# Google unveils advancements in AI with Ironwood TPU and Gemini models at Cloud Next



At the recent Cloud Next event held in Las Vegas, Google unveiled a series of significant advancements in its artificial intelligence (AI) strategy, positioning its platform among the most comprehensive available in the market.

A key highlight was the introduction of Ironwood, the seventh generation of Google's specialised Tensor Processing Unit (TPU). Designed specifically for inference tasks—executing pre-trained AI models—Ironwood represents a leap forward in both performance and energy efficiency. According to Google, a single Ironwood rack can deliver 42.5 exaFLOPS (floating-point operations per second), which is 24 times more powerful than the world’s most powerful supercomputer. Additionally, it is up to 30 times more energy efficient than the first generation of TPUs. Despite continuing to offer NVIDIA GPUs to its cloud customers, Google emphasised that its TPUs are engineered as a competitive alternative at the hardware level.

On the AI model front, Google provided further details about Gemini 2.5 Pro, its flagship AI model, along with the launch of Gemini 2.5 Flash, a more affordable and faster variant aimed at day-to-day tasks such as customer service and generating meeting summaries. A distinguishing feature of these models is their ability not only to deliver accurate responses but also to exhibit structured reasoning and multi-epoch thinking, described by Google as “the AI you think before answering.”

Google also integrated creative AI models into Vertex AI, the company’s developer platform. These additions include Imagine for image generation, Chirp for audio, I See for video, and Lyria for music. During the presentation, demonstrators showcased a video generated from a simple image of Las Vegas, enhanced by AI-created music and the removal of unwanted objects from the footage — a process likened to using Photoshop but for video.

Another major focus of Google’s announcement was on intelligent AI agents. Defined by Google as systems capable of reasoning, collaborating with agents from other companies, and performing complex tasks, these agents are supported by a new Agent Development Kit (ADK). This kit allows developers to create and deploy AI agents with fewer than 100 lines of code. Additionally, Google introduced Agent Space, a platform enabling employees to create custom AI agents simply through conversation, similar to configuring routines in voice assistants. For example, during the demo, a simulated bank manager created an AI agent that analysed portfolios, generated audio reports, and automatically scheduled customer meetings—all without any coding, just by typed instructions.

To support the interoperability of these agents and real-time data access, Google introduced two new protocols: the Model Context Protocol (MCP), which enables AI models to access data in real-time, and the Agent-to-Agent (A2A) protocol, which facilitates communication among agents across different platforms. This interoperability means that, for instance, a Salesforce agent can communicate seamlessly with an SAP agent, ensuring that a company’s data is accessible to all AI systems regardless of their origin or storage location.

Regarding Google Workspace, the company revealed that Gemini integration is now available to all business users at no extra cost. The update brings new functionalities such as automatic meeting summaries, audio generation from text, and natural language data analysis within Google Sheets.

On the cybersecurity front, Google announced specialised AI agents that monitor malware and security alerts in real-time. These agents will be available in on-premises environments through Google Distributed Cloud, even in configurations that are completely disconnected from external networks.

Overall, these announcements highlight Google’s comprehensive approach to advancing AI technology across hardware, software, and developer tools, as well as integrating AI more deeply into business workflows and security operations.

Source: [Noah Wire Services](https://www.noahwire.com)

## Bibliography

1. <https://cloud.google.com/blog/products/ai-machine-learning/vertex-ai-next-2023-announcements> - This article discusses updates and new AI models, though it does not specifically mention the AI models introduced at the event you described, such as Imagine, Chirp, Lyria, and I See. However, it does mention enhancements in AI offerings.
2. <https://cloud.google.com/blog/topics/google-cloud-next/next-2023-wrap-up> - This article provides an overview of Google Cloud Next 2023, highlighting new AI announcements, though it does not cover the specific details about Ironwood, Gemini models, or the Agent Development Kit.
3. <https://blog.google/products/google-cloud/cloud-next-gen-ai-vertex-ai-updates/> - This article mentions new generative AI tools for Vertex AI, which aligns with Google's focus on enhancing AI capabilities across various media types like video, image, and music.
4. <https://techwireasia.com/2023/08/what-went-down-at-google-cloud-next-2023/> - This article covers the focus on generative AI at Google Cloud Next 2023 but does not specifically address advancements like Ironwood TPUs, Gemini AI models, or intelligent AI agents.
5. <https://innovations.woolpert.com/google-cloud-next-2023-recap/> - This recap focuses on generative AI and other services like Duet AI, but it does not provide information about the specific AI developments mentioned in the article.
6. <https://en.wikipedia.org/wiki/Tensor_Processing_Unit> - This article provides background information on Tensor Processing Units (TPUs), which are crucial in Google's AI strategy and performance enhancements, but it does not cover the latest advancements like Ironwood specifically.
7. <https://news.google.com/rss/articles/CBMiwgFBVV95cUxPeGxlVl9wY2RfbTlSQ0dqdTF0ZFcwbVROWTM1UUJraUQwR1JFSEk4NjJDaFI3VlJCRk84bUpOQ29UN1pDSWtNZk1mdkJXRkRtTDVrTnJBSWhkU0oyVkQ0UVVuRUdLRTZER242VDAwNGFnVnhNM21aeDZWNHBjT1hYaE5oLXc3YXFjUHJHSVVWalZlS0gtcVFPNXNjRkVWRkhDMVRINXc1bFQxR0hGSmFUUkNyTVEzUzg0Uk9fNklaSTRKQQ?oc=5&hl=en-US&gl=US&ceid=US:en> - Please view link - unable to able to access data