

Allocation quotas

This document lists the *allocation quotas* that apply to Compute Engine.

Allocation quotas

Allocation quotas, also known as resource quotas, define the number of resources that your project has access to. Compute Engine enforces allocation quotas on resource usage for various reasons. For example, quotas help to protect the community of Google Cloud users by preventing unforeseen spikes in usage. Google Cloud also offers [free trial quotas](/free-trial/docs/free-trial-quotas) (/free-trial/docs/free-trial-quotas) that provide limited access for projects to help you explore Google Cloud on a free trial basis.

Not all projects have the same quotas. As you increasingly use Google Cloud over time, your quotas might increase accordingly. If you expect a notable upcoming increase in usage, you can proactively request quota adjustments from the [Quotas](https://console.cloud.google.com/iam-admin/quotas) (https://console.cloud.google.com/iam-admin/quotas) page in the Google Cloud console.

For information specific to quotas for rate limits for the Compute Engine API, see [API quota](/compute/api-quota) (/compute/api-quota).

Important:

- If you're using the [Google Cloud free trial](/free/docs/gcp-free-tier#free-trial) (/free/docs/gcp-free-tier#free-trial), you cannot request a change to your quota.
- If your project's billing service is disrupted or if you change your project's billing account, your quotas reset to their default values.

Quotas and resource availability

Allocation 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 are always available. If a resource is not available, or if the region you choose is out of the resource, you can't 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, a resource might not be available in a specific zone. For example, you might have quota to create VM instances in region `us-central1`, 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 the documentation for [troubleshooting resource availability](#).

(/compute/docs/troubleshooting/troubleshooting-vm-creation#resource_availability).

Allocation quotas

When planning your VM instance needs, you should consider several quotas 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, disk quotas, and CPU 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 `n2-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 quotas 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.

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. This quota is visible in the Google Cloud console on the **Quotas** page. Compute Engine automatically sets this quota to be 10 times your regular CPU quota. You do not need to request this quota. If you need quota for more VM instances, request more [CPUs](#) (`#cpu-quota`) because having more CPUs increases VM instance quota. The quota applies to both running and non-running VMs, and to normal and preemptible instances.

1. In the Google Cloud console, go to the **Quotas** page.

[Go to Quotas](https://console.cloud.google.com/iam-admin/quotas) (https://console.cloud.google.com/iam-admin/quotas)

2. Click  **Filter table** and select **Service**.
3. Choose **Compute Engine API**.
4. Choose **Quota: VM instances**.
5. To see a list of your VM instance quotas by region, click **All Quotas**. Your region quotas are listed from highest to lowest usage.
6. Click the checkbox of the region whose quota you want to change.
7. Click  **Edit Quotas**.
8. Complete the form.
9. Click **Submit Request**.

Instance groups

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

Service type	Service quota
Regional (multi-zone) managed instance group	Regional instance group managers
Zonal (single-zone) managed instance group	Both of: <ul style="list-style-type: none">• Instance group managers• Instance groups
Unmanaged (single-zone) instance group	Instance groups
Regional (multi-zone) autoscaler	Regional autoscalers
Zonal (single-zone) autoscaler	Autoscalers

Disk quotas

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

- **Local SSD per machine family (GB)**. This quota is the total combined size of local SSD (/compute/docs/disks/local-ssd) disk partitions you can attach to VMs in a region based on the machine type of each VM. 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 24 local SSD partitions can be attached to a single VM. In the gcloud CLI and the API, this quota is referred to as `LOCAL_SSD_TOTAL_GB_PER_VM_FAMILY`.



Note: `LOCAL_SSD_TOTAL_GB` quota has been deprecated. To view the local SSD quota usage and limits, you must use the quota metric `compute.googleapis.com/local_ssd_total_storage_per_vm_family`^{BETA} in your Cloud Monitoring dashboards, alerts, and queries. For more information, see [View and manage local SSD quota](/compute/docs/quotas/migrate-local-ssd-quota) (/compute/docs/quotas/migrate-local-ssd-quota).

- **Persistent disk standard (GB).** This quota is the total size of [standard persistent disks](/compute/docs/disks#pdspecs) (/compute/docs/disks#pdspecs) that can be created in a region. As described in [Optimizing persistent disk and local SSD performance](/compute/docs/disks/performance) (/compute/docs/disks/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 CLI and the API, this quota is referred to as `DISKS_TOTAL_GB`. This quota also applies to [regional standard persistent disks](/compute/docs/disks#repds) (/compute/docs/disks#repds), 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 quota is the total combined size of [SSD-backed persistent disks](/compute/docs/disks) (/compute/docs/disks) partitions that can be created in a region. SSD-backed persistent disks have multiple replicas and, as described in [Block storage performance](/compute/docs/disks/performance) (/compute/docs/disks/performance), offer higher IOPS and throughput than standard persistent disks. SSD-backed persistent disks are available indefinitely to attach to a VM within the same zone. In the gcloud CLI and the API, this quota is referred to as `SSD_TOTAL_GB`. This quota is separate from local SSD. This quota applies to the [disk types](/compute/docs/disks#disk-types) (/compute/docs/disks#disk-types) listed below. [Regional persistent disks](/compute/docs/disks#repds) (/compute/docs/disks#repds) consume twice the amount of quota per GB due to replication in two zones within a region:
 - Zonal and regional SSD persistent disk
 - Zonal and regional balanced persistent disk

CPU quota limits

CPU quota is the total number of virtual CPUs across all of your VM instances in a region. CPU quotas apply to running VMs and VM reservations. Both predefined and preemptible VMs (/compute/docs/instances/preemptible) consume this quota.

To help protect Compute Engine systems and other users, some new accounts and projects also have a global CPUs (All Regions) quota. That quota 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 reach the CPU (All Regions) quota and need to delete existing instances before you can launch new instances.

E2 and N1 machine types share a CPU quota pool. Unless otherwise noted, all other machine types have unique, separate CPU quota pools.

Note: M1 machine types started before November 2020 consume quota of type CPUS. M1 machine types created or restarted after November 2020 consume M1_CPUS quota.

If you are using committed use discounts for your VMs, you must have committed use discount quota before you purchase a committed use discount contract.

Machine type	Quota pool	CPU quota name	Committed CPU quota name
N1	shared pool	CPUS	Committed_CPUS
E2	shared pool	CPUS	Committed_E2_CPUS
N2	separate pool	N2_CPUS	Committed_N2_CPUS
N2D	separate pool	N2D_CPUS	Committed_N2D_CPUS
T2D	separate pool	T2D_CPUS	Committed_T2D_CPUS
T2A	separate pool	T2A_CPUS	Not available (N/A) for T2A
M1	separate pool	M1_CPUS	Committed_MEMORY-OPTIMIZED_CPUS
M2	separate pool	M2_CPUS	Committed_MEMORY-OPTIMIZED_CPUS
M3	separate pool	M3_CPUS	Committed_M3_CPUS
H3	separate pool	CPUS_PER_VM_FAMILY	Committed_H3_CPUS

Machine type	Quota pool	CPU quota name	Committed CPU quota name
C2	separate pool	C2_CPUS	Committed_C2_CPUS
C2D	separate pool	C2D_CPUS	Committed_C2D_CPUS
C3	separate pool	C3_CPUS	Committed_C3_CPUS
C3D	separate pool	CPUS_PER_VM_FAMILY	Committed_C3D_CPUS
Preemptible VMs	shared pool	PREEMPTIBLE_CPUS	Not available (N/A) for preemptible VMs

GPU quota

Similar to virtual CPU quota, GPU quota refers to the total number of virtual GPUs in all VM instances in a region. GPU quotas apply to running VMs and VM reservations. Both predefined and preemptible VMs (/compute/docs/instances/preemptible) consume this quota.

Check the **Quotas** page (<https://console.cloud.google.com/iam-admin/quotas>) 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. Request preemptible GPU quota to use those resources.

NVIDIAGPU quota name		Committed GPU quota name	Virtual workstation
A100 40GB	NVIDIA_A100_GPUS	COMMITTED_NVIDIA_A100_GPUS	N/A
A100 80GB	NVIDIA_A100_80GB_GPUS	COMMITTED_NVIDIA_A100_80GB_GPUS	N/A
L4	NVIDIA_L4_GPUS	COMMITTED_NVIDIA_L4_GPUS	NVIDIA_L4_VWS_GPUS
T4	NVIDIA_T4_GPUS	COMMITTED_NVIDIA_T4_GPUS	NVIDIA_T4_VWS_GPUS
V100	NVIDIA_V100_GPUS	COMMITTED_NVIDIA_V100_GPUS	N/A
P100	NVIDIA_P100_GPUS	COMMITTED_NVIDIA_P100_GPUS	NVIDIA_P100_VWS_GPUS
P4	NVIDIA_P4_GPUS	COMMITTED_NVIDIA_P4_GPUS	NVIDIA_P4_VWS_GPUS
K80	NVIDIA_K80_GPUS	COMMITTED_NVIDIA_K80_GPUS	N/A

Allocation 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 those respective resources.

You can [request](#) (#requesting_additional_quota) special preemptible quotas for **Preemptible CPUs**, **Preemptible GPUs**, or **Preemptible Local SSDs (GB)**. However, if your project does not have preemptible quota, and you have never requested preemptible quota, you can consume standard quota to launch preemptible resources.

After Compute Engine grants you preemptible quota in a region, all preemptible instances automatically count against preemptible quota. As this quota is depleted, you must request preemptible quota for those resources.

External IP addresses

You must have enough external IP addresses for every VM that needs to be directly 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 load balancers. Google Cloud offers different types of IP addresses, depending on your needs. For information about costs, refer to [External IP address pricing](#) (/vpc/network-pricing#ipaddress). For information about quota specifics, see [Quotas and limits](#) (/vpc/docs/quota).

- **In-use external IP addresses.** Includes both ephemeral and static IP addresses that are currently being used by a resource.

★ **Note:** If the same IP address is assigned to more than one forwarding rule, Google Cloud counts and adds each usage of the address towards the **IN_USE_ADDRESSES** quota rather than a unique count of IP address objects that are used.

- **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 let you reserve internal IP addresses from the internal IP range configured in the subnet. You can assign those reserved internal addresses to resources as needed.

Quota rollouts

Occasionally, Google Cloud changes the default quota for resources and APIs. These changes take place gradually. During the rollout of a new default quota, the maximum quota that appears in the Google Cloud console might not reflect the actual maximum quota that is available to you.

For example, suppose that Google Cloud changes the default maximum quota for firewall rules from 200 to 300, and you use the Google Cloud console to view your quota, you might see the new quota of 300, even though your actual quota is 200 until the rollout completes.

For information about ongoing quota rollouts, see [known issues](/compute/docs/troubleshooting/known-issues) (/compute/docs/troubleshooting/known-issues). If no issues are described, no quota rollouts are ongoing.

If a quota rollout is ongoing and you want to confirm the actual maximum quota that is available to you, [use the Google Cloud CLI to check your quota](#) (#check_your_quota). If you need more quota than you have access to, [submit a quota increase request](#) (/docs/quota_detail/view_manage#requesting_higher_quota).

What's next

- Read about [resource-based pricing](/compute/resource-based-pricing) (/compute/resource-based-pricing).
- Read about [VM instances pricing](/compute/vm-instance-pricing) (/compute/vm-instance-pricing).
- [View and manage quota](/docs/quota_detail/view_manage) (/docs/quota_detail/view_manage).
- Learn how to [set up quota usage alerts](/docs/quota_detail/monitor) (/docs/quota_detail/monitor)

Except as otherwise noted, the content of this page is licensed under the [Creative Commons Attribution 4.0 License](https://creativecommons.org/licenses/by/4.0/) (https://creativecommons.org/licenses/by/4.0/), and code samples are licensed under the [Apache 2.0 License](https://www.apache.org/licenses/LICENSE-2.0) (https://www.apache.org/licenses/LICENSE-2.0). For details, see the [Google Developers Site Policies](https://developers.google.com/site-policies) (https://developers.google.com/site-policies). Java is a registered trademark of Oracle and/or its affiliates.

Last updated 2023-11-29 UTC.