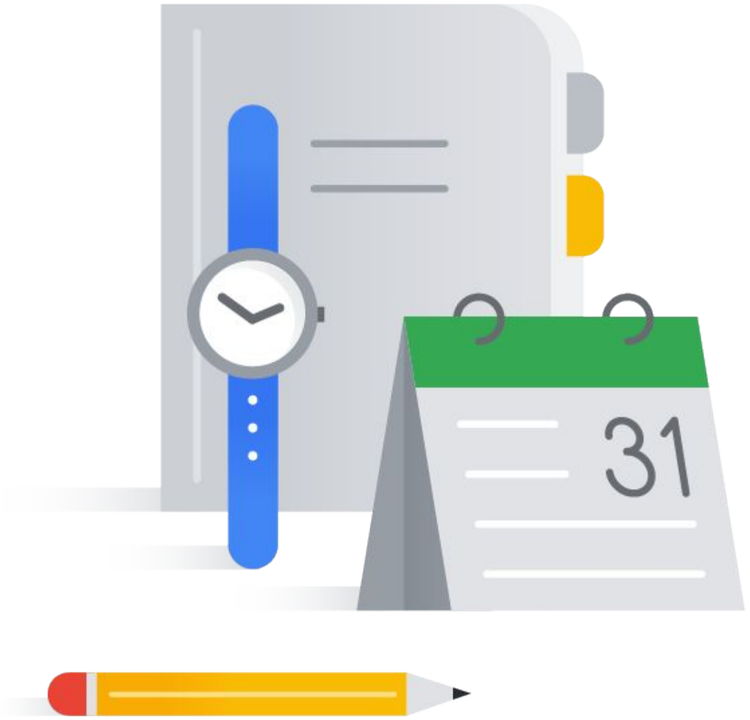
- Global External HTTP(S) Load Balancing
- External HTTP(S) Load Balancing (classic)
- External SSL Proxy Load Balancing
- External TCP Proxy Load Balancing

Regional

- Regional External HTTP(S) Load Balancing
- Internal HTTP(S) Load Balancing
- Internal TCP Proxy Load Balancing
- Internal TCP/UDP Load Balancing
- External TCP/UDP Network Load Balancing
- External Regional TCP Proxy Load Balancing

Agenda

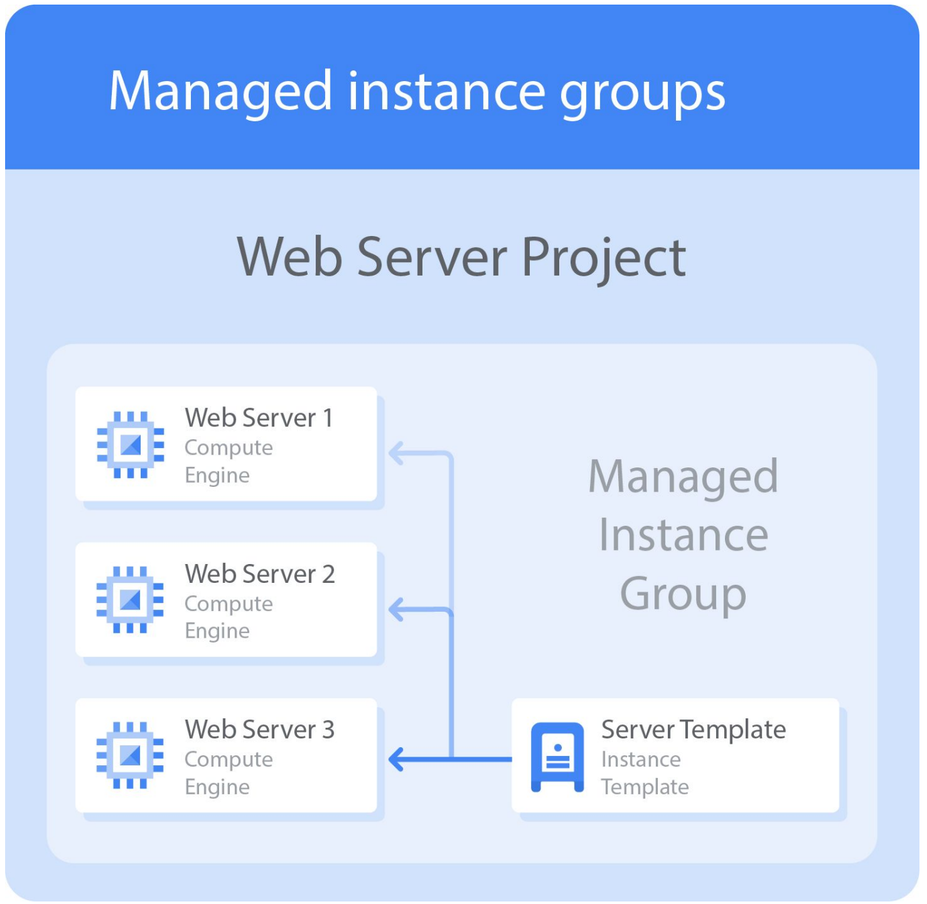
01	Managed Instance Groups
02	HTTP(S) Load Balancing Lab: Configuring an HTTP Load Balancer with Autoscaling
03	Cloud CDN
04	SSL Proxy/TCP Proxy Load Balancing
05	Network Load Balancing
06	Internal Load Balancing Lab: Configuring an Internal Load Balancer
07	Choosing a Load Balancer



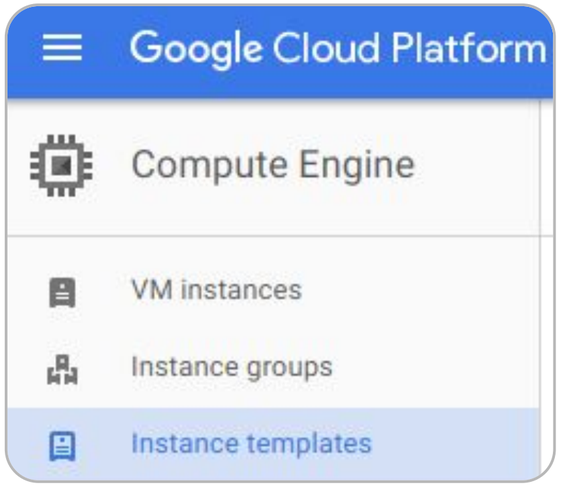
Managed Instance Groups

Managed instance groups

- Deploy identical instances based on instance template
- Instance group can be resized
- Manager ensures all instances are RUNNING
- Typically used with autoscaler
- Can be single zone or regional



Create an instance template



Name *
instance-template-1

Labels

+ ADD LABELS

Machine configuration

General purpose | Compute optimized | Memory optimized | GPUs

Machine types for common workloads, optimized for cost and flexibility

Series
E2

CPU platform selection based on availability

Machine type

Choose a machine type with preset amounts of vCPUs and memory that suit most workloads. Or, you can create a custom machine for your workload's particular needs. Learn more

PRESET | CUSTOM

e2-medium (2 vCPU, 4 GB memory)

vCPU	Memory
1-2 vCPU (1 shared core)	4 GB

Boot disk

Name	instance-template-1
Type	New balanced persistent disk
Size	10 GB
License type	Free
Image	Debian GNU/Linux 11 (bullseye)

CHANGE

Identity and API access

Service accounts

Service account
Compute Engine default service account

Requires the Service Account User role (roles/iam.serviceAccountUser) to be set for users who want to access VMs with this service account. Learn more

Access scopes

- (•) Allow default access
- ( ) Allow full access to all Cloud APIs
- ( ) Set access for each API

Firewall

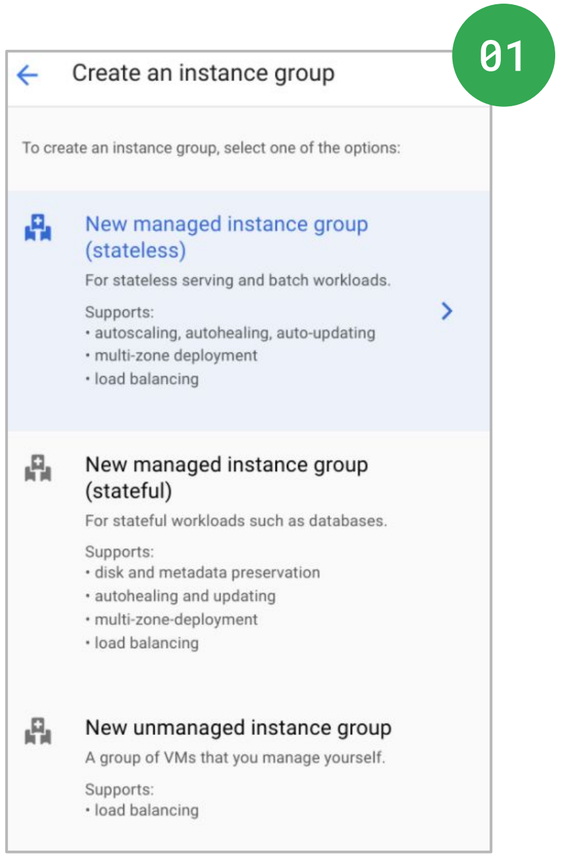
Add tags and firewall rules to allow specific network traffic from the Internet

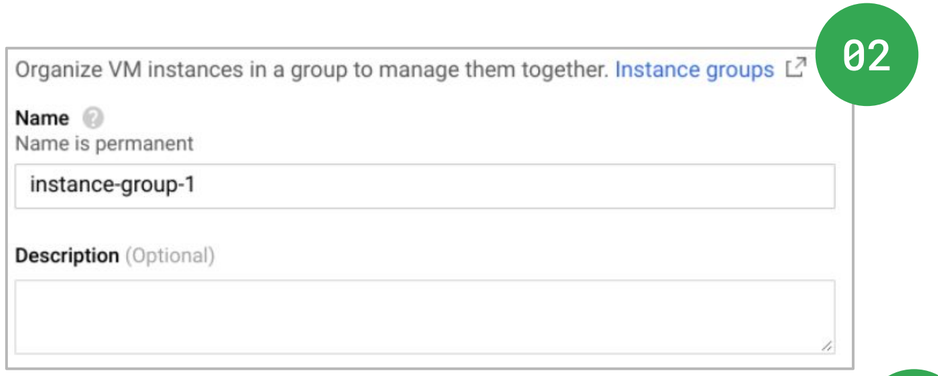
- [ ] Allow HTTP traffic
- [ ] Allow HTTPS traffic

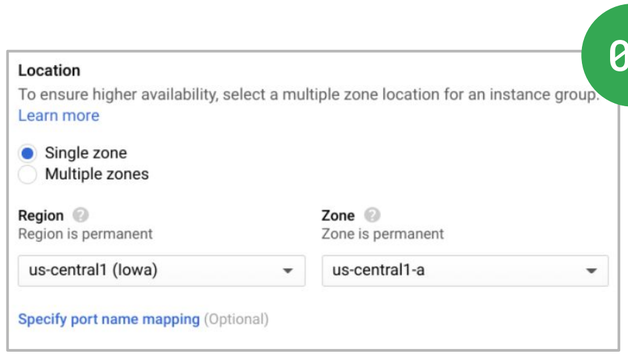
Advanced options

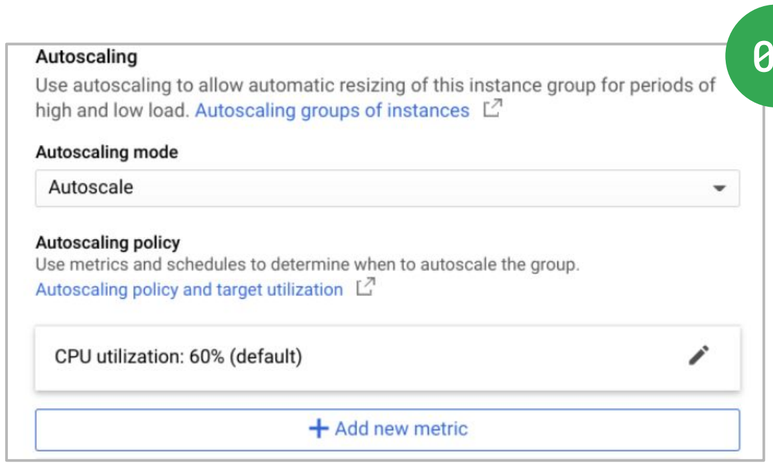
Networking, disks, security, management, sole-tenancy

Create a managed instance group









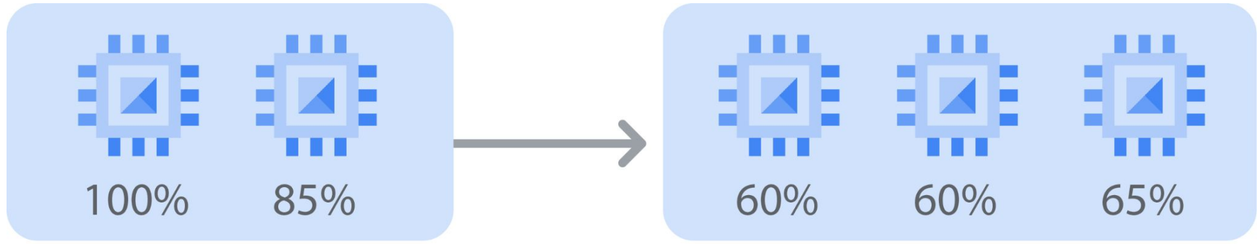
Managed instance groups offer autoscaling capabilities

Dynamically add/remove instances:

- Increases in load
- Decreases in load

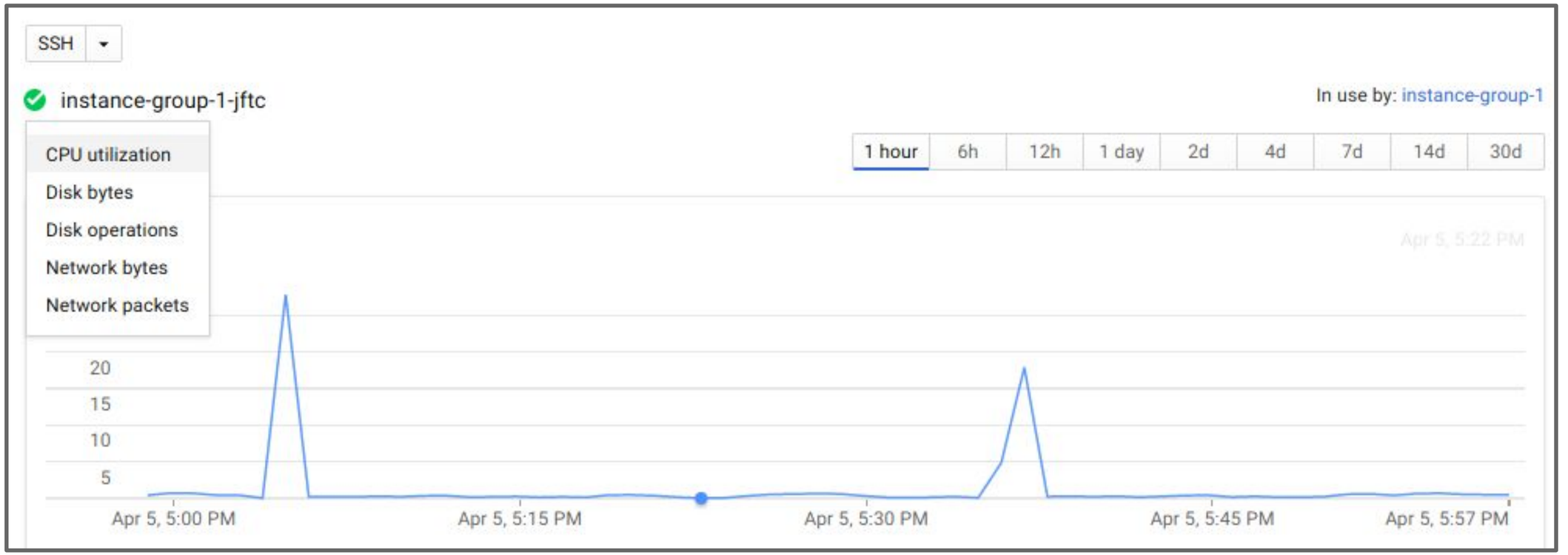
Autoscaling policy:

- CPU utilization
- Load balancing capacity
- Monitoring metrics
- Queue-based workload
- Schedule-based

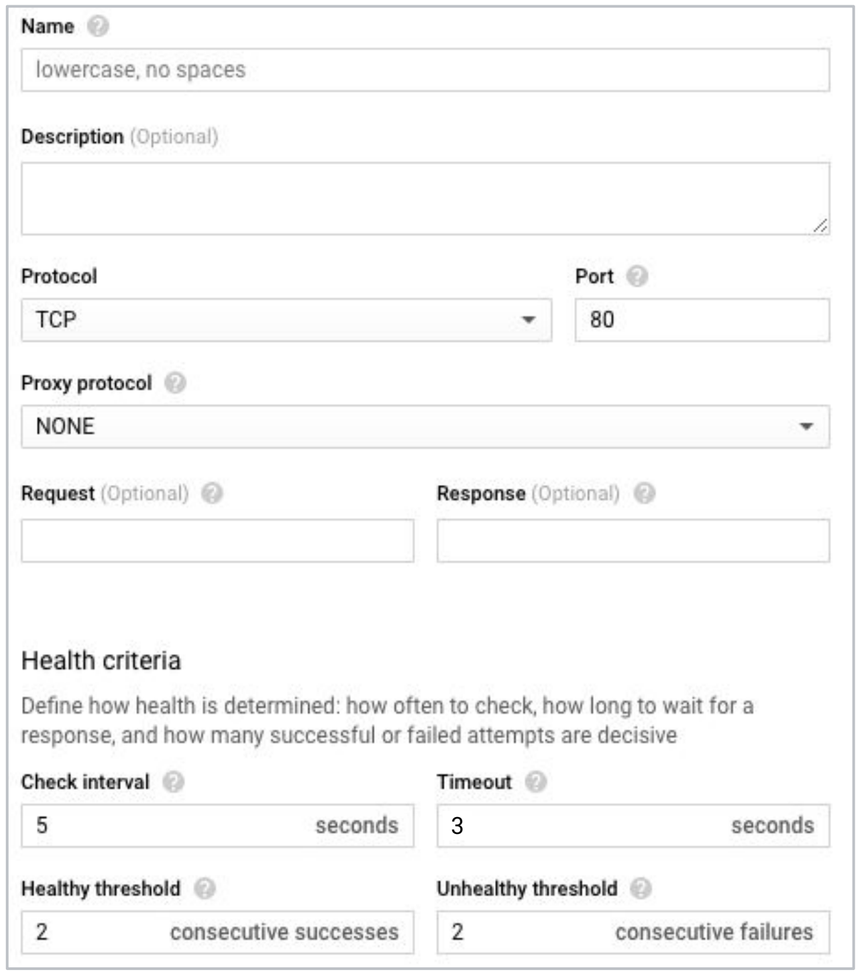


Target CPU utilization = 75%

VM graph helps set CPU utilization



Create a health check



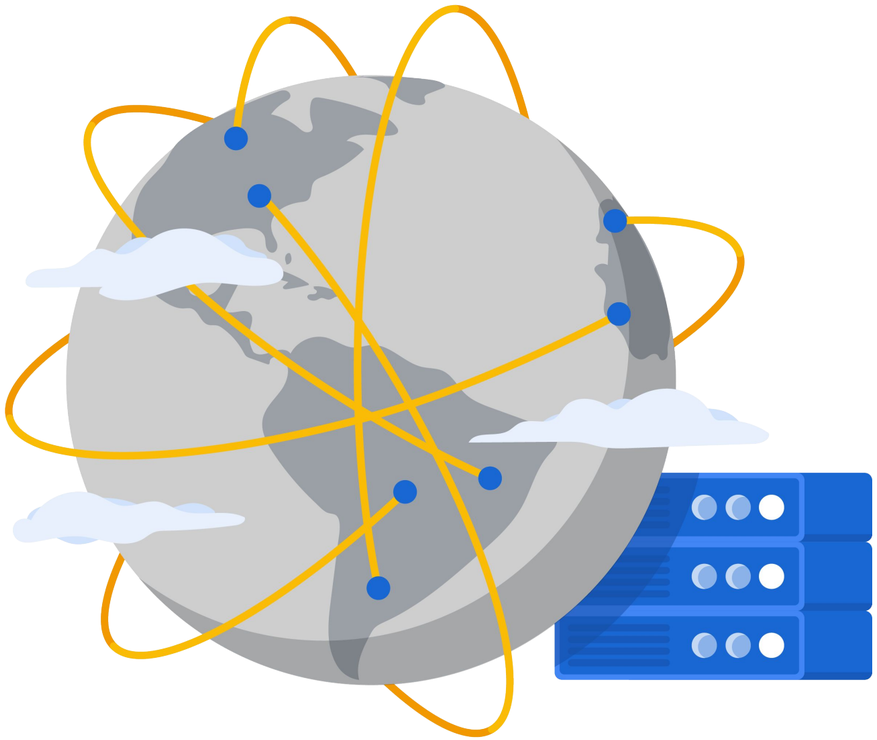
Elapsed time (seconds)	Event duration (seconds)	
1	1	wait
2	2	wait
3	3	wait
4	4	wait
5	5	wait
6	1	**health check #1 starts**
7	2	wait
8	3	wait
9		**health check #1 fails**
10		wait
11	1	**health check #2 starts**
12	2	wait
13	3	wait
14		**health check #2 fails**
15		**Unhealthy threshold reached**

HTTP(S) Load Balancing

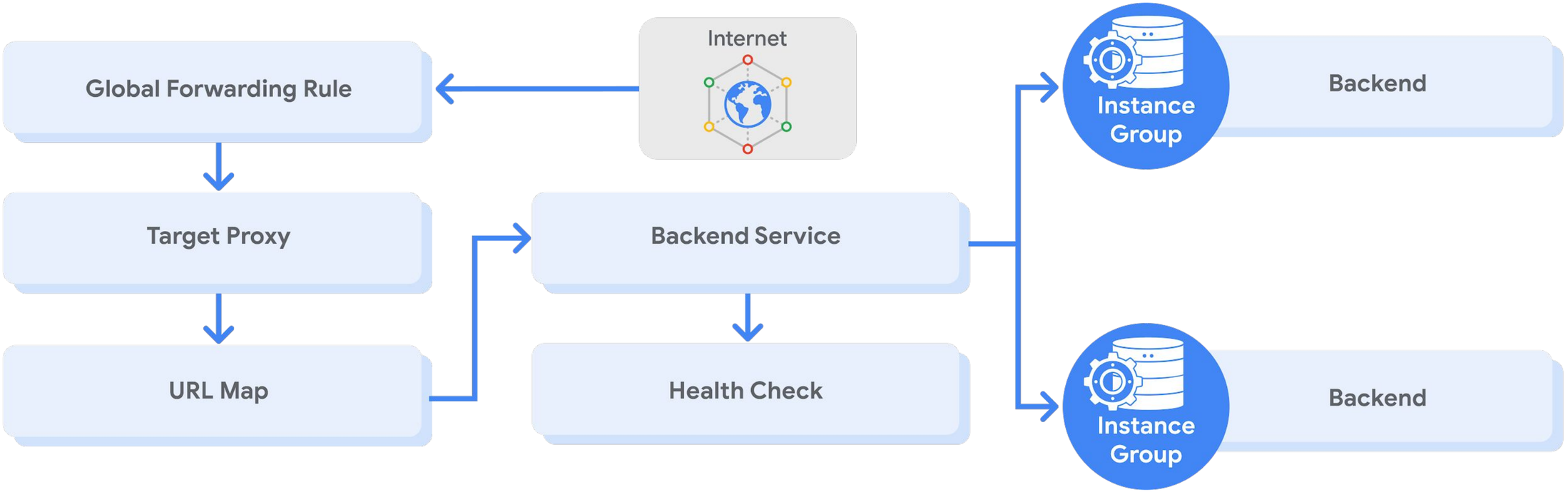
HTTP(S) load balancing

HTTP(S) Load Balancing

- Global load balancing
- Anycast IP address
- HTTP or port 80 or 8080
- HTTPs on port 443
- IPv4 or IPv6
- Autoscaling
- URL maps

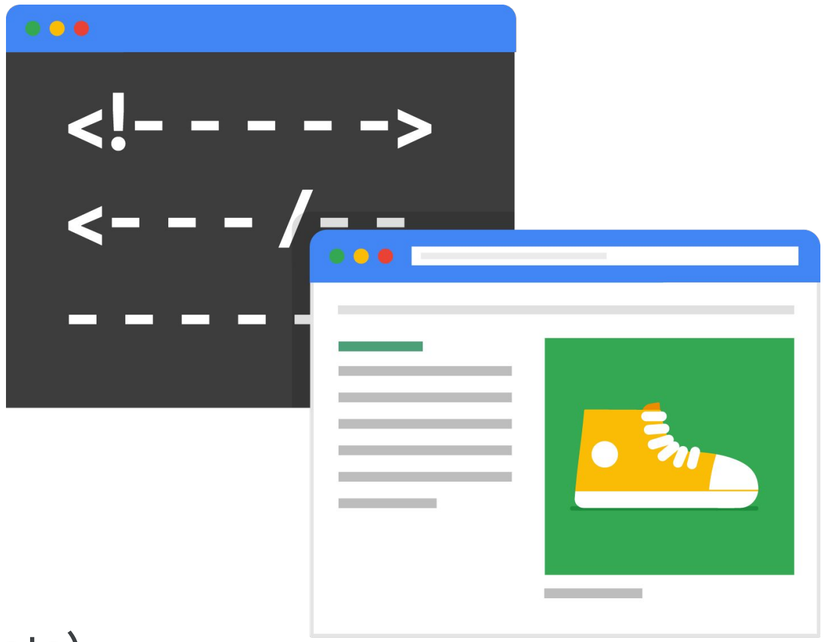


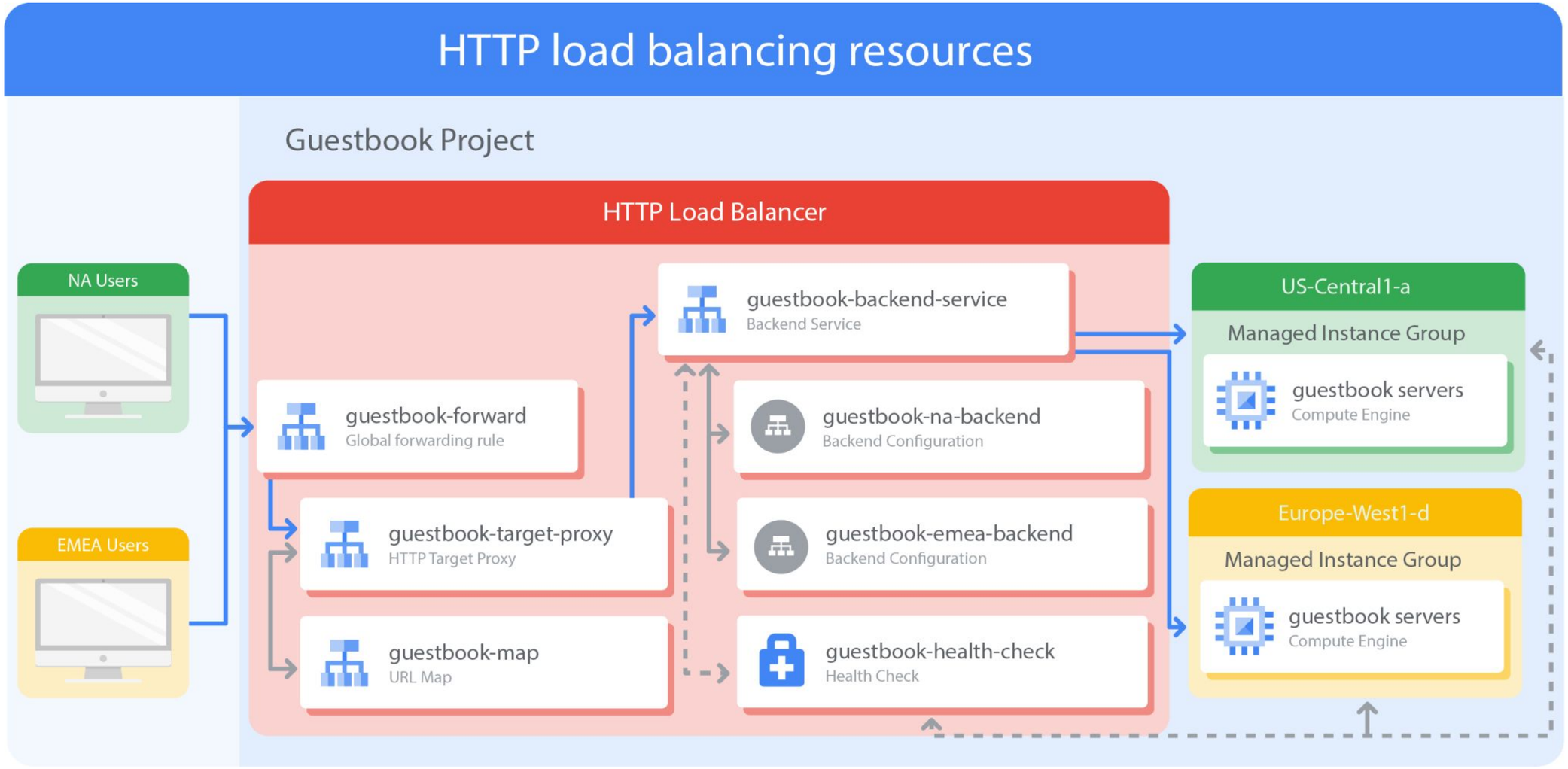
Architecture of an HTTP(S) load balancer



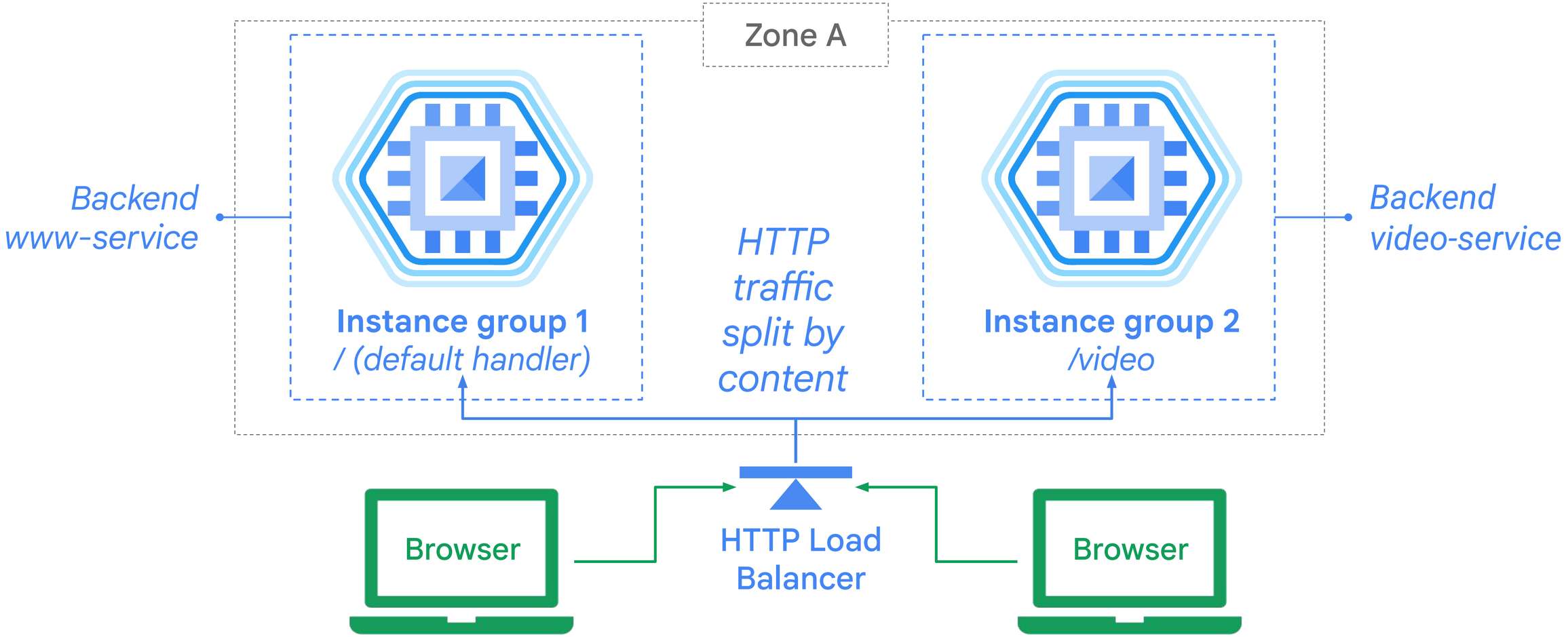
Backend services

- Health check
- Session affinity (optional)
- Time out setting (30-sec default)
- One or more backends
  - An instance group (managed or unmanaged)
  - A balancing mode (CPU utilization or RPS)
  - A capacity scaler (ceiling percentage of CPU/Rate targets)



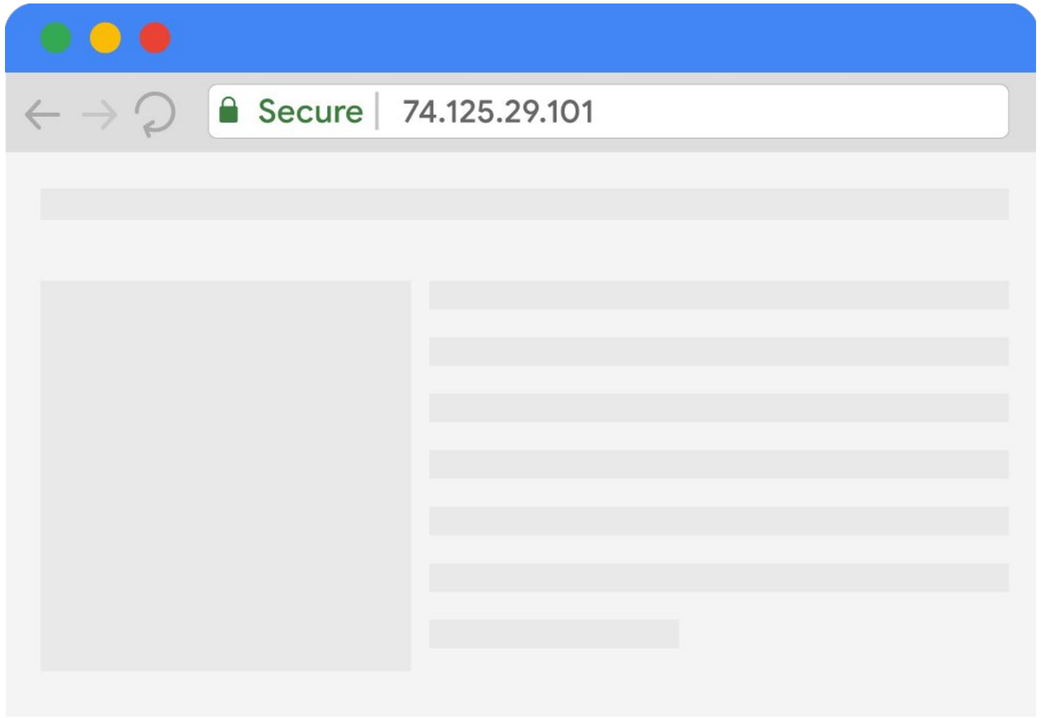


Example: Content-based load balancing



HTTP(S) load balancing

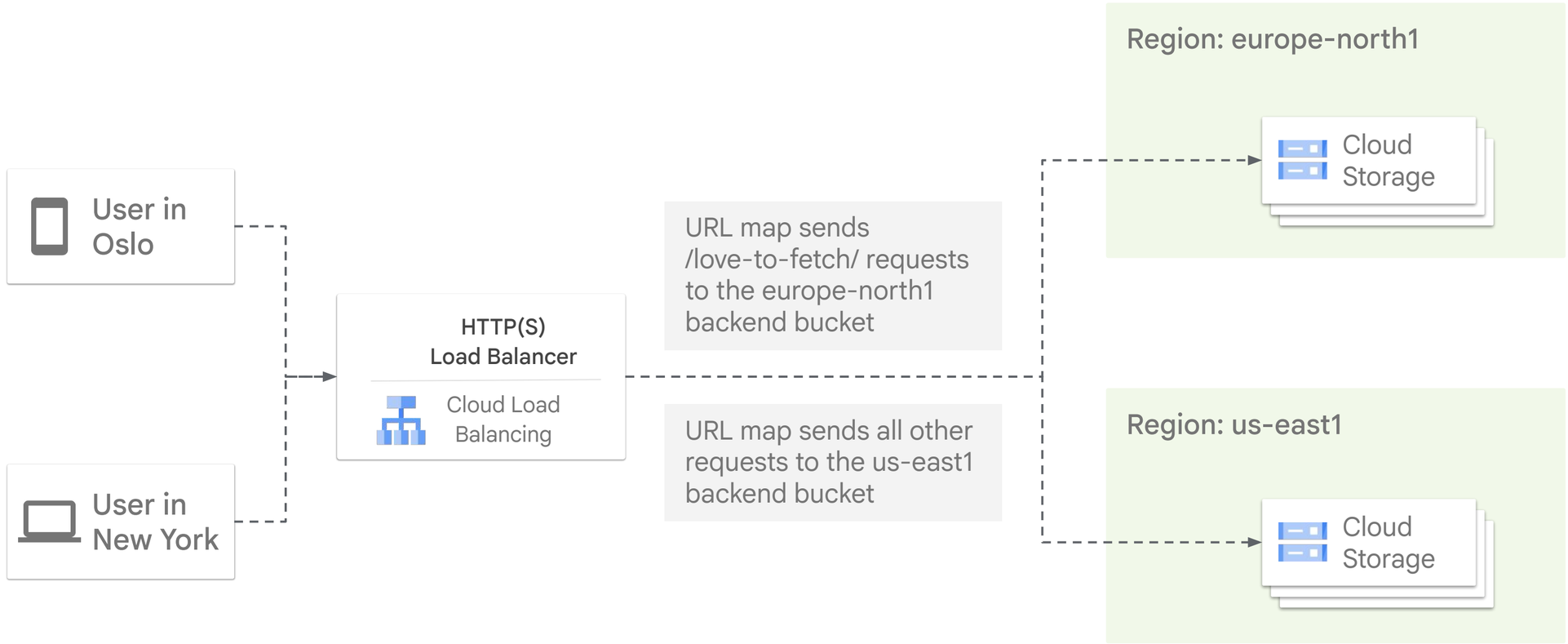
- Target HTTP(S) proxy
- One signed SSL certificate installed (minimum)
- Client SSL session terminates at the load balancer
- Support the QUIC transport layer protocol



SSL certificates

- Required for HTTP(S) load balancing
- Up to 15 SSL certificates (per target proxy)
- Create an SSL certificate resource

Backend buckets



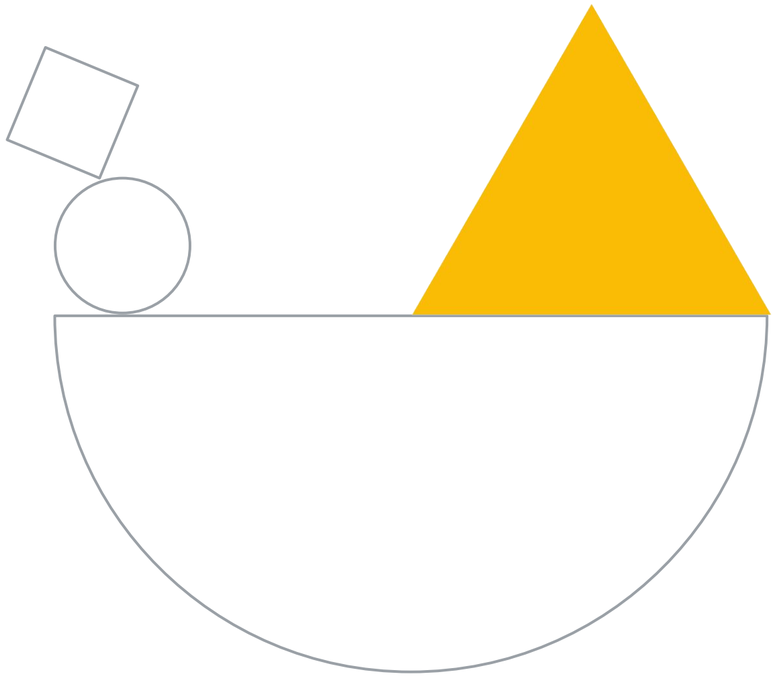
Network endpoint groups (NEG)

A network endpoint group (NEG) is a configuration object that specifies a group of backend endpoints or services.

There are four types of NEGs:

- Zonal
- Internet
- Serverless
- Hybrid connectivity

Lab: Configuring an HTTP Load Balancer with Autoscaling



Lab objectives

01 Create HTTP and health check firewall rules

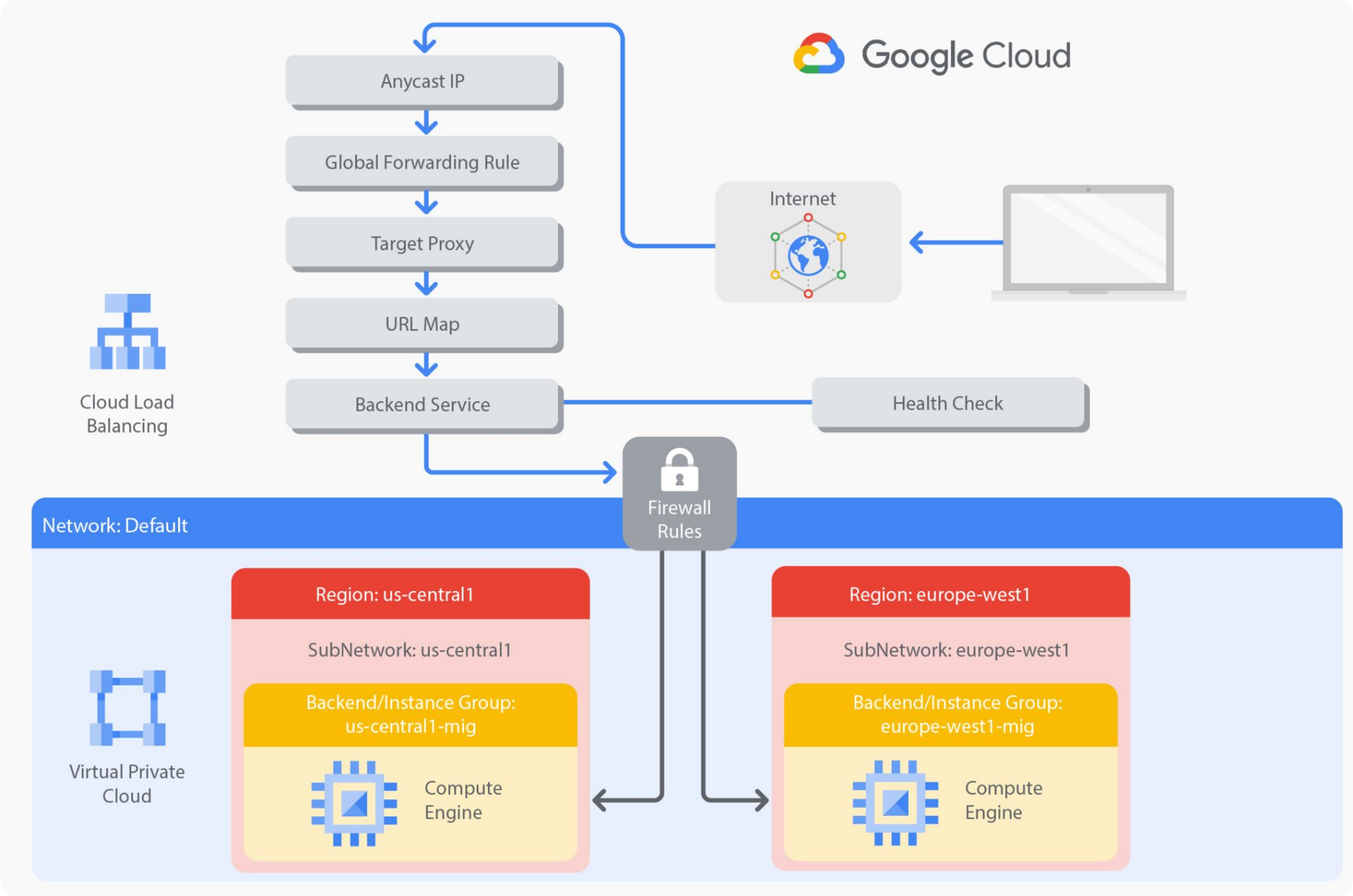
02 Create a custom image for a web server

03 Create an instance template based on the custom image

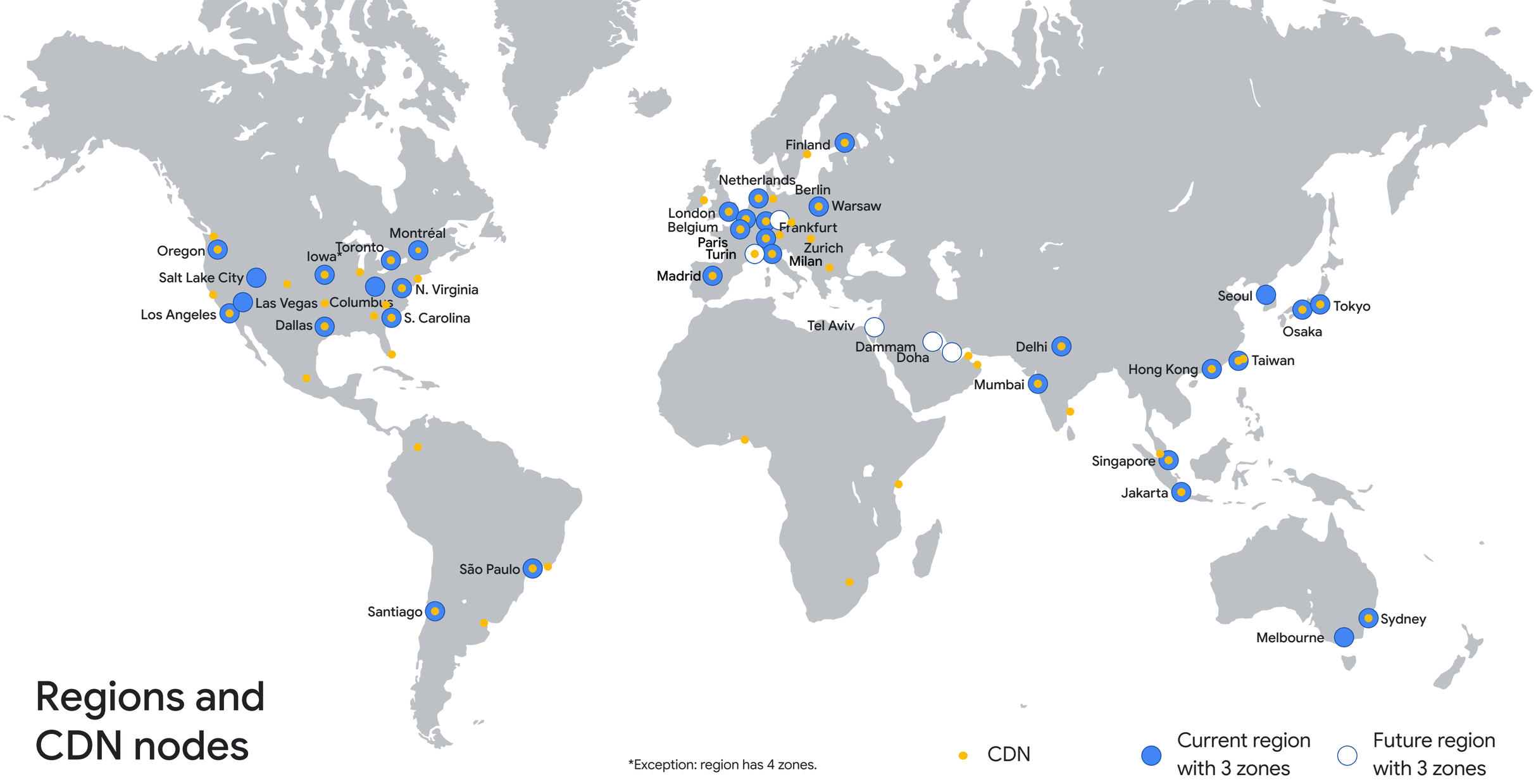
04 Create two managed instance groups

05 Configure an HTTP load balancer with IPv4 and IPv6

06 Stress test an HTTP load balancer



Cloud CDN



Regions and CDN nodes

*Exception: region has 4 zones.

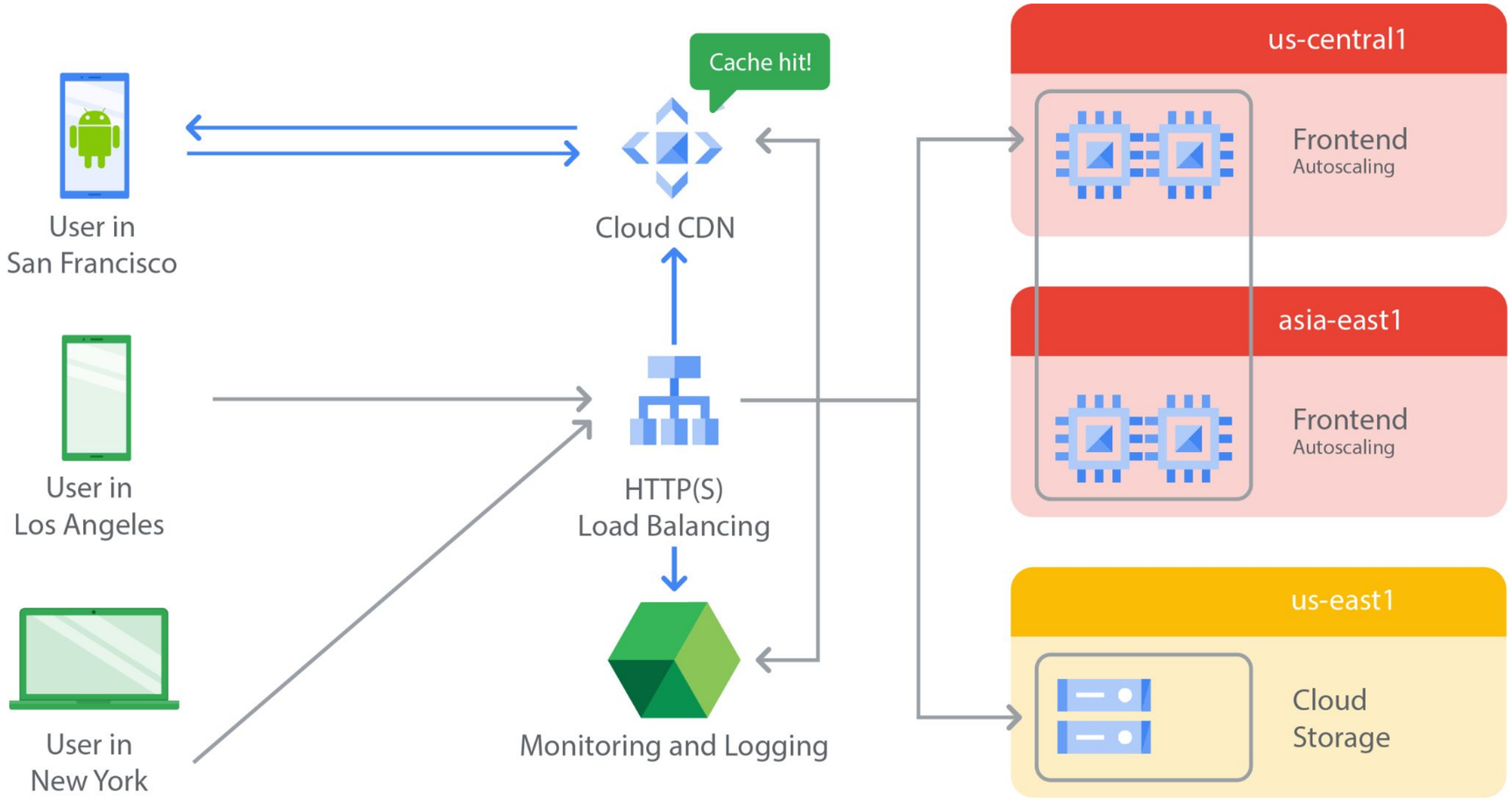
CDN

Current region with 3 zones

Future region with 3 zones

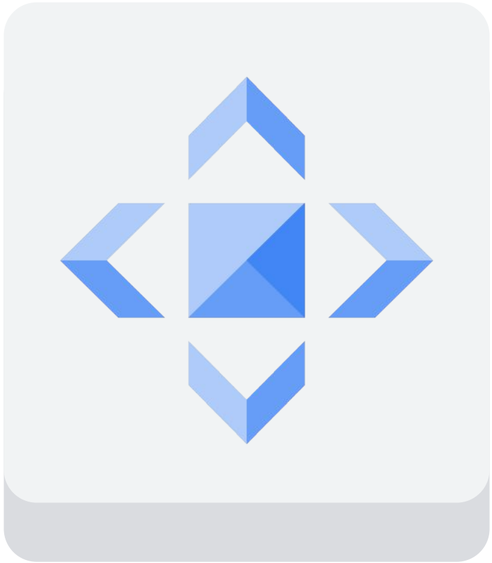
Google Cloud

Caching content with Cloud CDN



Cloud CDN cache modes

- Cache modes control the factors that determine whether or not Cloud CDN caches your content.
- Cloud CDN offers three cache modes:
  - USE_ORIGIN_HEADERS
  - CACHE_ALL_STATIC
  - FORCE_CACHE_ALL

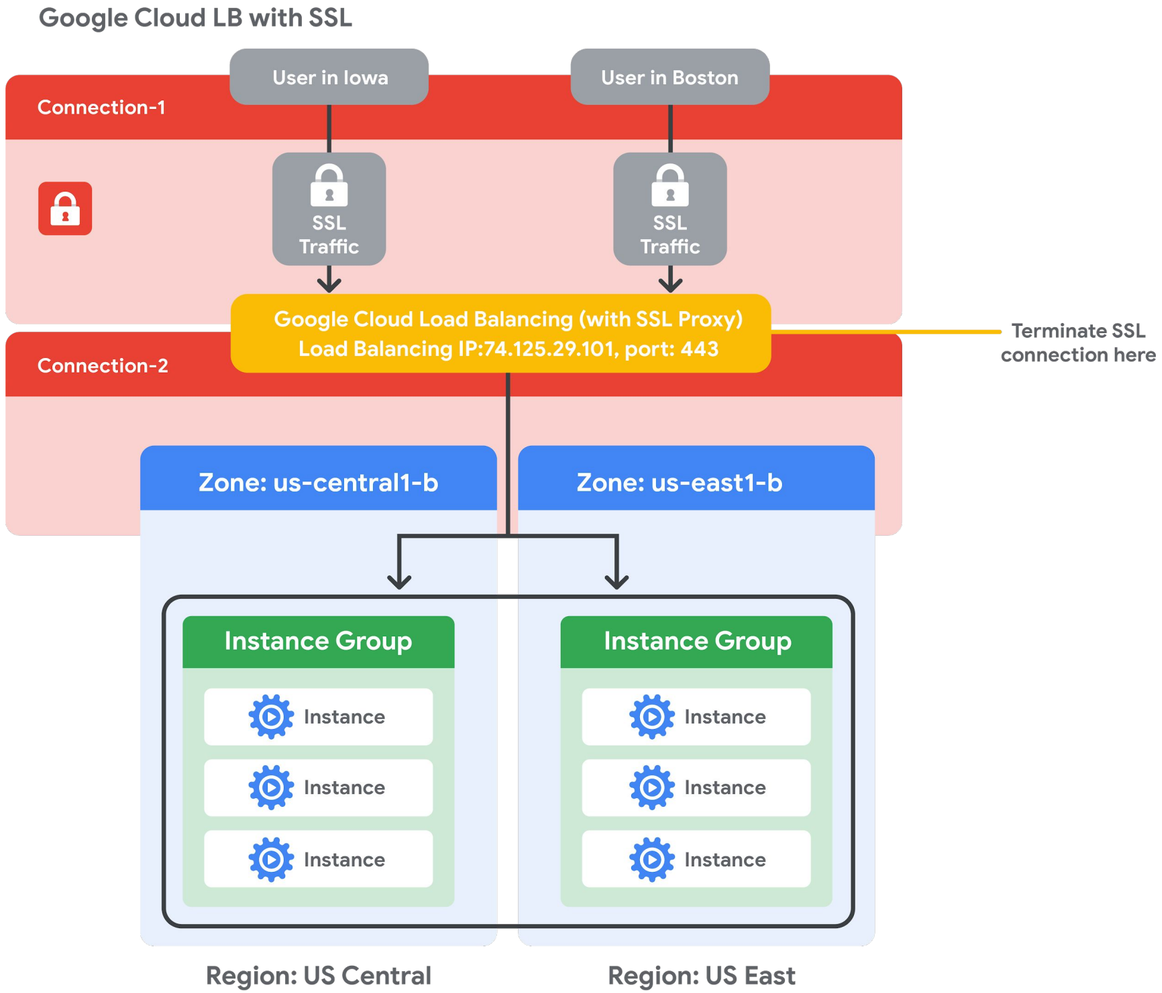


SSL Proxy/TCP Proxy Load Balancing

SSL proxy load balancing

- Global load balancing for encrypted, non-HTTP traffic
- Terminates SSL session at load balancing layer
- IPv4 or IPv6 clients
- Benefits:
  - Intelligent routing
  - Certificate management
  - Security patching
  - SSL policies

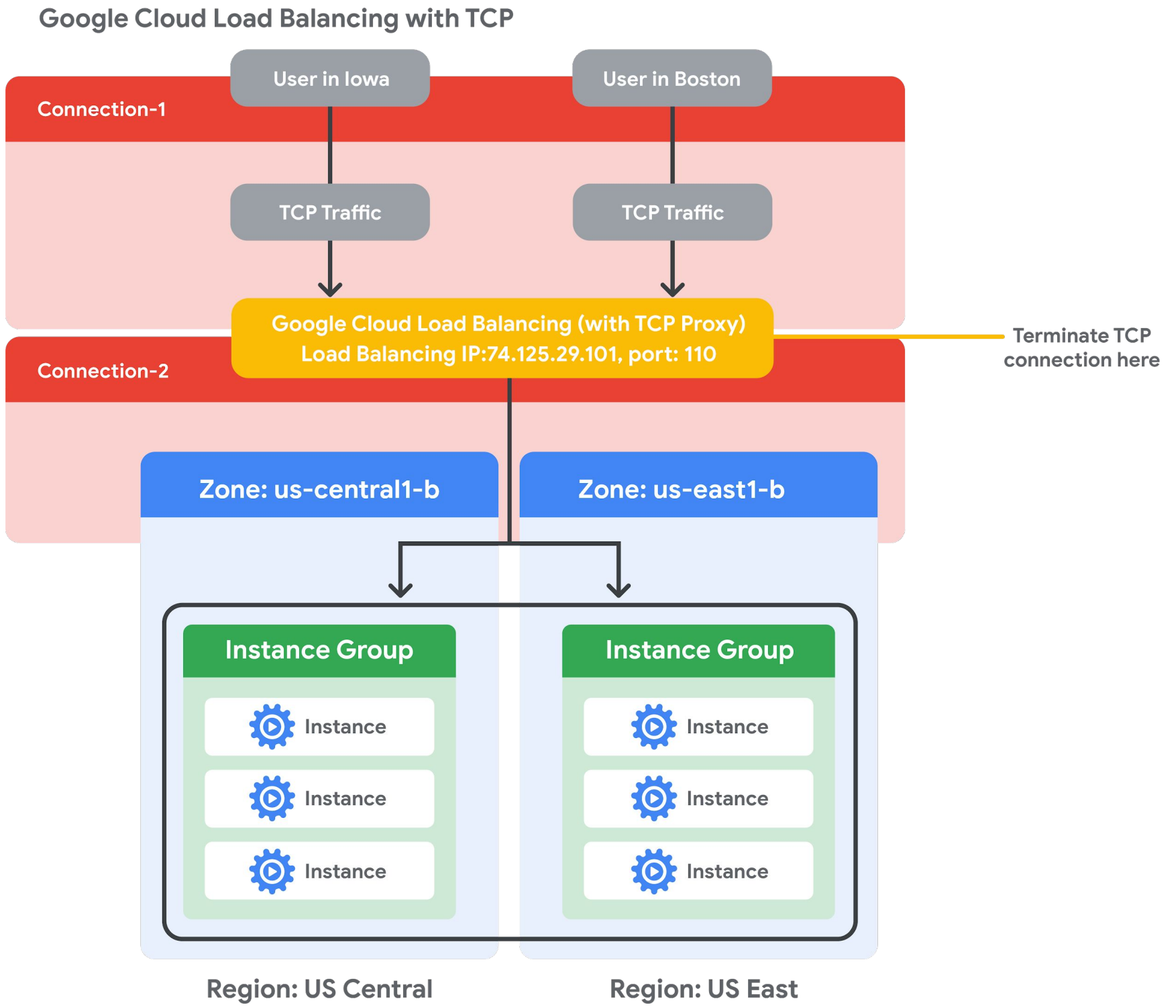
Example: SSL proxy load balancing



TCP proxy load balancing

- Global load balancing for unencrypted, non-HTTP traffic
- Terminates TCP sessions at load balancing layer
- IPv4 or IPv6 clients
- Benefits:
  - Intelligent routing
  - Security patching

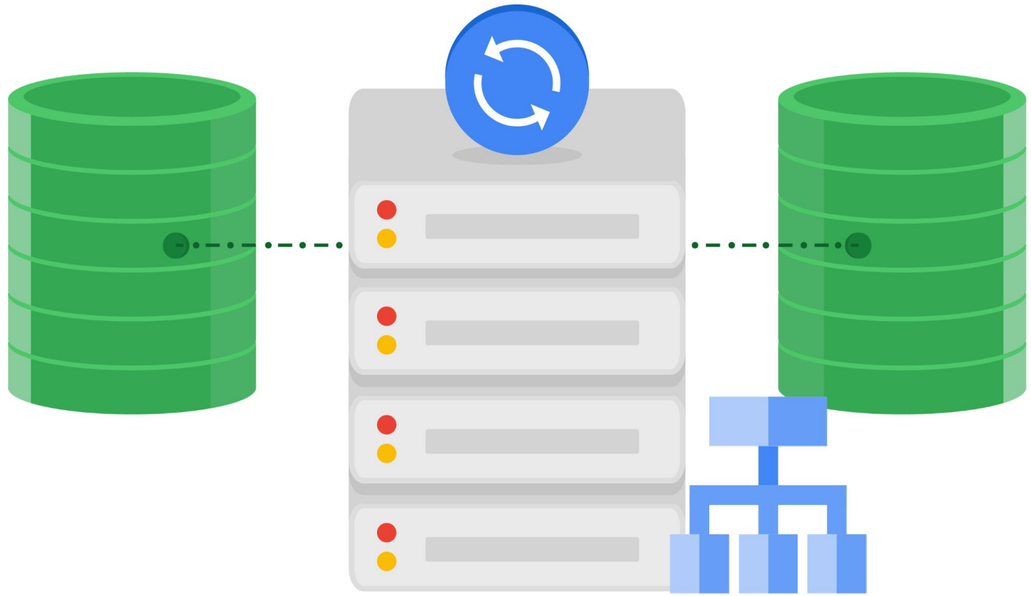
Example: TCP proxy load balancing



Network Load Balancing

Network load balancing

- Regional, non-proxied load balancer
- Forwarding rules (IP protocol data)
- Traffic:
  - UDP
  - TCP/SSL ports
- Architecture:
  - Backend service-based
  - Target pool-based



Backend service-based architecture

- Regional backend service
- Defines the behavior of the load balancer and how it distributes traffic to its backend instance groups
- Enables new features not supported with legacy target pools
  - Non-legacy health checks
  - Auto-scaling with managed instance groups
  - Connection draining
  - Configurable failover policy

Target pool-based architecture

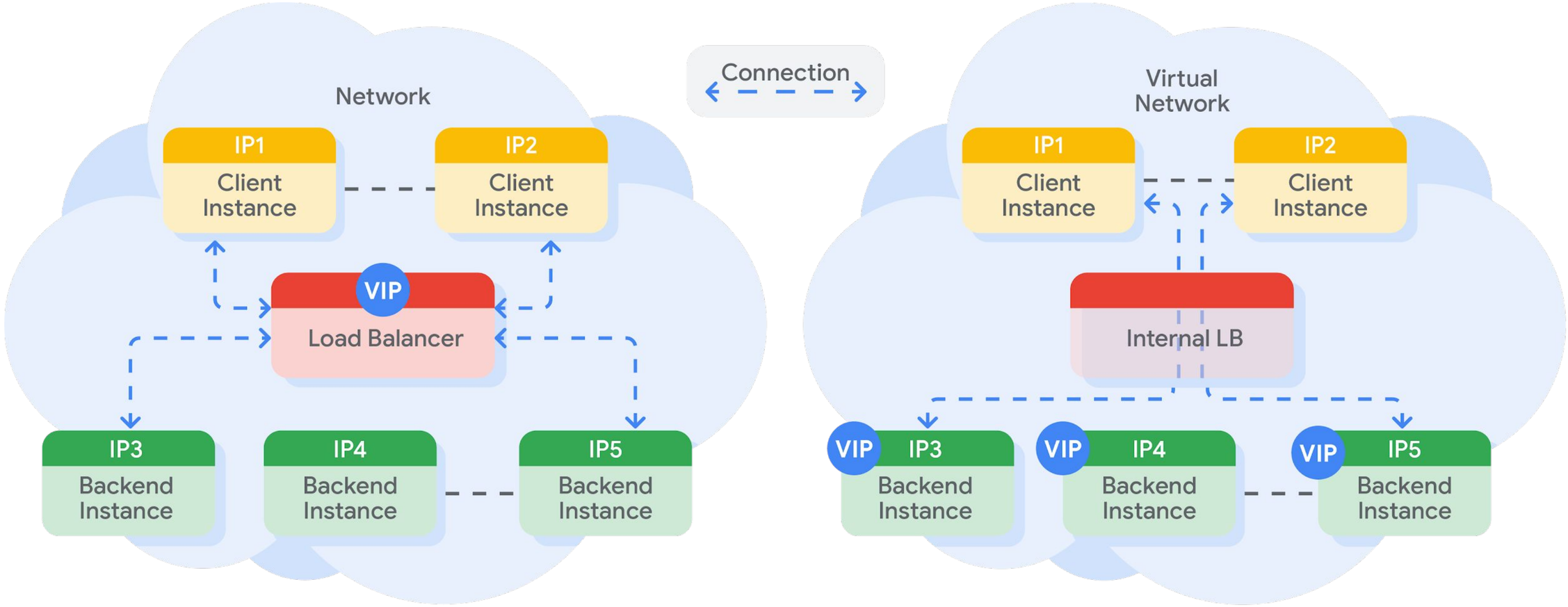
- Forwarding rules (TCP and UDP)
- Up to 50 per project
- One health check
- Instances must be in the same region

Internal Load Balancing

Internal TCP/UDP load balancing

- Regional, private load balancing
  - VM instances in same region
  - RFC 1918 IP addresses
- TCP/UDP traffic
- Reduced latency, simpler configuration
- Software-defined, fully distributed load balancing

Software-defined, fully distributed load balancing

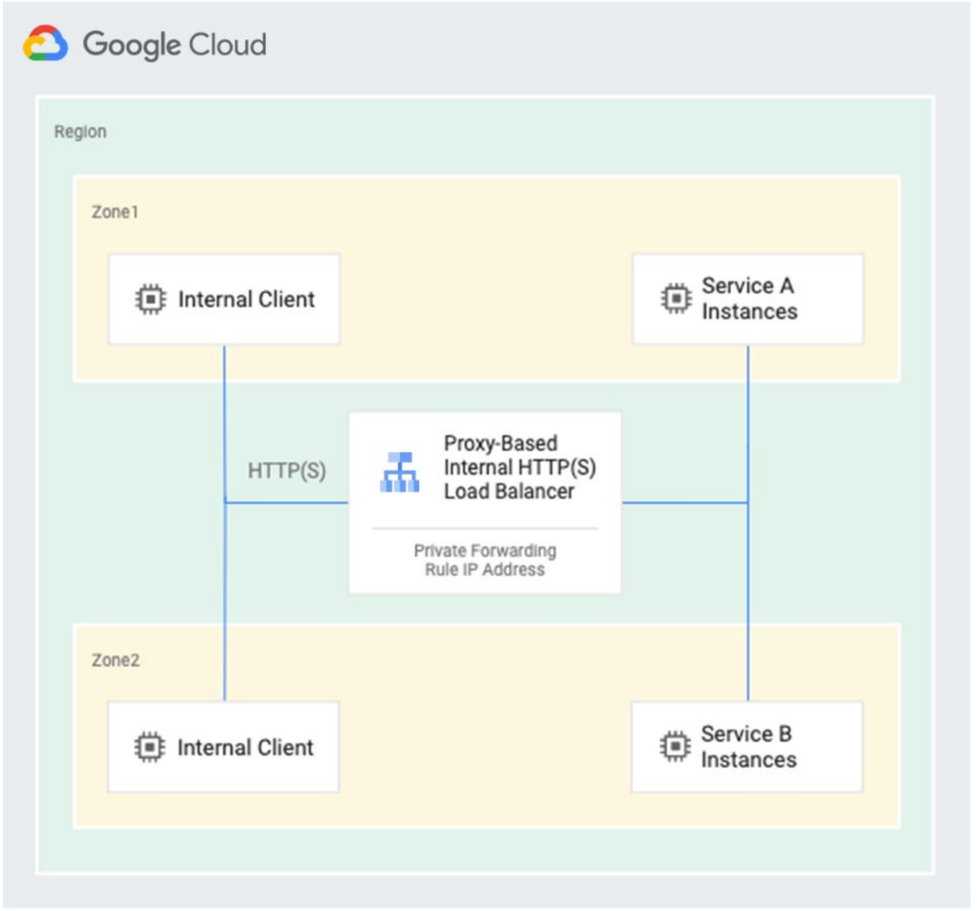


Traditional proxy model of internal load balancing

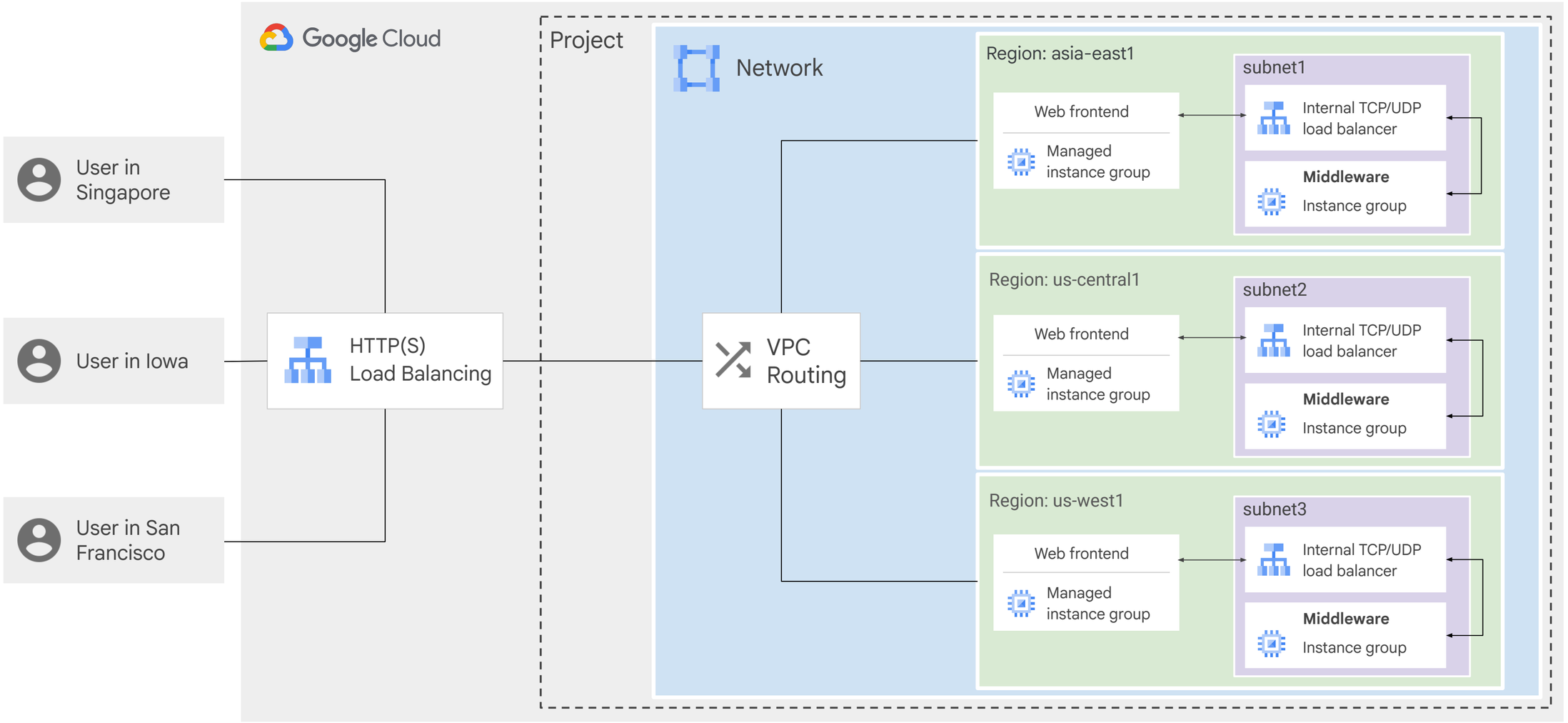
Internal TCP/UDP load balancer

Internal HTTP(S) load balancing

- Regional, private load balancing
  - VM instances in same region
  - RFC 1918 IP addresses
- HTTP, HTTPS, or HTTP/2 protocols
- Based on open source Envoy proxy

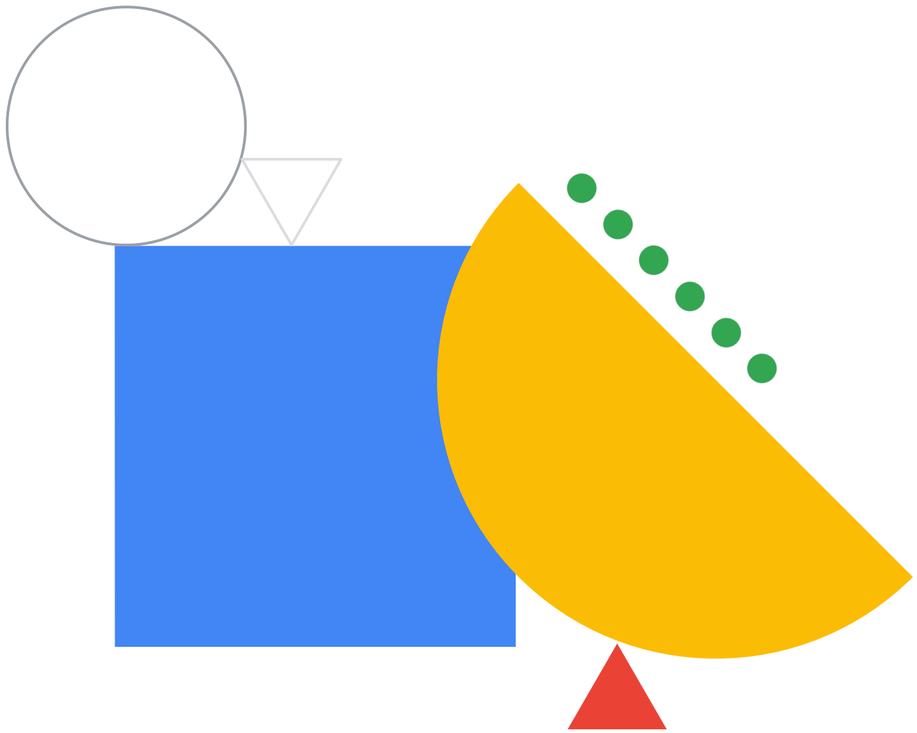


Internal load balancing supports 3-tier web services



Lab Intro

Configuring an Internal Load Balancer



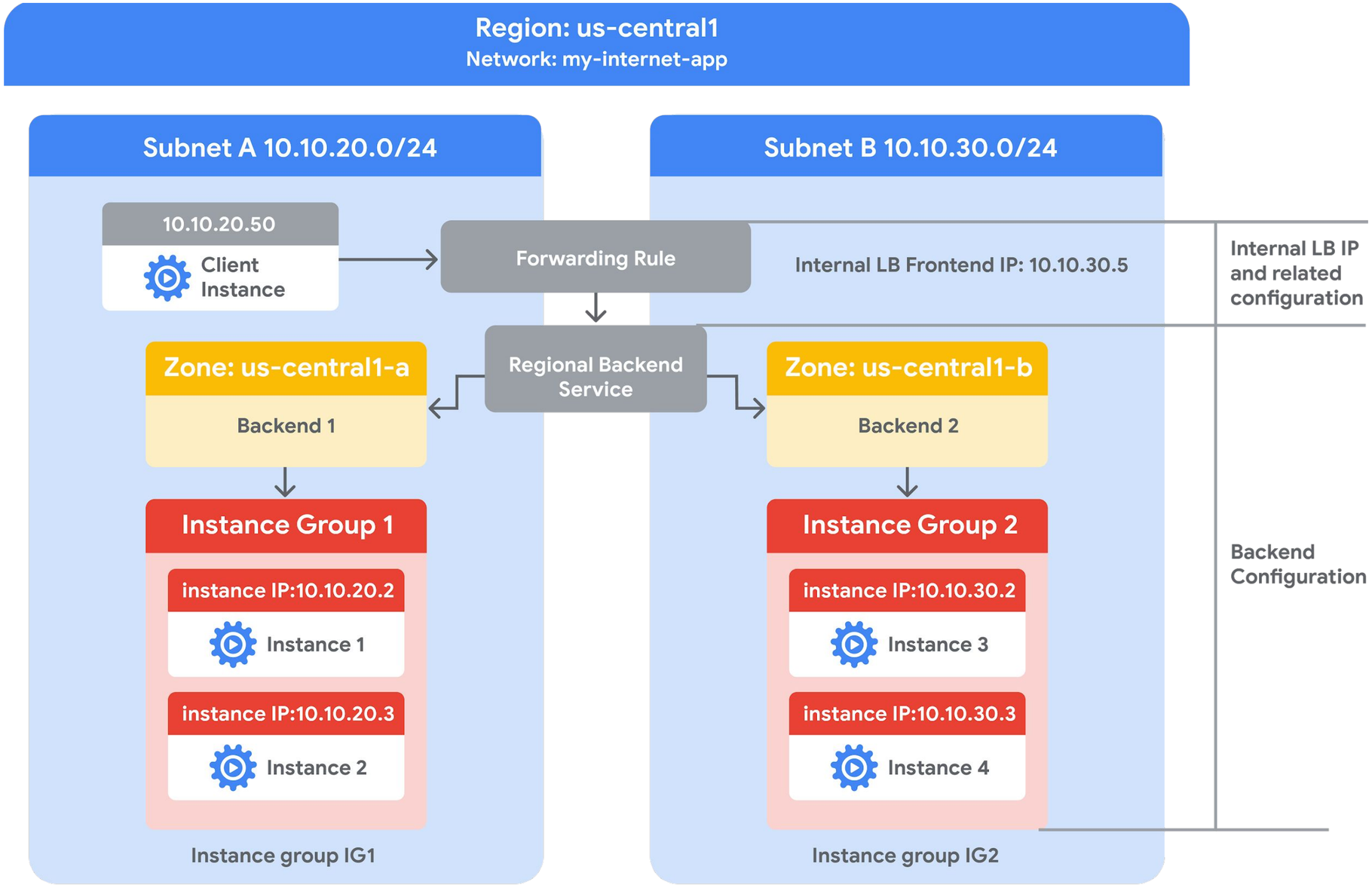
Lab objectives

01 Create HTTP and health check firewall rules

02 Configure two instance templates

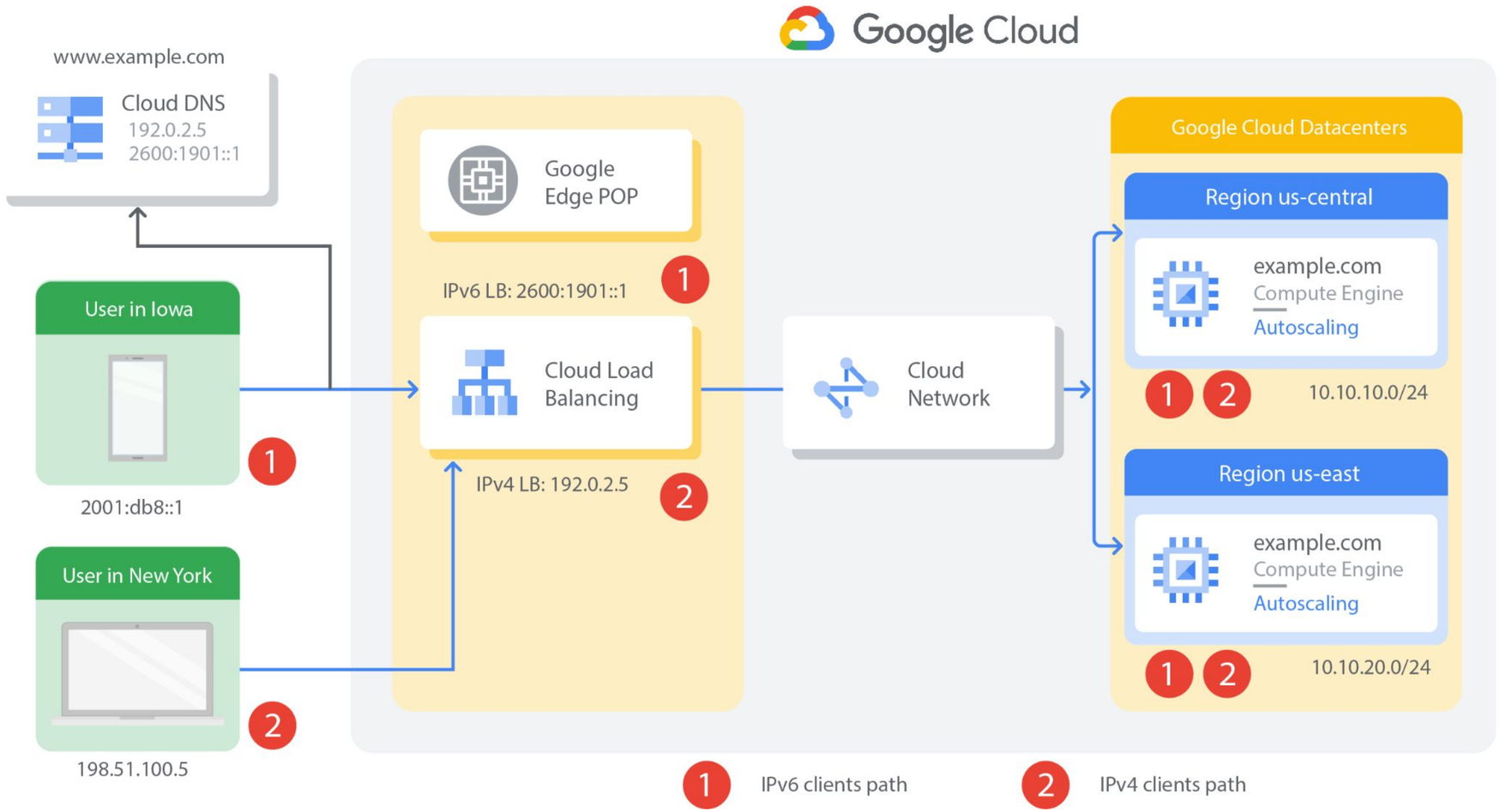
03 Create two managed instance groups

04 Configure and test an internal load balancer

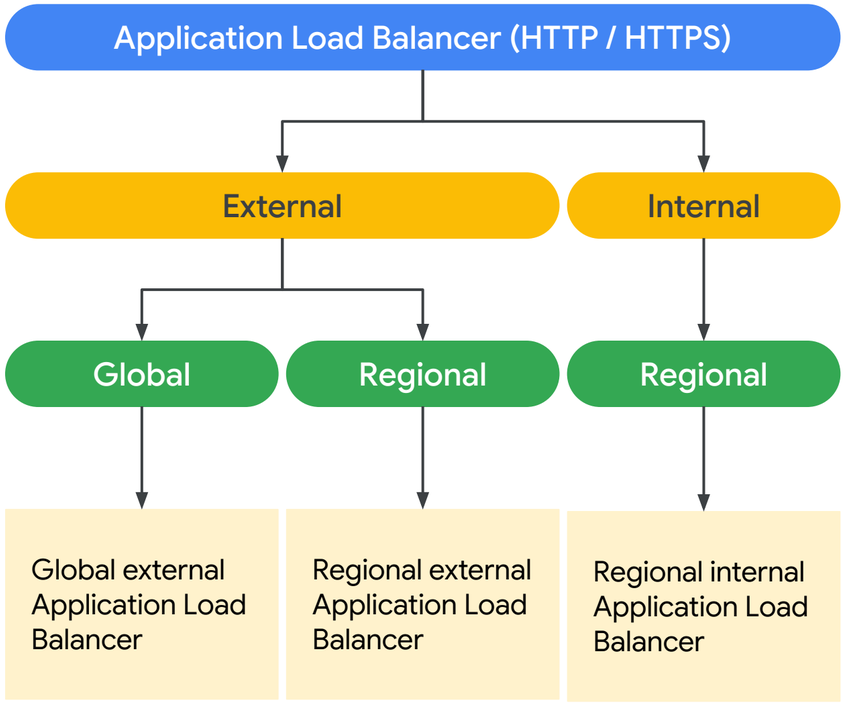


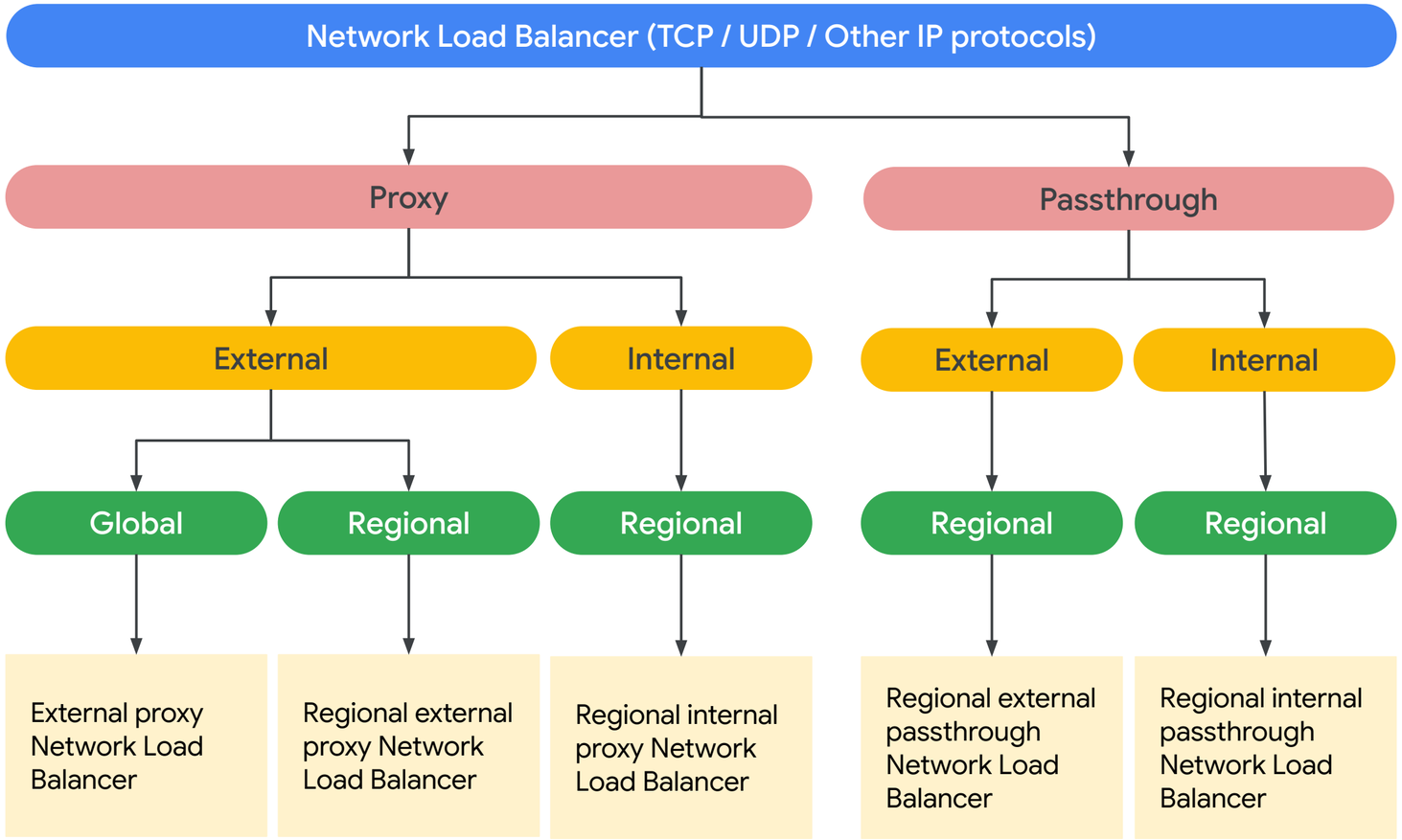
Choosing a Load Balancer

IPv6 termination for load balancing



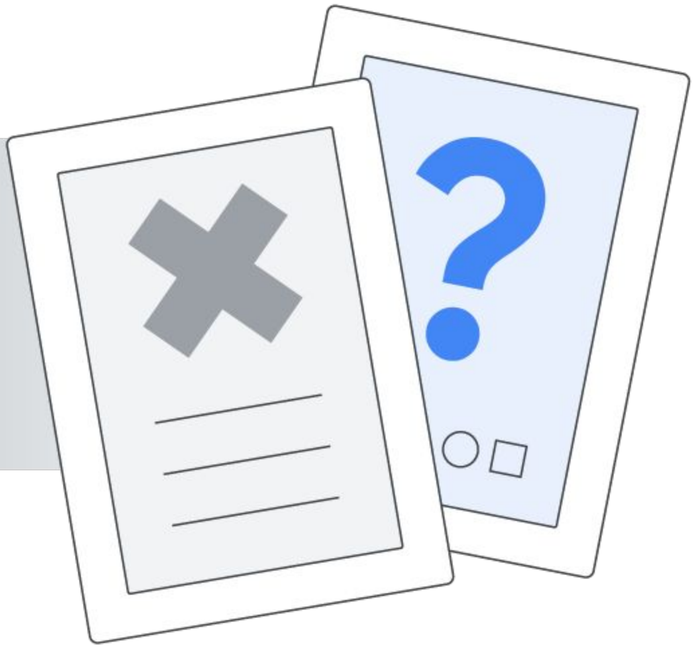
Deployment modes available for Cloud Load Balancing





Summary of Google Cloud load balancers

Load balancer	Deployment mode	Traffic type	Network Service Tier	Load-balancing scheme
Application Load Balancers	**Global external**	HTTP or HTTPS	Premium	EXTERNAL_MANAGED
	Regional external	HTTP or HTTPS	Standard	EXTERNAL_MANAGED
	Classic	HTTP or HTTPS	Global in Premium Regional in Standard	EXTERNAL
	Internal Always regional	HTTP or HTTPS	Premium	INTERNAL_MANAGED
Proxy Network Load Balancers	**Global external**	TCP with optional SSL offload	Global in Premium Regional in Standard	EXTERNAL
	Regional external	TCP	Standard only	EXTERNAL_MANAGED
	Internal Always regional	TCP without SSL offload	Premium only	INTERNAL_MANAGED
Passthrough Network Load Balancers	**External** Always regional	TCP, UDP, ESP, GRE, ICMP, and ICMPv6	Premium or Standard	EXTERNAL
	Internal Always regional	TCP or UDP	Premium only	INTERNAL



Quiz

Question #1

Question

Which of the following is not a Google Cloud load balancing service?

A. HTTP(S) load balancing

B. SSL proxy load balancing

C. TCP proxy load balancing

D. Hardware-defined load balancing

E. Network load balancing

F. Internal load balancing

Question #1

Answer

Which of the following is not a Google Cloud load balancing service?

A. HTTP(S) load balancing

B. SSL proxy load balancing

C. TCP proxy load balancing

D. Hardware-defined load balancing 

E. Network load balancing

F. Internal load balancing

Question #2

Question

Which three Google Cloud load balancing services support IPv6 clients?

A. HTTP(S) load balancing

B. SSL proxy load balancing

C. TCP proxy load balancing

D. Network load balancing

E. Internal load balancing

Question #2

Answer

Which three Google Cloud load balancing services support IPv6 clients?

A. HTTP(S) load balancing 

B. SSL proxy load balancing 

C. TCP proxy load balancing 

D. Network load balancing

E. Internal load balancing

Question #3

Question

Which of the following are applicable autoscaling policies for managed instance groups?

A. CPU utilization

B. Load balancing capacity

C. Monitoring metrics

D. Queue-based workload

Question #3

Answer

Which of the following are applicable autoscaling policies for managed instance groups?

A. CPU utilization 

B. Load balancing capacity 

C. Monitoring metrics 

D. Queue-based workload 

Review: Load Balancing and Autoscaling