

Google Cloud Load Balancing

Google Cloud Load Balancing can be used to load balance user requests among sets of instances.

Name	Internet	Internal	Single Region	Multi-Region
HTTP(S)	Yes	No	Yes	Yes
SSL Proxy	Yes	No	No	Yes
TCP Proxy	Yes	No	No	Yes
TCP LB	Yes	Yes	Yes	Use TCP Proxy
UDP LB	Yes	Yes	Yes	No

Features

- Global external load balancing.
 - HTTP(S) load balancing.
 - SSL Proxy load balancing.
 - TCP Proxy load balancing.
- Regional external load balancing (TCP/UDP).
- Regional internal load balancing (TCP/UDP).
- HTTP(S) will terminate on the load balancer (central SSL certificates).
- Can use TCP load balancer for HTTP(S) however it will terminate on the instance.
- Integrates with Managed Instance Groups.

Forwarding Rules

- Consist of name, region, IP address, protocol, ports, target-pool or target-instance.
- Managed with GCP Console, gcloud, or REST API.

Target Pools

- Max of 50 per project
- Instances can be in different zones within a region.
- SessionAffinity influences load distribution.
 - NONE: hash of source IP, source port, protocol, dest IP, dest port.
 - CLIENT_IP_PROT: specific protocols from the client end up on a single instance.
 - CLIENT_IP: all connections from the client end up on the same instance.

HTTP(S) Load Balancing

- Distributes traffic base on proximity to the user or URL or both.
- Supports multiple regions.
- Supports autoscalers.

HTTP(S) Load Balancing Backend

- Comprised of:
 - A health check.
 - Session affinity settings.
 - One or more backend services.
- Consists of:
 - An instance group (managed or unmanaged).

- A balancing mode (CPU or Rate in req/sec).
 - A capacity scaler (ceiling % or CPU/Rate targets).
- Supports up to 500 endpoints per zone.
- Supports URL mapping to service instances.
- Blocks illegal requests adding security.
- Writes request information into Stackdriver logs.

Managed Instance Groups

- Deploys identical instances based on an instance template.
- Instance group can be resized.
- Manager ensures all instances are in a running state.
- Used with autoscaler.
- Single zone or regional.
- Support connection draining:
 - Delays termination of an instance until existing connections are closed or timeout (1 to 3600 sec).
 - New connections are prevented.
- Triggered when an instance is removed from a group either by manual removal, resizing, autoscaling, etc.
- Instance Group Updater:
 - Zero downtime and staggered releases.
 - Use Instance Group Updater to apply a rolling update.
 - Apply canary updates with rollback.
- Autohealing:
 - Automated server monitoring and restarts.
 - If health check sees a failed service, re-create the instance.
- Balances instances across three zones in a region.

Regional Managed Instance Groups Best Practices

- Overprovision your services.
- Test using the "failure__zone" special tag.

Securing Load Balanced Servers

- A firewall rule must be created for the appropriate networks.
- Create firewall rules to only allow traffic from the GCP Load Balancer networks (130.211.0.0/22).
- Have no external IP addresses (use a bastion host for management purposes).