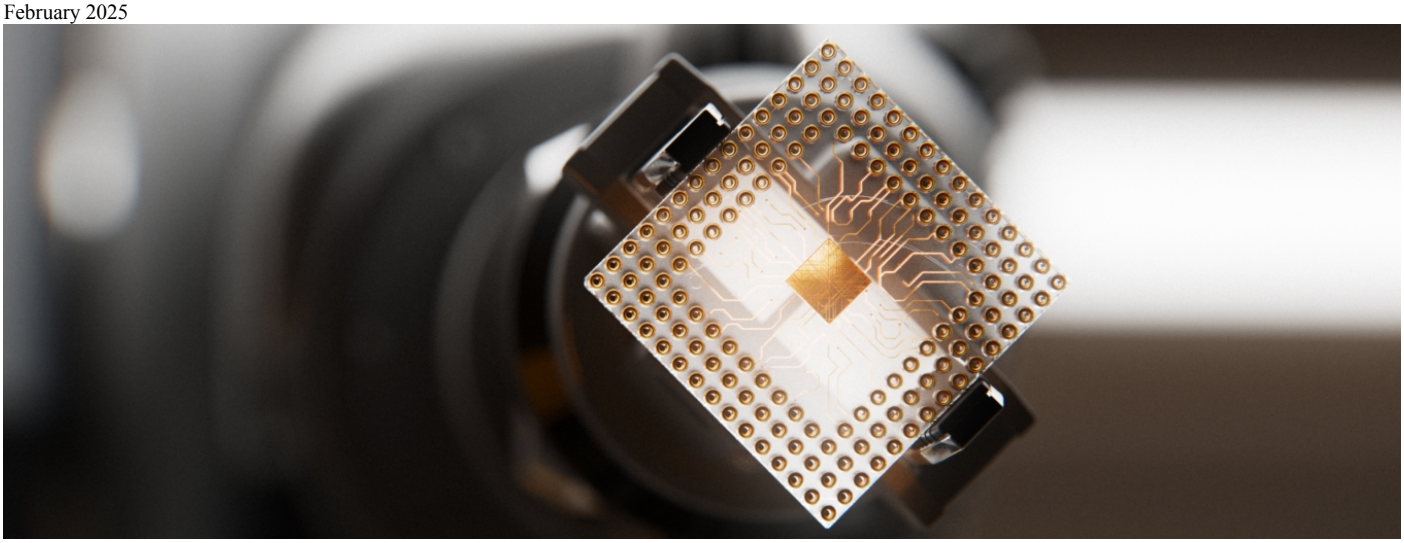
The Golden Age of Customized AI Chips

nb.com/en/us/insights/the-golden-age-of-customized-ai-chips

Article

 Yan Taw (YT) Boon, Head of Thematic – Asia
February 2025



As demand for training large-scale AI models shifts to delivering more precise inferencing capabilities, the race to build application-specific chips is heating up.

Imagine your morning coffee routine: For years you've relied on the local Starbucks, a dependable stalwart yet without that personal touch. Then up pops a cozy new shop, where the barista knows your name and exactly how you take your caffeine, all at a lower price.

So it goes with AI: As demand for training large-scale AI models shifts to delivering more precise inferencing capabilities, so is demand for generalized Graphics Processing Units (GPUs) shifting to Application-Specific Integrated Circuits (ASICs) to meet more specialized computing requirements.

Nvidia's AI GPUs have long been the standard for training large-language AI models in centralized data centers, thanks to their immense computational power and parallel-processing chops. Then came DeepSeek, a Chinese tech start-up that claims to have developed an impressive model powered by only a small fraction of the computing capacity required by more established competitors, such as OpenAI.

While that news rattled global markets, it also highlighted the intense underlying demand for more cost-efficient AI development and custom chips to support the rapid growth of AI agents—virtual assistants that can perform myriad tasks, from fielding customer service calls to whipping up computer code. (As of Thursday, January 30, shares of ASIC chipmakers—including Broadcom and Marvell—had gained back most of their losses from the sudden DeepSeek drawdown.)

We believe the rise of agentic AI is driving demand for optimized inferencing conducted within smaller, localized data centers or edge data centers positioned even closer to end users. This proximity reduces latency and accelerates processing speeds, enhancing real-time AI applications while saving costs. Barclays estimates that, by 2026, inferencing will account for more than 70% of general AI's computing needs—roughly 4.5 times more than training needs.[1]

Customized ASIC platforms are proliferating as cloud providers seek greater efficiencies. Consider:

- Amazon has introduced Trainium 2 in partnership with Marvell, achieving cost savings of 30% to 40% compared to Nvidia GPUs.[2] Amazon is also working with Marvell and Alchip on the next-generation Trainium 3 platform.
- Google recently unveiled its 6th generation of Tensor Processing Units (TPUs) through its ongoing collaboration with Broadcom.
- Meta is teaming up with Broadcom, Socionext and GUC to create various ASICs for the Meta Training and Inference Accelerator (MTIA).
- Microsoft is collaborating with Marvell and GUC on the Microsoft AI Accelerator (MAIA) to enhance the AI capabilities of its Azure cloud platform.
- Finally, U.S. President Donald Trump highlighted a joint venture with up to $500 billion invested in AI infrastructure through a partnership between OpenAI, Oracle and SoftBank.[3] We believe this initiative could herald a new ASIC ecosystem, as OpenAI is working with ARM and Broadcom on custom platforms.

Meanwhile, the race to build performance-enhancing clusters of 1 million custom XPUs—comprising CPUs, GPUs and ASICs across a single network—is heating up. Successful clusters rely on seamless data transfer which, in our view, gives companies with strong networking capabilities, such as Broadcom and Marvell, a competitive edge.

Despite the occasional caffeine jitters, we believe all of this activity will continue to present attractive opportunities for savvy thematic investors with a long-term view.

Latest Insights
SUBSCRIBE

Sources: 1) Guru Focus, AI Chip Demand to Surge by 2027, Barclays Predicts $300 Billion, December 10, 2024; 2) AIM Research, Amazon Bets Big on Trainium to Break Nvidia's AI Chip Dominance, December 12, 2024; 3) AP, Trump Highlights Partnership Investing $500 Billion in AI, January 22, 2025.