Special Analysis

Cloud Service Providers Poised to Expand AI Accelerator Options with Custom AI Chips

Jaclyn Ludema, Mark Nossokoff, and Earl Joseph
July 2024

HYPERION RESEARCH OPINION

The cloud AI accelerator market is experiencing a period of rapid growth and innovation, propelled by the increasing adoption of AI across various industries. Factors driving the adoption of cloud accelerators include the need for scalable computing resources, cost efficiency, and the ability to leverage advanced hardware without substantial upfront investment.

For years, the AI cloud accelerator market has been dominated by just a few powerful GPU offerings, enabling the rapid advancement of artificial intelligence and machine learning. However, the growing demand for efficient and specialized hardware to power complex AI workloads has led to the rise of a trend: the development of custom AI chips by cloud service providers (CSPs). CSPs, such as AWS, Google, Microsoft, and Alibaba. They have recognized the need for more tailored hardware solutions to address the unique requirements of their cloud-based AI services and customers. By developing their custom AI chips, these CSPs aim to diversify their accelerator offerings to users, optimize the hardware-software stack for their cloud environments, and gain a competitive edge in the rapidly evolving AI cloud ecosystem.

Among the CSPs' custom AI chip initiatives, AWS's Inferentia2, Trainium2 and Google’s TPU chips have particularly gained attention for their potential to disrupt the AI accelerator market. These chips are designed for high-performance inference and training workloads and have specialized architectural features, improved power efficiency, and tight integration with their respective cloud platforms and software ecosystems. This paper will explore the rise of custom AI chip development by CSPs, with a focus on the impact that AWS's Inferentia2 and Trainium2 chips may have on the AI accelerator market. By analyzing the unique capabilities of these custom AI chips, and the factors that may influence their adoption, this paper will assess the possibility of a shift in the AI hardware landscape.

THE CLOUD AI ACCELERATOR MARKET

Hyperion Research has closely monitored the rapid growth of moving scientific computing workloads to public clouds over the past several years. Access to advanced accelerator hardware, such as GPUs and specialized AI chips, has been identified as a key driver motivating many users to leverage cloud resources, especially amidst the rapid growth and adoption of AI technologies.

Oftentimes, AI users are attracted to the major CSPs because they offer access to cutting-edge accelerators. A recent Hyperion Research study found that when evaluating whether to use accelerators in the cloud, users consider a few key criteria:

- Cost
- Security
- Access to GPUs, or lack of, on-premises
- Ease of use in the cloud
- Less overhead to maintain GPUs in the cloud
- Scalability available in the cloud

CSPs are uniquely positioned to offer both general-purpose accelerators and CSP-specific accelerators, providing their users with additional options. While general-purpose accelerators remain highly popular, the CSP-specific offerings have their appeal for users with workloads that can benefit from the customized hardware and software stacks. The market size for custom GPUs is projected to grow significantly over the next few years, with more enterprises adopting AI technologies and seeking efficient solutions for their computational needs. It will be interesting to track the adoption of all types of cloud AI accelerators as the newest iterations of these chips enter the market.

RISE OF CUSTOM AI CHIPS

The growing demand for efficient and specialized hardware to power complex AI workloads has pushed CSPs to explore alternatives to off-the-shelf solutions. By developing their custom AI chips, CSPs can tailor the hardware and software stack to their specific needs, optimizing for performance on specific types of workloads, power efficiency, and cost-effectiveness. This increased control over the technology stack also reduces the reliance on third-party chip providers, allowing CSPs more autonomy and agility in responding to the evolving demands of the AI ecosystem. In addition, it can greatly reduce the costs of large-scale data centers by using their own in-house developed processors.

Moreover, the integration of custom AI chips with the CSPs' cloud-native software and service offerings can unlock significant advantages. Tight coupling between the hardware and software offers the promise of seamless optimization, streamlined deployment, and enhanced user experiences for cloud-based AI customers.

CSP Accelerator Initiatives

CSPs are responding to this surge in AI cloud migration by developing specialized hardware such as custom-designed chips that cater to the computational demands of modern AI workloads. Key initiatives by other major CSPs in the custom AI chip space include:

- **Google Cloud:** offers the Tensor Processing Unit (TPU), designed specifically for accelerating AI training and inference. TPUs are known for their high performance in handling TensorFlow operations, making them a strong competitor in the AI chip market.
- **Microsoft Azure:** offers the MAIA 100 chip designed for AI inference tasks, designed for high throughput and low latency, particularly in NLP (Natural Language Processing) and real-time applications. Microsoft also offers the Cobalt 100 chip, aimed at AI training, which delivers significant training speed and energy efficiency.
- **Alibaba Cloud:** has also entered the AI accelerator market with its Hanguang 800, a chip designed to optimize search algorithms, image processing, and video analysis, providing strong performance for AI inference tasks.

AWS Inferentia2 and Trainium2

AWS has recently captured market attention with its announcements and benchmarking of its Inferentia2 and Trainium2 custom AI chips. Inferentia2 was announced at the 2022 AWS re:Invent, and made generally available in April 2023, with the promise of delivering up to 4 times higher throughput and up to 10x lower latency than its predecessor. Benchmarking conducted by Hugging Face compared several popular NLPs and computer vision transformer models across Inferentia2, Inferentia1, and GPU instances. The results were compelling: Inferentia2 outperformed the other options, delivering around 4.5 times better latency on average compared to the GPU. Also, the new Amazon EC2 Inf2 instances, powered by Inferentia2 chips, provide up to 2.6 times better throughput, 8.1 times lower latency, and 50% better performance per watt than comparable GPU-based instances

AWS Trainium2, announced at 2023 AWS re:Invent, targets the high-performance training of foundation models and large language models (LLMs). A recent AWS press release describes Trainium2 performing up to 4 times faster training speeds and 3 times more memory capacity compared to the first-generation Trainium chips. As of this article, Trainium2 is not yet available for general AWS use but is expected to be available later this year.

Challenges and Adoption Considerations

While the potential disruptive impact of CSP custom AI chips on the AI accelerator market is significant, several challenges and adoption considerations may influence their widespread uptake.

One of the primary concerns is ecosystem lock-in and compatibility issues. Enterprises and AI developers may be reluctant to invest in custom AI chips that are tightly integrated with a specific cloud provider, as this can limit their ability to move workloads across different platforms or leverage existing expertise and software investments. Ensuring seamless integration with popular AI frameworks and software tools will be crucial for broader adoption.

Another factor is the potential performance and feature gaps compared to the latest GPU offerings. While the CSP offerings have demonstrated impressive capabilities, the rapid pace of innovation in the AI hardware space means that GPU vendors may continue to push the boundaries of performance, potentially outpacing the custom chips developed by CSPs.

Concerns around vendor lock-in and dependence on a single cloud provider may also hinder the adoption of custom AI chips. Enterprises may be hesitant to become overly reliant on a specific CSP's proprietary hardware solutions, as this can limit their flexibility and negotiating power in the long run.

CONCLUSION

The rise of custom AI chip development by cloud service providers, exemplified by AWS's Inferentia2 and Trainium2 chips, has the potential to disrupt the long-standing dominance of GPUs in the AI accelerator market. These specialized chips are designed to offer tailored performance, power efficiency, and tight integration with cloud-native software stacks, addressing the evolving needs of the AI ecosystem.

While the adoption of custom AI chips developed by CSPs may face some challenges, such as ecosystem lock-in and concerns around vendor dependence, the potential benefits in terms of cost-effectiveness and alignment with cloud-based AI workflows could drive significant interest and uptake among enterprises and AI developers.

As the AI accelerator market continues to evolve, the introduction of custom chips like Inferentia2 and Trainium2 signals a shift in the competitive landscape. The ability of these chips to effectively challenge general-purpose GPU market leaders will depend on factors such as their performance capabilities, the maturity of the supporting software and tooling, and the willingness of the AI community to embrace new hardware solutions tailored for specific cloud-based AI workloads.

Ultimately, the ongoing transformation in the AI accelerator market, driven by the development of custom chips by CSPs, represents an exciting and dynamic phase in the evolution of artificial intelligence technology. The potential disruptive impact of AWS's Inferentia2, Trainium2 and Google’s TPU chips highlights the increasing importance of specialized and cloud-optimized hardware solutions in the rapidly advancing field of artificial intelligence.

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