Resource quotas

Send feedback

Compute Engine enforces quotas on resource usage to prevent abuse and accidental usage, and to protect users from undesirable effects of other accounts. For example, quotas protect the community of Google Cloud users by preventing unforeseen spikes in usage. Google Cloud also offers free trial quotas that provide limited access for projects to help you explore Google Cloud on a free trial basis.

If you expect a notable upcoming increase in usage, you can proactively request quota adjustments from the **Quotas** page in the Cloud Console.

Important: If your project's billing service is disrupted, your quota is reset to its default value.

Checking your quota

Each project's quota is different. To check the available quota for a project, go to the **Quotas** page in the Google Cloud Console.

In the `gcloud` command-line tool, run the following command to check project-wide quotas. Replace `myproject` with your own project ID:

```
gcloud compute project-info describe --project myproject
```

Note that the results don't list per-region quotas. To list quotas in a region run this command, where `[REGION]` is the region where you want to list quota information:

```
gcloud compute regions describe [REGION]
```

Requesting an increase in quota

Request changes to quota from the **Quotas** page in the Cloud Console. There is no charge for requesting an increase in quota. Your costs increase only if you use more resources.

Permission for editing quota

To change your quotas, you must have `serviceusage.quotas.update` permission. This permission is included by default for the following predefined roles: Owner, Editor, and Quota Administrator.

Request a change to quota

1. Go to the **Quotas** page.

   Go to the Quotas page

2. In the **Quotas** page, select the quotas you want to change.
3. Click the **Edit Quotas** button on the top of the page.
4. Check the box of the service you want to edit.
5. Fill out your name, email, and phone number, and click **Next**.

6. Enter your request to increase your quota, and click **Next**.
7. Submit your request.
8. A request to decrease quota is rejected by default. If you must reduce your quota, reply to the support email with an explanation of your requirements. A support representative from the Compute Engine team will respond to your request within 24 to 48 hours.

Plan and request additional resources at least a few days in advance to ensure that there is enough time to fulfill your request.

Quotas and resource availability

Resource quotas are the maximum number of resources you can create of that resource type, if those resources are available. Quotas do not guarantee that resources will be available at all times. If a resource is not available, or if the region you choose is out of the resource, you will not be able to create new resources of that type, even if you have remaining quota in your region or project. For example, you might still have quota to create external IP addresses in `us-central1`, but there might not be available IP addresses in that region.

Similarly, even if you have a regional quota, it is possible that a resource might not be available in a specific zone. For example, you might have quota in region `us-central1` to create VM instances, but you might not be able to create VM instances in the zone `us-central1-a` if the zone is depleted. In such cases, try creating the same resource in another zone, such as `us-central1-f`. To learn more about your options if zonal resources are depleted, see General troubleshooting.

Understanding quotas

When planning your virtual machine (VM) instance needs, there are a number of quotas to consider that affect how many VM instances you can create.

Regional and global quotas

VM quotas are managed at the regional level. VM instance, instance group, CPU, and Disk quotas can be consumed by any VM in the region, regardless of zone. For example, CPU quota is a regional quota, so there is a different limit and

usage count for each region. To launch an `n1-standard-16` instance in any zone in the `us-central1` region, you need enough quota for at least 16 CPUs in `us-central1`.

Networking and load balancing quotas are required to create firewalls, load balancers, networks, and VPNs. These are global quotas that do not depend on a region. Any region can use a global quota. For example, in-use and static external IP addresses assigned to load balancers and HTTP(S) proxies consume global quotas.

CPUs

CPU quota is the total number of virtual CPUs across all of your VM instances in a region. CPU quotas apply to running instances and instance reservations. Both normal and preemptible instances consume this quota.

To protect Compute Engine systems and other users, some new accounts and projects also have a global `CPUs (All Regions)` quota that applies to all regions and is measured as a sum of all your vCPUs in all regions.

For example, if you have 48 vCPUs remaining in a single region such as `us-central1` but only 32 vCPUs remaining for the `CPUs (All Regions)` quota, you can launch only 32 vCPUs in the `us-central1` region, even though there is remaining quota in the region. This is because you will reach the `CPU (All Regions)` quota and need to delete existing instances before you can launch new instances.

GPUs

Similar to virtual CPU quota, GPU quota refers to the total number of virtual GPUs in all VM instances in a region. Check the quotas page to ensure that you have enough GPUs available in your project, and to request a quota increase. In addition, new accounts and projects have a global GPU quota that applies to all regions.

When you request a GPU quota, you must request a quota for the GPU models that you want to create in each region, and an additional global quota for the total number of GPUs of all types in all zones.

VM instances

The VM instances quota is a regional quota and limits the number of VM instances that can exist in a given region, regardless of whether the VM is running or not. This quota is *not* visible in the Google Cloud Console, but it is set automatically by Compute Engine to be 10x your regular CPU quota. You do not need to request this quota. If you need quota for more VM instances, request more CPUs because having more CPUs increases this quota. The quota applies to both running and non-running VMs, and to normal and preemptible instances.

Quotas for preemptible resources

To use preemptible CPUs or GPUs attached to preemptible VM instances, or to use local SSDs attached to preemptible VM instances, you must have available quota in your project for the respective resource.

You can request special preemptible quotas for: `Preemptible CPUs`, `Preemptible GPUs`, or `Preemptible Local SSDs (GB)`. However, if your project does not have preemptible quota, you can still use the regular quota to launch preemptible resources.

After Compute Engine grants you preemptible quota in a region, all preemptible instances automatically count against preemptible quota.

Disk quotas

The following persistent disk and local SSD quotas apply on a per-region basis:

- `Local SSD (GB)`. This is the total combined size of local SSD disk partitions that can be attached to VMs in a region. Local SSD is a fast, ephemeral disk that should be used for scratch, local cache, or processing jobs with high fault tolerance because the disk is not intended to survive VM instance reboots. Local SSD partitions are sold in increments of 375 GB and up to eight local SSD partitions can be attached to a single VM. In the `gcloud` tool and the API, this is referred to as `LOCAL_SSD_TOTAL_GB`.
- `Persistent disk standard (GB)`. This is the total size of standard persistent disks that can be created in a region. As described in Optimizing persistent disk and local SSD performance, standard persistent disks offer lower IOPS and throughput than SSD persistent disks or local SSD. It is cost effective when used as large durable disks for storage, as boot disks, and for serial write processes like logs. Standard persistent disks are durable and are available indefinitely to attach to a VM within the same zone. In the `gcloud` tool and the API, this is referred to as `DISKS_TOTAL_GB`. This quota also applies to regional standard persistent disks, but regional disks consume twice the amount of quota per GB due to replication in two zones within a region.
- `Persistent disk SSD (GB)`. This is the total combined size of SSD persistent disks partitions that can be created in a region. SSD persistent disks have multiple replicas and, as described in Optimizing persistent disk and local SSD performance, offers higher IOPS and throughput than standard persistent disks. SSD persistent disks are cost effective for durable storage with high I/O requirements. SSD persistent disks are available indefinitely to attach to a VM within the same zone. In the `gcloud` tool and the API, this is referred to as `SSD_TOTAL_GB`. This quota is separate from local SSD. This quota also applies to regional SSD persistent disks, but regional disks consume twice the amount of quota per GB due to replication in two zones within a region.

IP addresses

You must have enough IP addresses for every VM that needs to be reachable from the public internet. Regional IP quota is for assigning IPv4 addresses to VMs in that region. Global IP quota is for assigning IPv4 addresses to global networking resources such as HTTP proxies and load balancers. Google Cloud offers different types of IP addresses, depending on your needs. For information about costs, refer to External IP address pricing.

- **In-use IP address**. Includes both ephemeral and static IP addresses that are currently being used by a resource.
- **Static External IP addresses**: External IP addresses reserved for your resources that persist through machine restarts. You can register these addresses with DNS and domain provider services to provide a user-friendly address. For example, www.example-site.com.
- **Static Internal IP addresses:** Static internal IP addresses provide the ability to reserve internal IP addresses from the internal IP range configured in the subnet. You can assign those reserved internal addresses to resources as needed.

Instance groups

To use instance groups, you must have available quota for all of the resources that the group will use (for example, CPU quota) as well as available quota for the group resource itself. Depending on the type of group that you create, the following group resource quotas apply:

Service type	Service quota
Regional (multi-zone) managed instance group	`Regional instance group managers`
Zonal (single-zone) managed instance group	Both of: • `Instance group managers` • `Instance groups`
Unmanaged (single-zone) instance group	`Instance groups`
Regional (multi-zone) autoscaler	`Regional autoscalers`
Zonal (single-zone) autoscaler	`Autoscalers`

Was this page helpful?

☆ ☆ ☆ ☆ ☆

Send feedback

Last updated 2020-08-04 UTC.