Special Analysis

AI's Escalating Impact on GPU/Accelerator Design

Bob Sorensen
October 2024

HYPERION RESEARCH OPINION

The rapid emergence of AI and its unique and often demanding computational requirements are having a significant impact on the overall design and development trajectory of advanced HPC component accelerators. Prior to the rise of AI starting in 2019, such accelerators, particularly graphic processing units or GPUs, were largely seen as a complementary, specialized compute resource to facilitate offloading from the system's central processing unit (CPU) a narrow set of computationally-intensive operations integral to science and engineering workloads.

- The benefits of such accelerators were twofold: they could be specifically designed to address a small but important set of computationally intensive processes while freeing up the CPU to better accomplish more general-purpose compute operations including interpreting and executing program instructions, performing logic and less used arithmetic operations, and managing data and instruction flow among memory, processors, and storage.

With the emergence of AI, however, the role of accelerators, particularly GPUs at this time, have become more central, requiring significant changes in their design and in many cases becoming the defining compute engine for an increasingly large number of HPCs.

Key accelerator/GPU architectural features that differ significantly from counterpart CPUs include:

A large number of relatively simple computational engines (cores) within a single GPU that are wholly dedicated to performing key AI-specific operations

- Due to the relatively limited variety of such required operations, some leading-edge GPUs can contain upwards of hundreds if not thousands of individual processing cores. In contrast, most CPUs have counterpart, but more general-purpose and complex, cores, with typically less than 64 cores per CPU.

Faster and larger memories than counterpart CPUs

- Owing to GPU's high core counts combined with the ability of each core to demand and process large quantities of data, GPUs require significantly more performant and costly memory schemes than those typically found in a counterpart CPU.
- Indeed, most GPU offerings provide support for high bandwidth memory (HBM), a computer memory interface for 3D-stacked synchronous dynamic random-access memory (SDRAM) initially offered by Samsung, AMD, and SK Hynix. In contrast, the bulk of CPU designs call for more prosaic, slower, and less costly DDR5 memory.

Better and faster interconnect schemes that have high bandwidth, low latency GPU to GPU connectivity

- GPU networks are becoming increasingly performant and specialized, such as NVIDIA's NVLink, that are capable of supporting data and control code transfers between CPUs and GPUs or solely among GPUs.

Another major development within the accelerator/GPU architectural space is the increasing hardware support given to addressing mixed or low precision data formats versus the larger, more precise, but more data intensive formats typical on traditional science and engineering applications. Indeed, AI researchers and AI end users are finding that many AI applications do not need the higher precision formats, such as 64 bit floating points formats, which have been the sine qua non for HPC for the last 40 years. Instead, new formats that require only 32 bit, 16 bit, or even lower floating point or integer schemes can be adequate for many AI jobs.

- As a result, GPU designers are increasingly optimizing their chip and core designs to take advantage of this trend, configuring hardware to offer increased computational performance with lower memory overhead for these mixed and lower precision AI jobs.
- That said, there is no small concern within the traditional science and engineering community that some processor vendors will increasingly de-emphasize 64 bit floating point capabilities to support lower precision AI jobs, leaving many CPU-centric configurations underpowered for jobs that demand high precision formats.

In addition to AI-centric GPUs, there is a small but innovative class of alternate pure-play AI-centric processors on the market. Vendors include Graphcore, Cerebras, and SambaNova. Each of these companies offer distinct, custom AI processors that seek to offer alternatives to NVIDIA's dominant position in AI accelerators. Due to their concentration on AI-specific technical requirements, these components seek to offer better performance and lower power requirements at lower cost than GPU counterparts. However, to date, none of the AI-centric chip providers have realized significant progress in challenging NVIDIA’s leadership position.

A number of horizontally integrated chip makers have also developed or are already marketing AI-centric accelerators, including Intel with its Gaudi chips and AMD with its MI300. Likewise, the three main cloud service providers (CSPs) are also fielding AI accelerators such as Google's Trillium TPU, Microsoft's Maia 100, and AWS' Trainium and Inferentia offerings. These latter chips, however, likely are not targeted for direct commercial availability but instead will be used to support various CSP AI-centric cloud instances.

Finally, there is a small sector of custom-bult accelerators called Application Specific Integrated Circuits (ASICs) and Field Programmable Gate Arrays (FPGAs) that can be used for AI applications. These bespoke components are typically designed for a single target end use or end user. Despite their potential performance gains over more general-purpose components, their high development cost, limited range of application, and sometimes complex programming requirements bound their overall applicability for all but the most targeted applications.

About Hyperion Research, LLC

Hyperion Research provides data-driven research, analysis and recommendations for technologies, applications, and markets in high performance computing and emerging technology areas to help organizations worldwide make effective decisions and seize growth opportunities. Research includes market sizing and forecasting, share tracking, segmentation, technology, and related trend analysis, and both user & vendor analysis for multi-user technical server technology used for HPC and HPDA (high performance data analysis). Hyperion Research provides thought leadership and practical guidance for users, vendors, and other members of the HPC community by focusing on key market and technology trends across government, industry, commerce, and academia.

Headquarters

365 Summit Avenue
St. Paul, MN 55102
USA
612.812.5798
www.HyperionResearch.com and www.hpcuserforum.com