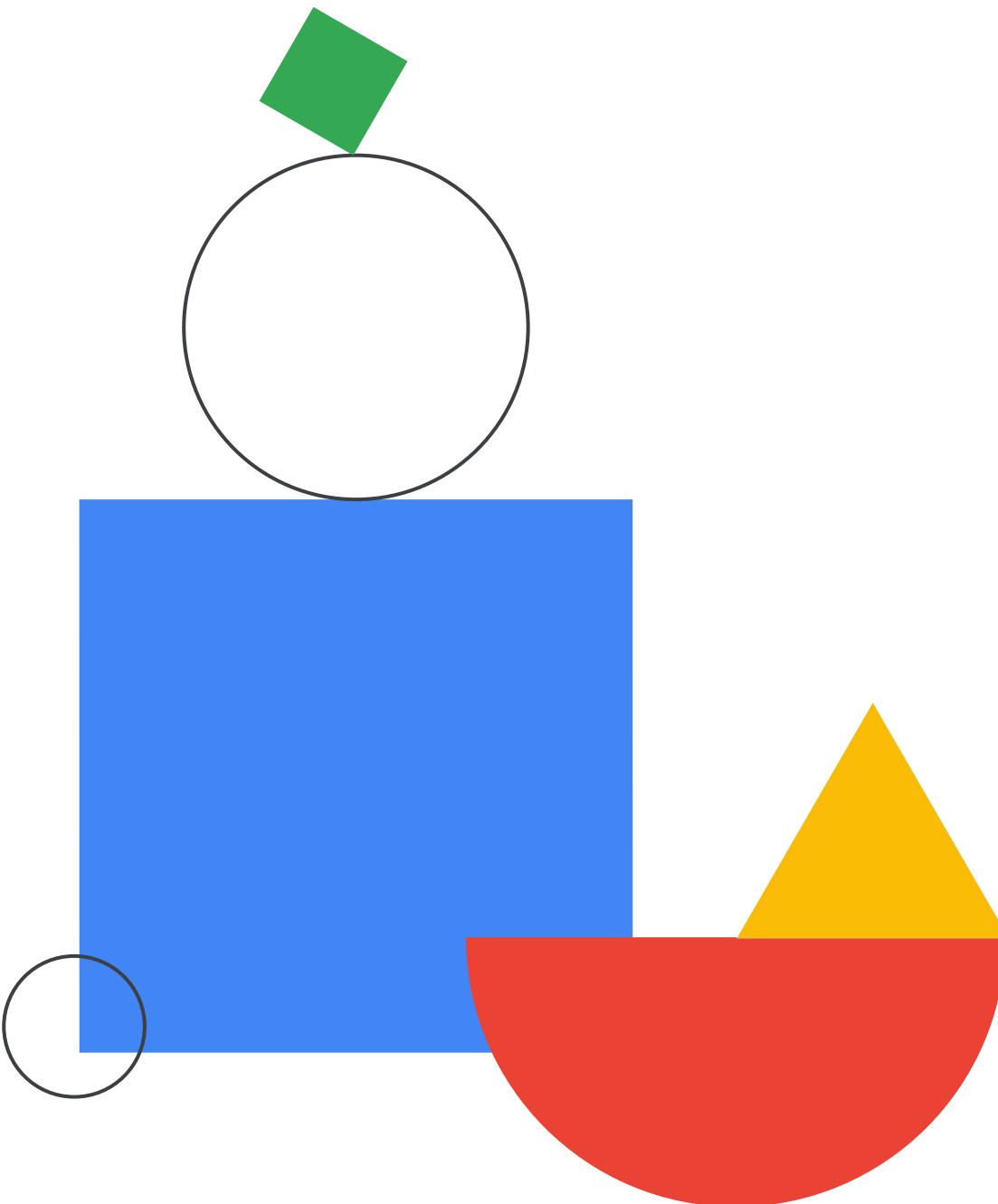
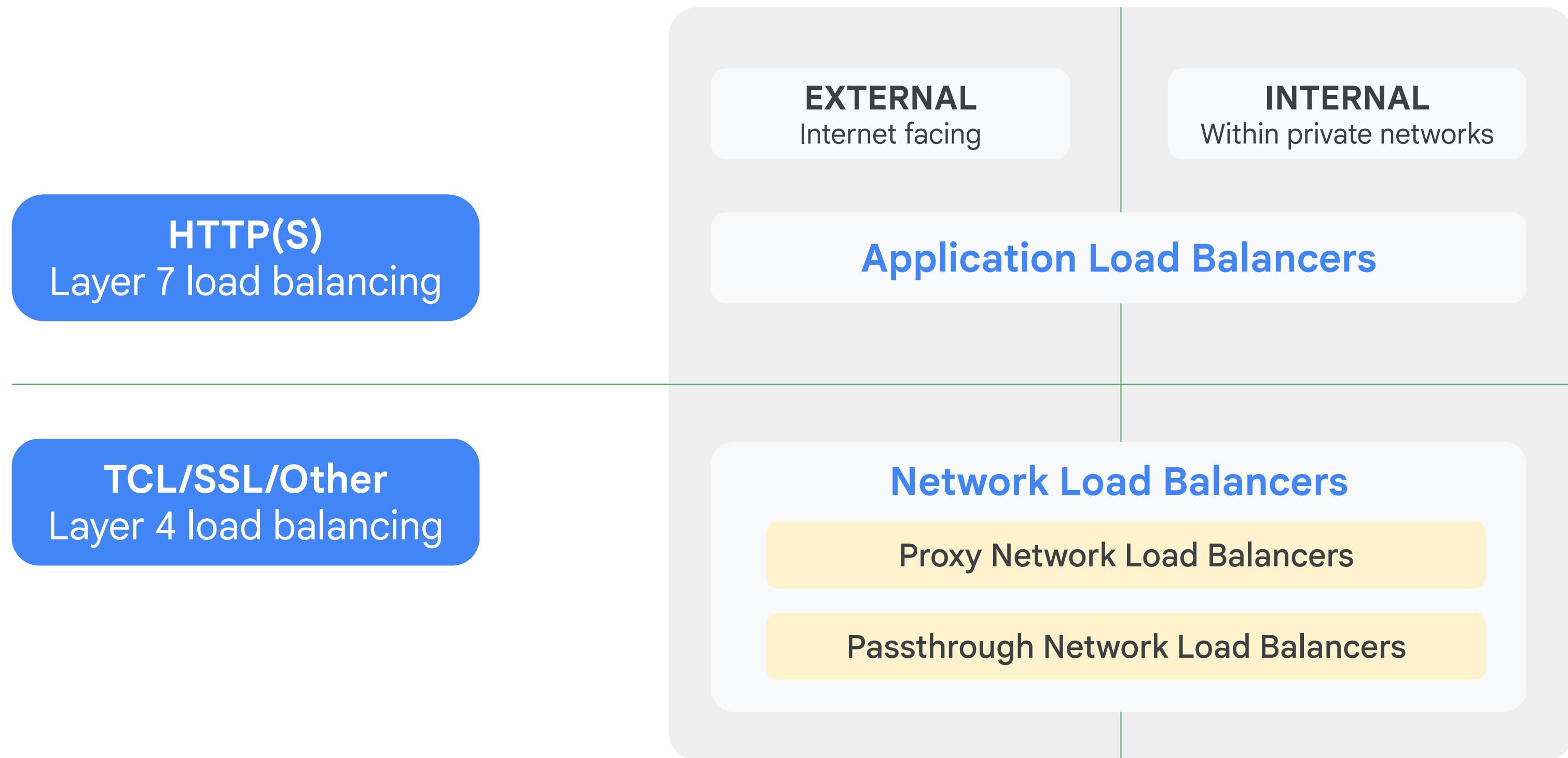




Load Balancing and Autoscaling

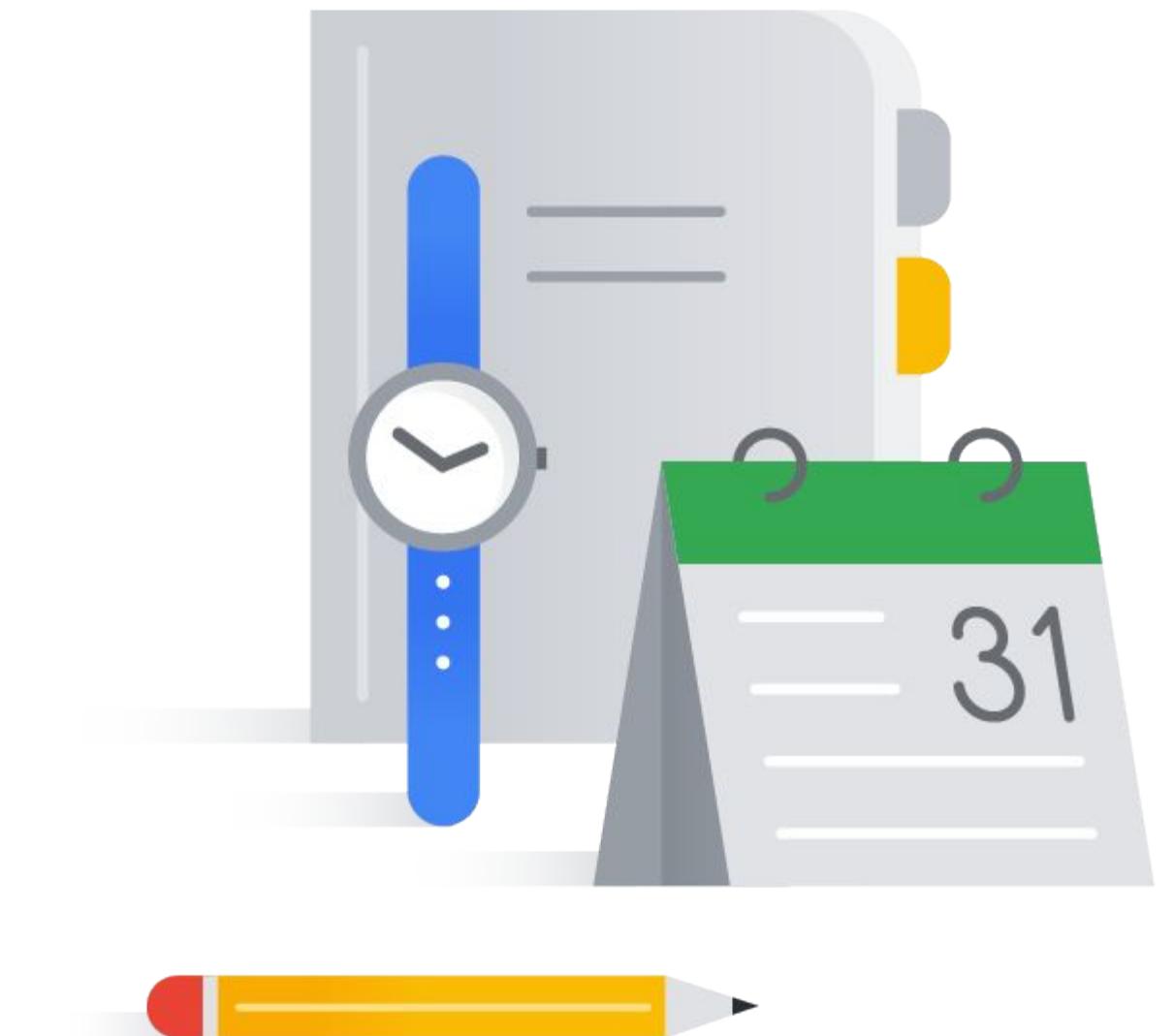


Types of load balancers



Agenda

-
- 01 Managed Instance Groups
 - 02 Application Load Balancers
 - Lab: Configure an Application Load Balancer (HTTP) with Autoscaling
 - 03 Cloud CDN
 - 04 Network Load Balancing
 - 05 Internal Load Balancing
 - Lab: Configure an Internal Network Load Balancer
 - 06 Choosing a Load Balancer

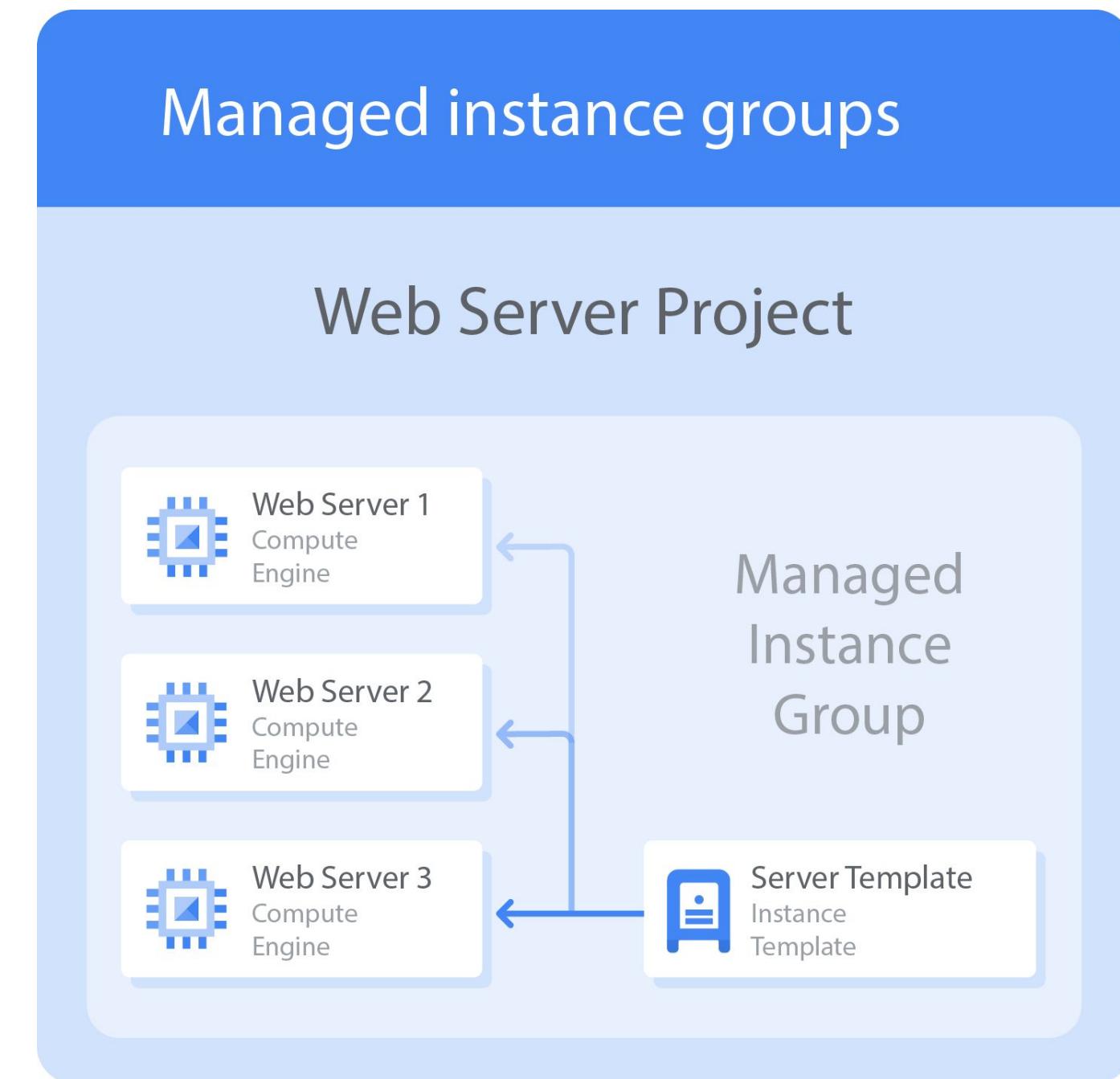




Managed Instance Groups

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

The screenshot shows the Google Cloud Compute Engine dashboard. On the left, there's a sidebar with 'Virtual machines' expanded, showing 'VM instances' and 'Instance templates'. A red arrow points from the 'CREATE INSTANCE TEMPLATE' button in the main content area to the 'Instance templates' link in the sidebar.

The screenshot shows the 'Create instance template' dialog box. At the top, it says 'Name *' followed by 'mywebserver-template'. Below that is a section for 'MANAGE TAGS AND LABELS'. Under 'Machine configuration', it says 'NEW: General-purpose machine series in Preview' and lists 'General purpose', 'Compute optimized', 'Memory optimized', 'Storage optimized', 'NEW', and 'GPUs'. The 'General purpose' tab is selected. It shows a table of machine types:

Series	Description	vCPUs	Memory	Platform
C3	Consistently high performance	4 - 176	8 - 1,408 GB	Intel Sapphire Rapids
C3D	Consistently high performance	4 - 360	8 - 2,880 GB	AMD Genoa
E2	Low cost, day-to-day computing	0.25 - 32	1 - 128 GB	Based on availability
N2	Balanced price & performance	2 - 128	2 - 864 GB	Intel Cascade and Ice Lake
N2D	Balanced price & performance	2 - 224	2 - 896 GB	AMD EPYC
T2A	Scale-out workloads	1 - 48	4 - 192 GB	Ampere Altra Arm
T2D	Scale-out workloads	1 - 60	4 - 240 GB	AMD EPYC Milan
N1	Balanced price & performance	0.25 - 96	0.6 - 624 GB	Intel Skylake

Below this is a 'Machine type' section with a note about choosing a machine type with preset amounts of vCPUs and memory. It shows a 'PRESET' tab selected, showing 'e2-medium (2 vCPU, 1 core, 4 GB memory)'. It also shows 'vCPU' (1-2 vCPU (1 shared core)) and 'Memory' (4 GB). At the bottom, there's a 'ADVANCED CONFIGURATIONS' section.

On the right side of the dialog box, there are several sections: 'Boot disk', 'Identity and API access', 'Firewall', and 'Advanced options'. Each section contains configuration fields and notes.

Create a managed instance group

01 Create Instance Group

New managed instance group (stateless)
Automatically manage groups of VMs that do stateless serving and batch processing.

New managed instance group (stateful)
Automatically manage groups of VMs that have persistent data or configurations (such as databases or legacy applications).

New unmanaged instance group
Manually manage groups of load balancing VMs.

02

Set up automatic management for a group of stateless VMs, including updates, regional deployments, load balancing, autoscaling, and autohealing. [Learn more](#)

Name * instance-group-1

Name is permanent

Description

03 Location

For higher availability, select multiple zones in a region instead of a single zone. [Learn more](#)

Single zone

Multiple zones

Region * us-west1 (Oregon)

Zones us-west1-b, us-west1-c, and ...

Target distribution shape Even

04

Instance template * mywebserver-template

f1-micro, mywebserver, global

05 Autoscaling

Use autoscaling to automatically add and remove instances to the group for periods of high and low load. [Learn more](#)

Autoscaling mode On: add and remove instances to the group

Minimum number of instances * 1

Maximum number of instances * 2

06 Autohealing

Autohealing recreates VM instances if your application cannot be reached by the health check. [Learn more](#)

Health check

Compute Engine will recreate VM instances only when they're not running.

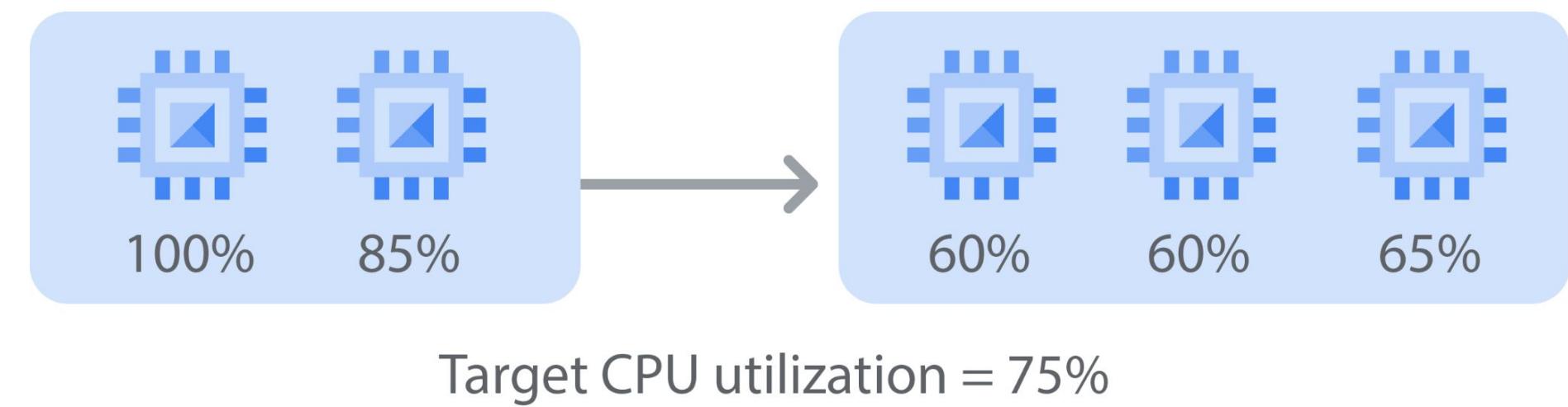
Managed instance groups offer autoscaling capabilities

Dynamically add/remove instances:

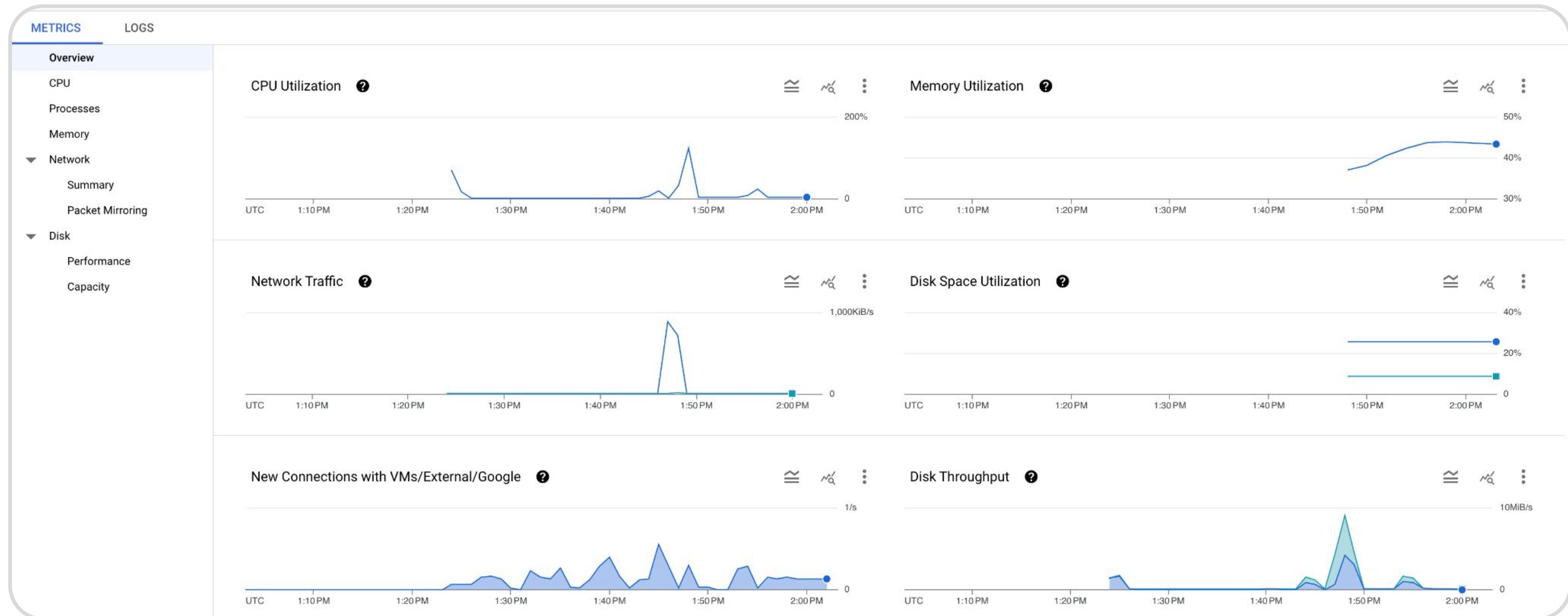
- Increases in load
- Decreases in load

Autoscaling policy:

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



VM graph helps set CPU utilization



Create a health check

Health checking mechanisms determine whether VM instances respond properly to traffic. You cannot create a legacy health check using this page. For more information, refer to the [Health Checks Concepts](#) documentation.

Name *	<input type="text"/>	?
Lowercase, no spaces.		
Description		
Scope		
<input checked="" type="radio"/> Global	<input type="radio"/> Regional	
Protocol	TCP	Port * 80 ?
Proxy protocol	NONE	
Request	<input type="text"/>	Response <input type="text"/>
Logs	<input type="radio"/> On <small>Turning on Health check logs can increase costs in Logging.</small> <input checked="" type="radio"/> Off	
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 ?	

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

Configuring stateful IP addresses

Preserve the unique state of each MIG VM instance on machine restart, recreation, auto-healing, or update event. Useful in the following scenarios:

-  IP address to remain static after it has been assigned.
-  Configuration depends on specific IP addresses.
-  Server is accessed through a dedicated static IP address.
-  Migrate workloads without changing network configuration.

-  Configure IP addresses as stateful for all existing and future instances in the group.
-  Update the existing stateful configuration for IP addresses.

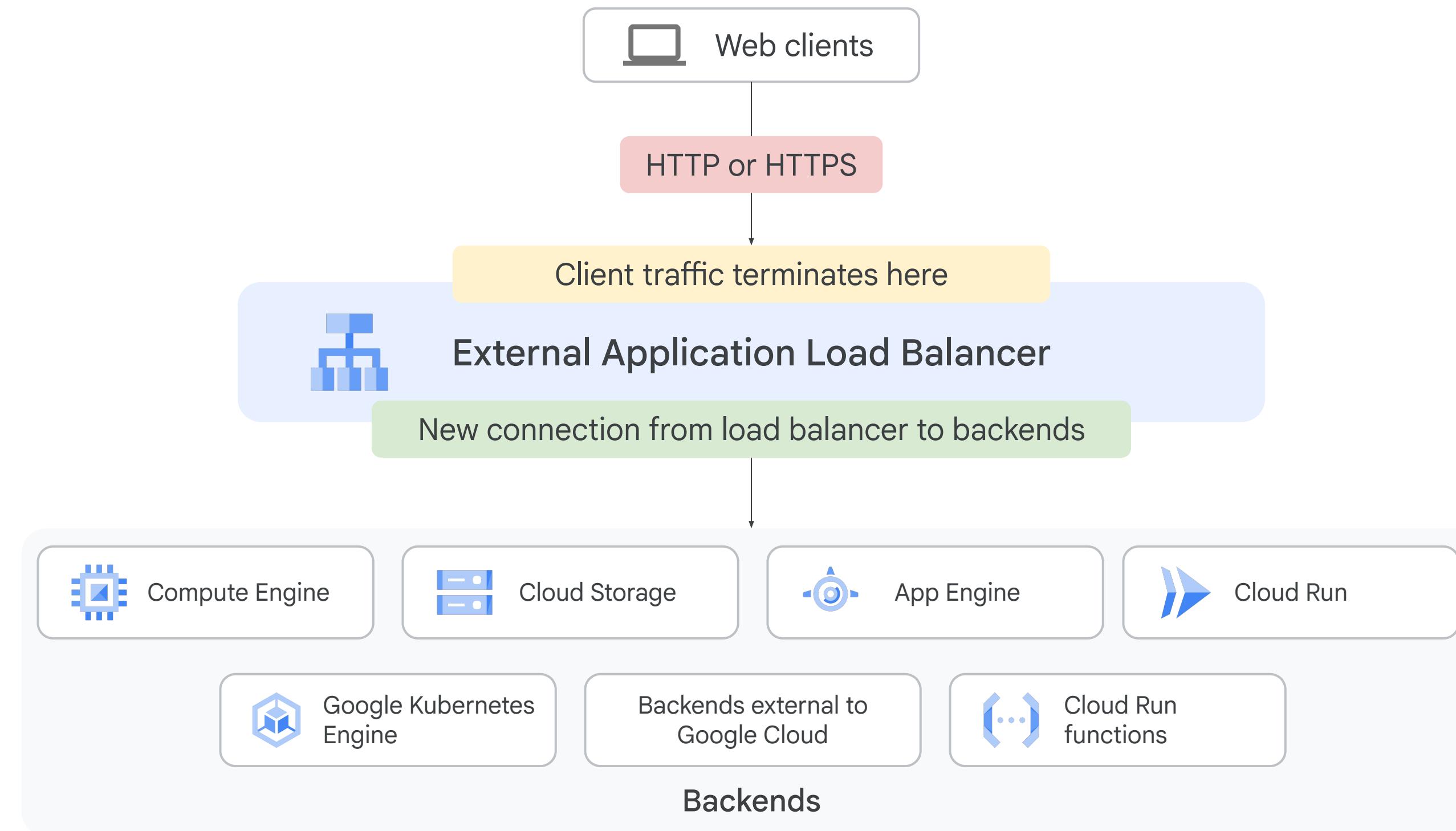


Application Load Balancing

Overview of an Application Load Balancer

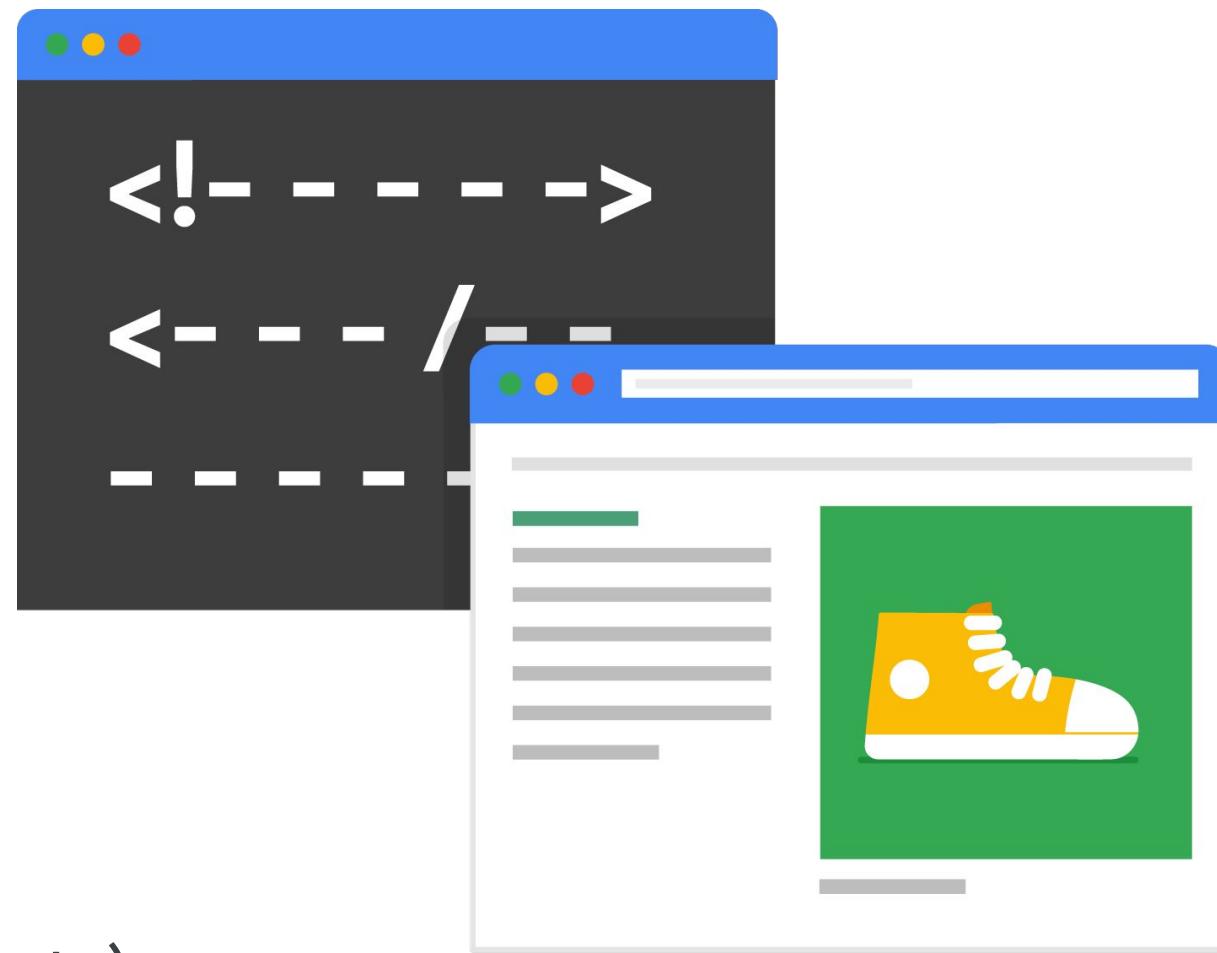
Deployment mode	Network service tier	Load balancing scheme	IP address	Frontend ports
Global external	Premium Tier	EXTERNAL_MANAGED	IPv4 IPv6	Can reference exactly one port from 1-65535
Regional external	Premium or Standard Tier	EXTERNAL_MANAGED	IPv4	
Classic	Global in Premium Tier Regional in Standard Tier	EXTERNAL	IPv4 IPv6 (requires Premium Tier)	

Architecture of an external Application Load Balancer

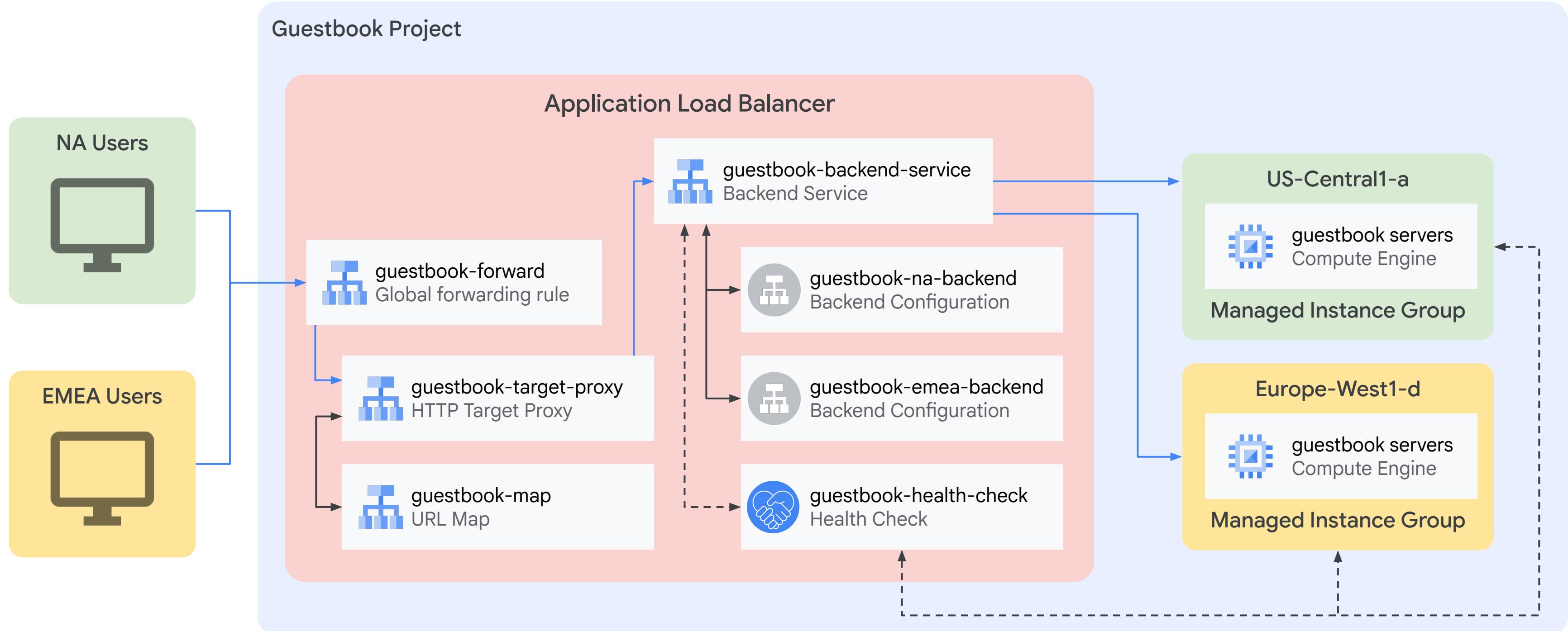


Backend services

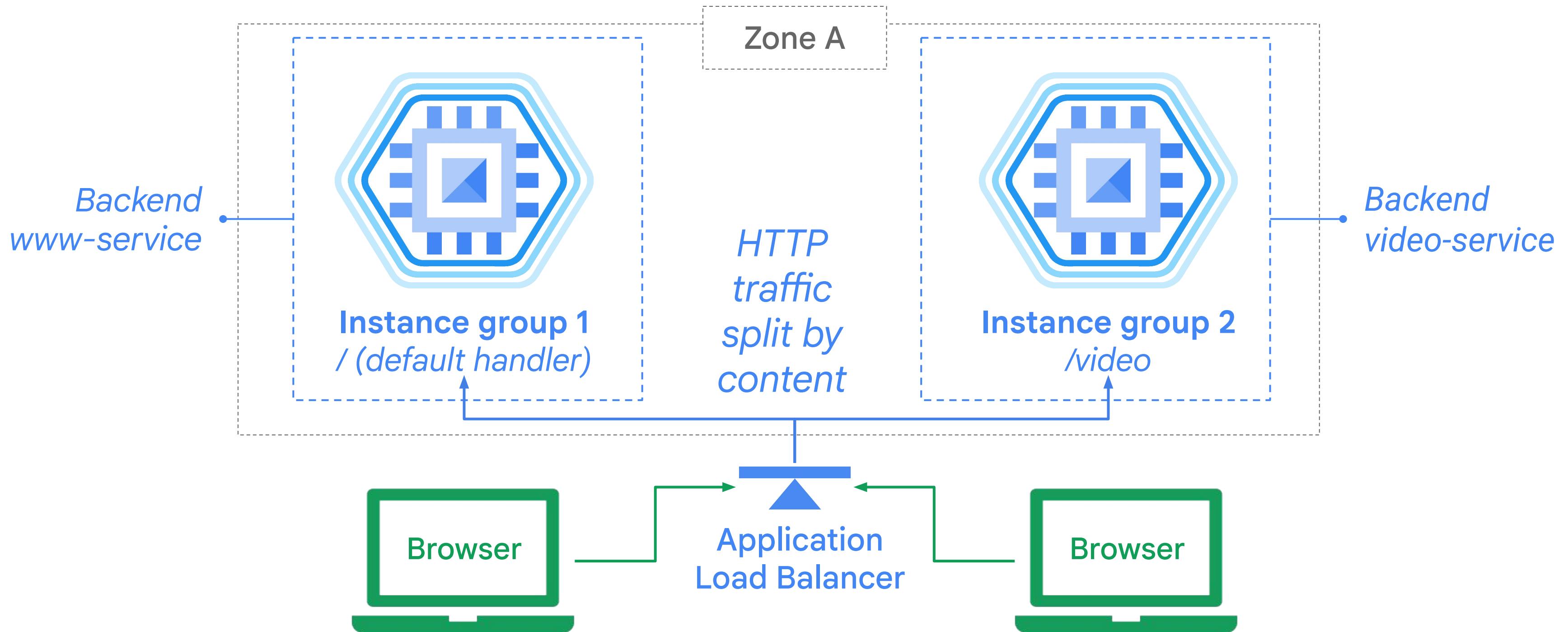
- Health check
- Session affinity (optional)
- Time out setting (30-sec default)
- One or more backends
 - An instance group (managed or unmanaged)
 - A balancing mode (CPU utilization or RPS)
 - A capacity scaler (ceiling percentage of CPU/Rate targets)



Application Load Balancing resources

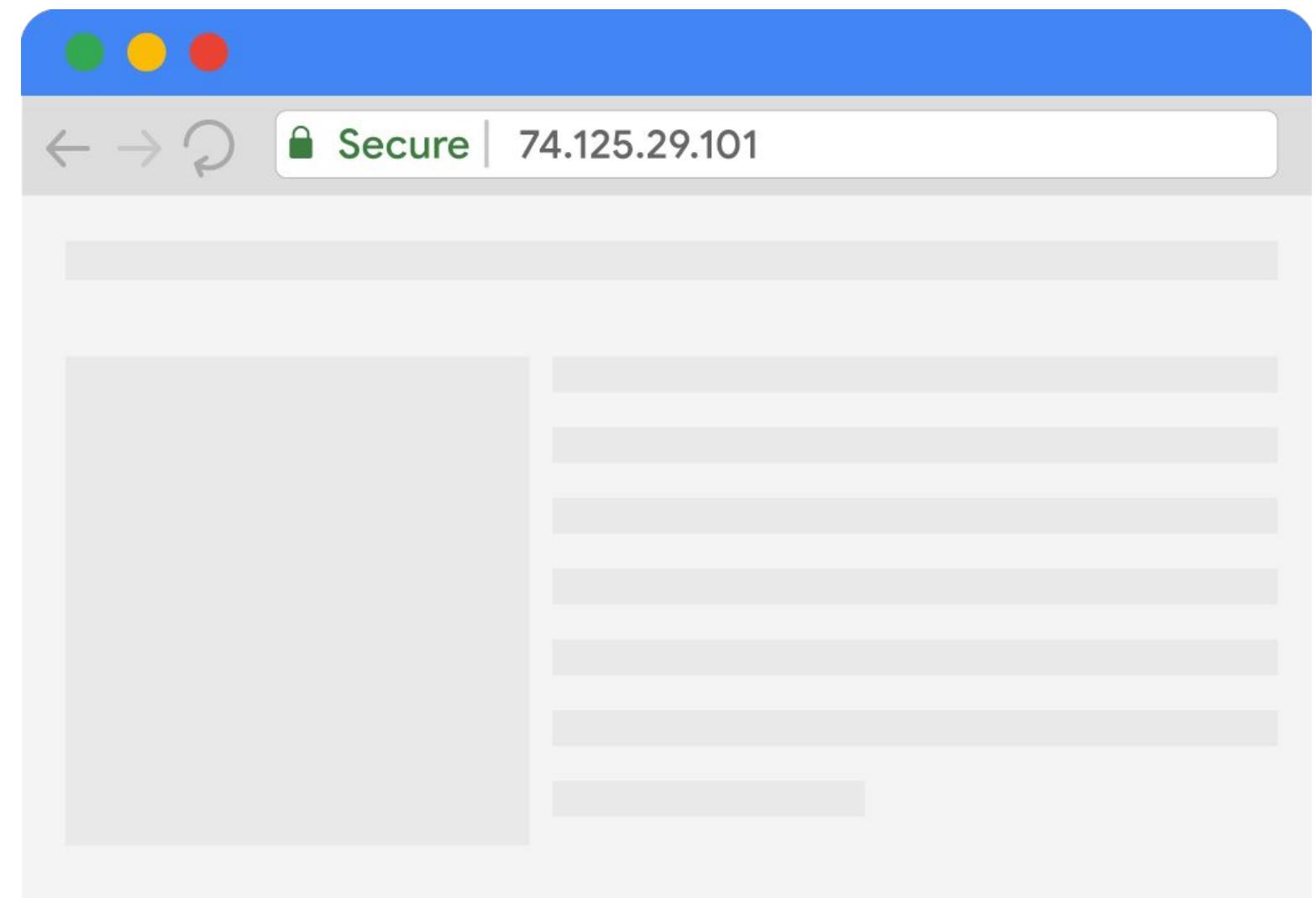


Example: Content-based load balancing



Application Load Balancing - Target HTTPS proxy

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

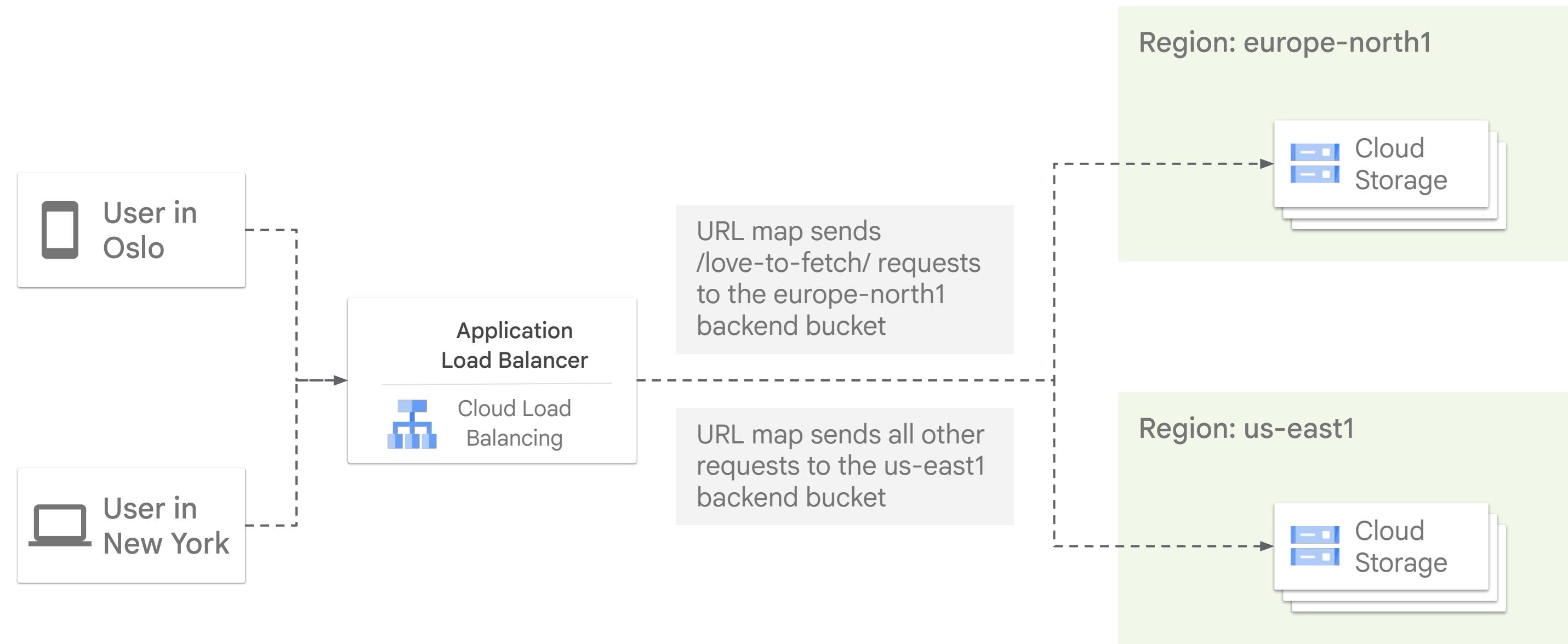


SSL certificates

- Required for Application Load Balancing
- Up to 15 SSL certificates (per target proxy)
- Create an SSL certificate resource



Backend buckets



Network endpoint groups (NEG)

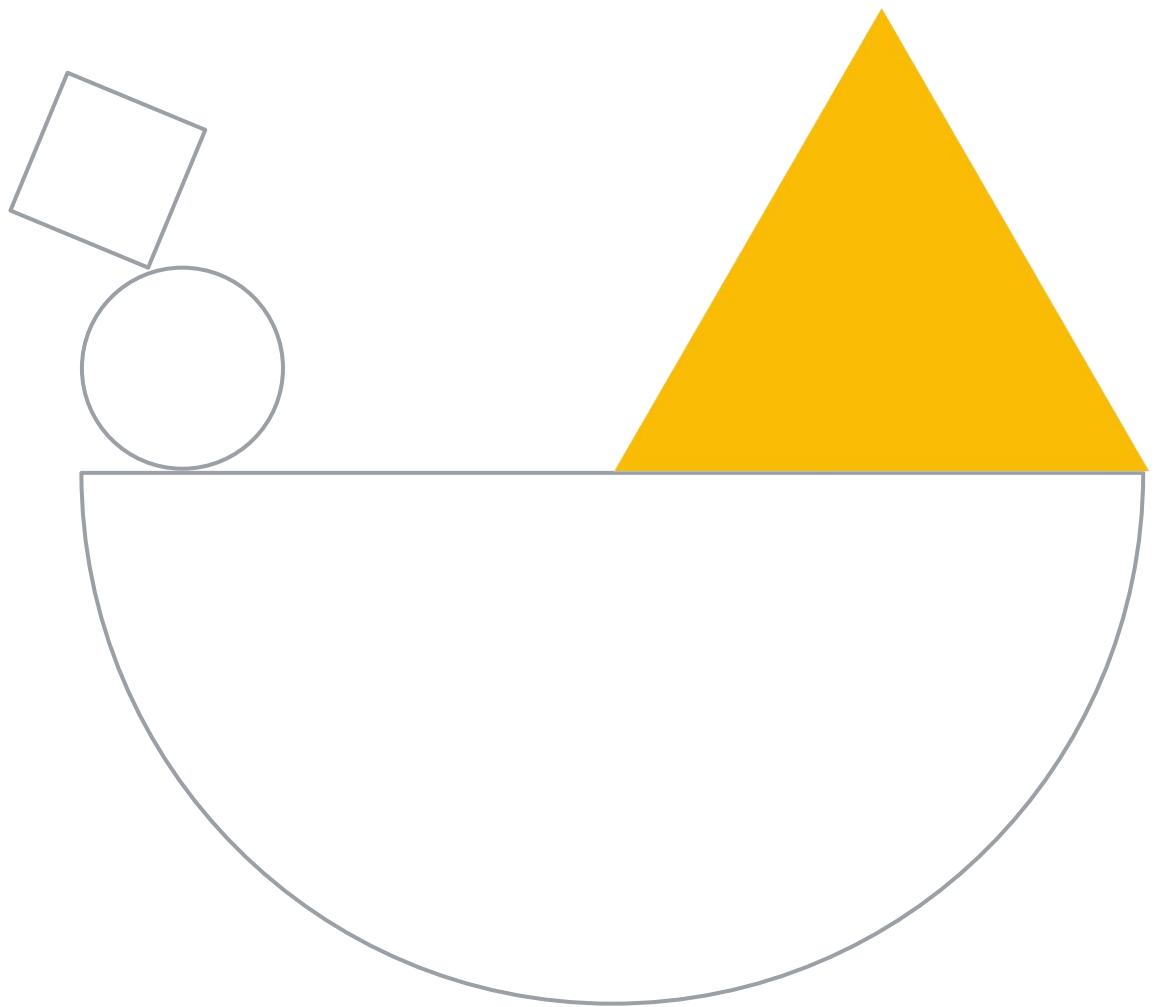
A network endpoint group (NEG) is a configuration object that specifies a group of backend endpoints or services.

There are four types of NEGs:

- Zonal
- Internet
- Serverless
- Hybrid connectivity

Lab Intro

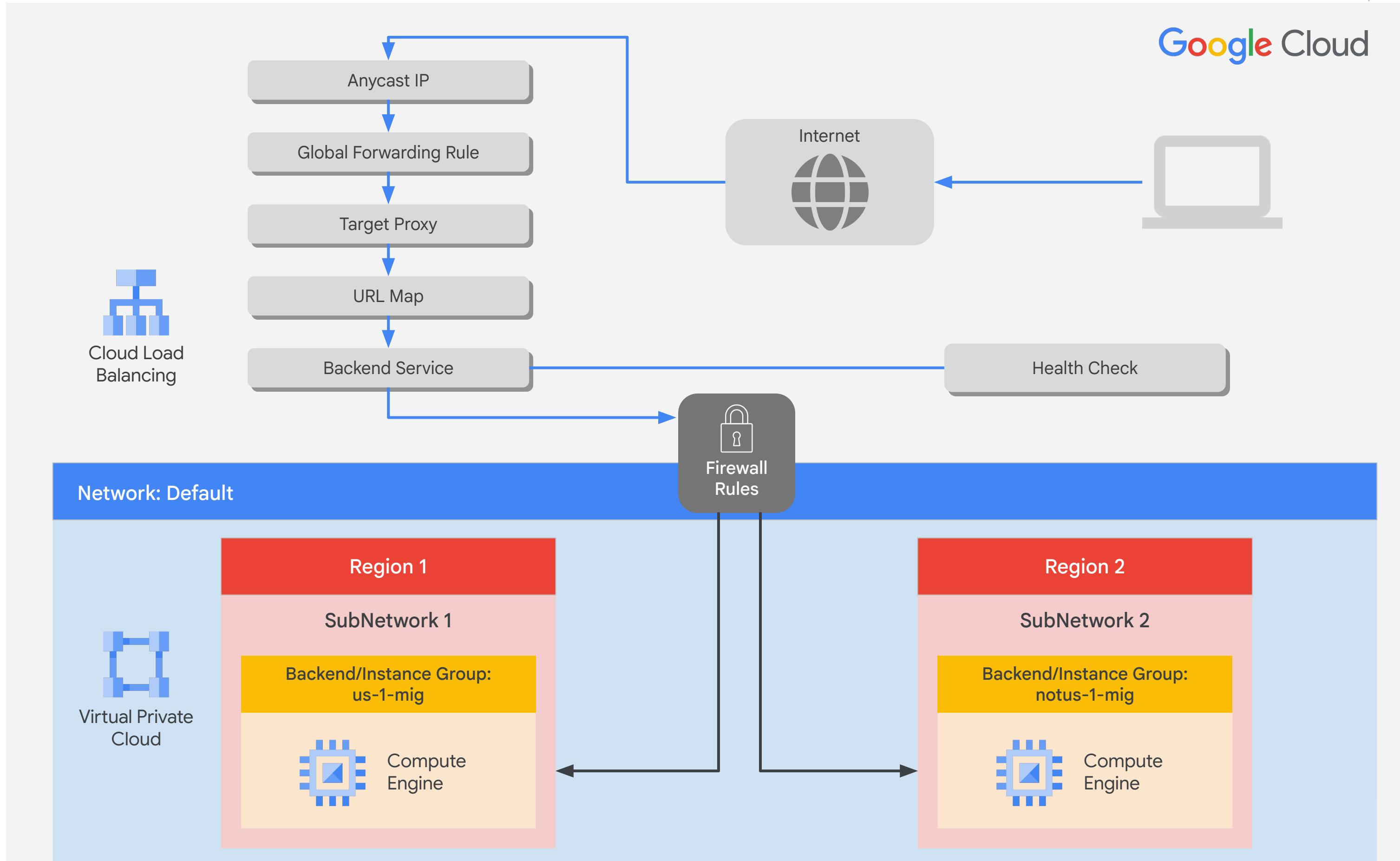
Configure an Application Load
Balancer (HTTP) with Autoscaling



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 Application Load Balancer with IPv4 and IPv6
- 06 Stress test an Application Load Balancer







Cloud CDN



Regions and CDN nodes

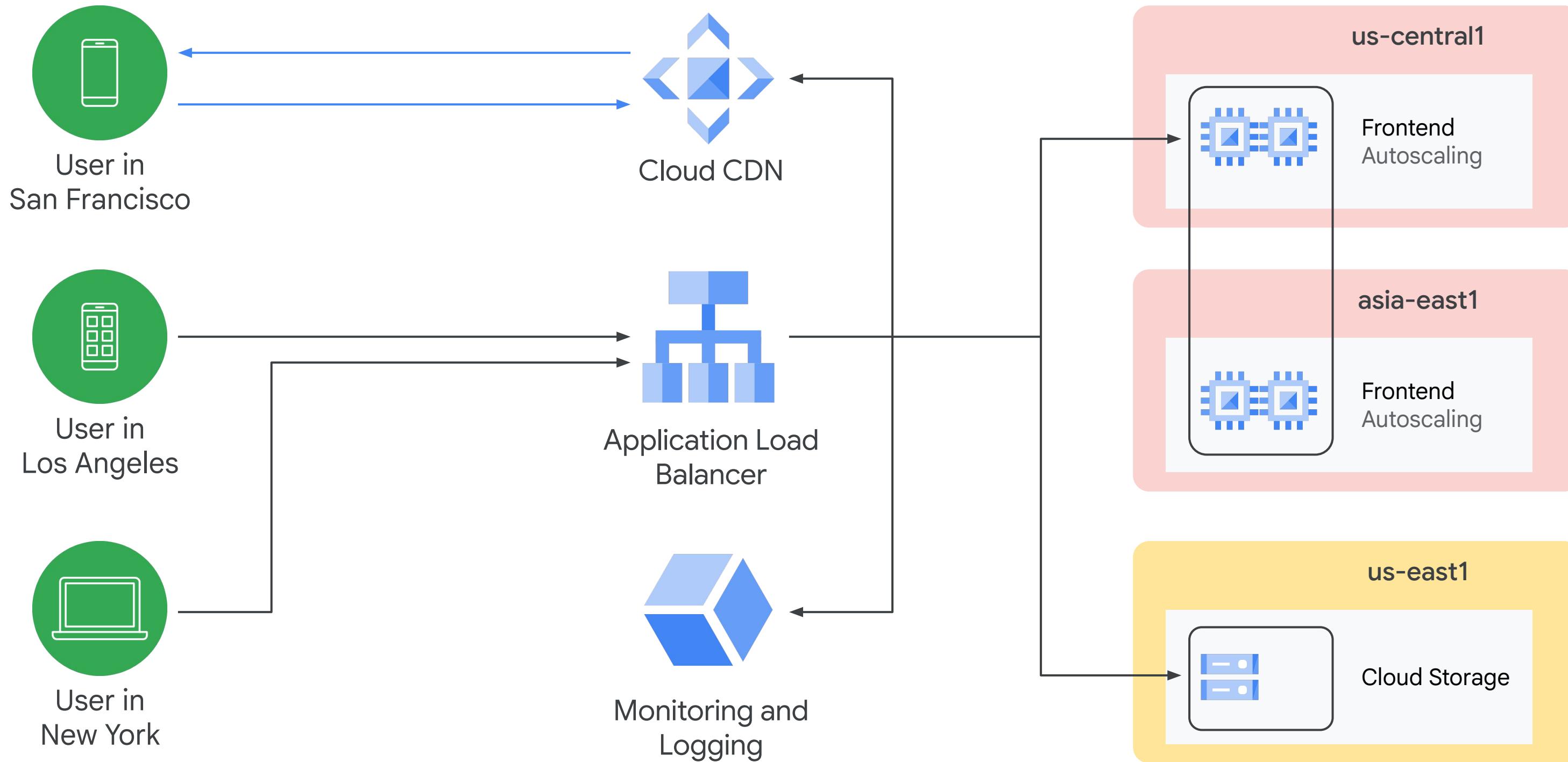
*Exception: region has 4 zones.

CDN

Current region
with 3 zones

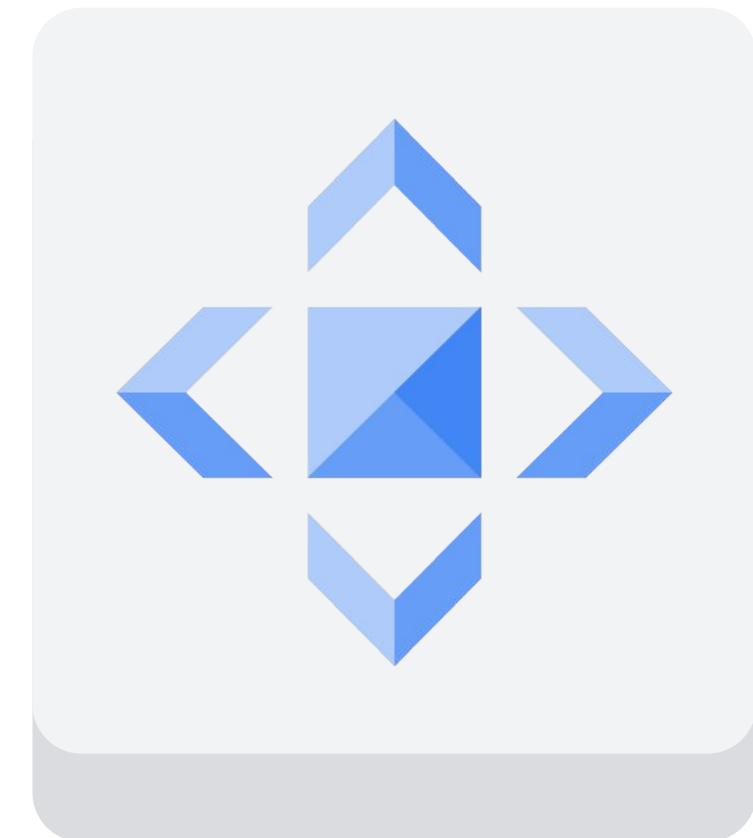
Future region
with 3 zones

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

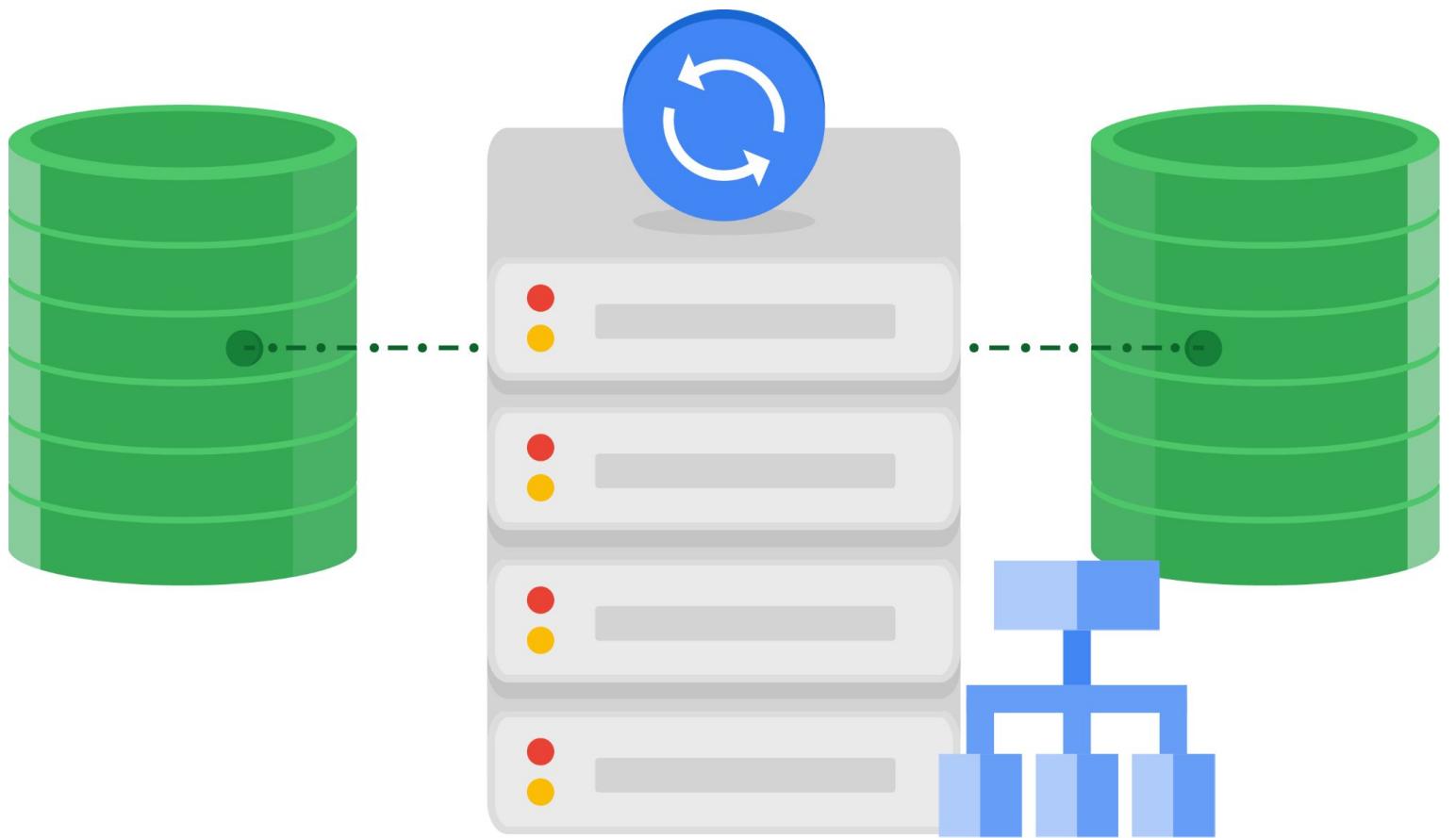




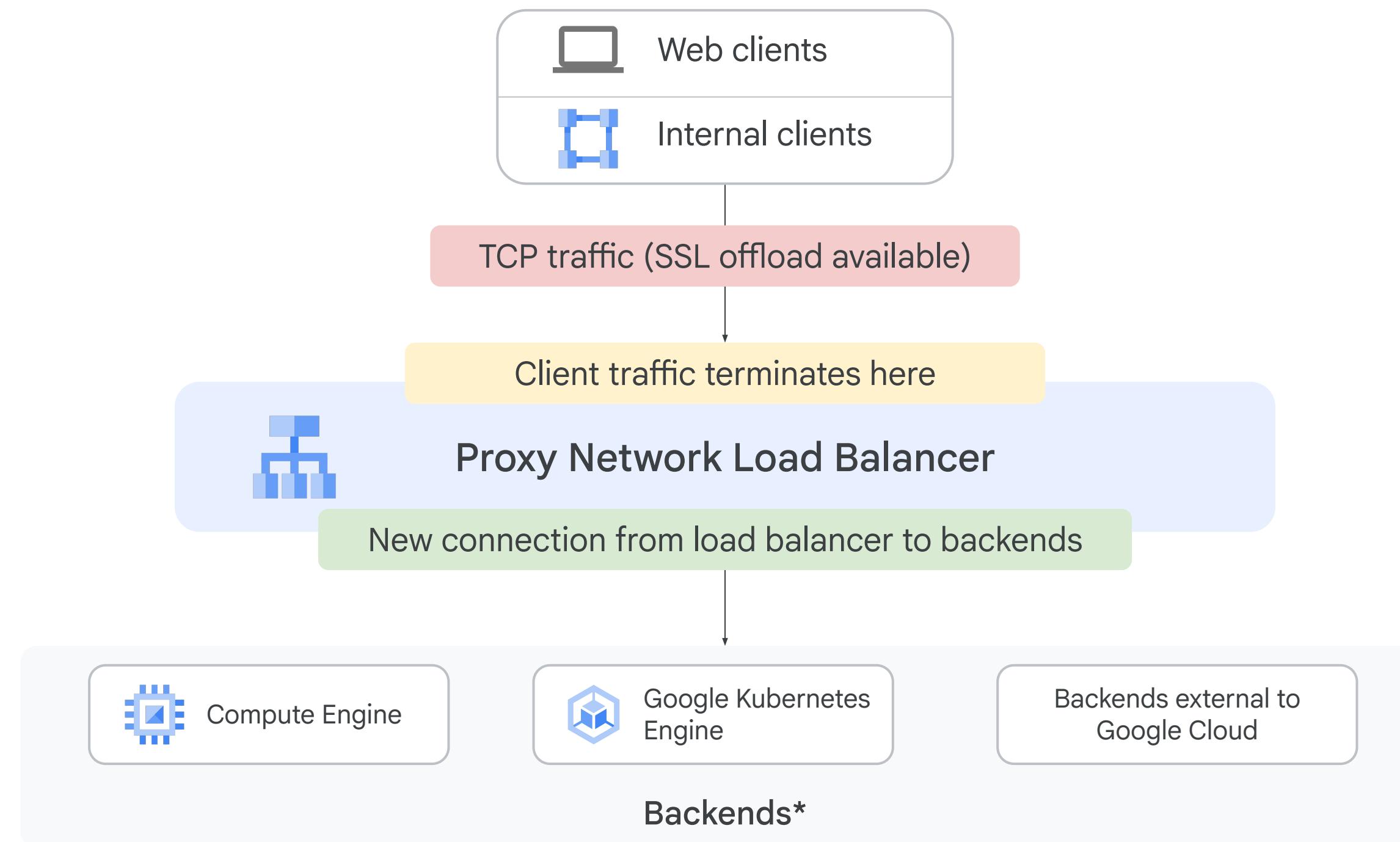
Network Load Balancing

Network load balancing

- Architecture:
 - Proxy
 - Passthrough
- Traffic:
 - TCP/SSL ports
 - UDP, ESP, GRE
 - ICMP, ICMPv6

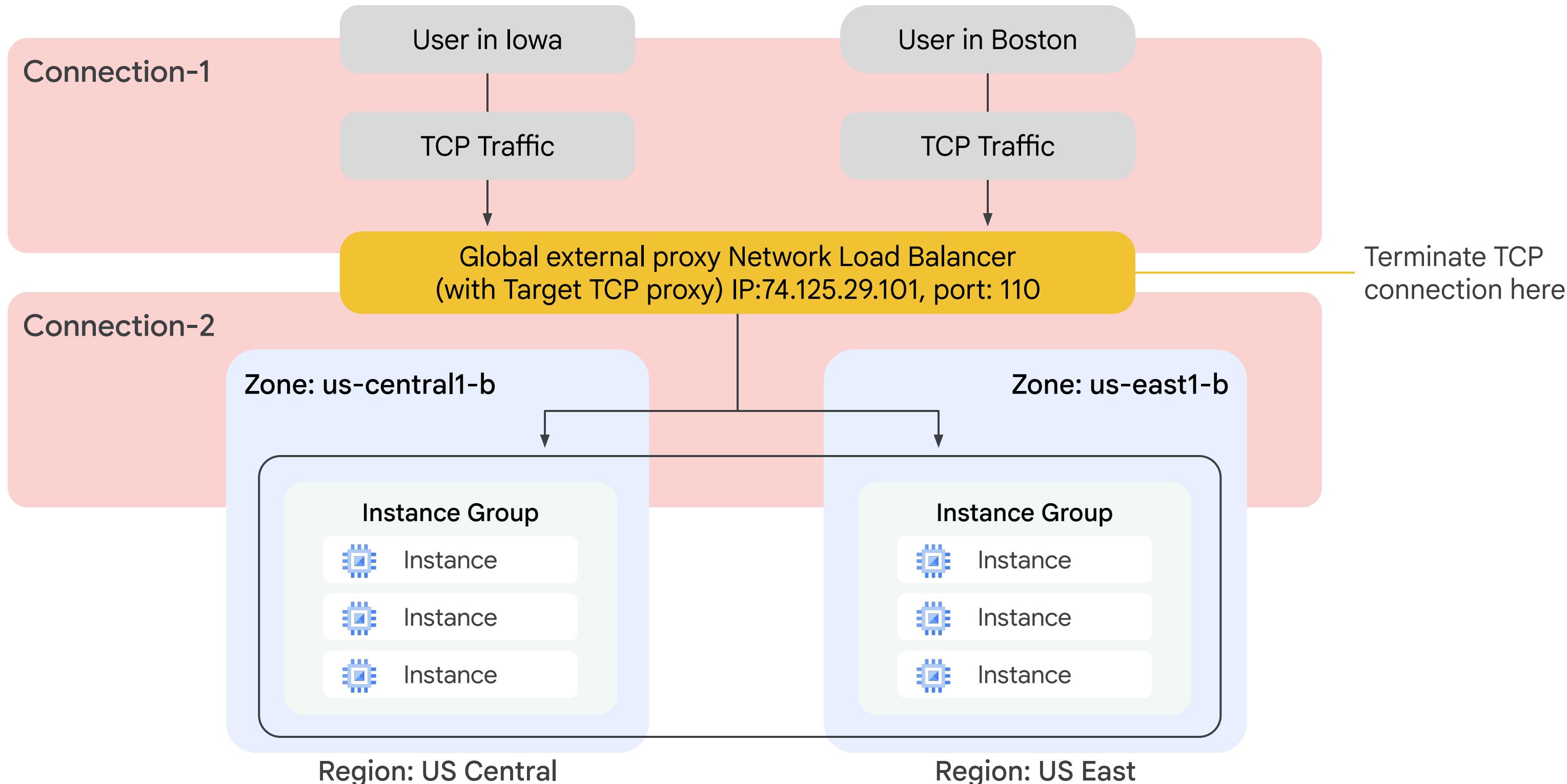


Architecture of a Proxy Network Load Balancer

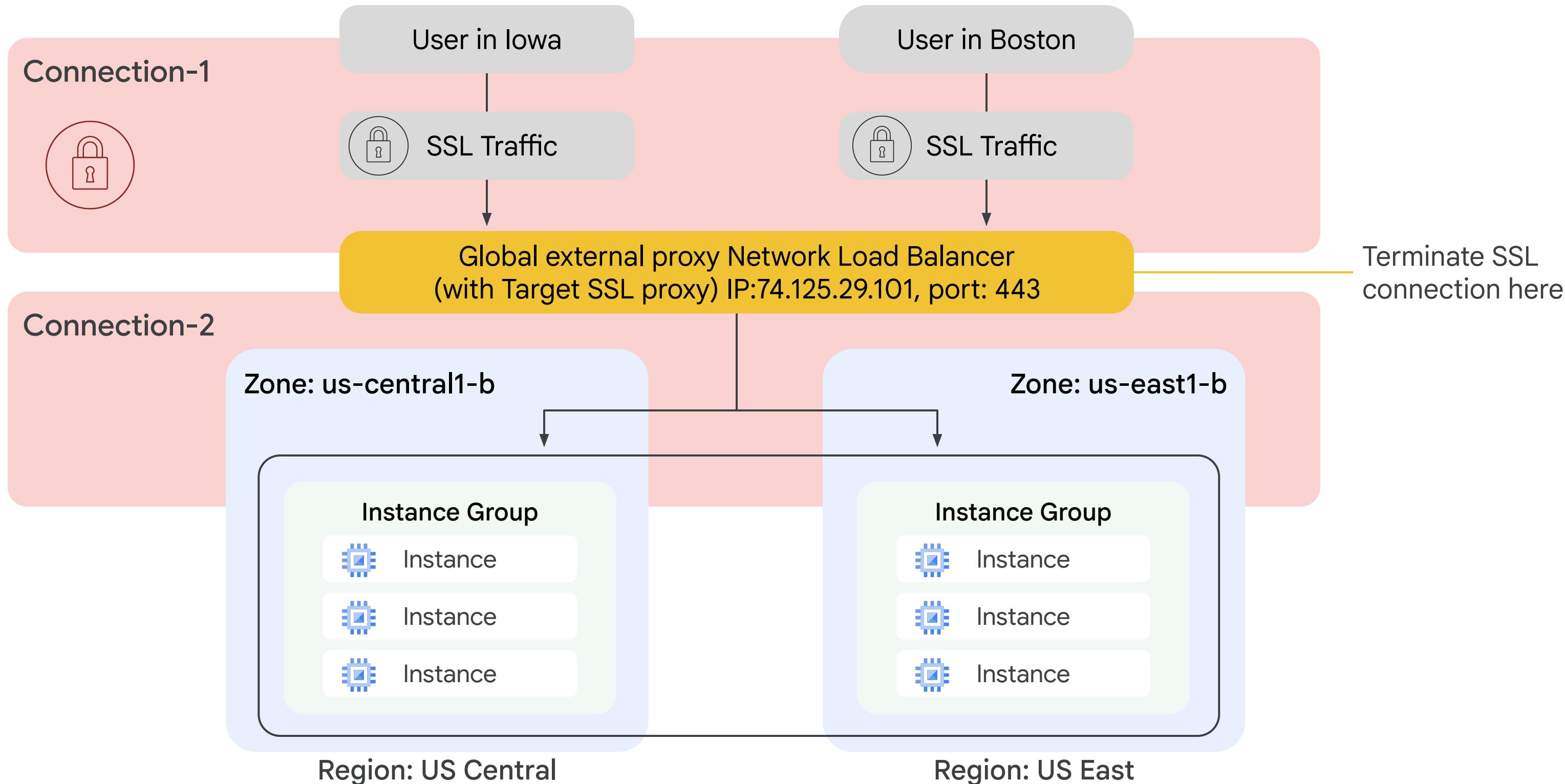


*Backend support differs depending on the deployment mode of the load balancer (internal or external, global or regional).

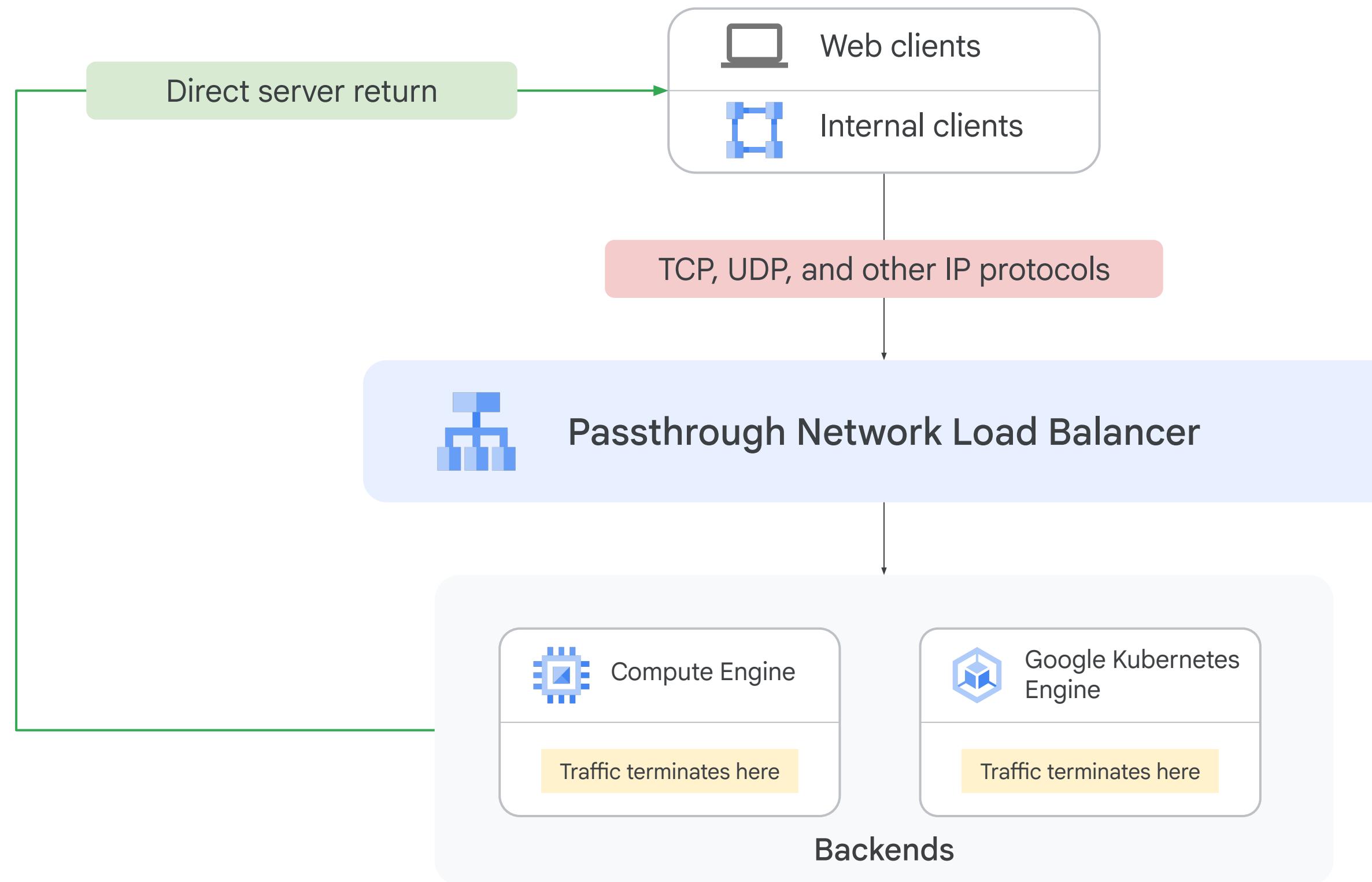
Proxy Network Load Balancer - Target TCP proxy



Proxy Network Load Balancer - Target SSL proxy



Architecture of a passthrough Network Load Balancer



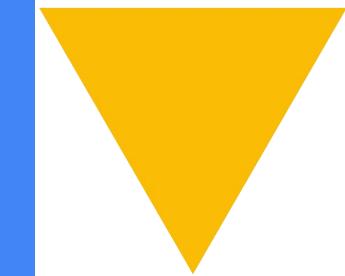
Backend service-based architecture

- Regional backend service
- Defines the behavior of the load balancer and how it distributes traffic to its backend instance groups
- Support for IPv4 and IPv6 traffic
- Multiple protocols
- Managed and unmanaged instance groups
- Non legacy health checks

Target pool-based architecture

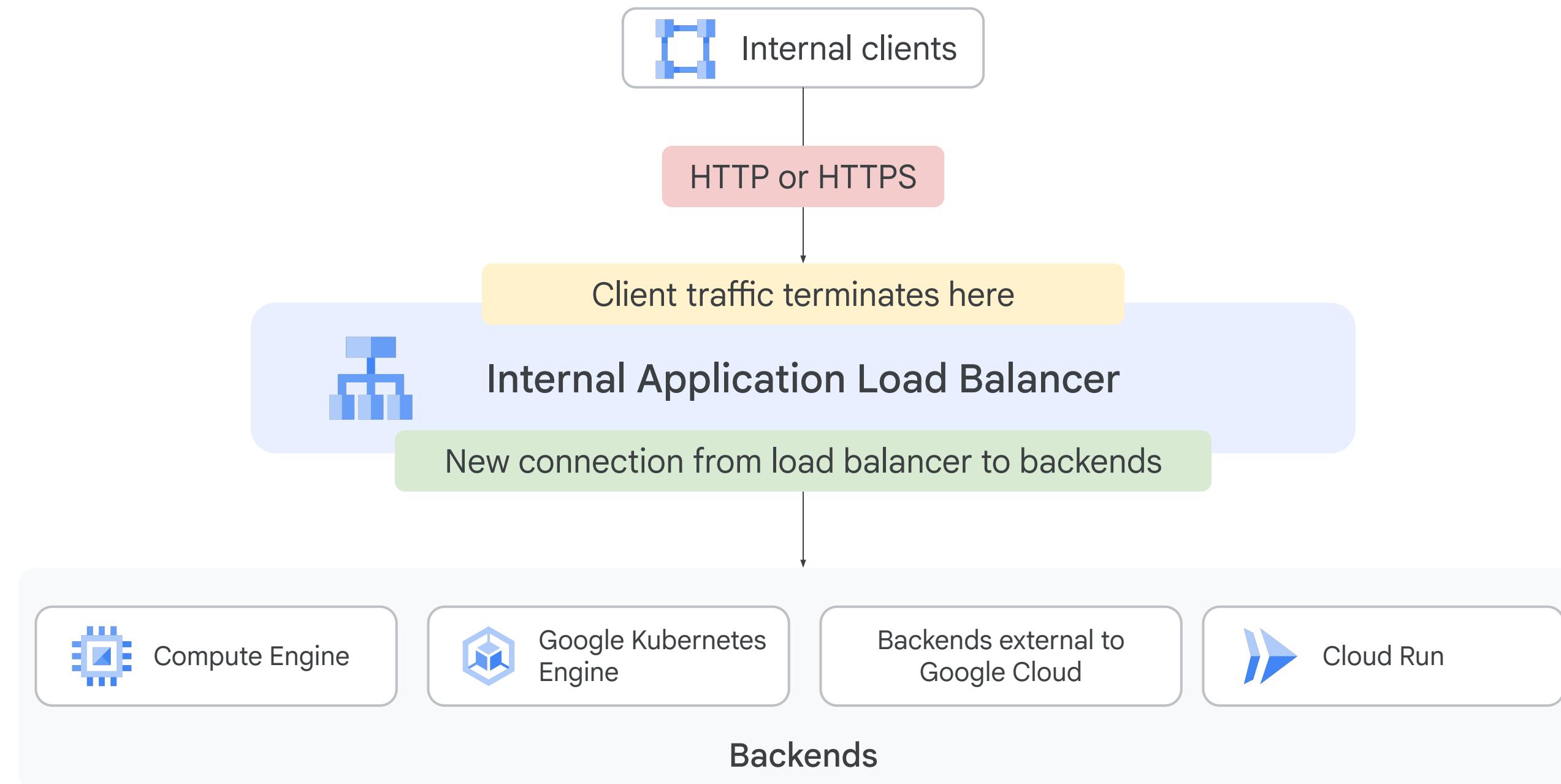
- Forwarding rules (TCP and UDP)
- Up to 50 per project
- One health check
- Instances must be in the same region

105



Internal Load Balancing

Architecture of an internal Application Load Balancer



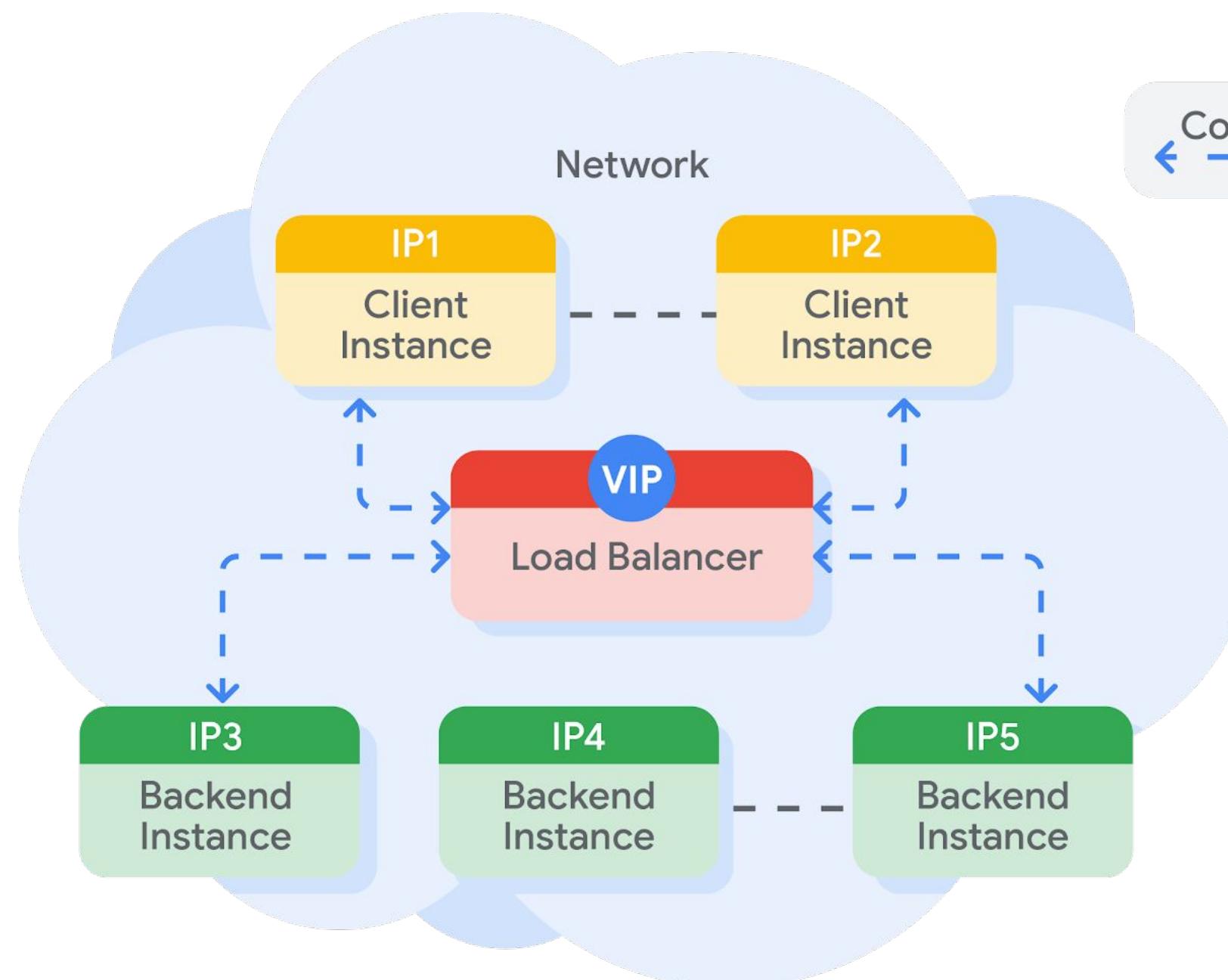
Internal Application Load Balancers

Deployment mode	Network service tier	Load balancing scheme	IP address	Frontend ports
Regional internal	Premium Tier	INTERNAL_MANAGED	IPv4	Can reference exactly one port from 1-65535
Cross-region internal	Premium Tier	INTERNAL_MANAGED	IPv4	

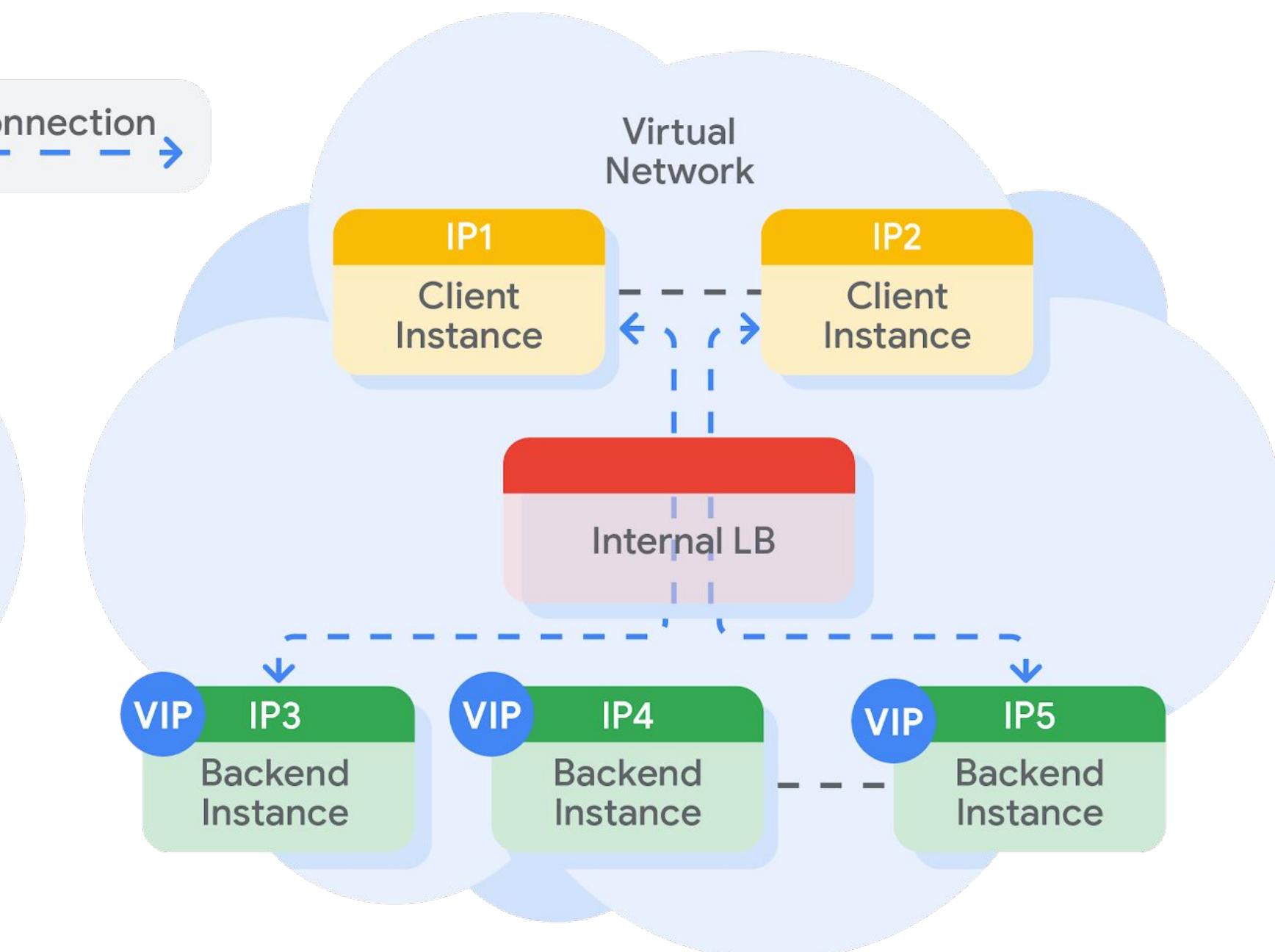
Internal passthrough Network Load Balancers

- Regional, private load balancing
 - VM instances in same region
 - RFC 1918 IP addresses
- TCP, UDP, ICMP, ICMPv6, SCTP, ESP, AH, and GRE traffic
- Reduced latency, simpler configuration
- Software-defined, fully distributed load balancing

Software-defined, fully distributed load balancing



Traditional proxy model of internal load balancing

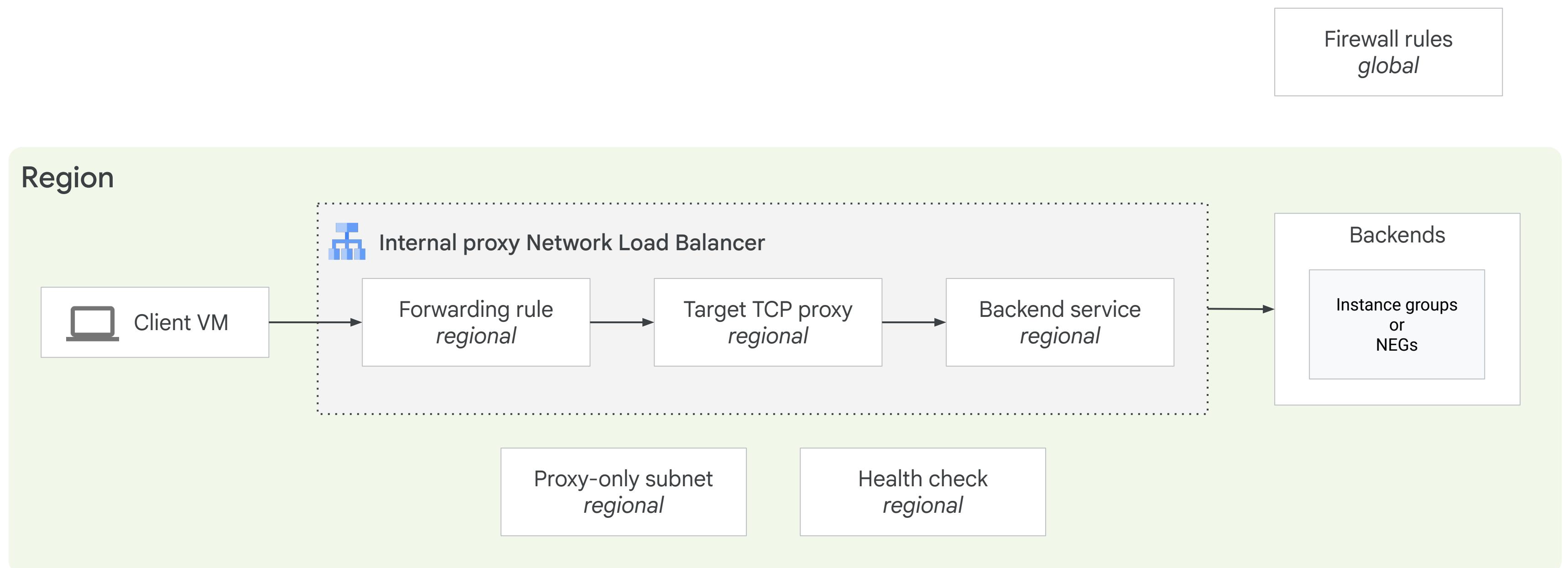


Internal passthrough Network Load Balancer

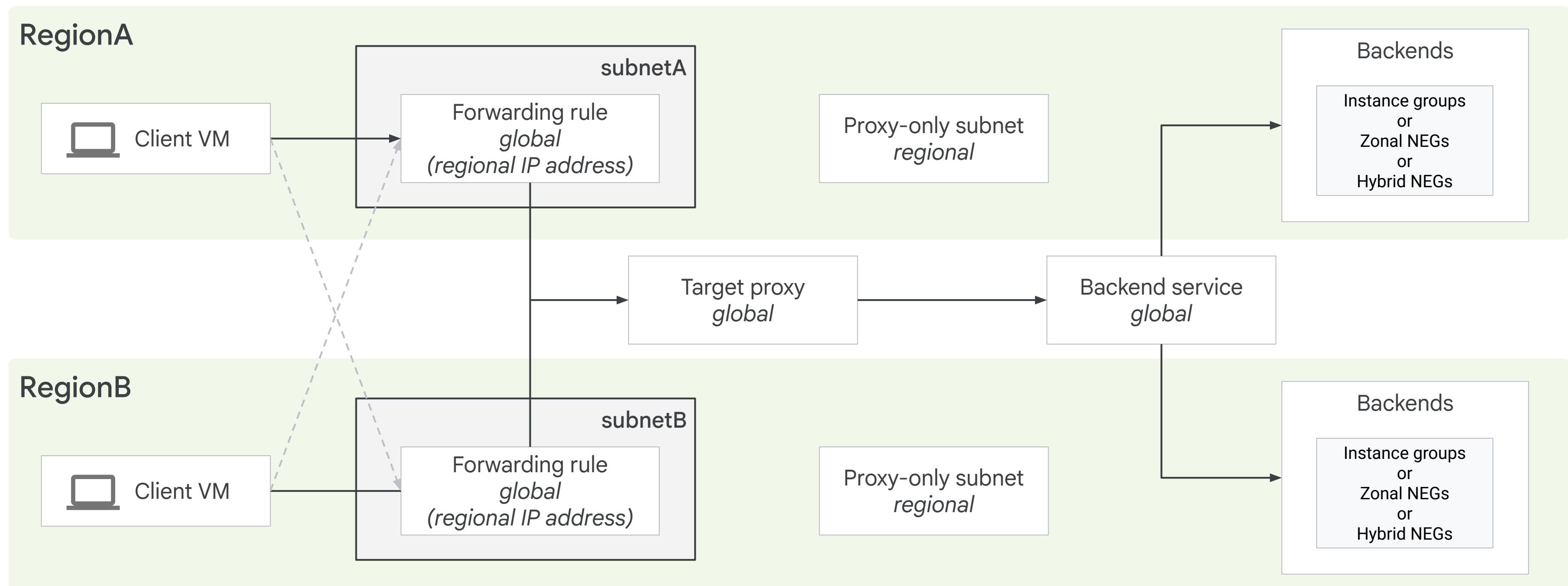
Internal proxy Network Load Balancers

- Proxy-based load balancer
- Balances traffic within your VPN network
- Regional or Cross-region
- Software-defined, fully distributed load balancing

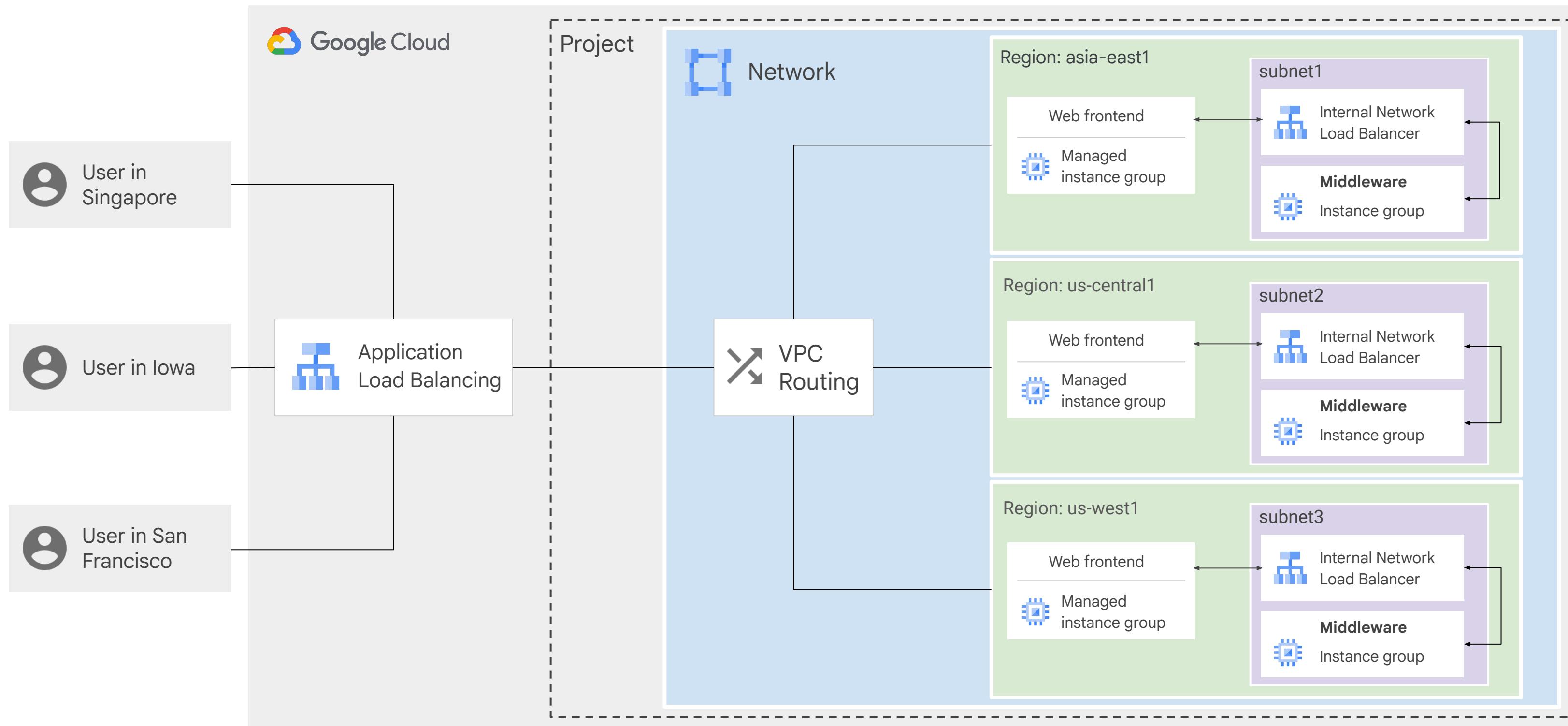
Architecture of a regional internal proxy Network Load Balancer



Architecture of a cross-region internal proxy Network Load Balancer

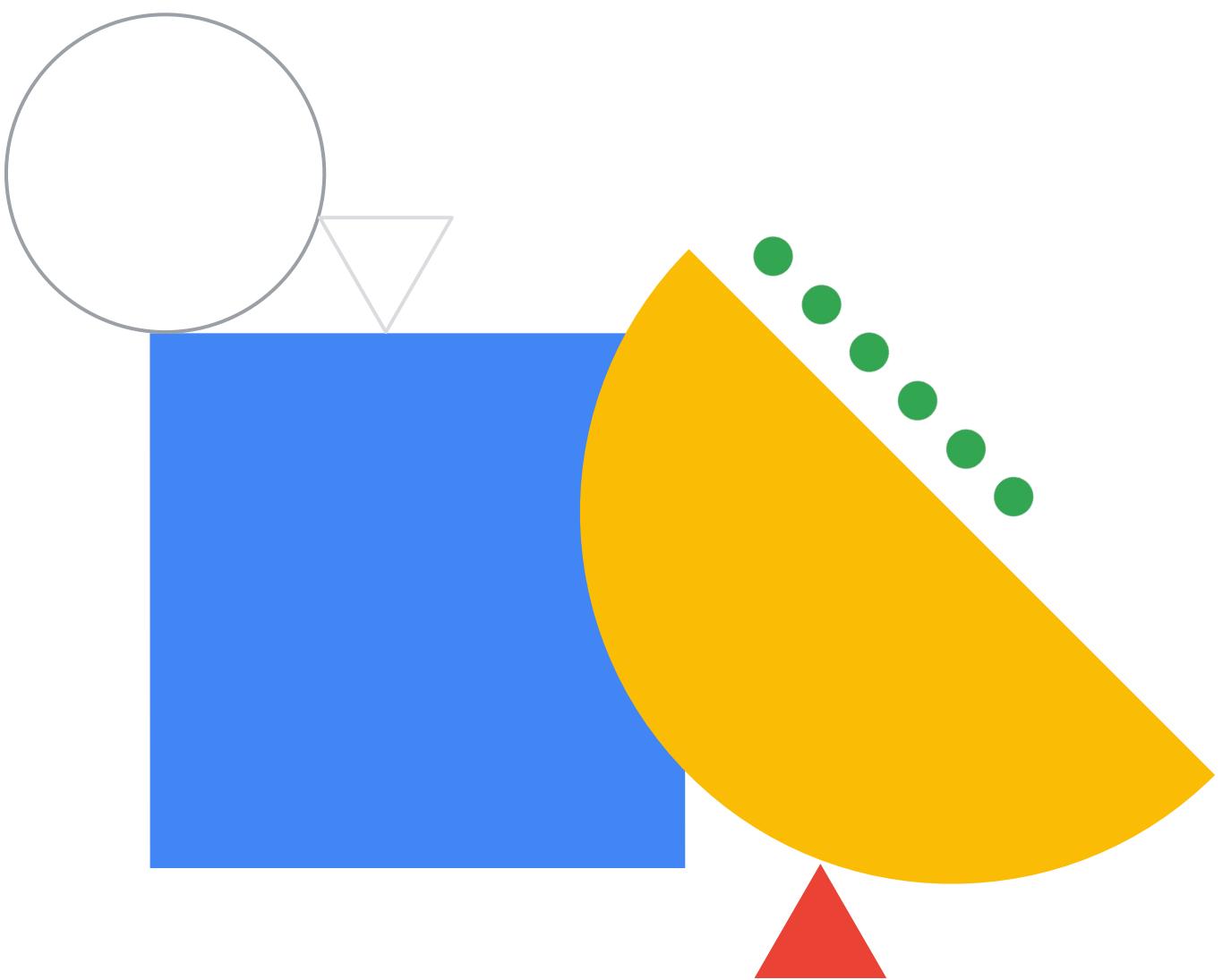


Internal load balancing supports 3-tier web services



Lab Intro

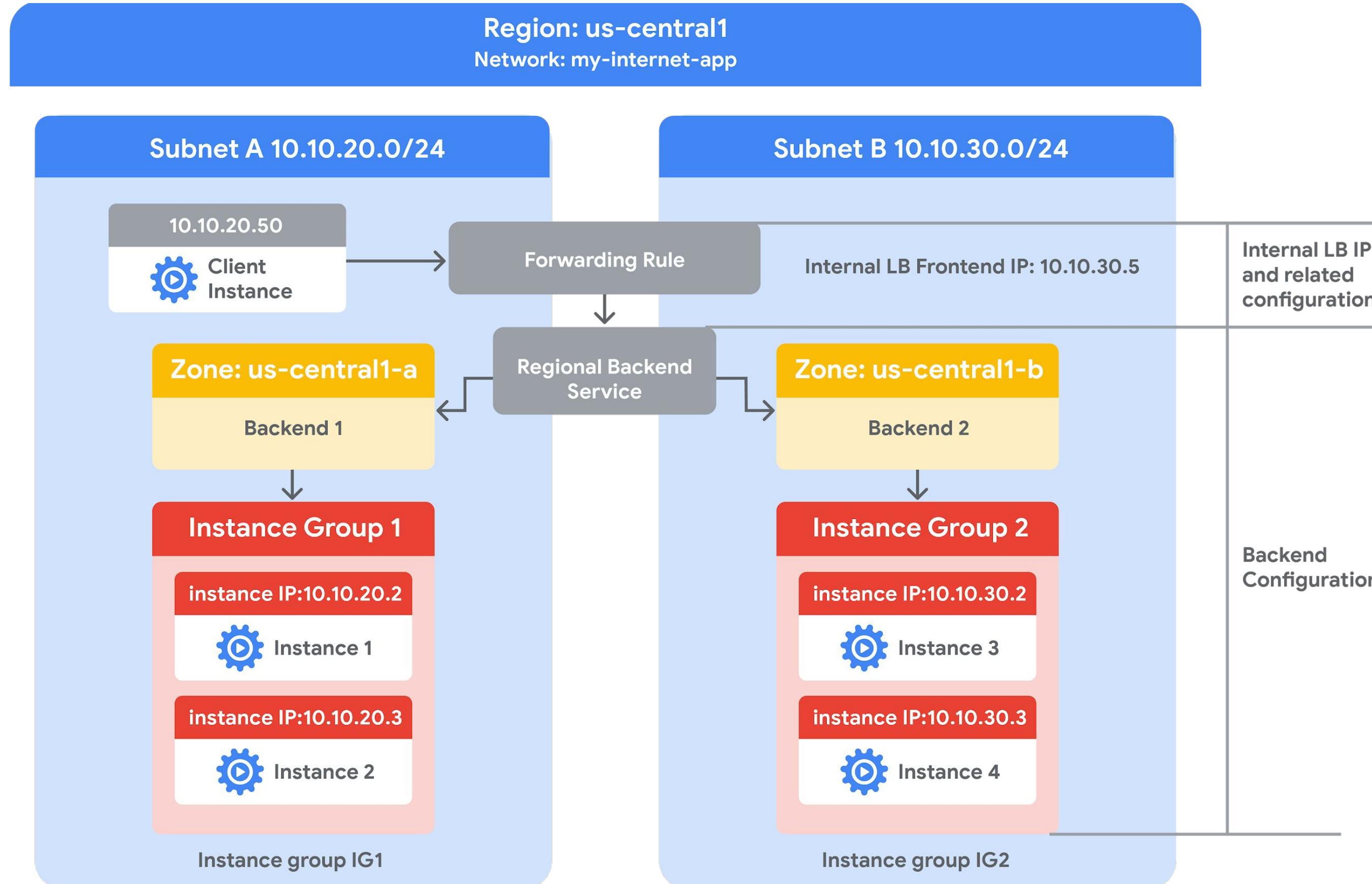
Configure an Internal Network
Load Balancer



Lab objectives

- 01 Create HTTP and health check firewall rules
- 02 Configure two instance templates
- 03 Create two managed instance groups

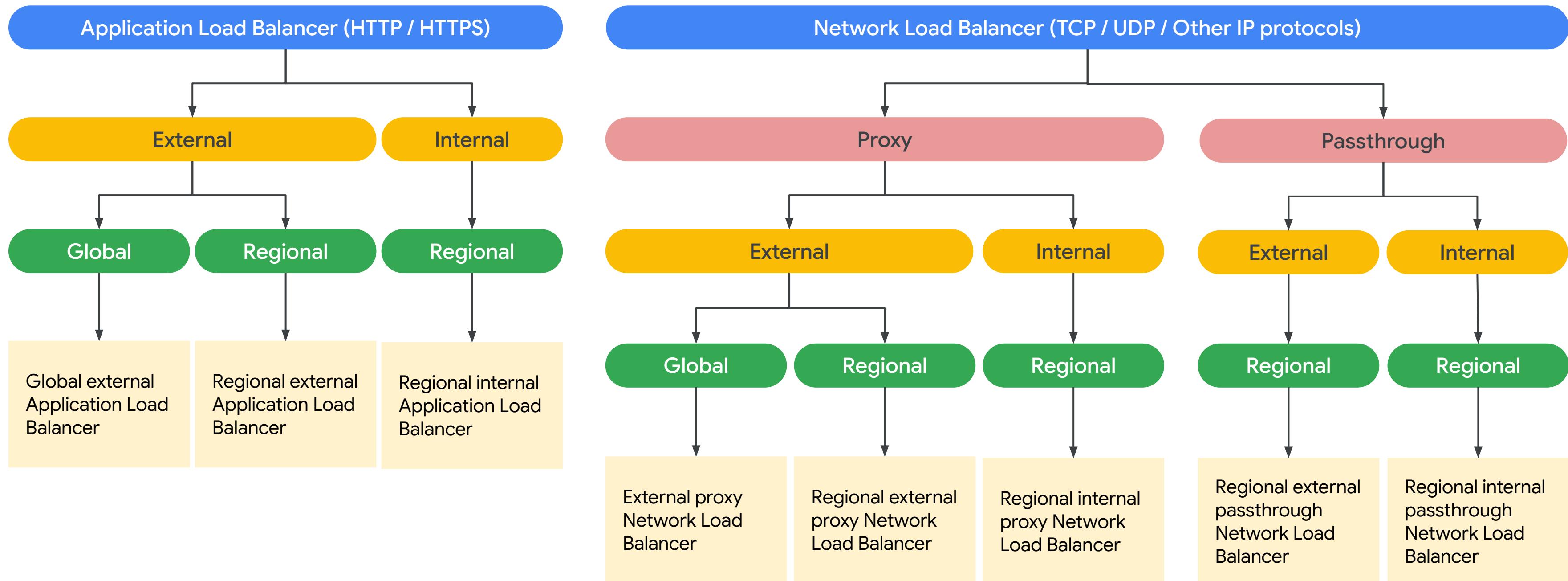






Choosing a Load Balancer

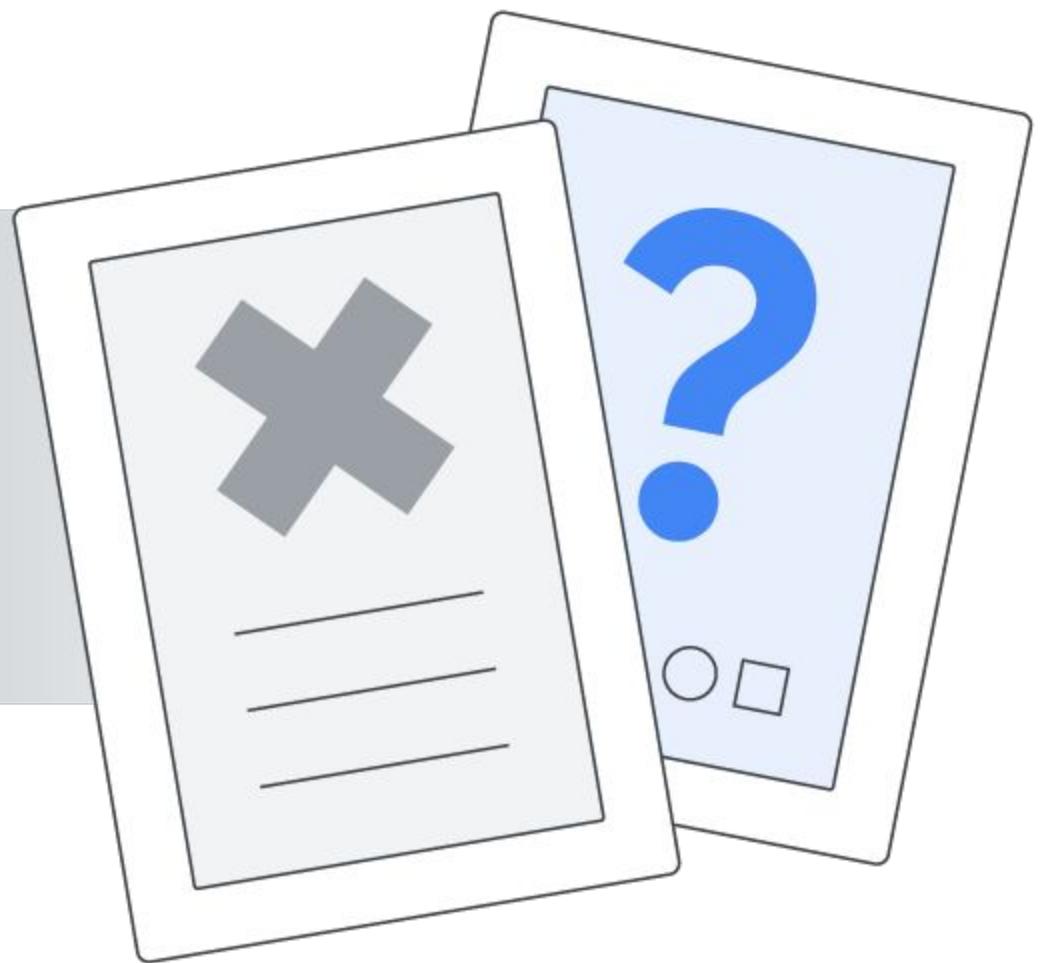
Deployment modes available for Cloud Load Balancing



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	EXTERNAL
			Regional in Standard	
	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. Global external Application Load Balancer
- B. External proxy Network Load Balancer
- C. Regional external proxy Network Load Balancer
- D. Global hardware-defined Load Balancer
- E. Regional external passthrough Network Load Balancer
- F. Regional internal Application Load Balancer

Question #1

Answer

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

- A. Global external Application Load Balancer
- B. External proxy Network Load Balancer
- C. Regional external proxy Network Load Balancer
- D. Global hardware-defined Load Balancer
- E. Regional external passthrough Network Load Balancer
- F. Regional internal Application Load Balancer



Question #2

Question

Which load balancer is recommended for HTTP(S) traffic?

- A. Regional internal passthrough Network Load Balancer
- B. Regional internal proxy Network Load Balancer
- C. Global external Application Load Balancer
- D. Global external proxy Network Load Balancer
- E. Regional external passthrough Network Load Balancer

Question #2

Answer

Which load balancer is recommended for HTTP(S) traffic?

- A. Regional internal passthrough Network Load Balancer
- B. Regional internal proxy Network Load Balancer
- C. Global external Application Load Balancer**
- D. Global external proxy Network Load Balancer
- E. Regional external passthrough Network Load Balancer



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

