Google Skills

Consumption options

You've learned about the powerful architecture of Cloud TPUs and how they scale within Google Cloud. Now, let's explore the crucial financial and logistical aspects of **accessing and utilizing these compute resources**, and the various consumption options available to you.

Think of consumption options as different rental agreements for your TPU compute power—each one has unique terms that determine how quickly you get capacity, how long you can keep it, and what it costs.

When selecting an option, consider these key factors:

- **Speed:** How quickly do you need the TPU capacity to be available?
- **Duration:** How long will you need the capacity for your workload?
- **Flexibility versus fixed time:** Are you running batch or online inference?
- **Preemption tolerance:** Can your workload handle being unexpectedly stopped (preempted) by Google Cloud?
- **Pricing:** What's your budget?

Understanding quota: Your compute allowance

Before we dive into the options, let's talk about quota. Quota is like your pre-approved allowance for using Cloud TPU cores.

Google Cloud uses **quotas** to ensure fair resource distribution and prevent usage spikes, helping you avoid cost overruns. A quota limits how much of a specific Google Cloud resource your project can consume—applying to hardware, software, and network components.

For instance, quotas might restrict API call volumes, the number of concurrent load balancers, or the total projects you can create. This system protects all Google Cloud users by preventing service overload, and it helps you manage your own resource consumption effectively.

When working with Cloud TPUs, your quota requirements will vary based on whether you're using:

- **APIs:** You'll need an on-demand or preemptible quota for the number of Cloud TPU cores you want to use—different TPU versions have different default quotas.
- **Google Kubernetes Engine:** You'll use Compute Engine API quota, which is a different quota system.

See Cloud TPU quotas for more information.

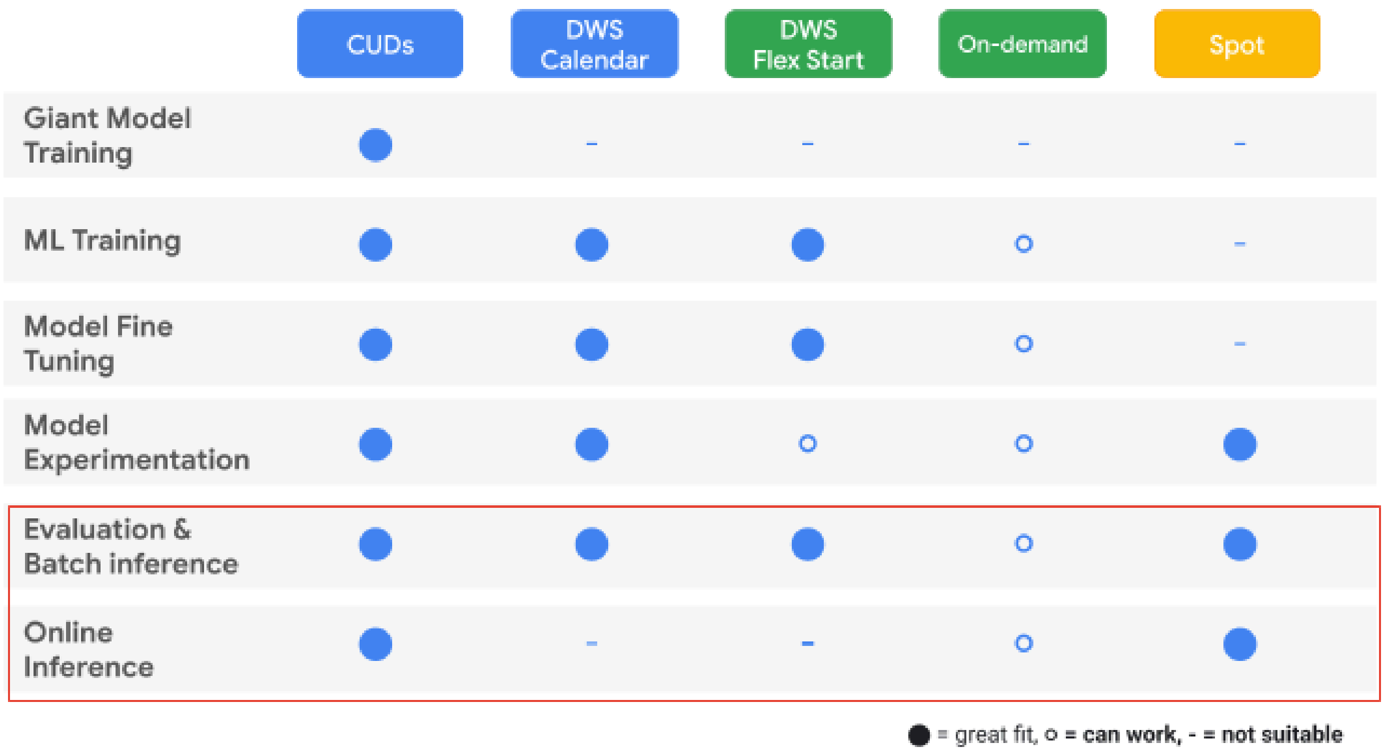
Exploring your options

A consumption option is the way that you get and use compute resources. Choose the option that best fits your workload, its duration, and your budget.

Each consumption option will specify:

- How you access capacity to create VMs or clusters.
- The underlying provisioning model, which determines the obtainability, lifespan, and pricing of your VMs.

For inference workloads, you'll have several options to choose from:



	CUDs	DWS Calendar	DWS Flex Start	On-demand	Spot
Giant Model Training	●	-	-	-	-
ML Training	●	●	●	○	-
Model Fine Tuning	●	●	●	○	-
Model Experimentation	●	●	○	○	●
Evaluation & Batch inference	●	●	●	○	●
Online Inference	●	-	-	○	●

● = great fit, ○ = can work, - = not suitable

Let’s look at how the four options work

1. On-demand

This is the basic, non-discounted price you can get for a TPU VM instance—you only pay for the resources you use, with no long-term contracts or upfront commitments. While this approach gives you the most flexibility, it also means you have no guarantee that the capacity will be available when you need it.

2. Long-term reservations

This allows you to request and reserve TPU resources in advance for an extended period of time—typically one year or longer. This option is best used for stable inference workloads that require consistent, uninterrupted compute power.

Choose it for:

- The highest level of guarantee for dedicated capacity
- Up to 70% off on-demand resources in exchange for your long-term commitment (committed use discount)

2. Spot VMs

This allows you to request TPU resources that could be preempted (shut down) by Google Cloud at any time if capacity is needed elsewhere. While there's no limit on runtime duration, the possibility of preemption is always present.

Choose it for:

- Scheduling fault-tolerant workloads
- Significant discounts in comparison to on-demand resources

Here’s more information on
Managing TPU Spot VMs.

4. Dynamic Workload Scheduler

A little different to the others, DWS is a resource management and capacity scheduling platform that provides cost-effective and simple options to consume TPUs. It offers two modes:

Flex-start mode
This helps you obtain capacity in-bulk with a single request. DWS continuously monitors and provisions TPU resources simultaneously. Choose it for batch inference jobs that have start time flexibility. These requests are not pre-emptible, so once you get your capacity, your request will be served until the job is done.

Calendar mode
This helps you plan for known events and assure capacity using calendar-based future reservations. It provides short-term ML capacity—up to 90 days of reserved capacity—without requiring long-term commitments. Choose it when you have a known spike in resource requirements on the horizon—like a big event that will require additional capacity. Similar to a flight or hotel booking experience, Calendar mode makes it easy to search for and reserve ML capacity. Simply define your resource type, number of instances, expected start date and duration, and in a few seconds, you’ll be able to see the available capacity and reserve it.

Learn more about
Dynamic Workload Scheduler.

Using multiple consumption models together

Combining multiple pricing and consumption models in a single workflow can significantly reduce your costs by aligning the right pricing structure with the appropriate workload. This means you can pay less for predictable, baseline resource usage, while retaining the ability to scale with on-demand resources for variable or short-term workloads.

For GKE users, this is achieved through custom Compute Classes—allowing you to control the properties of different node pools with rules. For instance, one node pool could use cost-effective Spot VMs for fault-tolerant workloads, while another might use performance-optimized instances for critical services. This is not a direct combination of different pricing models within a single virtual machine; It's a strategic allocation of workloads to differently priced node pools within the same cluster.

If you're working directly with your VMs in Google Compute Engine (GCE), you can combine committed use discounts (CUDs) with on-demand instances. This allows you to purchase CUDs for predictable, long-running workloads, then use on-demand instances to handle unexpected traffic spikes or short-term processing needs.

How to avoid unwanted charges

Once you've provisioned a TPU, you have two options to stop incurring charges: Stop or delete.

Using the Google Cloud console:

1. Navigate to the TPUs page in the Google Cloud Console.
2. Select the checkbox next to the TPU you want to stop.
3. Click "Stop'.' This stops the TPU but saves its configuration and the attached virtual machine (VM). You will still be charged for any persistent disk storage, but not for theTPU usage itself.
4. Alternatively, click 'Delete' to permanently remove the TPU and all its associated resources.

Using the gcloud command-line tool:

To stop a TPU, use the following command:

```
gcloud compute tpus tpu-vm stop
<TPU_NAME> --zone=<ZONE>
```

To delete a TPU, use the following command:

```
gcloud compute tpus tpu-vm delete
<TPU_NAME> --zone=<ZONE>
```

Queued resources: You cannot stop TPU slices or TPUs allocated through the queued resources API. To avoid subsequent charges for these resources, you must delete them.

Using Google Colab:

You can terminate the session through the user interface. You don't need to manually shut down the TPU, as Colab sessions will automatically disconnect after a period of inactivity.

To manually terminate your Colab session:

1. Click ‘Runtime’ in the top menu.
2. Select ‘Manage sessions.’
3. In the panel that appears, find your active session and click ‘Terminate’ next to it.

Alternatively, you can perform a "factory reset" of the runtime, which also terminates the session.

1. Click ‘Runtime’ in the top menu.
2. Select ‘Factory reset runtime.’

During a Python script's execution:

If a Python script using a TPU appears to be unresponsive, you can force it to stop. To force-exit a script, press Ctrl+C. In TensorFlow, this sends a SIGQUIT and forces Python to exit immediately.

Key takeaways

- Google Cloud offers various TPU consumption options: Long-term reservations, on-demand, and spot; each option is tailored to different needs regarding cost, availability, and workload tolerance for interruptions.
- Long-term reservations provide guaranteed capacity and significant discounts for stable, critical, and long-running AI workloads. On demand offers maximum flexibility and instant access for urgent, unpredictable tasks without preemption risk.
- Spot VMs are ideal for fault-tolerant, lower-priority workloads due to their much lower price, though they are subject to preemption and use a specific preemptible quota.
- Dynamic Workload Scheduler provides a new option, allowing for capacity requests in-bulk (Flex Start mode) or on a specified date (Calendar mode).