When you create TPU nodes to handle your machine learning workloads, you must select a TPU type. The TPU type defines the TPU version, the number of TPU cores, and the amount of TPU memory that is available for your machine learning workload.

For example, the `v2-8` TPU type defines a TPU node with 8 TPU v2 cores and 64 GiB of total TPU memory. The `v3-2048` TPU type defines a TPU node with 2048 TPU v3 cores and 32 TiB of total TPU memory.

To learn about the hardware differences between TPU versions and configurations, read the System Architecture (/tpu/docs/system-architecture) documentation.

To see pricing for each TPU type in each region, see the Pricing (/tpu/pricing) page.

A model that runs on one TPU type can run with no TensorFlow code changes for another TPU type. For example, `v2-8` code can run without changes on a `v3-8`. However, scaling from a `v2-8` or `v3-8` to a larger TPU type, such as `v2-32` or `v3-128`, requires significant tuning and optimization.

The main differences between each TPU type are price, performance, memory capacity, and zonal availability.

Google Cloud Platform uses regions, subdivided into zones, to define the geographic location of physical computing resources. For example, the `us-central1` region denotes a region near the geographic center of the United States that has the following zones: `us-central1-a`, `us-central1-b`, `us-central1-c`, and `us-central1-f`. When you create a TPU node, specify the zone in which you want to create it.

You can configure your TPU nodes with the following TPU types:

TPU types with higher numbers of cores are available only in limited quantities. TPU types with lower core counts are more likely to be available.

To see pricing for each TPU type in each region, see the Pricing (/tpu/pricing) page.

To learn about the hardware differences between TPU versions and configurations, read the System Architecture (/tpu/docs/system-architecture) documentation.

To decide which TPU type you want to use, you can do experiments using a Cloud TPU tutorial (/tpu/docs/tutorials) to train a model that is similar to your application.

Run the tutorial for 5 - 10% of the number of steps you will use to run the full training on a `v2-8` and a `v3-8` TPU type. The result tells you how long it takes to run that number of steps for that model on each TPU type.

Because performance on TPU types scales linearly, if you know how long it takes to run a task on a `v2-8` or `v3-8` TPU type, you can estimate how much you can reduce task time by running your model on a larger TPU type with more cores.

For example, if a `v2-8` TPU type takes 60 minutes to 10,000 steps, a `v2-32` node should take approximately 15 minutes to perform the same task.

To determine the difference in cost within your region between the different TPU types for Cloud TPU and the associated Compute Engine VM, see the TPU pricing page (/tpu/docs/pricing). When you know the approximate training time for your model on a few different TPU types, you can weigh the VM/TPU cost against training time to help you decide your best price/performance tradeoff.

You specify a TPU type when you create a TPU node (/tpu/docs/managing-vm-tpu-resources). For example, you can select a TPU type using one of the following methods:

- Learn more about TPU architecture in the system architecure (/tpu/docs/system-architecture) page.
- See When to use TPUs (/tpu/docs/tpus#when_to_use_tpus) to learn about the types of models that are well suited to Cloud TPU.
- If you plan to run on Kubernetes or ML Engine, see Deciding on a TPU service (/tpu/docs/deciding-tpu-service).