

Google Cloud Platform (GCP): TPU (Tensor Processing Unit)

Latest TPU Version: Ironwood

Source: Michael Nuñez (<https://venturebeat.com/>)

(Google's new Ironwood chip is 24x more powerful than the world's fastest supercomputer
<https://venturebeat.com/ai/googles-new-ironwood-chip-is-24x-more-powerful-than-the-worlds-fastest-supercomputer/>)




	 TPU v4	 TPU v5p	 Ironwood
	2022	2023	2025
Pod Size (chips)	4896	8960	9216
HBM Bandwidth/ Capacity	32 GB @ 1.2 TBs HBM	95 GB @ 2.8 TBs HBM	192 GB @ 7.4 TBs HBM
Peak Flops per chip	275 TFLOPS	459 TFLOPS	4614 TFLOPS

Figure 2. Side by side comparison of technical specifications of the 3D torus version of Cloud TPU products including the latest generation Ironwood. FP8 peak TFlops emulated for v4 and v5p, but natively supported for Ironwood.

[Google Cloud](#) unveiled its seventh-generation [Tensor Processing Unit](#) (TPU) called [Ironwood](#) on Wednesday, a custom AI accelerator that the company claims delivers more than 24 times the computing power of the world's fastest supercomputer when deployed at scale.

The new chip, announced at [Google Cloud Next '25](#), represents a significant pivot in Google's decade-long AI chip development strategy. While previous generations of TPUs were designed primarily for both training and inference workloads, Ironwood is the first purpose-built specifically for inference — the process of deploying trained AI models to make predictions or generate responses.

“Ironwood is built to support this next phase of generative AI and its tremendous computational and communication requirements,” said Amin Vahdat, Google's Vice President and General Manager of ML, Systems, and Cloud AI, in a virtual press conference ahead of the event. “This is what we call the ‘age of inference’ where AI agents will proactively retrieve and generate data to collaboratively deliver insights and answers, not just data.”

Shattering computational barriers: Inside Ironwood's 42.5 exaflops of AI muscle

The technical specifications of [Ironwood](#) are striking. When scaled to 9,216 chips per pod, Ironwood delivers 42.5 exaflops of computing power — dwarfing [El Capitan](#)'s 1.7 exaflops, currently the world's fastest supercomputer. Each individual Ironwood chip delivers peak compute of 4,614 teraflops.

Ironwood also features significant memory and bandwidth improvements. Each chip comes with 192GB of High Bandwidth Memory (HBM), six times more than [Trillium](#), Google's previous-generation TPU announced last year. Memory bandwidth reaches 7.2 terabits per second per chip, a 4.5x improvement over Trillium.

Perhaps most importantly in an era of power-constrained data centers, [Ironwood](#) delivers twice the performance per watt compared to [Trillium](#), and is nearly 30 times more power efficient than Google's first Cloud TPU from 2018.

“At a time when available power is one of the constraints for delivering AI capabilities, we deliver significantly more capacity per watt for customer workloads,” Vahdat explained.



Up to **9,216 chips** in a single pod
with **42.5 ExaFLOPS** of compute per pod

From model building to ‘thinking machines’: Why Google’s inference focus matters now

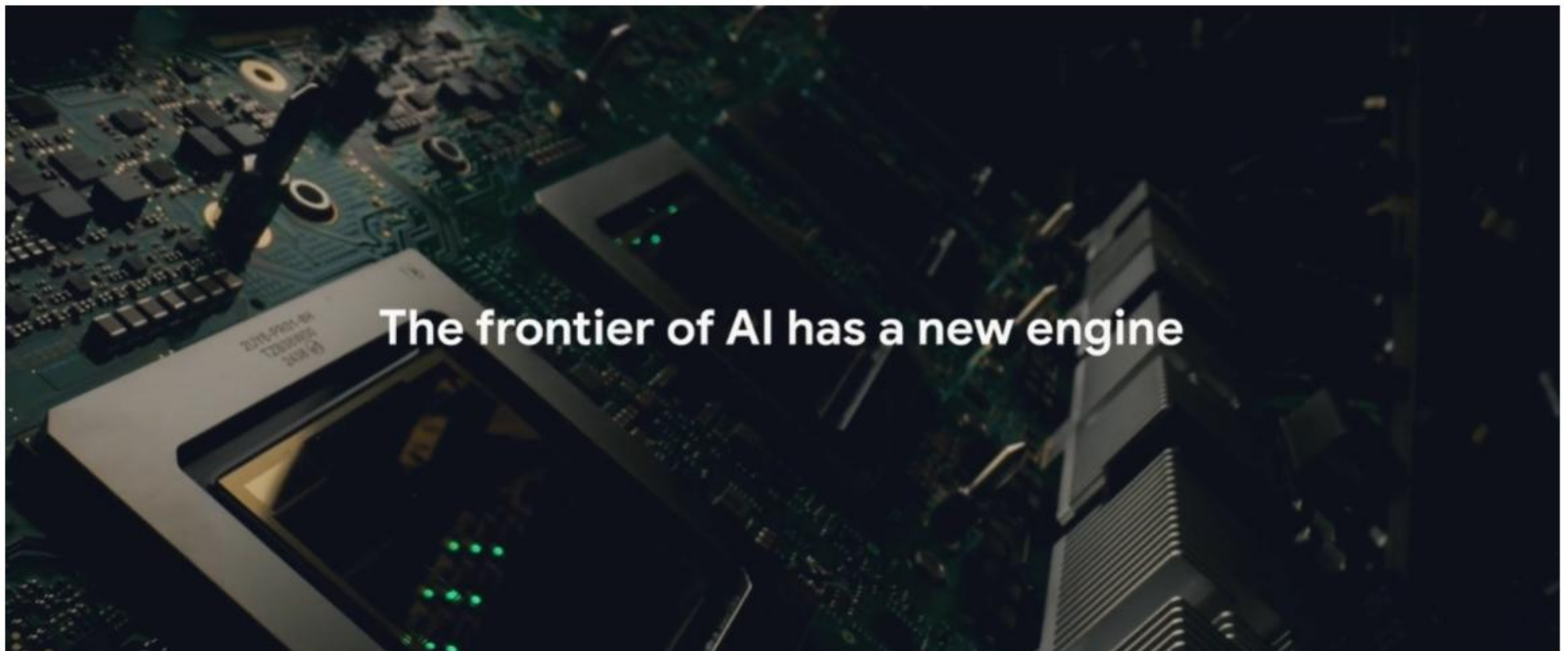
The emphasis on inference rather than training represents a significant inflection point in the AI timeline. For years, the industry has been fixated on building increasingly massive foundation models, with companies competing primarily on parameter size and training capabilities. Google’s pivot to inference optimization suggests we’re entering a new phase where deployment efficiency and reasoning capabilities take center stage.

This transition makes sense. Training happens once, but inference operations occur billions of times daily as users interact with AI systems. The economics of AI are increasingly tied to inference costs, especially as models grow more complex and computationally intensive.

During the press conference, Vahdat revealed that Google has observed a 10x year-over-year increase in demand for AI compute over the past eight years — a staggering factor of 100 million overall. No amount of [Moore's Law](#) progression could satisfy this growth curve without specialized architectures like Ironwood.

What's particularly notable is the focus on “thinking models” that perform complex reasoning tasks rather than simple pattern recognition. This suggests Google sees the future of AI not just in larger models, but in models that can break down problems, reason through multiple steps, and essentially simulate human-like thought processes.

Gemini's thinking engine: How Google's next-gen models leverage advanced hardware



Google is positioning Ironwood as the foundation for its most advanced AI models, including [Gemini 2.5](#), which the company describes as having “thinking capabilities natively built in.”

At the conference, Google also announced [Gemini 2.5 Flash](#), a more cost-effective version of its flagship model that “adjusts the depth of reasoning based on a prompt’s complexity.” While Gemini 2.5 Pro is designed for complex use cases like drug discovery and financial modeling, Gemini 2.5 Flash is positioned for everyday applications where responsiveness is critical.

The company also demonstrated its full suite of generative media models, including text-to-image, text-to-video, and a newly announced text-to-music capability called [Lyria](#). A demonstration showed how these tools could be used together to create a complete promotional video for a concert.

Beyond silicon: Google’s comprehensive infrastructure strategy includes network and software

[Ironwood](#) is just one part of Google’s broader AI infrastructure strategy. The company also announced [Cloud WAN](#), a managed wide-area network service that gives businesses access to Google’s planet-scale private network infrastructure. “Cloud WAN is a fully managed, viable and secure enterprise networking backbone that provides up to 40% improved network performance, while also reducing total cost of ownership by that same 40%,” Vahdat said.

[THUAN L NGUYEN]

Google has been showing a keen interest in the Space technologies for a while now and since the last week, the news media was focusing on a possible deal between SpaceX and Google. The deal was confirmed yesterday, Google together with Fidelity, an investment firm have invested \$ 1 billion in SpaceX and now own a little less than 10% of the company. (Read: Google and Fidelity invest \$1 billion in SpaceX).

(Source: <https://geoawesome.com/eo-hub/google-and-fidelity-invest-1-billion-in-spacex/>)

Artificial Intelligence (AI) & Technology Advancements:

Big Data Infrastructure: Google's Optical Networks & Satellites – a Double Internet?

