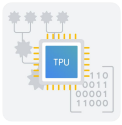
Google Cloud

AI Infrastructure: Cloud TPUs

Congratulations on completing the course. This course summary is your review guide. Print it for a handy reference.

TPU stands for Tensor Processing Unit. These are custom-developed, application-specific integrated circuits (ASICs) built by Google specifically to accelerate the intensive computations found in machine learning. Here's when they're your best bet:



Training massive deep learning models

Models relying on embeddings

Scientific and healthcare AI

Google TPU evolution

2015	2018	2020	2022	2023	2024	2025
V1 **1x/chip Inference** Internal inference accelerator	**V2** **1x/chip** **1x/pod** Distributed shared memory	**V3** **3x/chip** **12x/pod** Liquid cooled	**V4** **6.6x/chip** **100x/pod** Optically reconfigurable	**v5p** **21x/chip** **750x/pod** Most flexible AI accelerator	**Trillium** **100x v2 performance** Enabling the next frontier of AI models	**Ironwood** **TPU 7x: 9,216 chips/pod** **TPU 7: 256 chips/pod** • Cutting-edge chip • Largest pod

GPU and TPU interoperability:

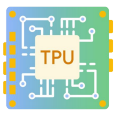
- To achieve seamless GPU and TPU switching, use a dual-container approach within a single pod.
- If the appropriate accelerator (GPU or TPU) is present, that container's vLLM server starts; otherwise, it sleeps. This ensures only the correct vLLM server is active based on the underlying hardware.

Dynamic Workload Scheduler (DWS)

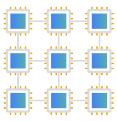
- Flex mode allows users to request hardware for specified periods (from 1 minute to 7 days).
- Calendar mode enables users to create future reservations for hardware they know they will need in advance.

Cloud TPU consumption options

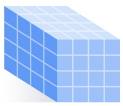
- Long-term reservations allow you to request and reserve TPU resources in advance for an extended period.
- On-demand allows you to request TPU resources to be used as soon as possible, and you can keep them for as long as you want.
- Spot allows you to request TPU resources that could be preempted (shut down) by Google Cloud at any time if capacity is needed elsewhere.



A TPU pod is a collection of TPUs that are physically grouped together and connected by a specialized, high-speed network.



A slice is a subset of chips within a single TPU pod, all connected by incredibly fast Inter-Chip Interconnects (ICI).



TPU cube

A 4x4x4 topology of interconnected TPU chips.

Best practices for model development

Principle 1: Layout for efficiency

Principle 2: Fixed shapes for predictable performance

Principle 3: Avoiding unnecessary padding

Additional Resources

1. AI Hypercomputer documentation
2. Ironwood: The first Google TPU for the age of inference
3. About TPUs in GKE

Google Cloud

Google Cloud

AI Infrastructure: Cloud TPUs

Congratulations on completing the course. This course summary is your review guide. Print it for a handy reference.

TPU stands for Tensor Processing Unit. These are custom-developed, application-specific integrated circuits (ASICs) built by Google specifically to accelerate the intensive computations found in machine learning. Here's when they're your best bet:

- Training massive deep learning models
- Models relying on embeddings
- Scientific and healthcare AI

Google TPU evolution

2015	2018	2020	2022	2023	2024	2025
V1	V2	V3	V4	v5p	Trillium	Ironwood
1x/chip Inference	**1x/chip** **1x/pod**	**3x/chip** **12x/pod**	**6.6x/chip** **100x/pod**	**21x/chip** **750x/pod**	**100x v2 performance**	**TPU 7x: 9,216 chips/pod**
Internal inference accelerator	Distributed shared memory	Liquid cooled	Optically reconfigurable	Most flexible AI accelerator	Enabling the next frontier of AI models	**TPU 7: 256 chips/pod** • Cutting-edge chip • Largest pod

GPU and TPU interoperability:

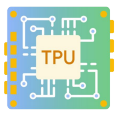
- To achieve seamless GPU and TPU switching, use a dual-container approach within a single pod.
- If the appropriate accelerator (GPU or TPU) is present, that container's vLLM server starts; otherwise, it sleeps. This ensures only the correct vLLM server is active based on the underlying hardware.

Dynamic Workload Scheduler (DWS)

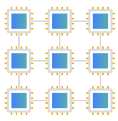
- Flex mode allows users to request hardware for specified periods (from 1 minute to 7 days).
- Calendar mode enables users to create future reservations for hardware they know they will need in advance.

Cloud TPU consumption options

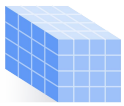
- Long-term reservations allow you to request and reserve TPU resources in advance for an extended period.
- On-demand allows you to request TPU resources to be used as soon as possible, and you can keep them for as long as you want.
- Spot allows you to request TPU resources that could be preempted (shut down) by Google Cloud at any time if capacity is needed elsewhere.



A TPU pod is a collection of TPUs that are physically grouped together and connected by a specialized, high-speed network.



A slice is a subset of chips within a single TPU pod, all connected by incredibly fast Inter-Chip Interconnects (ICI).



TPU cube

A 4x4x4 topology of interconnected TPU chips.

Best practices for model development

Principle 1: Layout for efficiency

Principle 2: Fixed shapes for predictable performance

Principle 3: Avoiding unnecessary padding