NAND RESEARCH

RESEARCH NOTE

GOOGLE TRILLIUM TPU

STEVE MCDOWELL, CHIEF ANALYST
MAY 15, 2024

CONTEXT

The Trillium TPU, Google's sixth-generation TPU, was announced at Google I/O. It promises unprecedented compute performance, memory capacity, and energy efficiency for generative AI training and inference.

GOOGLE TRILLIUM TPU

Google's new Trillium TPU delivers a 4.7X increase in peak compute performance per chip as compared to its predecessor, the TPU v5e. This increase is crucial for training and fine-tuning the most capable AI models and serving them globally with reduced latency and lower costs.

Key Enhancements of Trillium TPU include:

1. **Increased Compute Performance**: Trillium offers a 4.7X improvement in compute performance per chip. This is achieved through expanded matrix multiply units (MXUs) and increased clock speeds. Third-generation SparseCore accelerators are also introduced to optimize ultra-large embeddings, which are essential for advanced ranking and recommendation algorithms.

2. **Enhanced Memory and Bandwidth**:

    - **High-bandwidth memory (HBM)**: The capacity and bandwidth of HBM have been doubled, allowing for larger models with more weights and broader key-value caches. This next-generation HBM enhances memory throughput and power efficiency.

    - **Interchip Interconnect (ICI) Bandwidth**: Doubling the ICI bandwidth facilitates scaling training and inference jobs across tens of thousands of chips. This is supported by custom optical ICI interconnects and Google's Jupiter Networking, enabling scalability to hundreds of pods.

3. **Energy Efficiency**: Trillium TPUs are over 67% more energy-efficient than their predecessors, aligning with Google's commitment to sustainability in AI advancements.

Trillium can scale up to 256 TPUs in a single pod, offering high-bandwidth and low-latency connections. Beyond single pods, Trillium employs multi-slice technology and Titanium Intelligence Processing Units (IPUs) to enable scaling to building-scale supercomputers.

This network interconnects tens of thousands of chips through a multi-petabit-per-second data center network, which is crucial for deploying foundation models and other complex AI applications.

ANALYSIS

Google's Trillium TPU delivers notable advancements over its predecessor, the TPU v5e, particularly in areas crucial for the demands of modern generative AI applications.

The Trillium TPU demonstrates a 4.7X increase in peak compute performance per chip, a key indicator of its ability to handle AI models' increasingly complex computational demands. This upgrade is critical for reducing latency and enhancing the efficiency of AI model training and deployment on a global scale.

Additionally, Google has doubled the capacity and bandwidth of both the HBM) and the Interchip Interconnect (ICI). These improvements significantly enhance the Trillium TPU's capability to process and manage larger and more complex datasets and AI models.

Energy efficiency is another area where the Trillium TPU has made strides, with a promised 67% increase in efficiency over the TPU v5e, which is increasingly important as the energy demands of large-scale AI computations grow.

Google's Trillium TPU is a nice update to its TPU portfolio, providing much-needed (and much-desired) choice to users. The TPU's enhanced performance, greater memory capacity, and improved energy efficiency support not only Google’s AI initiatives but also provide substantial benefits to enterprises relying on high-performance AI computing infrastructure.