Leveraging Load Balancing Options on the GCP

UNDERSTANDING LOAD BALANCING OPTIONS ON THE GCP

Vitthal Srinivasan
CO-FOUNDER, LOONYCORN

www.loonycorn.com

Overview

Introducing load balancers on the GCP

Global and regional load balancers

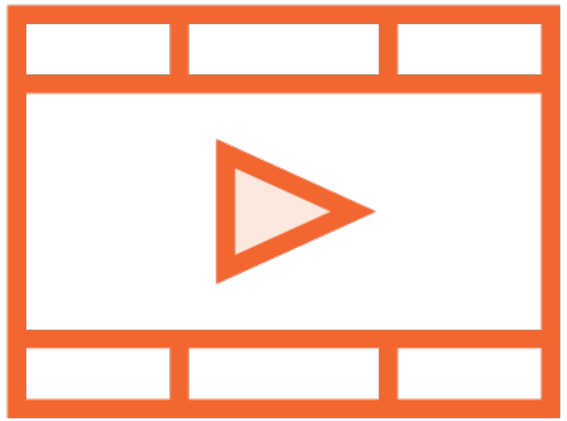
External and internal load balancers

Types of load balancers: HTTP(S), SSL proxy, TCP proxy, network and internal

Choosing the right load balancer

Prerequisites and Course Outline

Prerequisites: Basic Cloud Computing



Choosing and Implementing Google Cloud Compute Engine Solutions

Building Scalable Compute Solutions Using Managed Instance Groups

Course Outline

Load balancing options on the GCP

- Global and regional, external and internal
- Types of load balancers
- Choosing the right load balancer

Implementing HTTP(S) load balancing

- Unmanaged and managed instance groups
- HTTP(S) load balancing components
- HTTP(S) load balancing with autoscaling

Configuring other types of load balancers

- Understanding and implementing other load balancers on the GCP
- SSL proxy, TCP proxy, network and internal load balancing

Scenarios: SpikeySales.com

Hypothetical online retailer

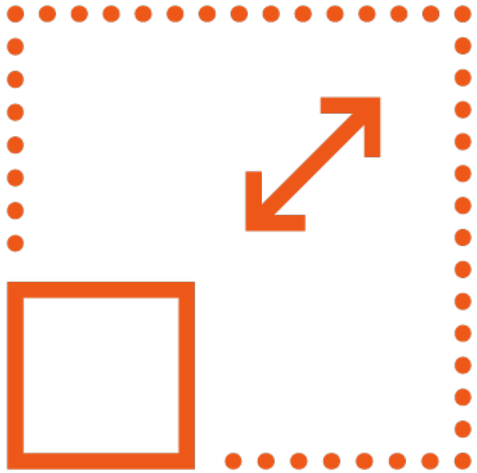
- Flash sales of trending products
- Spikes in user traffic

SpikeySales on the GCP

- Cloud computing fits perfectly
- Pay-as-you-go
- No idle capacity during off-sale periods

Introducing Load Balancing

Attractions of Cloud Computing



Autoscaling

Compute capacity automatically changes with changing need

Autohealing

Platform ensures health of compute resources

Cloud VM Instances

Individual VM instances do not provide either advantage

Some higher level abstraction is needed to do so

Two Types of Scaling

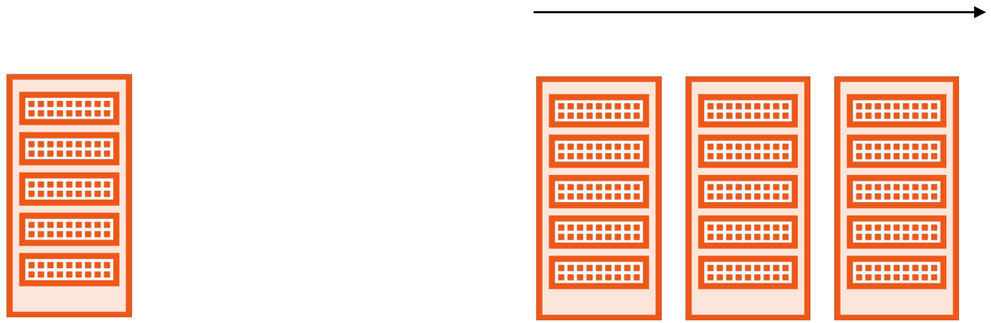
Vertical

Make server more powerful

As need for compute grows

Horizontal

Add more servers (each still generic)



As need for compute grows

Managed Instance Groups are a horizontally scaled IaaS offering with autohealing and autoscaling

Managed Instance Group

Group of identical GCE VM instances, created from the same instance template that are managed by the platform

Instance Template

A specification of machine type, boot disk (or container image), zone, labels and other instance properties that can be used to instantiate either individual VM instances or a Managed Instance Group

Instance Template

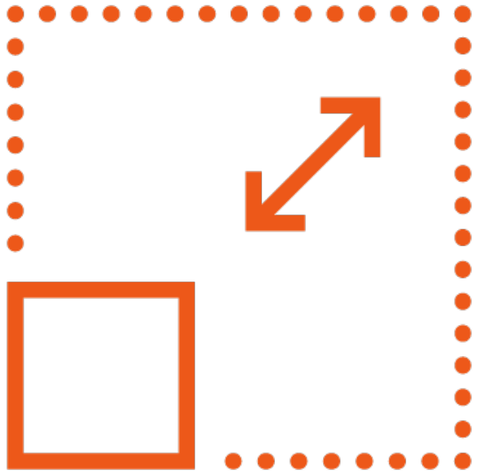
VM Instance
configured a certain way

Instance Template
captures config specs

New Individual VM Instance

Managed Instance Group

Attractions of Cloud Computing



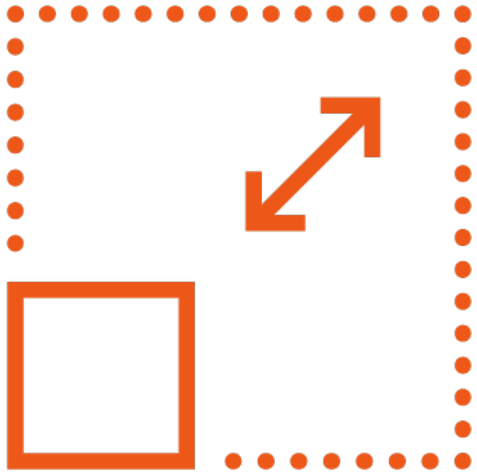
Autoscaling

Compute capacity automatically changes with changing need

Autohealing

Platform ensures health of compute resources

Applying to MIGs



Autoscaling

Associate autoscaling policy with MIG

Autohealing

Associate health check and autohealing policy with MIG

Attractions of Cloud Computing

Autoscaling

Associate autoscaling policy with MIG

Autohealing

Associate health check and autohealing policy with MIG

Health Checks

Health Check

Associate Health Check with MIG

Instance Template

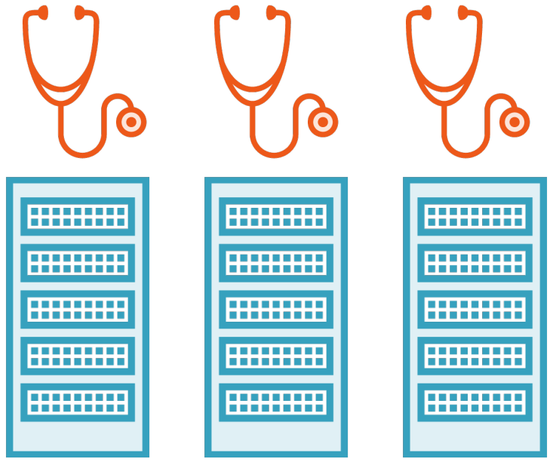
Managed Instance Group

Health Checks

Health Check

Sends probes to check health of each instance in MIG

Instance Template



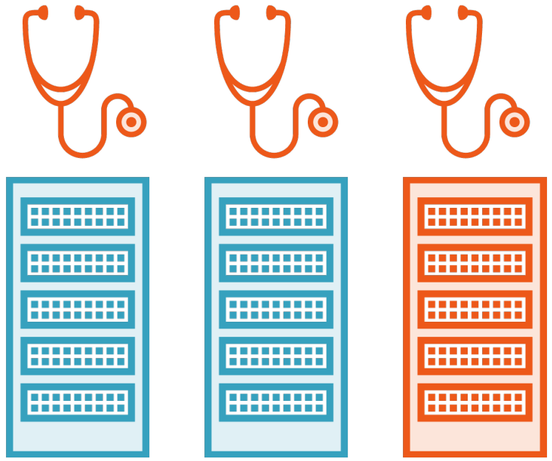
Managed Instance Group

Health Checks

Health Check

Instance Template

Probes identify any unhealthy instance



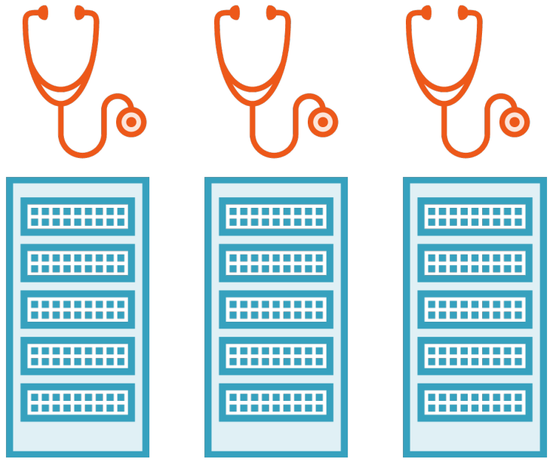
Managed Instance Group

Health Checks

Health Check

Instance Template

MIG then replaces it with a healthy one



Managed Instance Group

Scalable Compute with MIGs

User Traffic

Incoming requests from users during Black Friday sale

Backend Service

Managed Instance Group to serve those incoming requests

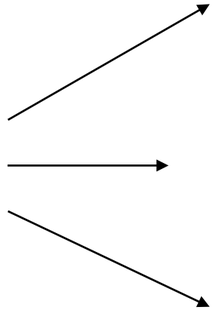
Still two missing pieces of puzzle

Scalable Compute with MIGs

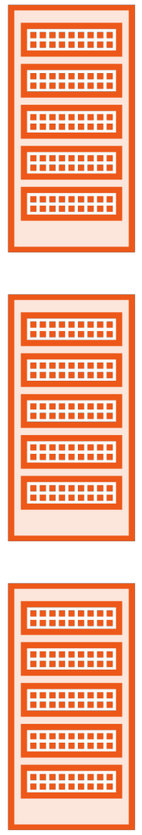
User Request

What IP address to access?

Which specific instance services request?



Backend Service



Still two missing pieces of puzzle

Two Missing Pieces of the Puzzle

What IP Address?

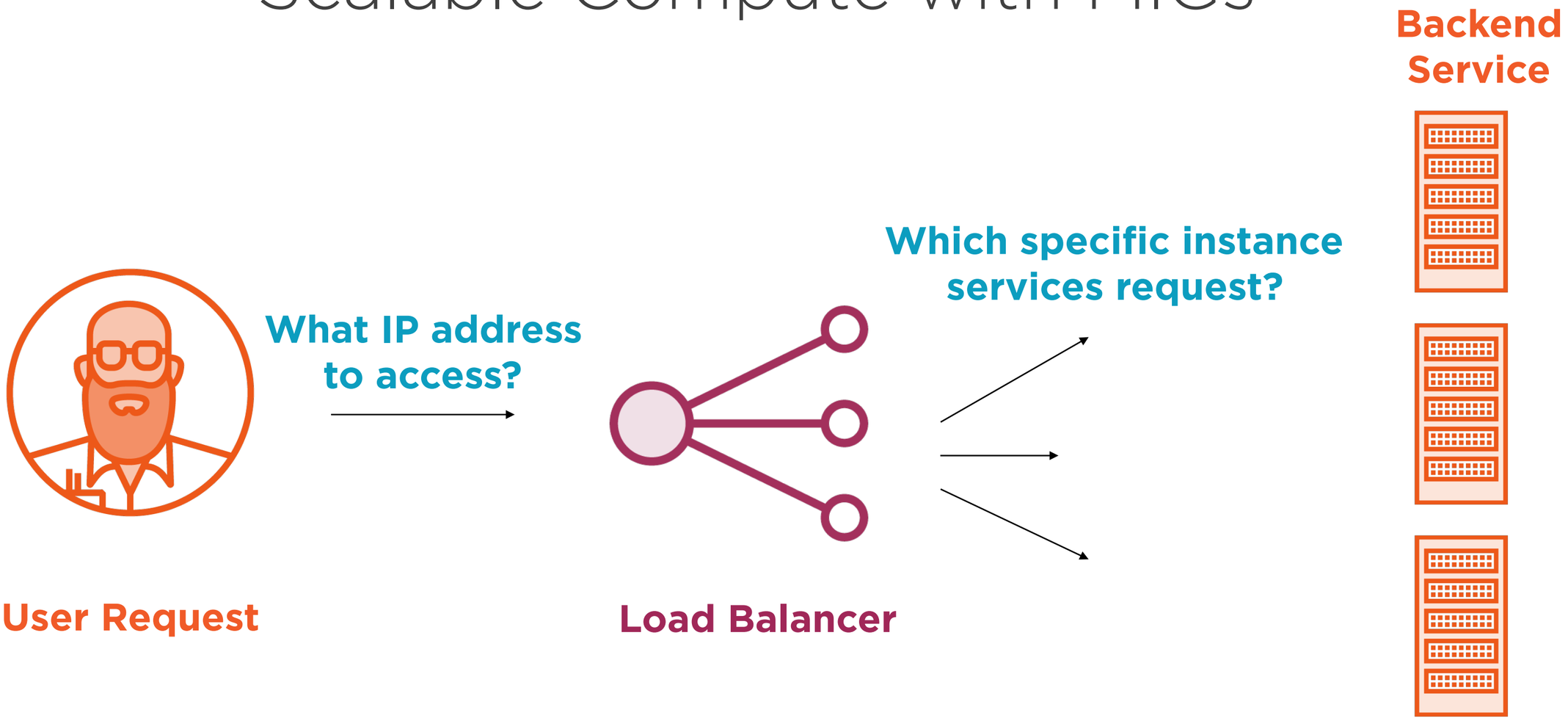
Individual VM IP addresses are ephemeral; traffic needs stable front-end IP

Which specific instance?

Individual VMs will come and go, and experience varying load; load needs to be balanced

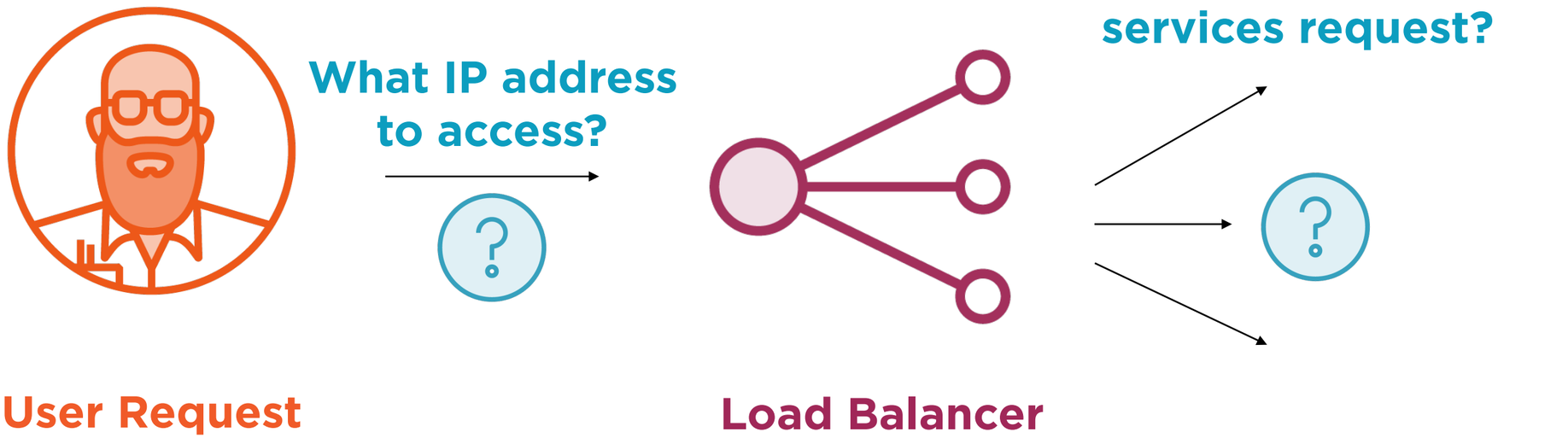
Load Balancer to the rescue

Scalable Compute with MIGs

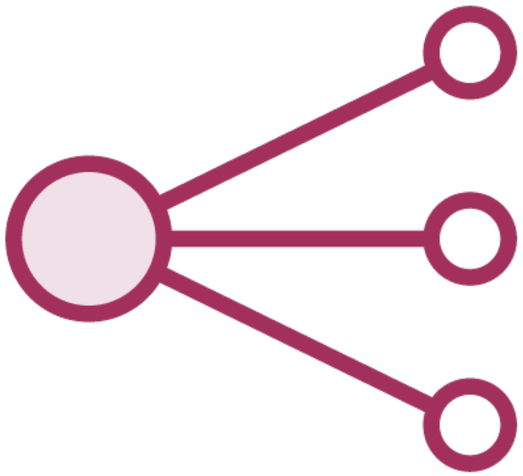


Load Balancer to the rescue

Scalable Compute with MIGs



Load Balancers



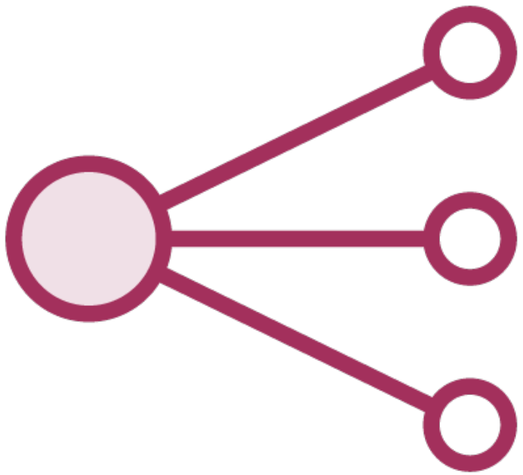
Complex service with many moving parts

Basic idea

- Stable front-end IP
- Forwarding rules to funnel traffic
- Connect to backend service
- Distribute load intelligently
- Health checks to avoid unhealthy instances

The primary purpose of load balancers is to **distribute** traffic to resources close to users and meet **high-availability** requirements

Load Balancers on the GCP



Fully managed, software-defined, redundant and highly available

Supports > 1 million queries per second with high performance and low latency

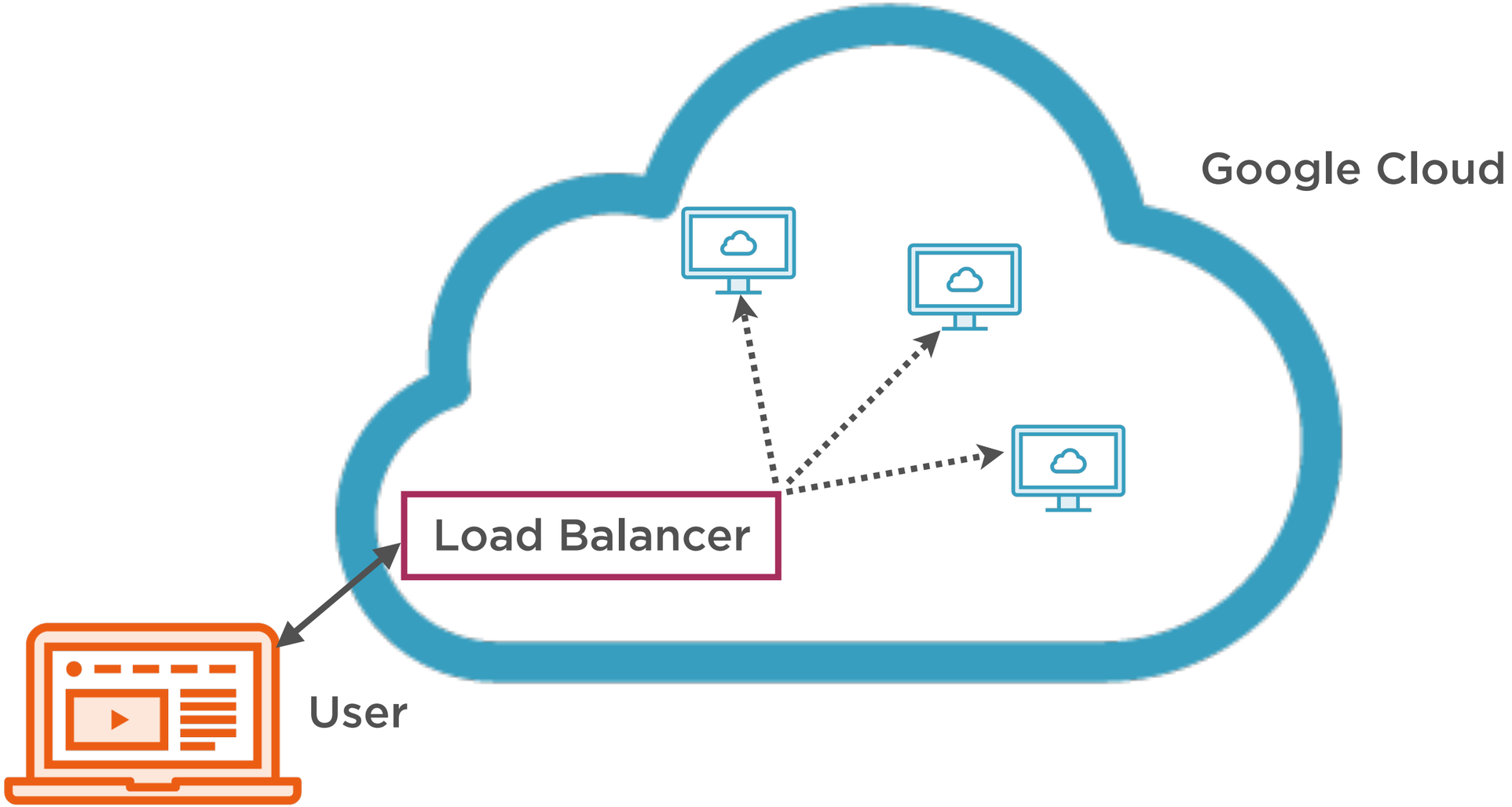
Autoscaling with no pre-warming to scale to increased traffic

Route traffic to closest VM

Load balancers on the GCP can also work with **unmanaged instance groups** which offer **no** autoscaling and autohealing properties

Types of Load Balancing on the GCP

Load Balancing



Global Load Balancing

Use when your users and instances are globally distributed, Provides IPv4 and IPv6 termination

Regional Load Balancing

Use when instances and users are concentrated in one region and only IPv4 termination is needed

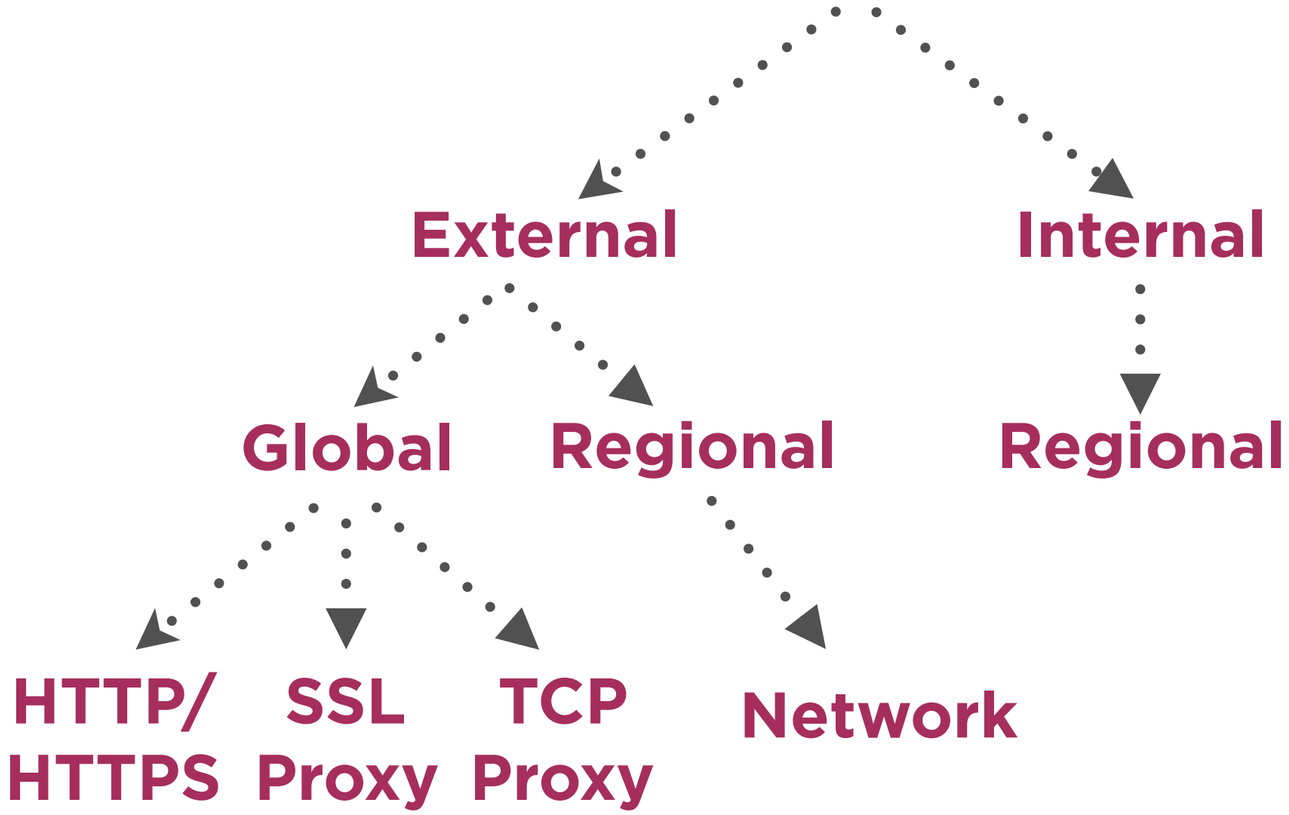
External Load Balancing

Distributes traffic from the internet to a GCP network

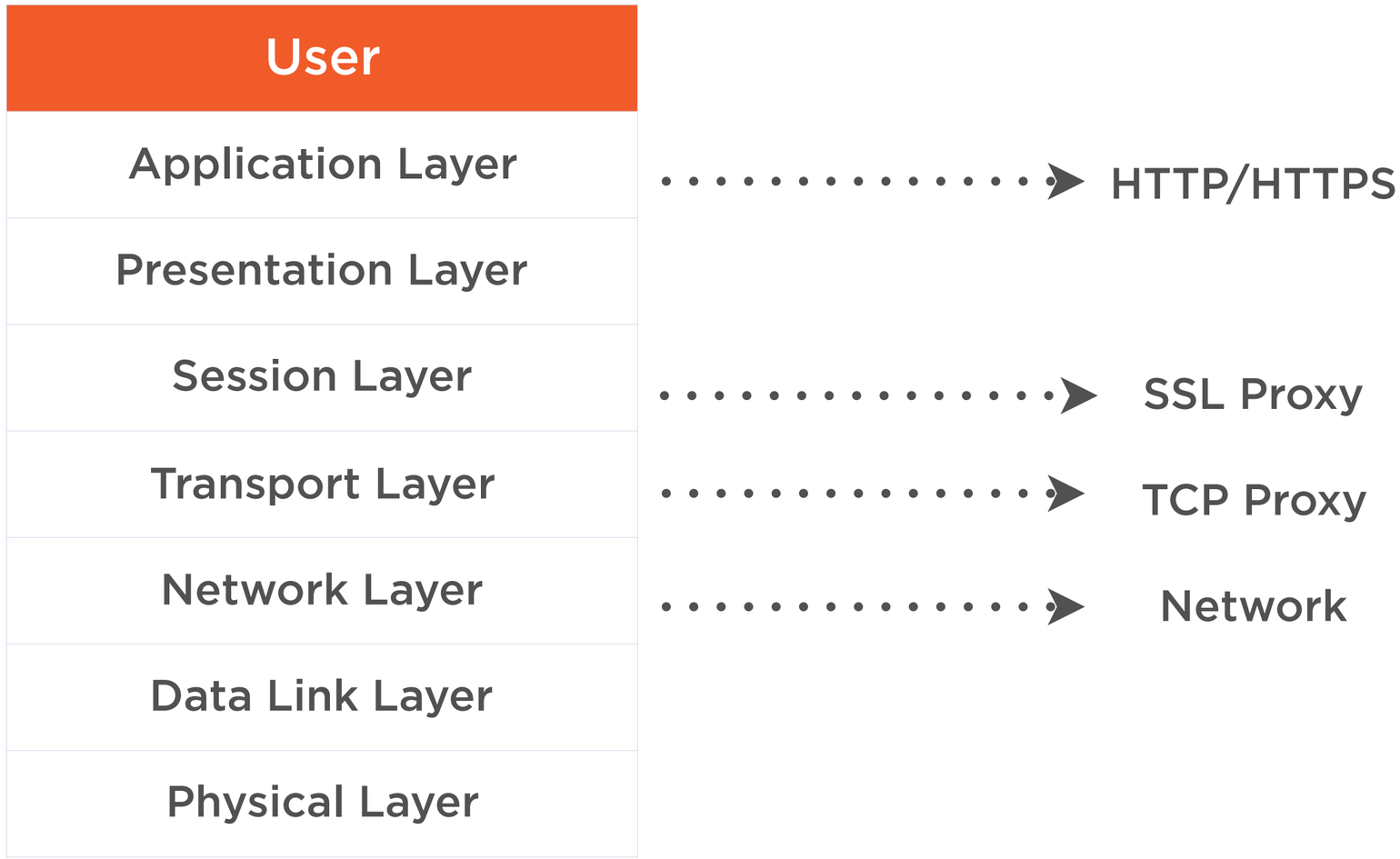
Internal Load Balancing

Distributes traffic only within a GCP network

Load Balancing



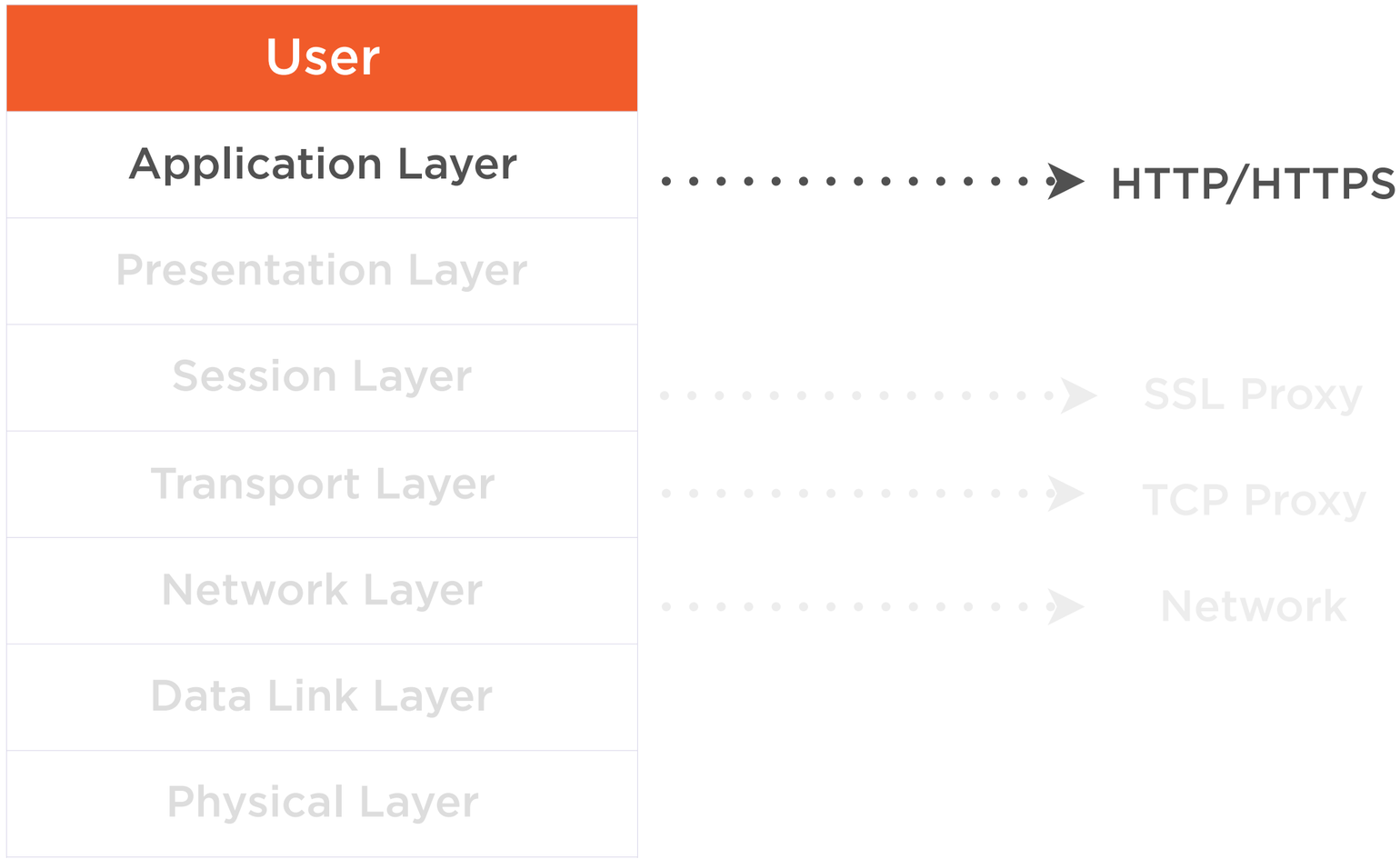
OSI Network Stack



The different load balancers operate at different layers of the OSI network stack

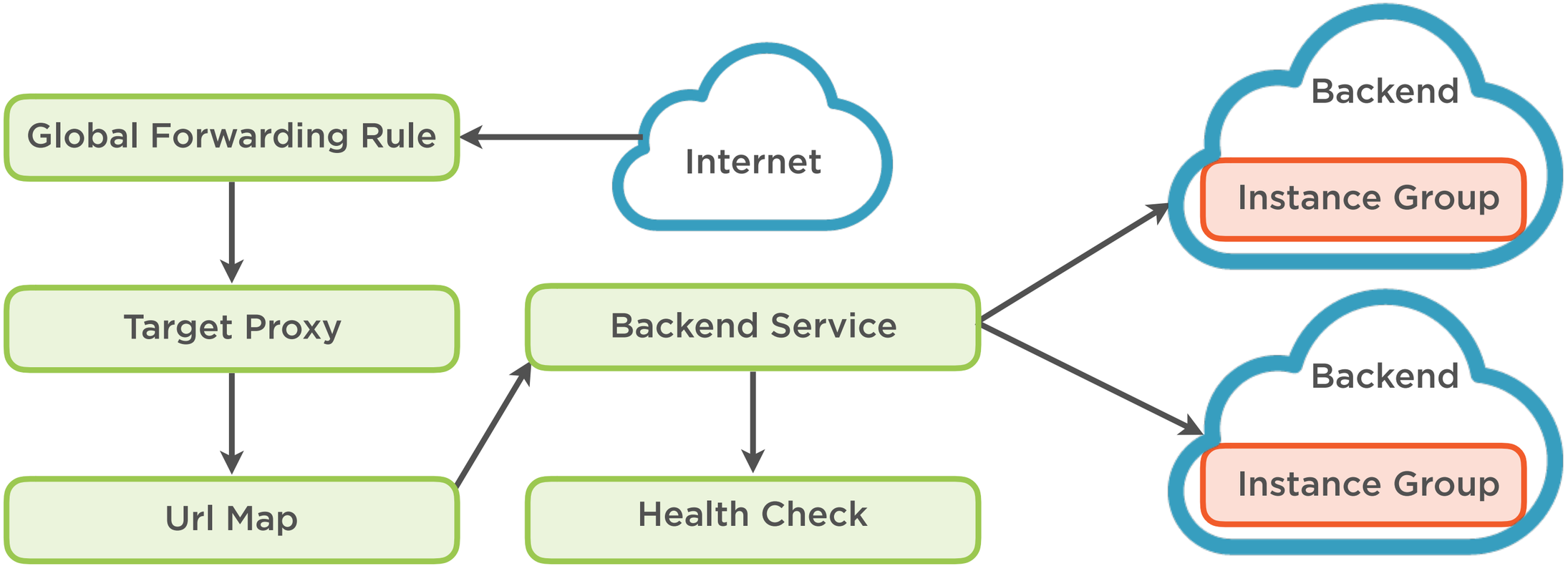
Load Balancing

HTTP(S) Load Balancing



HTTP(S) is used to balance global, external traffic

HTTP(S) Load Balancing



A global, external load balancing service offered on the GCP

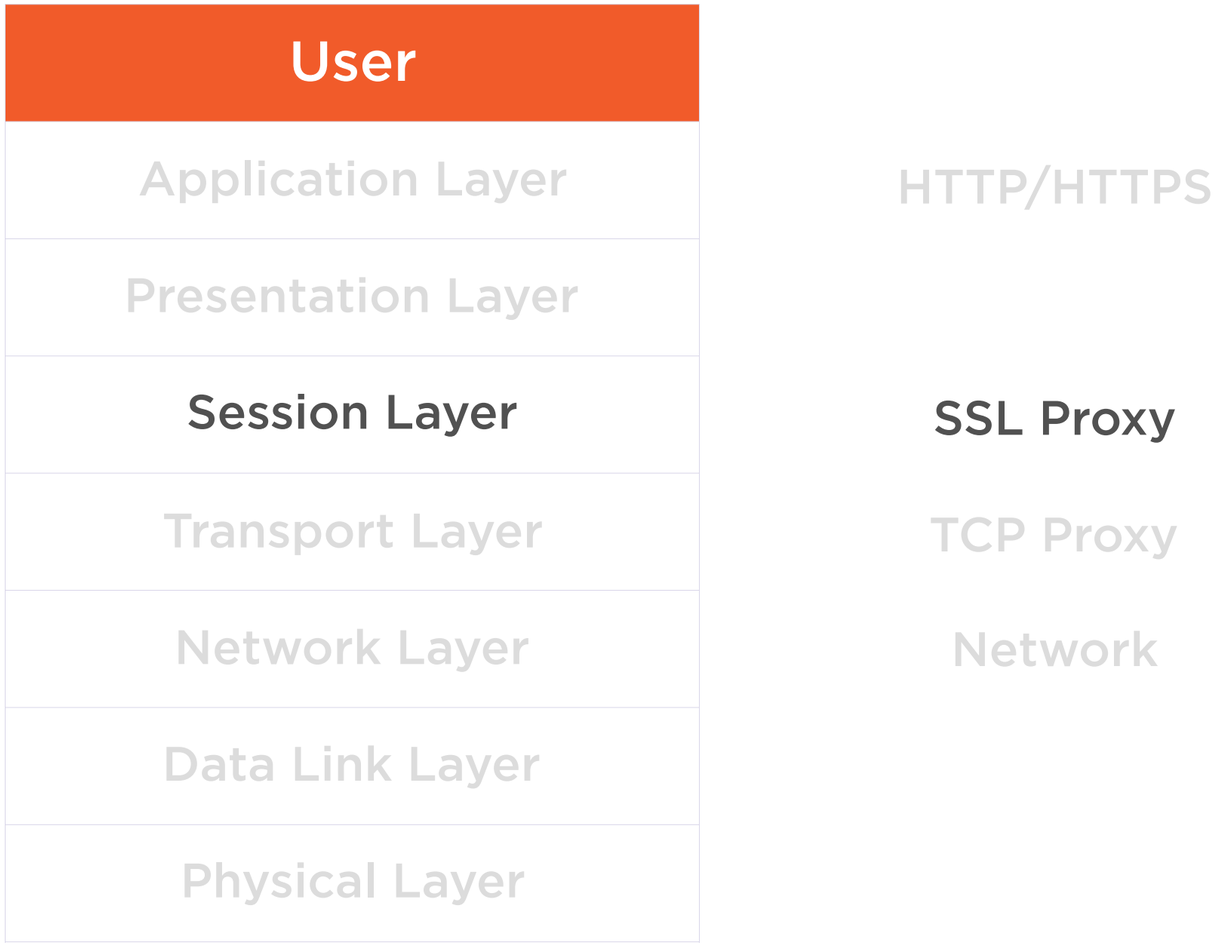
HTTP(S) Load Balancing

Distributes HTTP(S) traffic among groups of instances based on:

- Proximity to the user
- Requested URL
- Or both.

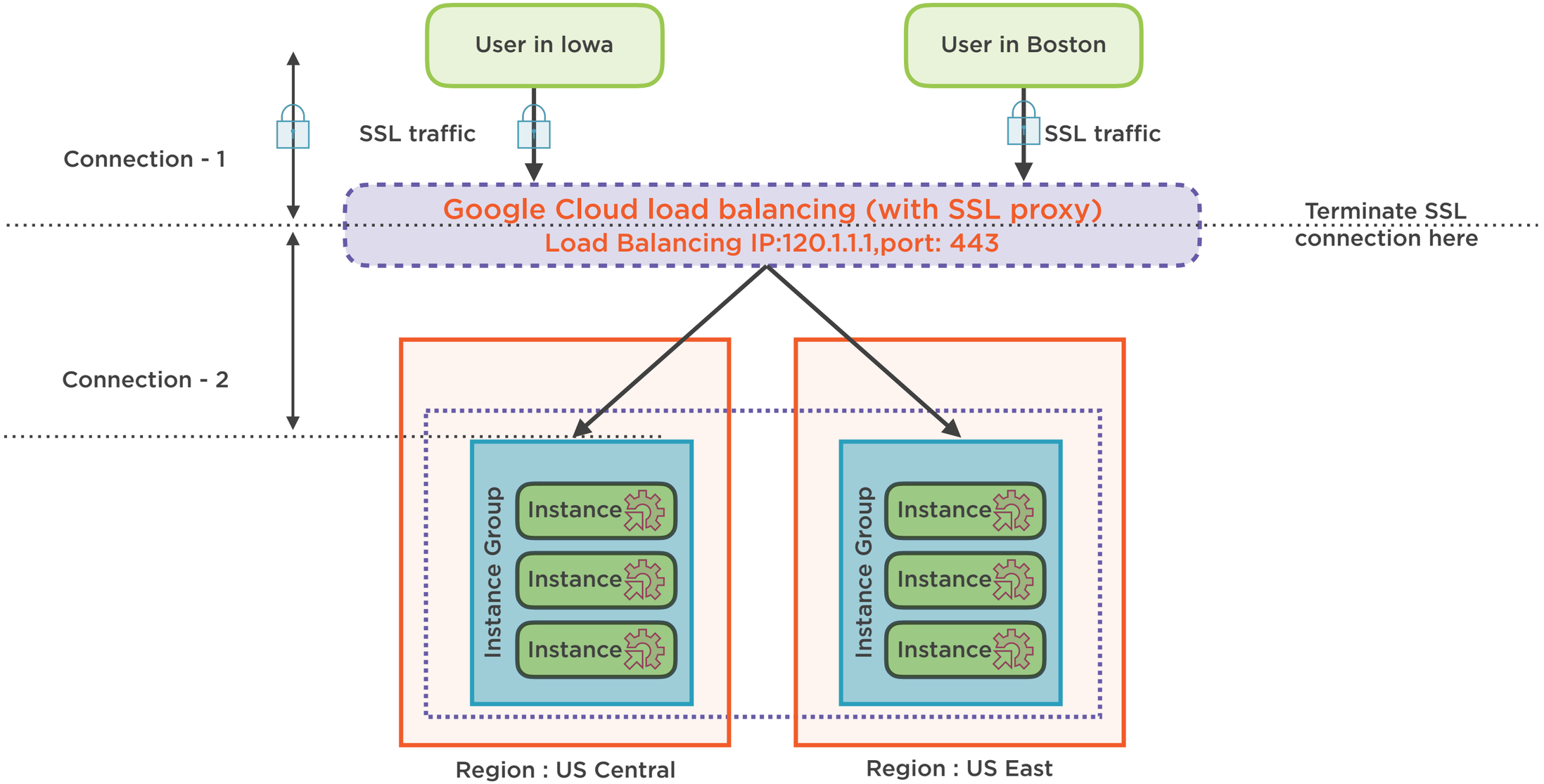
Load Balancing

SSL Proxy Load Balancing



SSL operates in the session layer

SSL Proxy Load Balancing



SSL Proxy Load Balancing

Use only for non-HTTP(S) SSL traffic

For HTTP(S), just use HTTP(S) load balancing

SSL connections are terminated at the global layer

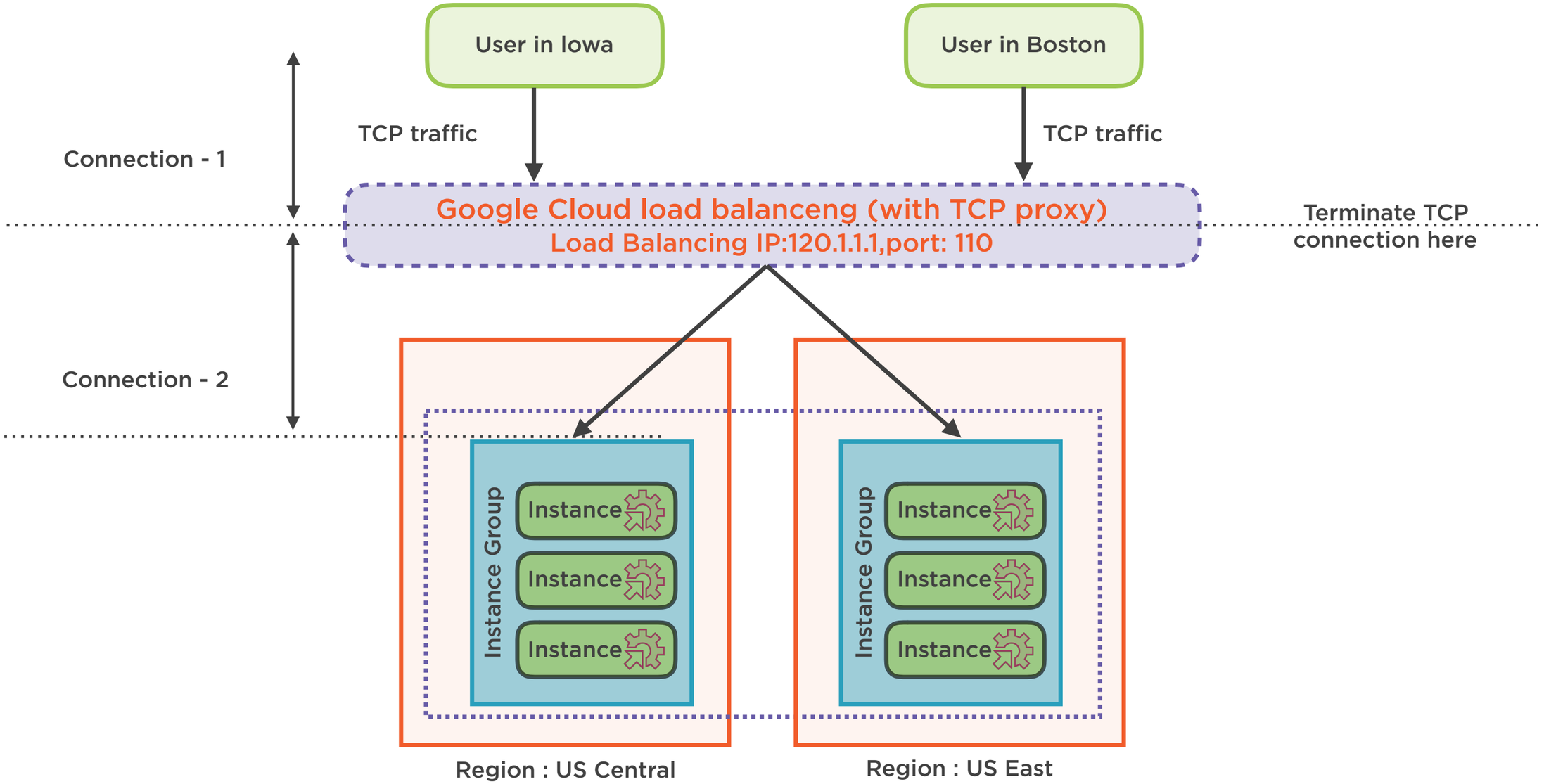
Then proxied to the closest available instance group

Load Balancing

TCP Proxy Load Balancing

User	
Application Layer	HTTP/HTTPS
Presentation Layer	
Session Layer	SSL Proxy
Transport Layer	**TCP Proxy**
Network Layer	Network
Data Link Layer	
Physical Layer	

TCP Proxy Load Balancing



TCP Proxy Load Balancing

Allows you to use a single IP address for all users around the world

Automatically routes traffic to the instances that are closest to the user

More intelligent routing than network load balancing

Better security, TCP vulnerabilities patched at the load balancer

Load Balancing

Network Load Balancing

User	
Application Layer	HTTP/HTTPS
Presentation Layer	
Session Layer	SSL Proxy
Transport Layer	TCP Proxy
Network Layer	**Network**
Data Link Layer	
Physical Layer	

Network Load Balancing

Based on incoming IP protocol data, such as address, port, and protocol type

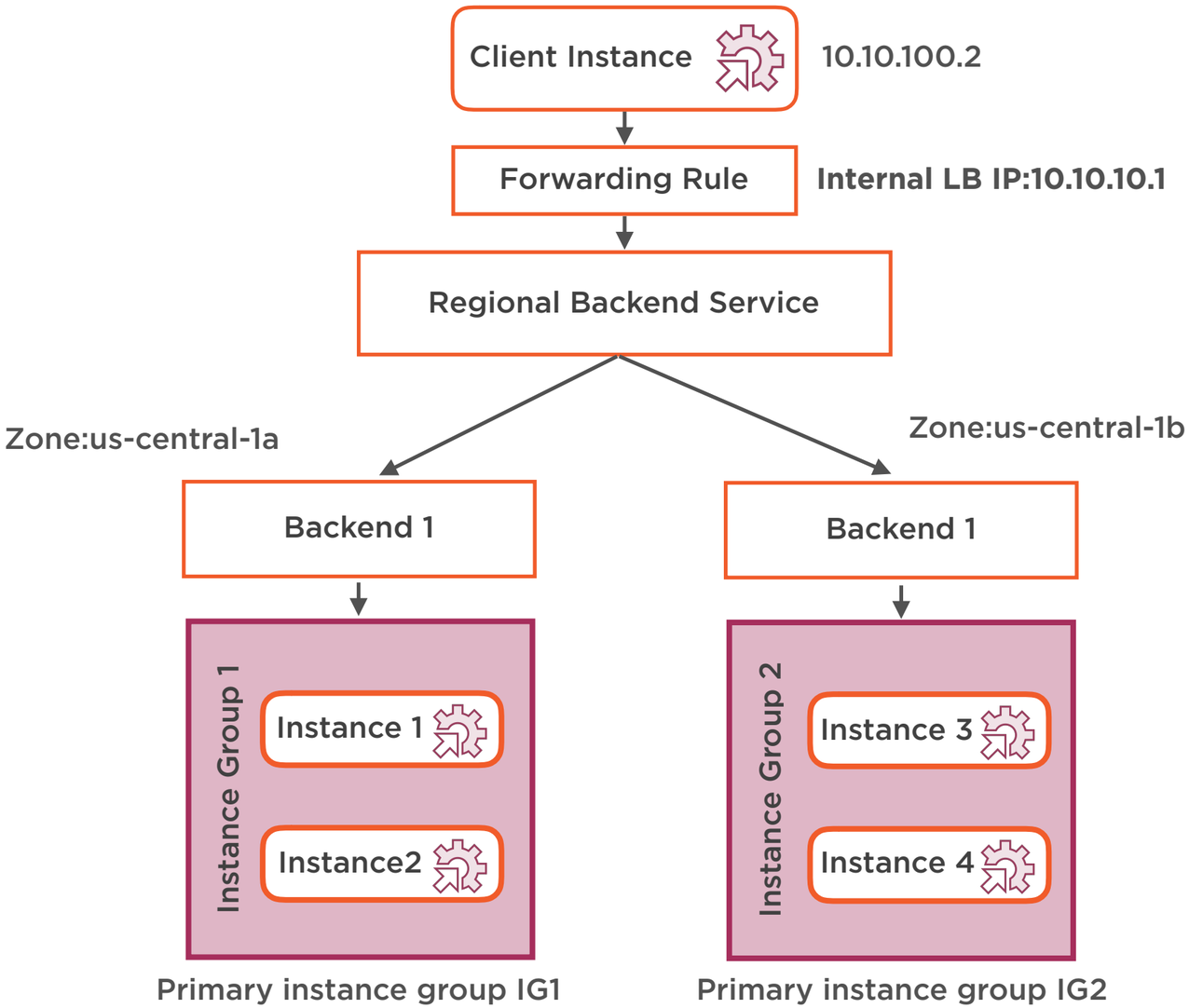
Pass-through, regional load balancer - does not proxy connections from clients

Use it to load balance UDP traffic, and TCP and SSL traffic

Load balances traffic on ports that are not supported by the SSL proxy and TCP proxy load balancers

Load Balancing

Internal Load Balancing



Internal Load Balancing

Private load balancing IP address that only your VPC instances can access

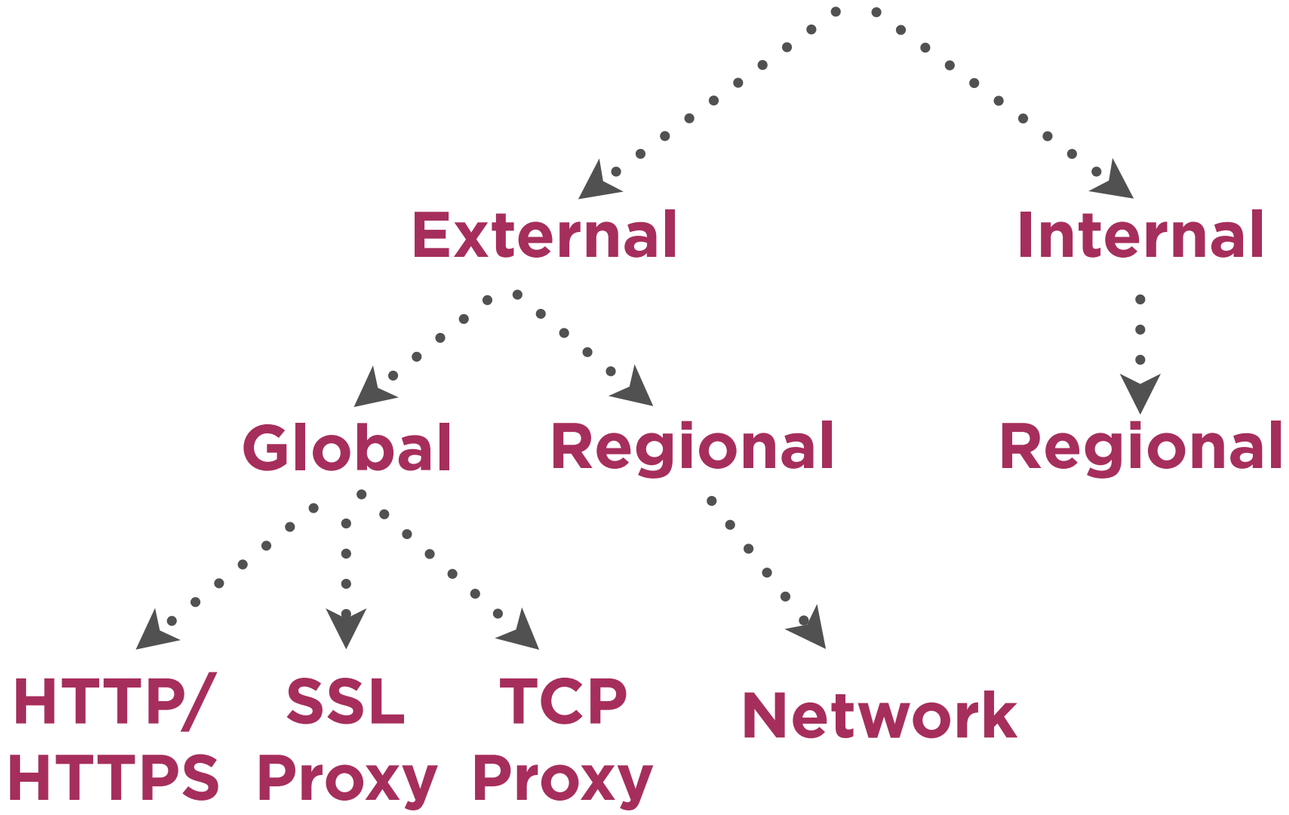
VPC traffic stays internal - less latency, more security

No public IP address needed

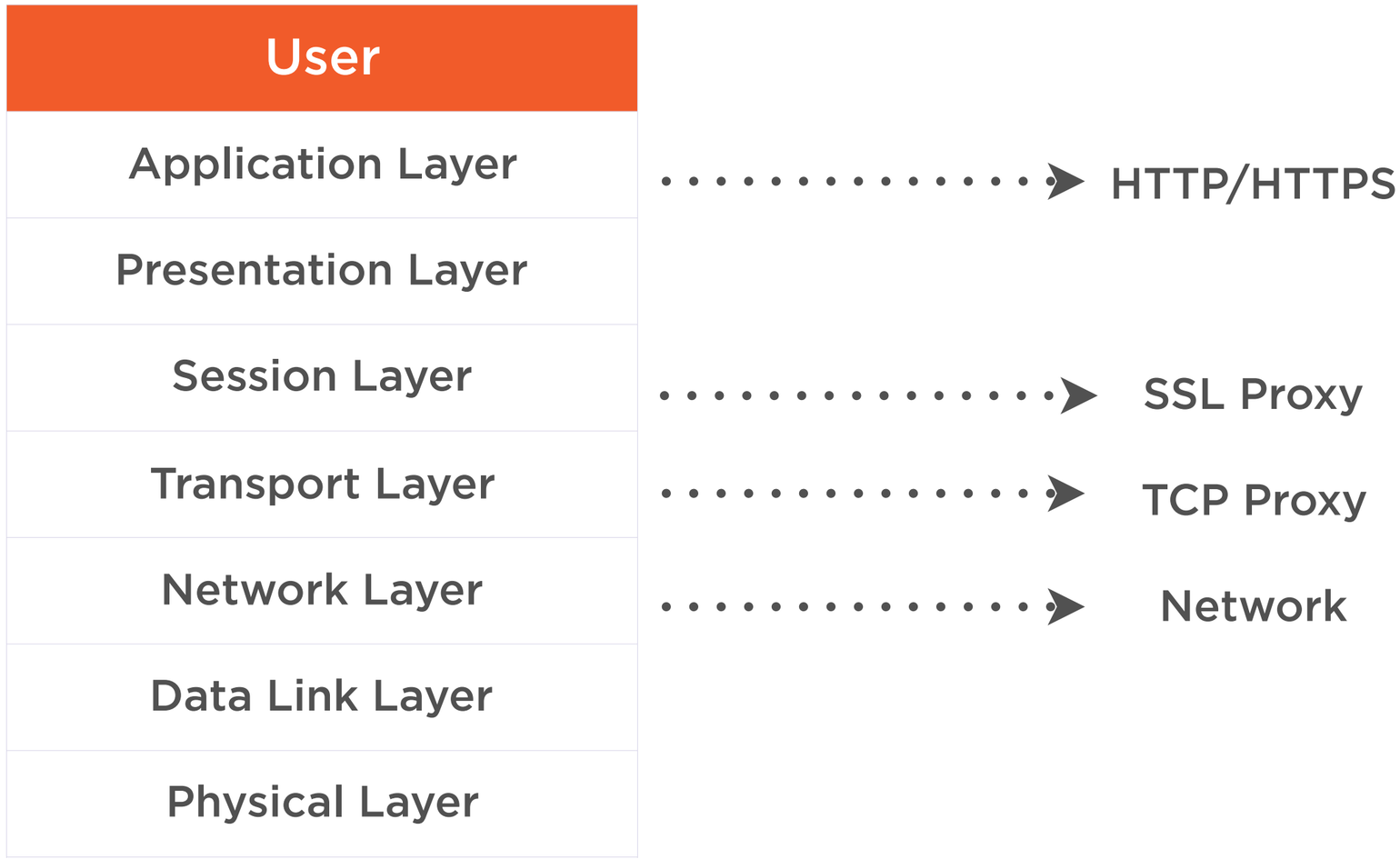
Useful to balance requests from your frontend to your backend instances

Choosing the Right Load Balancing Option

Load Balancing



OSI Network Stack



Which load balancer is the right one for you?

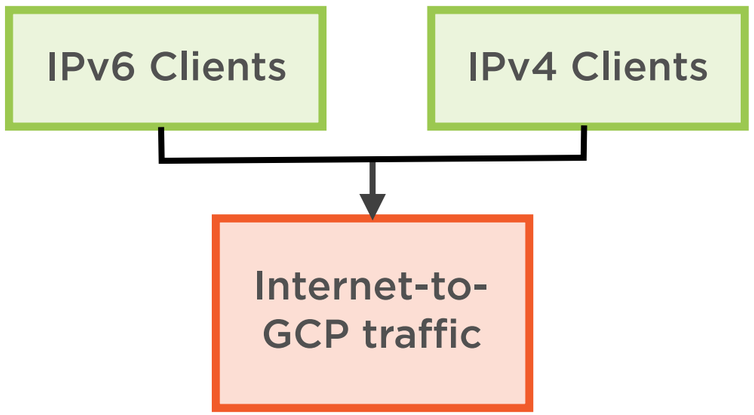
Which kind of load balancer is the right one for your case?

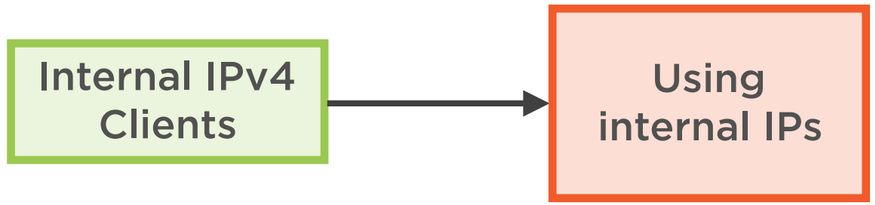
OSI Network Stack

User	
Application Layer	HTTP/HTTPS
Presentation Layer	
Session Layer	SSL Proxy
Transport Layer	TCP Proxy
Network Layer	Network
Data Link Layer	
Physical Layer	

Rule-of-thumb: Load balancer in the highest layer possible

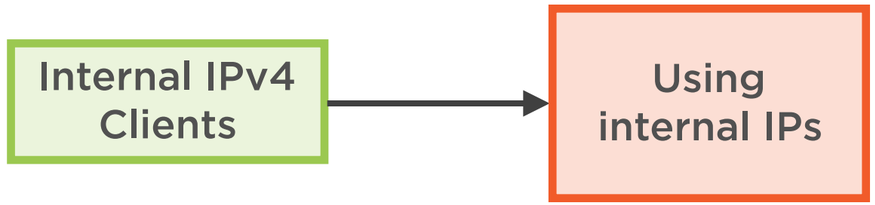
Choosing the Right Load Balancer





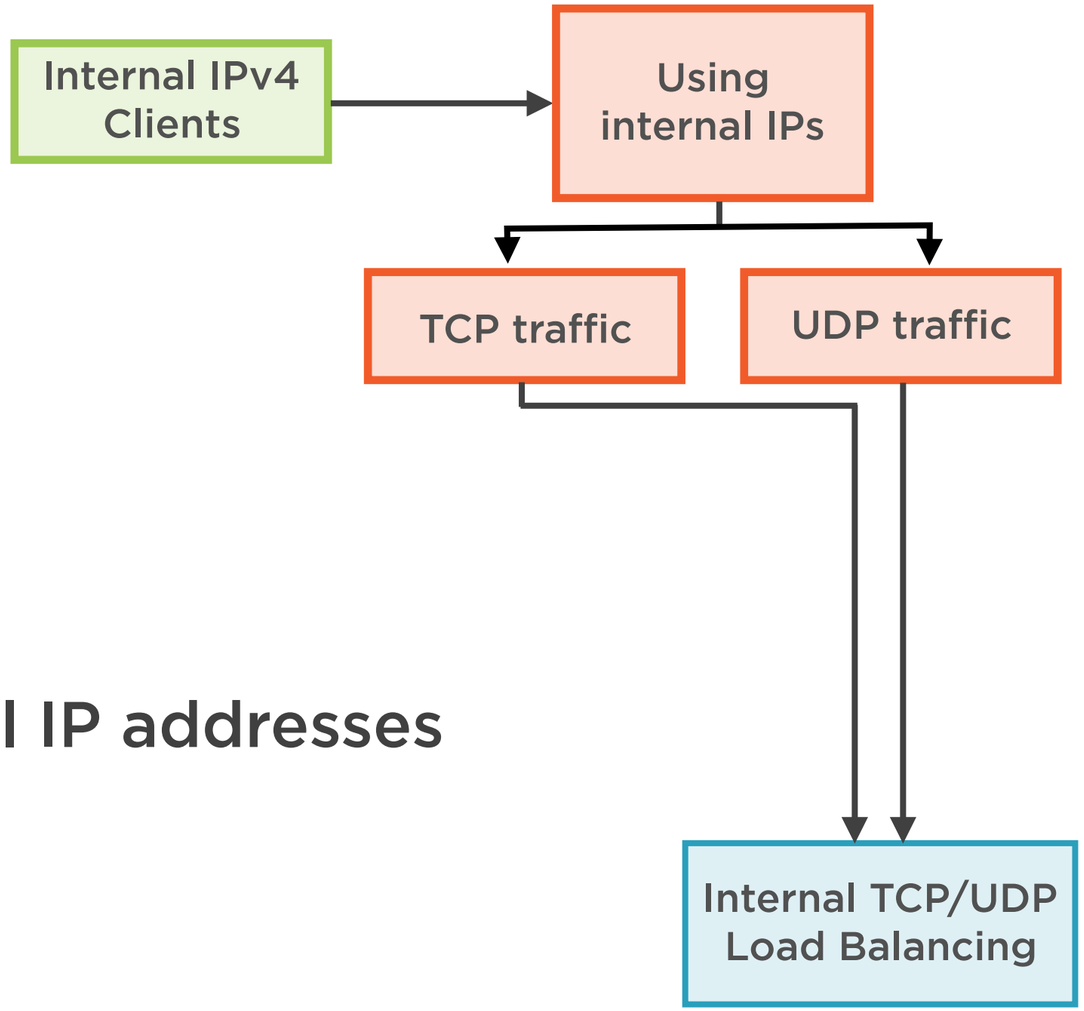
Where does the traffic to your network come from? External clients or from services and VMs which are on the same network?

Choosing the Right Load Balancer



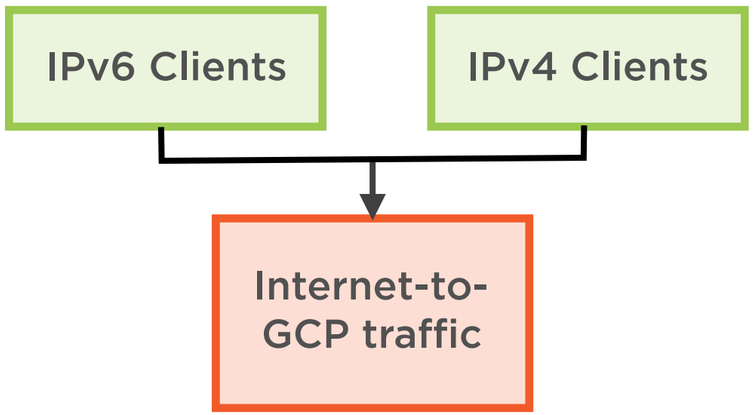
Where does the traffic to your network come from? External clients or from services and VMs which are on the same network?

Internal Load Balancing

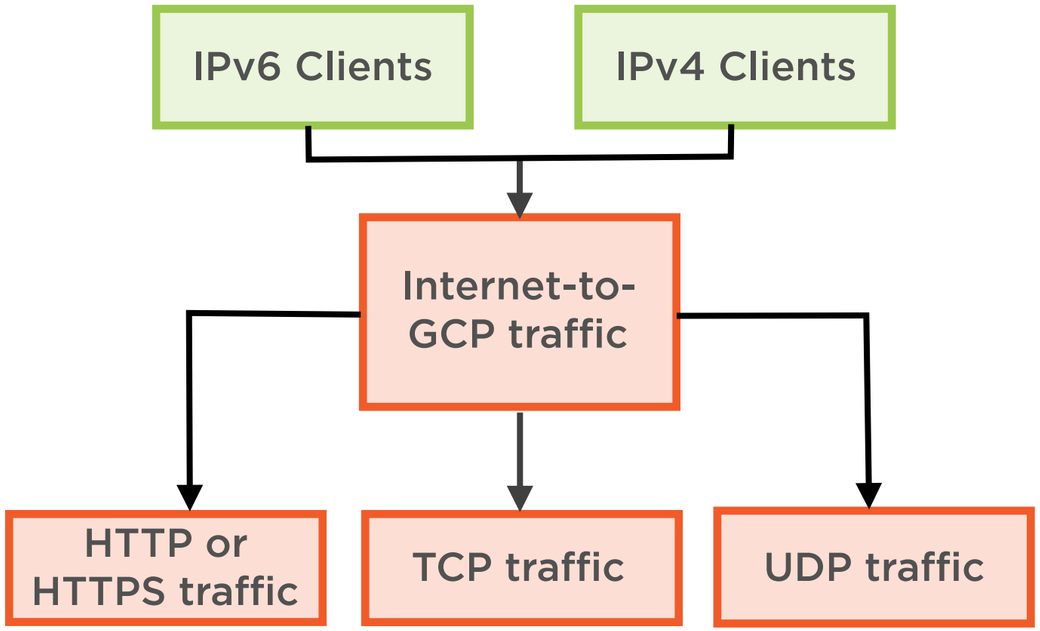


Works with internal IP addresses

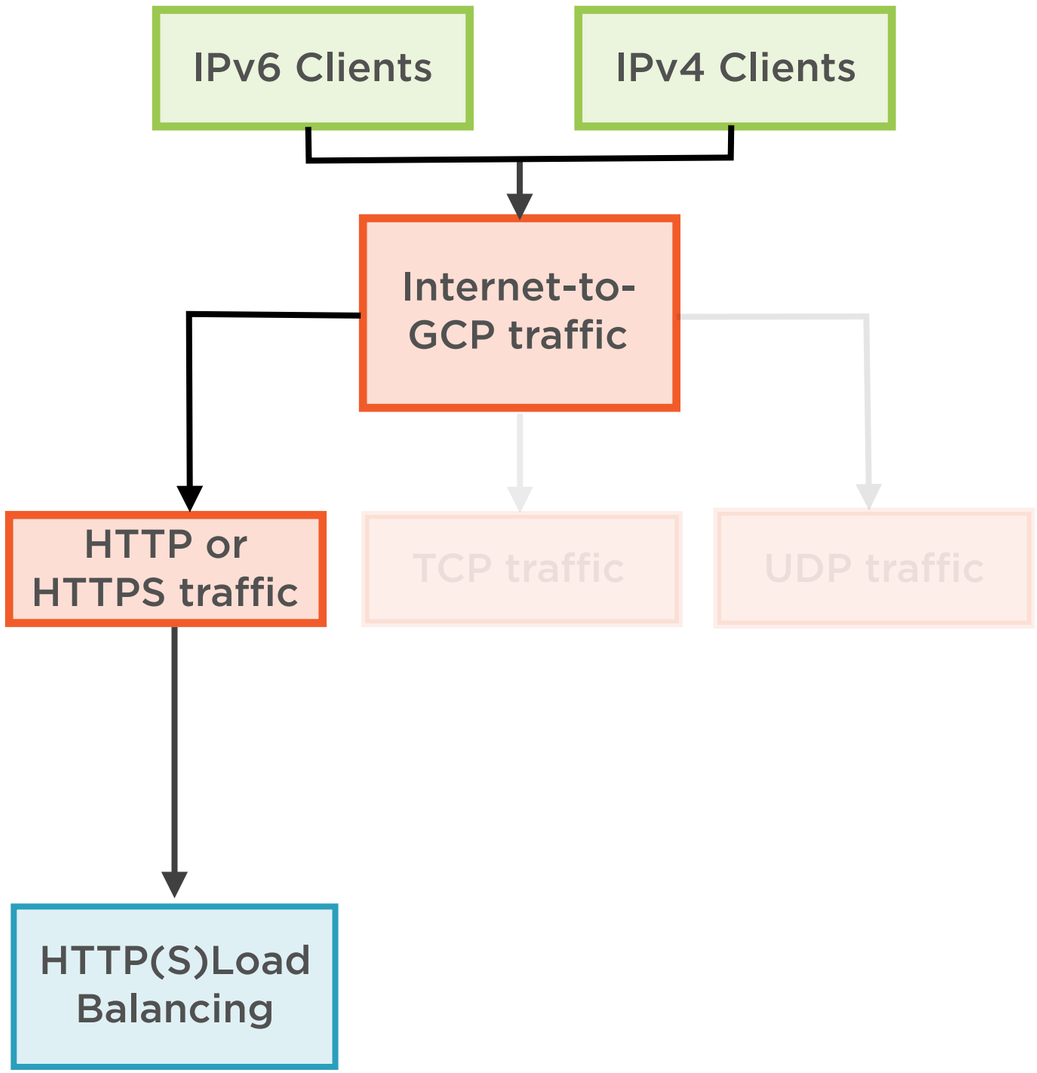
Choosing the Right Load Balancer



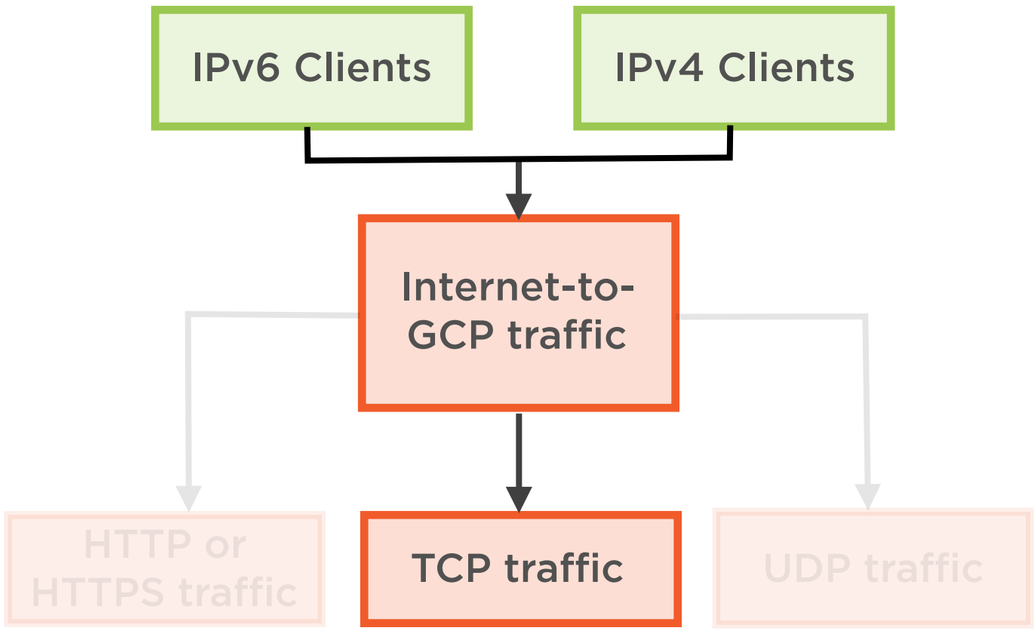
What Kind of External Traffic?



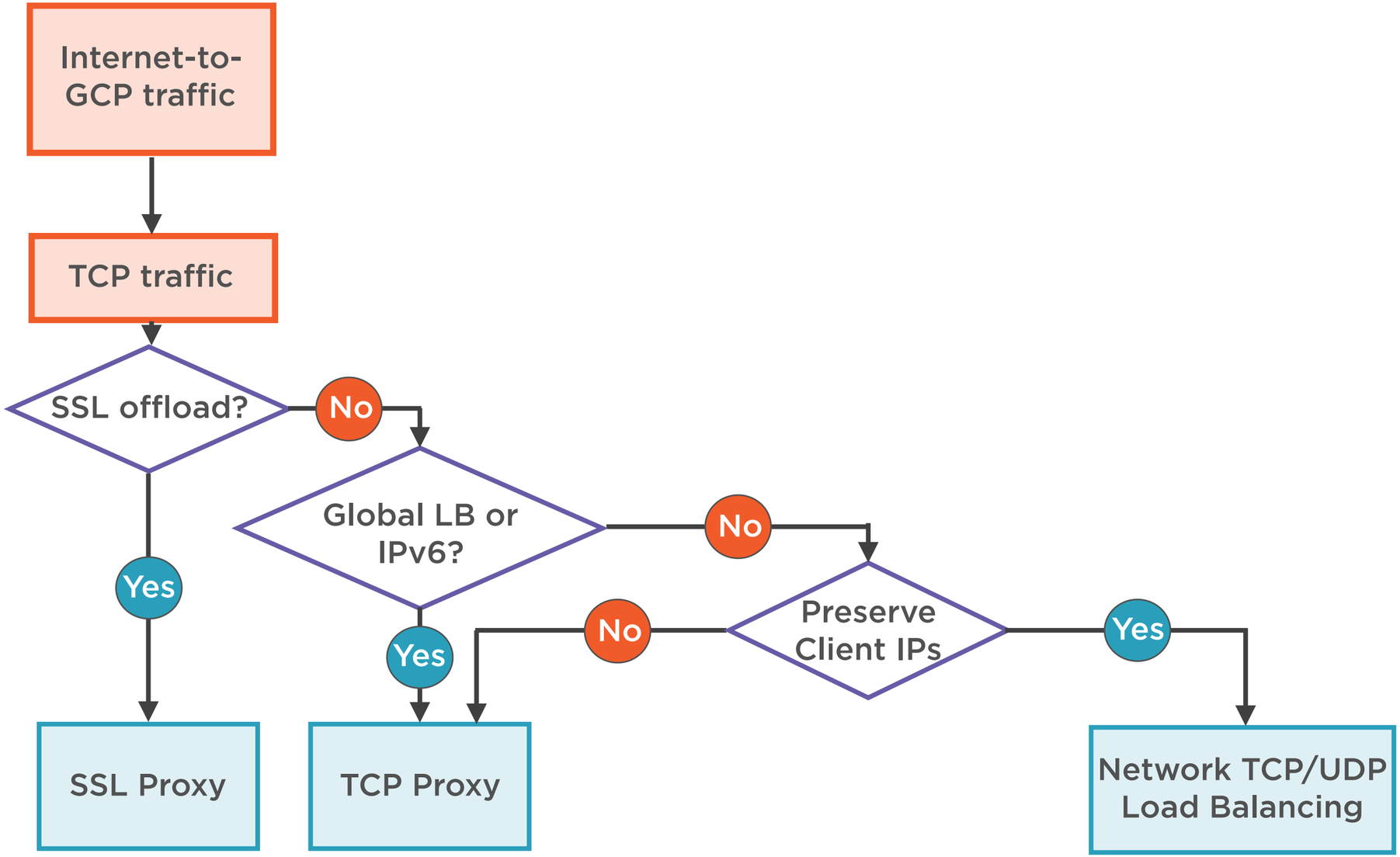
HTTP(S) Load Balancing



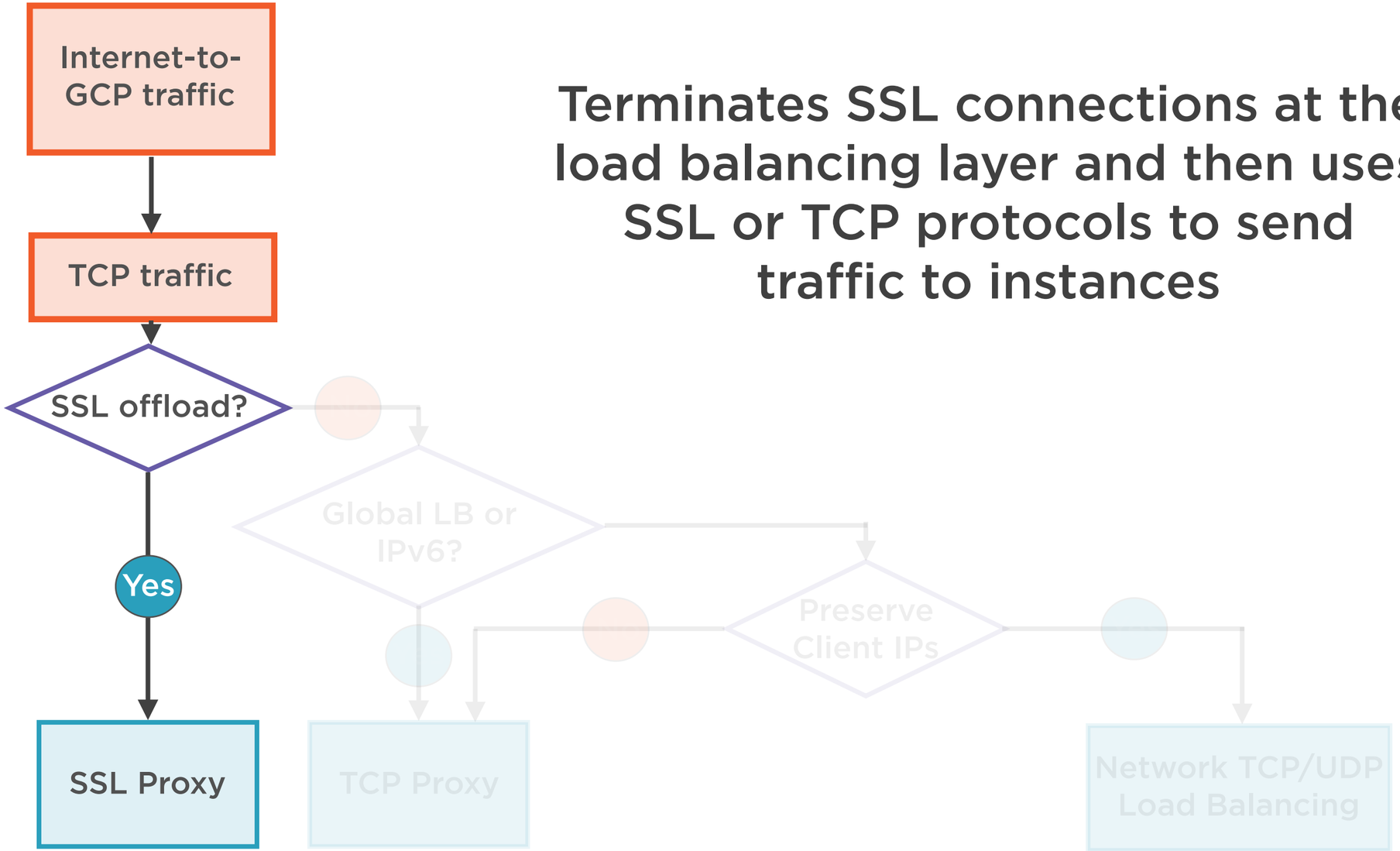
What Kind of External Traffic?



More Decisions

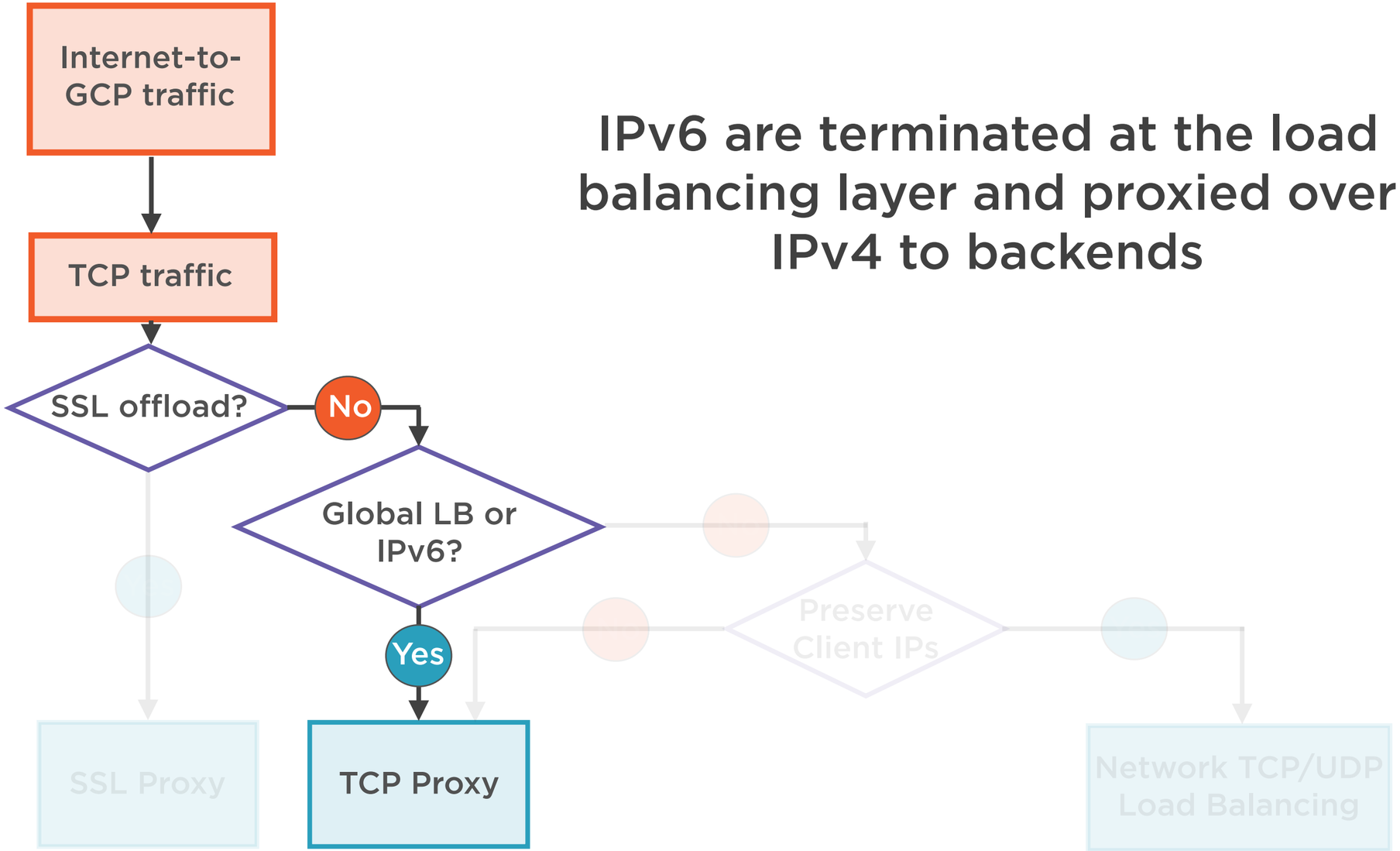


SSL Proxy



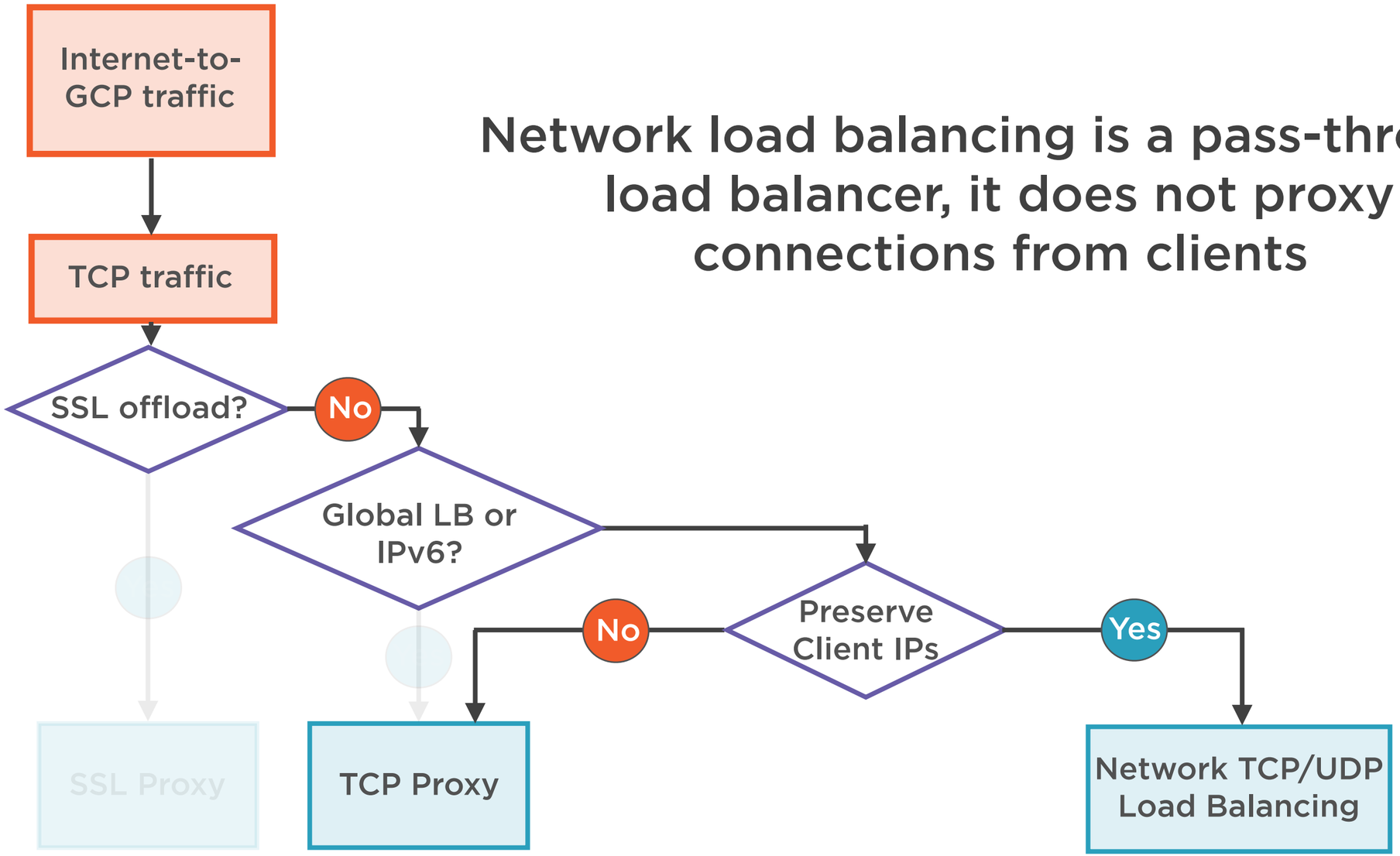
Terminates SSL connections at the load balancing layer and then uses SSL or TCP protocols to send traffic to instances

TCP Proxy



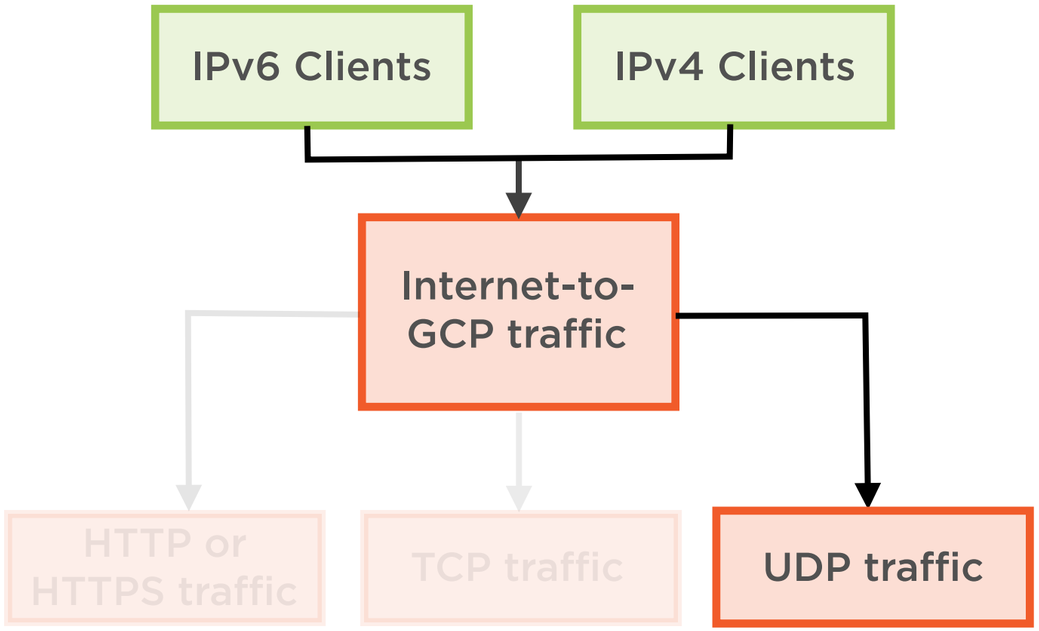
IPv6 are terminated at the load balancing layer and proxied over IPv4 to backends

TCP Proxy or Network Load Balancing



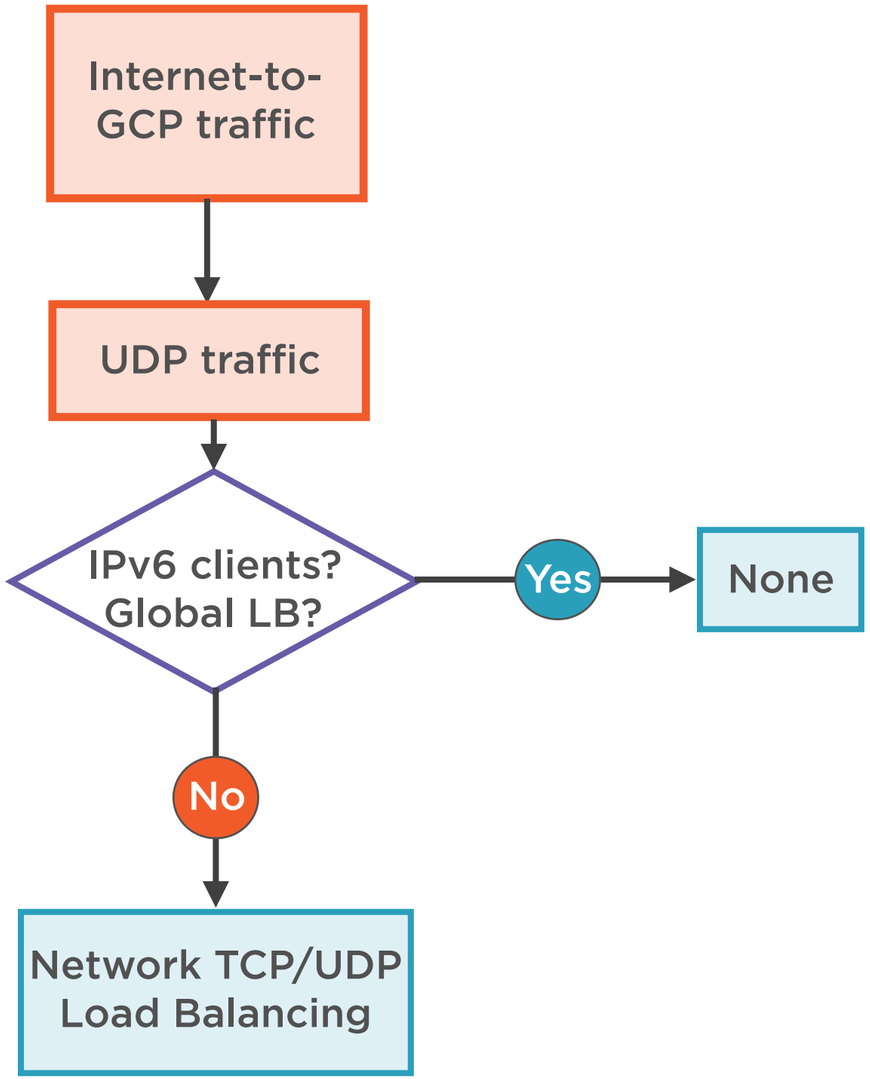
Network load balancing is a pass-through load balancer, it does not proxy connections from clients

What Kind of External Traffic?

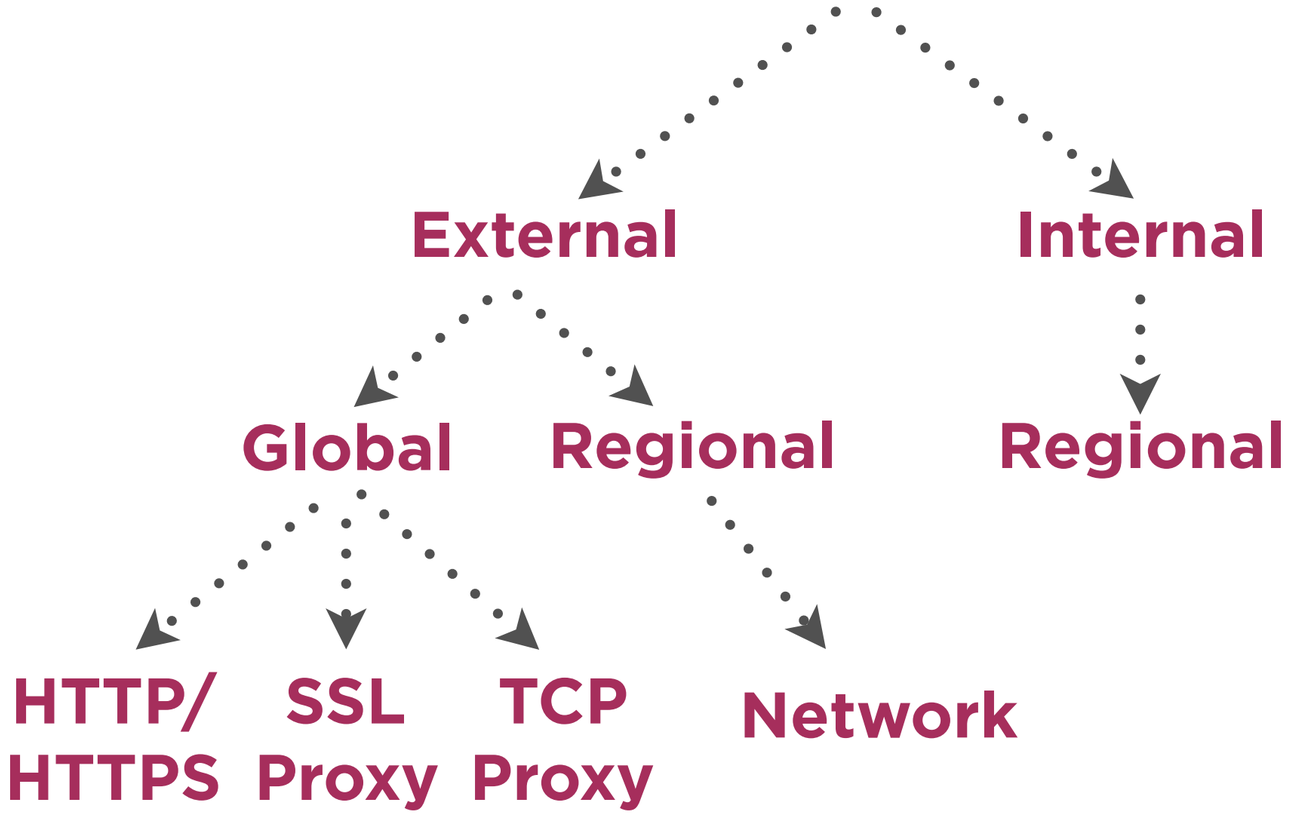


Network Load Balancing

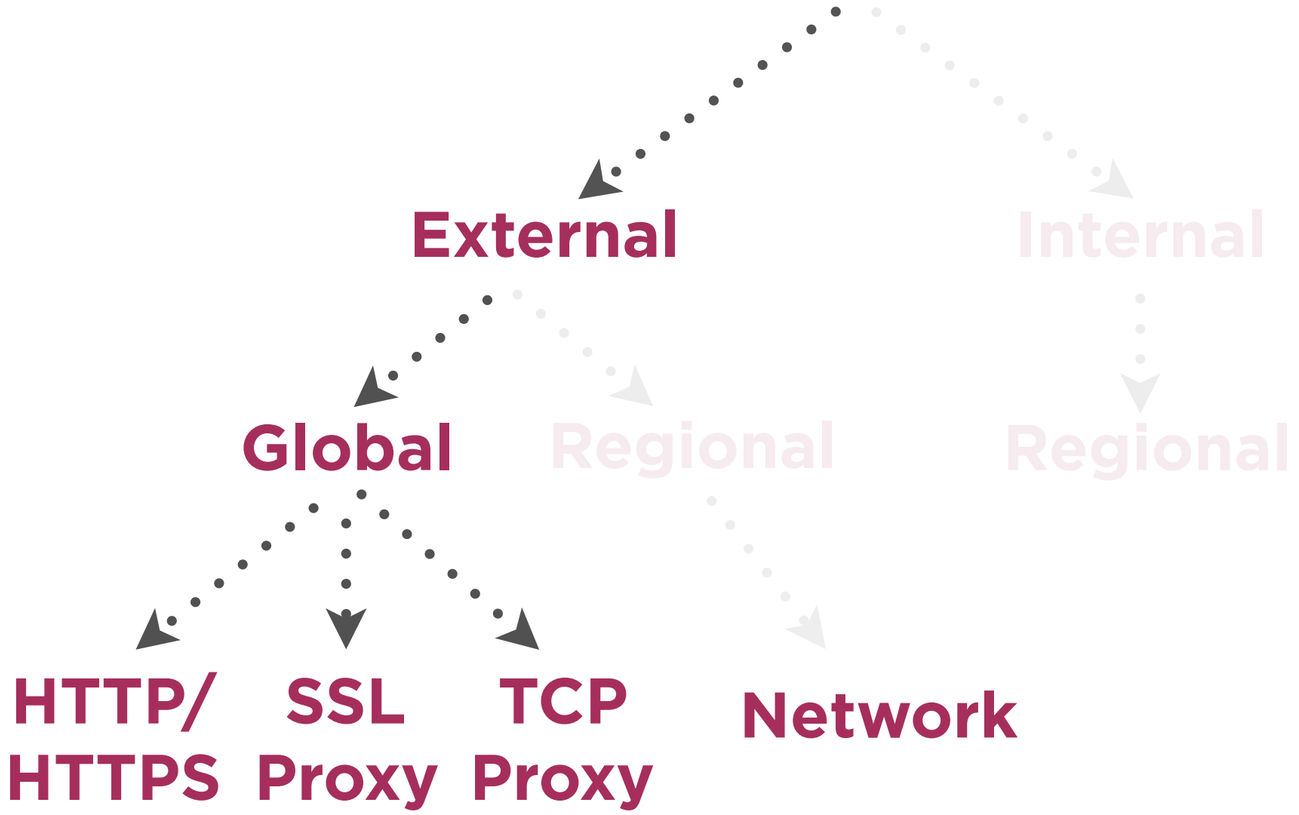
Network load balancing is a non-proxied load balancer for traffic not supported by other types of load balancers



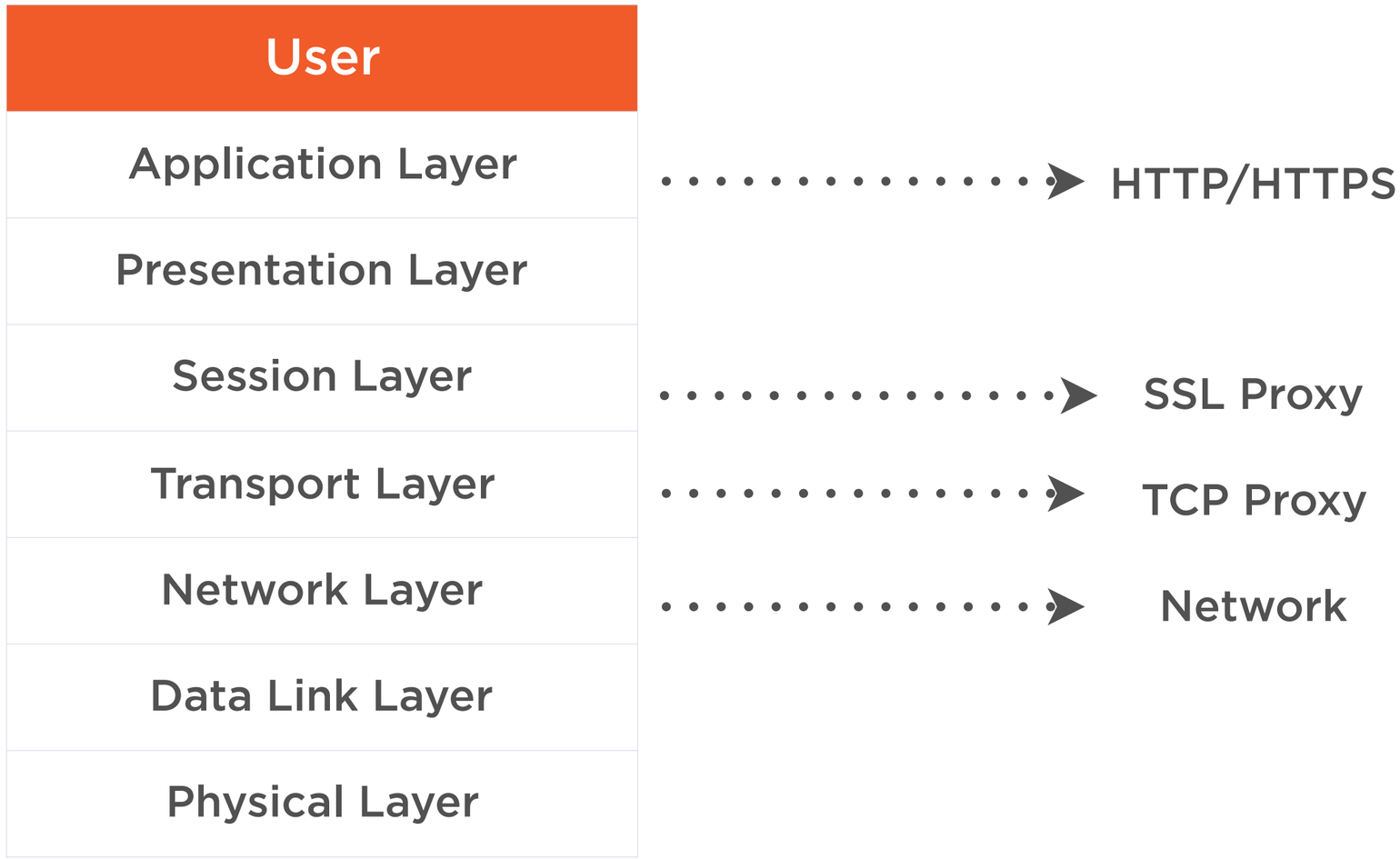
Load Balancing



External, Global



OSI Network Stack



OSI Network Stack

User	
Application Layer	HTTP/HTTPS
Presentation Layer	
Session Layer	SSL Proxy
Transport Layer	TCP Proxy
Network Layer	Network
Data Link Layer	
Physical Layer	

Rule-of-thumb: Load balancer in the highest layer possible

OSI Network Stack

User	
Application Layer	HTTP/HTTPS
Presentation Layer	
Session Layer	SSL Proxy
Transport Layer	TCP Proxy
Network Layer	Network
Data Link Layer	
Physical Layer	

HTTP(S) load balancing is the “smartest”

Pricing

Load Balancing and Forwarding Rules

Item	Price per Unit (USD)	Pricing Unit
First 5 forwarding rules	$0.025	Per Hour
Per additional forwarding rule	$0.010	Per Hour
Ingress data processed by load balancer	$0.008	Per GB

https://cloud.google.com/compute/pricing#lb

Egress Charges

Normal egress rates are charged for traffic outbound from a load balancer

There is no additional load balancer egress cost beyond normal egress rates

Load Balancing Pricing

Compute Engine charges for:

- Load balancing
- Forwarding rules

Forwarding rules charges

Pricing:

- 5 forwarding rules = $0.025/hour
- Each additional forwarding rule = $0.01/hour

$0.025/hour

Forwarding Rules Charges

Five forwarding rules

$0.025/hour for 5 rules + (5 additional rules * $0.01/hour) = $0.075/hour

---

Forwarding Rules Charges

Ten forwarding rules

Traffic Through External IP Addresses

Pricing:

- Egress between zones in the same region
- Egress between regions within the US
- Internet egress pricing.

Summary

Introducing load balancers on the GCP

Global and regional load balancers

External and internal load balancers

Types of load balancers: HTTP(S), SSL proxy, TCP proxy, network and internal

Choosing the right load balancer