Add or remove GPUs

LINUX WINDOWS

Compute Engine provides graphics processing units (GPUs) that you can add to your virtual machine instances (VMs). You can use these GPUs to accelerate specific workloads on your VMs such as machine learning and data processing.

If you did not attach GPUs (/compute/docs/gpus/create-vm-with-gpus) during VM creation, you can add GPUs to your existing VMs to suit your application needs as they arise.

If you attached GPUs during or after VM creation, you can detach these GPUs from these VMs when you no longer need them.

Overview

In summary, the process to add or remove a GPU from an existing VM is as follows:

1. Check that your VM has a boot disk size of at least 40 GB.
2. Prepare your VM (#prep-vm) for the modification.
3. Stop the VM.
4. Add or remove the GPU.
5. If you are adding a GPU, you need to complete the following steps:
    - Modify the host maintenance setting for the VM. VMs with GPUs cannot live migrate (/compute/docs/instances/setting-instance-scheduling-options#live_migrate) because they are assigned to specific hardware devices. For more information, see GPU restrictions (/compute/docs/gpus#restrictions).
    - Change the machine type. GPUs are only supported on select machine types (/compute/docs/machine-types#gpus).
    - Install a GPU driver on your VM (/compute/docs/gpus/install-drivers-gpu), so that your system can use the device.

Before you begin

- If you want to use the command-line examples in this guide, do the following:

1. Install or update to the latest version of the Google Cloud CLI (/compute/docs/gcloud-compute).
2. Set a default region and zone (/compute/docs/gcloud-compute#set_default_zone_and_region_in_your_local_client).

- If you want to use the API examples in this guide, set up API access (/compute/docs/api/prereqs).
- Read about GPU pricing on Compute Engine (/compute/gpus-pricing#gpus) to understand the cost to use GPUs on your VMs.
- Read about restrictions for VMs with GPUs (/compute/docs/gpus/about-gpus#restrictions).
- Check your GPU quota (#check-quota).

Checking GPU quota

To protect Compute Engine systems and users, new projects have a global GPU quota, which limits the total number of GPUs you can create in any supported zone.

Use the `regions describe` command (/sdk/gcloud/reference/compute/regions/describe) to ensure that you have sufficient GPU quota in the region where you want to create VMs with GPUs.

```
gcloud compute regions describe REGION
```

Replace ***REGION*** with the region (/compute/docs/regions-zones) that you want to check for GPU quota.

Note: Some regions might display quotas even though GPUs are not currently available in that region. Ensure that the region that you are requesting quotas for support GPUs. For a list of regions with GPUs, see GPUs regions and zone availability (/compute/docs/gpus/gpu-regions-zones).

If you need additional GPU quota, request a quota increase (/compute/quotas#requesting_additional_quota). When you request a GPU quota, you must request a quota for the GPU types 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.

If your project has an established billing history, it will receive quota automatically after you submit the request.

Preparing your VM

When a GPU is added to a VM, the order of the network interface can change.

Most public images on Compute Engine do not have persistent network interface names and adjust to the new order.

However, if you are using either SLES or a custom image, you must update the system setting to prevent the network interface from persisting. To prevent the network interface from persisting, run the following command on your VM:

```
rm /etc/udev/rules.d/70-persistent-net.rules
```

Modifying the GPU count for existing A2 VMs

This section covers how to increase or decrease your A100 GPU count by switching between A2 machine types.

If you are using an A2 machine and no longer require GPUs, you need to change your machine from A2 to another machine type. For more information, see Changing the machine type of a VM instance (/compute/docs/instances/changing-machine-type-of-stopped-instance).

Limitations (A100)

- `a2-megagpu-16g` machine types are not supported on Windows operating system. When using Windows operating systems, choose a different machine type. For a list of machine types, see NVIDIA® A100 GPUs (/compute/docs/gpus#a100-gpus).
- For Windows VMs that use A2 machine types, you cannot do a quick format of the attached local SSDs. To format these local SSDs, you must do a full format by using the diskpart utility (https://docs.microsoft.com/windows-server/administration/windows-commands/diskpart) and specifying `format fs=ntfs label=tmpfs`.

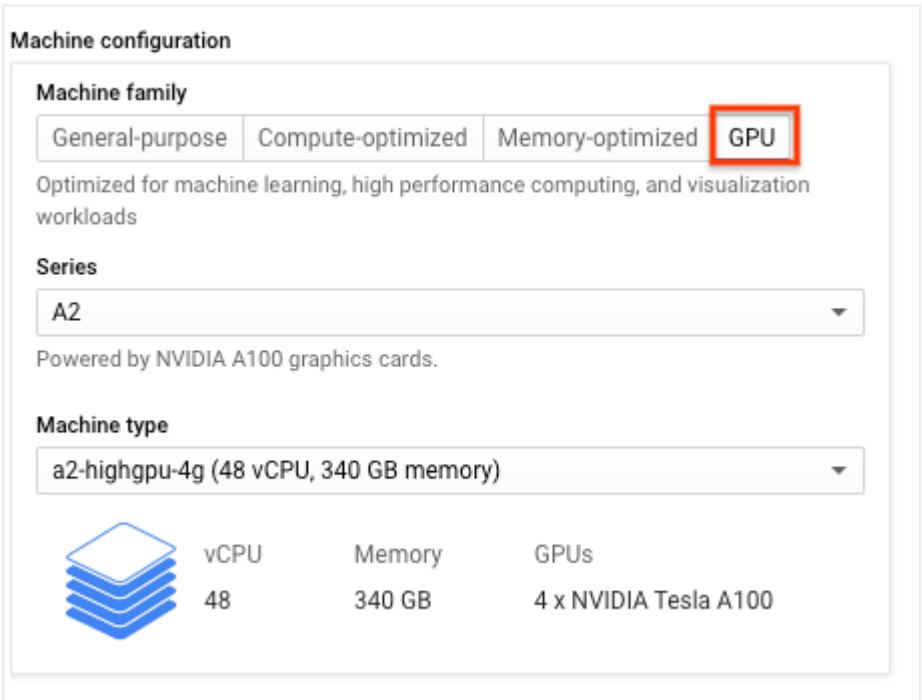
Console (#console) API (#api)

You can add GPUs to your VM by stopping the VM and editing the VM configuration.

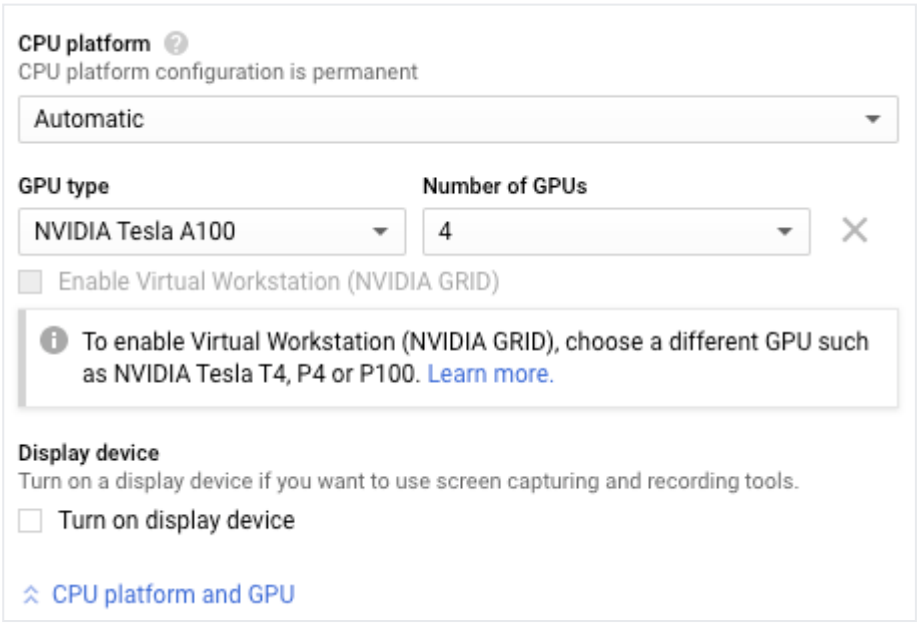
1. Verify that all of your critical applications are stopped on the VM.
2. In the Google Cloud console, go to the **VM instances** page to see your list of VMs.

   Go to VM instances (https://console.cloud.google.com/compute/instances)

3. Click the name of the VM where you want to add GPUs. The **VM instance details** page opens.
4. On the **VM instance details** page, complete the following steps:
   a. Click **Stop** to stop the VM. You can check the notification panel to see when the instance is stopped.
   b. On the stopped VM, click **Edit** to change the VM properties.
   c. From **Machine configuration**, complete the following steps.
      i. Under **Machine family**, click **GPU**.
      ii. Under **Series**, select **A2**.
      iii. Under **Machine type**, select the A2 machine type (/compute/docs/gpus#a100-gpus) that you want.



iv. Expand the **CPU platform and GPU** section.



v. Under **CPU platform and GPU**, use the **Number of GPUs** field to increase or decrease the GPU count.

★ **Note:** Each A2 machine type has a fixed GPU count, vCPU count, and memory size. If you adjust the **Number of GPUs**, the **Machine type** changes.

d. Click **Save** to apply your changes.

e. Click **Start/Resume** to restart the VM.

Adding GPUs to existing VMs

This section covers how to add the following GPU types to existing VMs.

NVIDIA GPUs:

- NVIDIA T4: `nvidia-tesla-t4`
- NVIDIA P4: `nvidia-tesla-p4`
- NVIDIA P100: `nvidia-tesla-p100`
- NVIDIA V100: `nvidia-tesla-v100`
- NVIDIA K80: `nvidia-tesla-k80`

NVIDIA RTX (formerly known as NVIDIA GRID) virtual workstation GPUs:

- NVIDIA T4 Virtual Workstation: `nvidia-tesla-t4-vws`
- NVIDIA P4 Virtual Workstation: `nvidia-tesla-p4-vws`
- NVIDIA P100 Virtual Workstation: `nvidia-tesla-p100-vws`

  For these virtual workstations, an NVIDIA RTX Virtual Workstation license is automatically added to your VM.

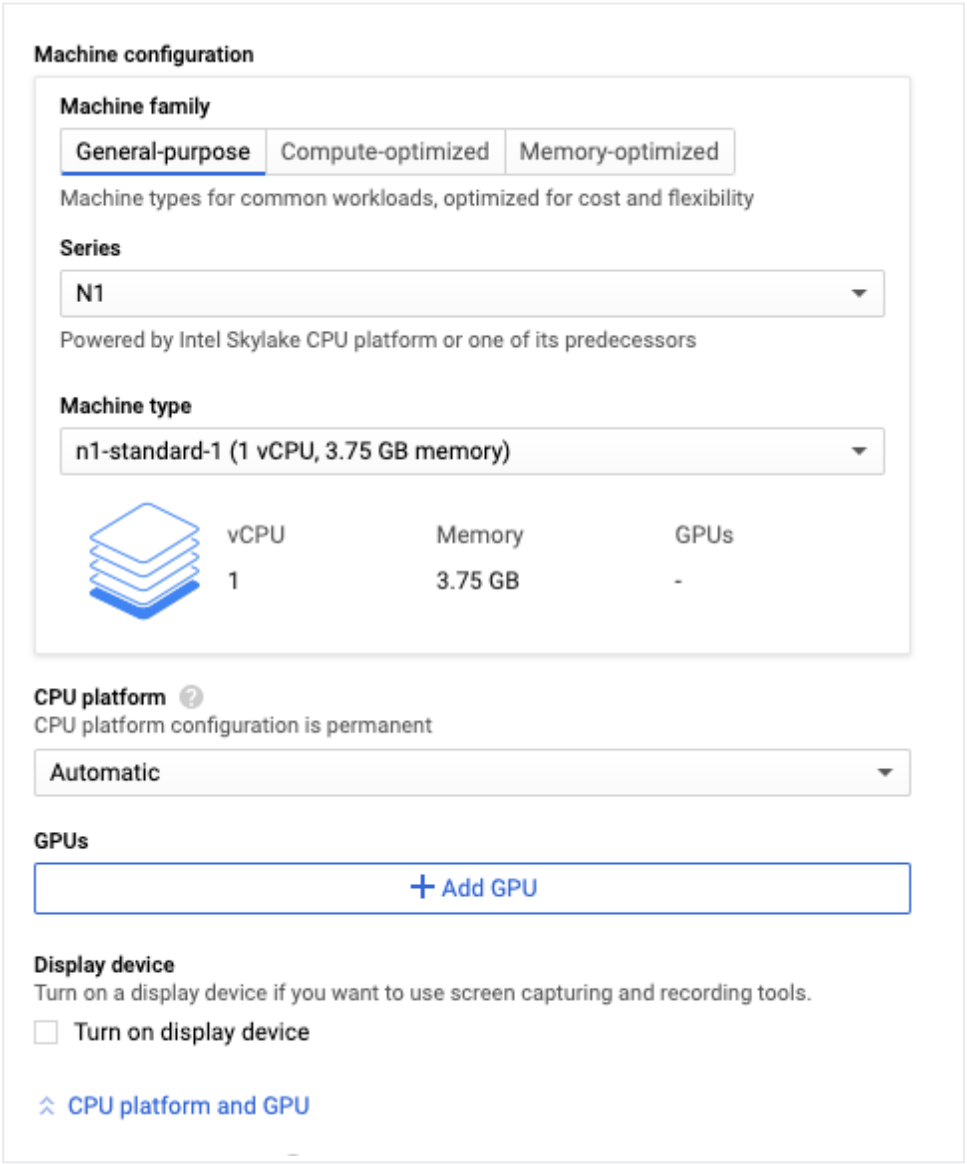
ConsoleAPI (#api)
(#console)

You can add or remove GPUs from your VM by stopping the VM and editing the VM configuration.

1. Verify that all of your critical applications are stopped on the VM.
2. In the Google Cloud console, go to the **VM instances** page to see your list of VMs.

   Go to VM instances (https://console.cloud.google.com/compute/instances)

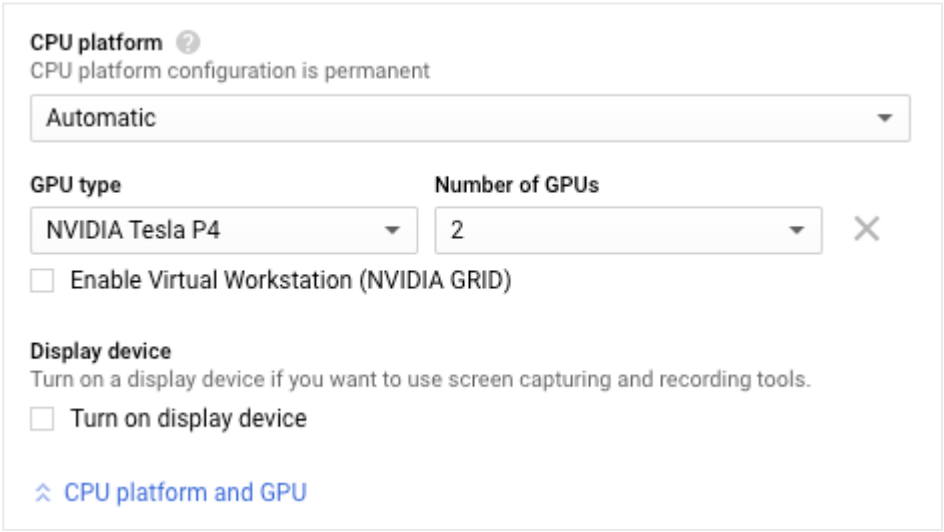
3. Click the name of the VM where you want to add GPUs. The **VM instance details** page opens.
4. Complete the following steps from the **VM instance details** page.
   a. Click **Stop** to stop the VM. You can check the notification panel to see when the instance is stopped.
   b. On the stopped VM, click **Edit** and complete the following steps:
   c. From the **Machine configuration** section, complete the following steps.
      i. Under **Series**, select **N1**.
      ii. Under **Machine type**, select the N1 machine type (/compute/docs/machine-types#n1_machine_types) that you want.
      iii. Expand the **CPU platform and GPU** section.
      iv. Click **Add GPU**.



v. Specify the **GPU type** and **Number of GPUs**.

vi. If your GPU model supports virtual workstations, and you plan on running graphics-intensive workloads on this VM, select **Enable Virtual Workstation (NVIDIA GRID)**.

For information about NVIDIA RTX virtual workstations, see NVIDIA RTX virtual workstations for graphics workloads (formerly known as NVIDIA GRID) (/compute/docs/gpus#gpu-virtual-workstations).



d. Scroll to the **On host maintenance** section. When you add GPUs to a VM, the host maintenance setting is automatically set to **Terminate VM instance**. See Handling GPU host maintenance events (/compute/docs/gpus/gpu-host-maintenance).

e. Click **Save** to apply your changes.

f. Click **Start/Resume** to restart the VM.

Next: To ensure that your system can use the GPUs, complete the following steps:

- Install the GPU drivers (/compute/docs/gpus/install-drivers-gpu).
- If you enabled NVIDIA RTX virtual workstation, install a driver for the virtual workstation (/compute/docs/gpus/install-grid-drivers).

Removing or modifying GPUs

This section covers how to remove the following GPU types from an existing VM.

NVIDIA GPUs:

- NVIDIA T4: `nvidia-tesla-t4`
- NVIDIA P4: `nvidia-tesla-p4`
- NVIDIA P100: `nvidia-tesla-p100`
- NVIDIA V100: `nvidia-tesla-v100`
- NVIDIA K80: `nvidia-tesla-k80`

NVIDIA RTX (formerly known as NVIDIA GRID) virtual workstation GPUs:

- NVIDIA T4 Virtual Workstation: `nvidia-tesla-t4-vws`
- NVIDIA P4 Virtual Workstation: `nvidia-tesla-p4-vws`
- NVIDIA P100 Virtual Workstation: `nvidia-tesla-p100-vws`

  For these virtual workstations, an NVIDIA RTX Virtual Workstation license is automatically added to your VM.

You can use the Google Cloud console (https://console.cloud.google.com/) to remove GPUs from an existing VM, or modify the number or type of GPU that you have attached. To remove or modify GPUs, complete the following steps:

1. Verify that all of your critical applications are stopped on the VM.
2. In the Google Cloud console, go to the **VM instances** page to see your list of VMs.

   Go to VM instances (https://console.cloud.google.com/compute/instances)

3. Click the name of the VM where you want to remove or modify GPUs. The **VM instance details** page opens.
4. Complete the following steps from the **VM instance details** page.
   a. Click **Stop** to stop the VM. You can check the notification panel to see when the instance is stopped.
   b. On the stopped VM, click **Edit**.
   c. Under **Machine configuration**, expand the **CPU platform and GPU** section.
   d. Remove or modify the GPUs as follows:
      - To modify the GPUs, adjust the **Number of GPUs** or the **GPU Type** as needed.
      - To remove all GPUs, click **Delete GPU**.
   e. Optional: Modify the VM host maintenance policy setting. When you add GPUs to a VM, the host maintenance setting is automatically set to **Terminate VM instance**. With no GPUs attached, you now have the option to live migrate during host maintenance. For more information about setting VM host maintenance policy, see Set VM host maintenance policy (/compute/docs/instances/setting-instance-scheduling-options).
   f. Click **Save** to apply your changes.

g. Click **Start/Resume** to restart the VM.

What's next?

- Learn more about GPU platforms (/compute/docs/gpus).
- Add Local SSDs to your instances (/compute/docs/disks/local-ssd). Local SSD devices pair well with GPUs when your apps require high-performance storage.
- Create groups of GPU instances using instance templates (/compute/docs/gpus/gpu-instance-groups).
- To monitor GPU performance, see Monitoring GPU performance (/compute/docs/gpus/monitor-gpus).
- To optimize GPU performance, see Optimizing GPU performance (/compute/docs/gpus/optimize-gpus).
- To handle GPU host maintenance, see Handling GPU host events (/compute/docs/gpus/gpu-host-maintenance).
- Try the Running TensorFlow Inference Workloads at Scale with TensorRT5 and NVIDIA T4 GPU (/compute/docs/tutorials/ml-inference-t4) tutorial.

Last updated 2022-08-25 UTC.