Load balancing and scaling

Google Cloud offers load balancing and autoscaling for groups of instances.

Load balancing

Google Cloud offers server-side load balancing so you can distribute incoming traffic across multiple virtual machine (VM) instances. Load balancing provides the following benefits:

- Scale your app
- Support heavy traffic
- Detect and automatically remove unhealthy VM instances using health checks. Instances that become healthy again are automatically re-added.
- Route traffic to the closest virtual machine

Google Cloud load balancing uses forwarding rule resources to match certain types of traffic and forward it to a load balancer. For example, a forwarding rule can match TCP traffic destined to port 80 on IP address `192.0.2.1`, then forward it to a load balancer, which then directs it to healthy VM instances.

Google Cloud load balancing is a managed service, which means its components are redundant and highly available. If a load balancing component fails, it is restarted or replaced automatically and immediately.

Google Cloud offers several different types of load balancing that differ in capabilities, usage scenarios, and how you configure them. See Google Cloud load balancing documentation for descriptions.

Autoscaling

Compute Engine offers autoscaling to automatically add or remove VM instances from a managed instance group based on increases or decreases in load. Autoscaling lets your apps gracefully handle increases in traffic, and it reduces cost when the need for resources is lower. After you define the autoscaling policy, the autoscaler performs automatic scaling based on the measured load.

Policies

When you create an autoscaler, you must specify at least one autoscaling policy. You can choose a policy based on CPU utilization, load balancing serving capacity, or Cloud Monitoring metrics. If you use multiple policies, the autoscaler scales an instance group based on the policy that provides the largest number of VM instances in the group.

The following sections discuss the autoscaling policies in general. For more information about how to set up a specific autoscaling policy, see the respective policy documentation.

CPU utilization

CPU utilization is the most basic autoscaling that you can perform. This policy tells the autoscaler to watch the average CPU utilization of a group of VM instances and add or remove instances from the group to maintain your desired utilization. This is useful for configurations that are CPU intensive but might fluctuate in CPU usage.

For more information, see Scaling based on CPU utilization.

Load balancing serving capacity

When you set up an autoscaler to scale based on load balancing serving capacity, the autoscaler watches the serving capacity of an instance group and scales when the VM instances are over or under capacity.

The serving capacity of an instance can be defined in the load balancer's backend service and can be based on either utilization or requests per second.

For more information, see Scaling based on the serving capacity of an external HTTP(S) load balancer.

Monitoring metrics

If you export or use Cloud Monitoring metrics, you can set up autoscaling to collect data of a specific metric and perform scaling based on your desired utilization level. You can scale based on standard metrics provided by Monitoring or by using any custom metrics you create.

For more information, see Scaling based on Monitoring metrics.

What's next

- Learn more about instance groups.
- Learn how to autoscale managed instance groups based on:
  - CPU utilization
  - The serving capacity of an external HTTP(S) load balancer
  - Monitoring metrics
- Learn how to choose a load balancer and
  - How to set up HTTP(S) load balancing
  - How to set up network load balancing