Types of Cloud Load Balancing

Cloud load balancers can be divided up as follows:

* Global versus regional load balancing
* External versus internal load balancing
* Traffic type
* HTTPS, HTTP, or TCP/SSL
* Single anycast IP address
* Instances globally distributed
* Health checks
* IP address and cookie-based affinity
* IPv6 and IPv4 client termination
* Connection draining
* Autoscaling
* Monitoring and logging
* Load balancing for cloud storage
* Cross-region overflow and failover
* Requires Premium Tier of [Network Service Tiers](https://cloud.google.com/network-tiers/docs/overview)

### **External versus internal load balancing**

GCP's load balancers can be divided into external and internal load balancers. External load balancers distribute traffic coming from the internet to your GCP network. Internal load balancers distribute traffic within your GCP network.

### **Traffic type**

The type of traffic you need your load balancer to handle is another factor in determining which load balancer to use.

* HTTP and HTTPS traffic require global, external load balancing.
* TCP traffic can be handled by global, external load balancing; external, regional load balancing; or internal, regional load balancing.
* UDP traffic can be handled by external regional load balancing or internal regional load balancing.

SEND FEEDBACK

* [Load Balancing](https://cloud.google.com/load-balancing/)

* [Documentation](https://cloud.google.com/load-balancing/docs/)

# **Overview of Load Balancing**

This document provides you with an overview of the different load balancing solutions that are available on the Google Cloud Platform.

Google Cloud Platform Load Balancing gives you the ability to distribute load-balanced compute resources in single or multiple regions, to meet your high availability requirements, to put your resources behind a single anycast IP and to scale your resources up or down with intelligent Autoscaling. Cloud Load Balancing is fully integrated with Cloud CDN for optimal content delivery.

Using Cloud Load Balancing, you can serve content as close as possible to your users, on a system that can respond to over 1 million queries per second. Cloud Load Balancing is a fully distributed, software defined, managed service. It is not instance or device based, so you do not need to manage a physical load balancing infrastructure.

## Types of Cloud Load Balancing

Cloud load balancers can be divided up as follows:

* Global versus regional load balancing
* External versus internal load balancing
* Traffic type

The sections that follow describe these different types of load balancers.

### **Global versus regional load balancing**

Use global load balancing when your users and instances are globally distributed, your users need access to the same applications and content, and you want to provide access using a single anycast IP address. Global load balancing can also provide IPv6 termination.

Use regional load balancing when your users and instances are concentrated in one region and you only require IPv4 termination.

Global and regional load balancing types(click to enlarge)

Global load balancing requires that you use the Premium Tier of [Network Service Tiers](https://cloud.google.com/network-tiers/docs/overview). For regional load balancing, you can use Standard Tier.

#### **Global Load Balancing**

Google global load balancing is implemented entirely in software, done by Google Front Ends (GFEs). The GFEs are distributed globally and load balance traffic in sync with each other by working with Google’s other software-defined systems and global control plane. Your traffic is directed to a single anycast IP address. GFEs terminate your user traffic as close as possible to your users and direct load balanced traffic to the closest healthy backend that has capacity.

* HTTPS, HTTP, or TCP/SSL
* Single anycast IP address
* Instances globally distributed
* Health checks
* IP address and cookie-based affinity
* IPv6 and IPv4 client termination
* Connection draining
* Autoscaling
* Monitoring and logging
* Load balancing for cloud storage
* Cross-region overflow and failover
* Requires Premium Tier of [Network Service Tiers](https://cloud.google.com/network-tiers/docs/overview)

Global Load Balancing (click to enlarge)

#### **Regional Load Balancing**

Google regional load balancing is implemented entirely in software. Your instances are in a single GCP region and traffic is distributed to instances within a single region. Use network TCP/UDP Load Balancing to load balance external traffic. Use internal TCP/UDP Load Balancing to load balance internal traffic.

* Internal TCP/UDP Load Balancing
* UDP or TCP/SSL traffic
* Instances in one region
* Single IP address per region
* Health checks
* Session affinity
* IPv4 only
* Autoscaling
* Standard Tier of [Network Service Tiers](https://cloud.google.com/network-tiers/docs/overview)

The following illustration shows regional load balancing.

Regional Load Balancing (click to enlarge)

### **External versus internal load balancing**

GCP's load balancers can be divided into external and internal load balancers. External load balancers distribute traffic coming from the internet to your GCP network. Internal load balancers distribute traffic within your GCP network.

External and internal load balancing types (click to enlarge)

The following diagram illustrates a common use case: how to use external and internal load balancing together. In the illustration, traffic from users in San Francisco, Iowa, and Singapore is directed to an external load balancer, which distributes that traffic to different regions in a GCP network. An internal load balancer then distributes traffic between the us-central-1a and us-central-1b zones.

How external and internal load balancing work together (click to enlarge)

#### **External load balancing**

Use external load balancing when you need to distribute traffic from the Internet to a GCP network.

GCP external load balancing offers the following:

* HTTP or HTTPS traffic: global HTTP(S) Load Balancing
* TCP traffic with SSL offload: global SSL Proxy Load Balancing
* TCP traffic without SSL offload: global TCP Proxy Load Balancing
* UDP traffic: regional Network TCP/UDP Load Balancing
* IPv4 or IPv6 clients
* Global or regional load balancing

Global load balancing requires that you use the Premium Tier of [Network Service Tiers](https://cloud.google.com/network-tiers/docs/overview). For regional load balancing, you can use Standard Tier.

#### **Internal Load Balancing**

Use Internal Load Balancing when you need to distribute traffic to instances within a GCP network.

GCP Internal Load Balancing offers the following:

* TCP or UDP traffic
* RFC 1918 load balancing
* Client IP address is preserved
* Health checks
* Autoscaling without prewarming
* Session affinity
* Regional load balancing

### **Traffic type**

The type of traffic you need your load balancer to handle is another factor in determining which load balancer to use.

* HTTP and HTTPS traffic require global, external load balancing.
* TCP traffic can be handled by global, external load balancing; external, regional load balancing; or internal, regional load balancing.
* UDP traffic can be handled by external regional load balancing or internal regional load balancing.

## A closer look at Cloud load balancers

This section provides information on each type of GCP load balancer, including links to documentation.

### **HTTP(S) Load Balancing**

HTTP(S) Load Balancing provides the following benefits:

* Global load balancing: Your applications are available to your customers at a single global IP address, which simplifies your DNS setup. HTTP(S) Load Balancing balances HTTP and HTTPS traffic across multiple backend instances and across multiple regions.
* HTTP or HTTPS traffic
* IPv4 or IPv6 clients
* HTTP requests are load balanced on port 80 or 8080.
* HTTPS requests are load balanced on port 443.
* Autoscaling: HTTP(S) Load Balancing is scalable, requires no pre-warming, and enables content-based and cross-region load balancing.
* URL maps direct requests based on rules: You can configure URL maps that route some URLs to one set of instances and route other URLs to other instances. Requests are generally routed to the instance group that is closest to the user. If the closest instance group does not have sufficient capacity, the request is sent to the next closest instance group that does have capacity.

### **HTTP Load Balancing**

A complete HTTP load balancer is structured as follows:

1. A [global forwarding rule](https://cloud.google.com/load-balancing/docs/https/global-forwarding-rules) directs incoming requests to a [target HTTP proxy](https://cloud.google.com/load-balancing/docs/target-proxies).

The global forwarding rule provides a single global IPv4 or IPv6 address that you can use in DNS records for your site. You can use more than one forwarding rule with a given proxy, so you can have one for IPv4 traffic and another for IPv6 traffic.

Target proxies are referenced by one or more [global forwarding rules](https://cloud.google.com/load-balancing/docs/https/global-forwarding-rules). In the case of [HTTP(S)](https://cloud.google.com/load-balancing/docs/https) load balancing, proxies route incoming requests to a [URL map](https://cloud.google.com/load-balancing/docs/https/url-map). In the case of [SSL proxy](https://cloud.google.com/load-balancing/docs/ssl/) and [TCP proxy](https://cloud.google.com/load-balancing/docs/tcp/) load balancing, target proxies route incoming requests directly to [backend services](https://cloud.google.com/load-balancing/docs/backend-service).

Global forwarding rules route traffic by IP address, port, and protocol to a load balancing target proxy, which in turn forwards the traffic to an [instance group](https://cloud.google.com/compute/docs/instance-groups) containing your virtual machine instances.