Authored by:

Tejas Dessai

Date: June 23, 2023
Topic: Thematic

# Artificial Intelligence Opens a Generational Opportunity for Chipmakers

As artificial intelligence (AI) applications expand in scope, existing computational infrastructure both in the data center and at the edge, will require rewiring to support emerging needs for data intensive computing. This shift had been in the works, but the rapid proliferation and grand potential of large language models (LLMs) are likely to accelerate the timeline. By the end of this decade, annual spending on AI chips is projected to grow at a compound annual growth rate (CAGR) of over 30% to nearly $165 billion.[1] Graphic processing unit (GPU) based chip providers will benefit the most and remain the primary choice, driven by superior performance and developer preference. Other associated hardware, including memory and networking gear, could benefit from this wave as well.

NVIDIA'S DATA CENTER SEGMENT BUSINESS HAS GROWN TO $15 BILLION IN ANNUALIZED REVENUE
Sources: Nvidia Investor Relations, Quarterly Revenue Trend, FQ1 2024. Accessed on June 15, 2023.
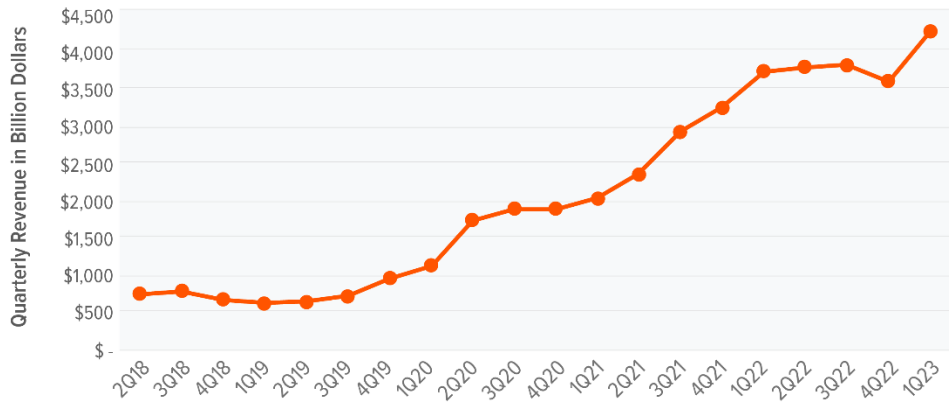
Key Takeaways

- AI applications are expanding in scope and will require new computing infrastructure, leading to significant investment in AI chips, memory, and networking gear.

- While the majority of AI chips go into the data center and on edge networks, device-based AI processors could become common as well, opening new markets for chip providers.

- AI infrastructure providers are looking at investments and partnerships to secure themselves against shortages.

Infrastructure Spending Strong as Demand for AI Services Rises

Large language models went mainstream in 2023. Foundational models like GPT-4, available as both application programming interfaces (APIs) and via chat-based interfaces, are catalyzing a whirlwind of user and developer activity, making AI a definitive battleground for technology and non-technology enterprises.

OUR ETFs
RESEARCH
ABOUT
CONTACT
NEWS
PRIVACY POLICY

1

As adoption picks up steam, so too is the demand for AI infrastructure providers. Nvidia, one of the primary suppliers of chips to train AI models, grew its data center segment revenues by nearly 14% Year-over-Year (YoY) in Q1 2023, and revised guidance to the upside by over 50% YoY for Q2.[2] Similarly, Microsoft management claimed that even though growth in their broad cloud broadly is cooling off, they expect AI-as-service and machine learning workloads to boost growth in near term.[3]

The demand for AI raises concerns about the sustained availability of AI hardware necessary to train, test, and deploy models.[4] We believe that we're likely at the beginning of a potential long capital investment cycle that drives record growth for AI hardware and infrastructure providers. A major part of that spending will go towards replacing legacy computing stacks in data centers and on the edge with hardware that can effectively support data-intensive computing and the broad adoption of AI.

Hyperscale cloud services providers grew capital infrastructure at better than 20% growth rates over the last five years, supporting broad enterprise digital transformation mandates.[5] And growth in AI-based infrastructure upgrades could tick upwards over the next five years, creating a vibrant opportunity for datacenter semiconductor vendors.

Artificial Intelligence Makes GPUs a Must-Have

Traditional processors, also known as CPUs, were designed to perform general purpose computing tasks efficiently in a single cycle.[6] GPUs are purpose-designed to process less-complex but multiple logic-based computations in parallel, which makes them incredibly efficient for data heavy processing, and thus popular for AI training and inference tasks.[7]

HIGH LEVEL ARCHITECTURAL DIFFERENCES BETWEEN CPUS AND GPUS
Sources: Global X ETFs with information derived from FS Community Blog. (2022, June 1). Deep Comparison Between Server CPU and GPU.
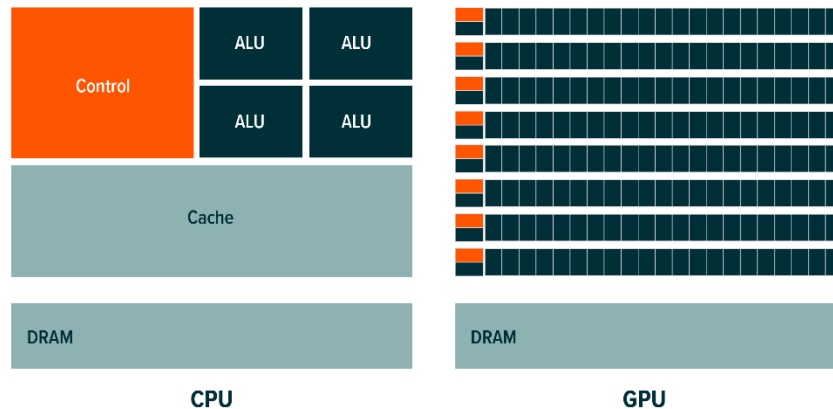

Nearly $16 billion worth of GPUs went into AI acceleration-related use cases worldwide in 2022.[8] Nvidia is the dominant supplier, owning nearly 80% of the market share in GPU-based AI acceleration.[9] Nvidia's investments in software frameworks, including the CUDA architecture ensures that developers and engineers can harness computational efficiency from the GPUs, giving the company a sustained edge over its competition, which lacks similar software support.

Outside of GPUs, field-programmable gate arrays (FPGAs) and application-specific integrated circuits (ASICs) make up the rest of the AI chip market. Together, FPGAs and ASICs resulted in roughly $19.6 billion sales for 2022, growing an estimated 50% YoY.[10] Over the next eight years, combined spending is expected to grow at a nearly 30% annual growth rate to top $165 billion by 2030.[11]

## AI CHIPS COULD GROW TO NEARLY $165 BILLION IN REVENUES BY 2030

Sources: Global X Estimates with information derived from StratView Research. Artificial Intelligence Chip Market. Accessed on June 15, 2023. *Indicates forecast.
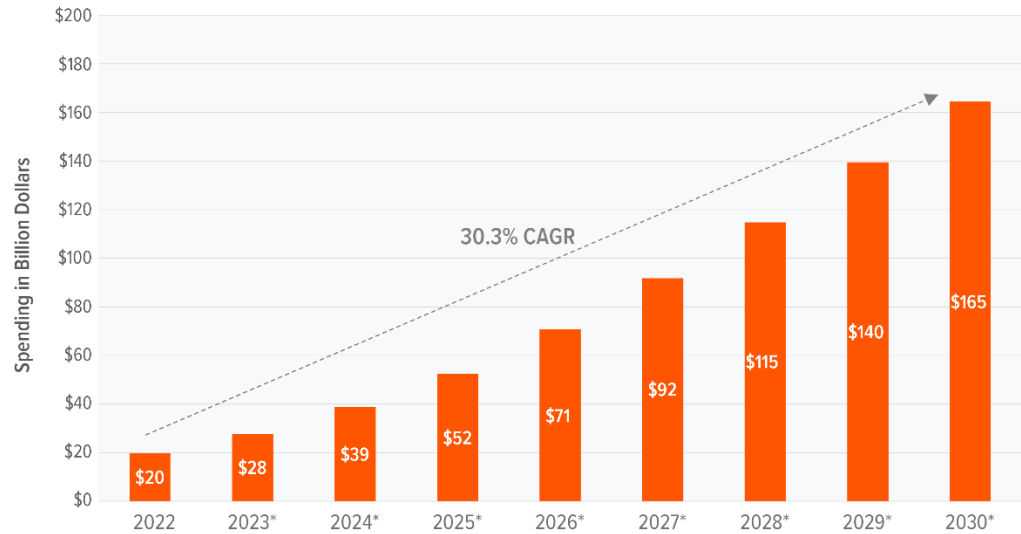
As AI chips gain prominence inside the data center, we believe spending on premium CPUs could slow as CPUs get demoted from the primary processing unit to a mere control unit, essentially directing and distributing workloads between clusters of GPUs. This dynamic would put pressure on premium data center suppliers like Intel, which built multi-billion franchises selling CPUs over the past three decades while tilting that market in favor of low-cost CPU providers such as Qualcomm and Marvell.

WORLDWIDE SPENDING ON DATA CENTER SYSTEMS IS PROJECTED TO TOP $224 BILLION IN 2023

Sources: Gartner Press Release. (2023, April 6). Garner Forecasts Worldwide IT Spending To Grow 5.5% in 2023. *indicates forecast.
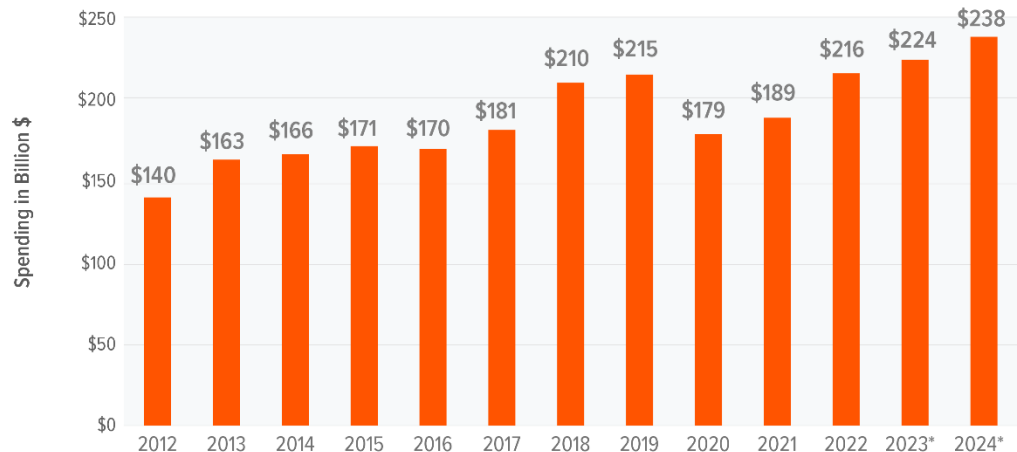
IT buyers spend close to $224 billion on data center systems annually, including processors, memory and storage, networking components, and other hardware products.[12] Assuming data center systems are

replaced every four years, more than a trillion dollars' worth of computational infrastructure will likely be up for replacement over the next few years, presenting a generational opportunity for AI-first hardware to gain ground amid the rapid adoption of generative AI solutions.

Outside the data center and high-performance computing use cases, AI chips in vehicles for autonomous driving, edge computing, and mobile devices could further boost demand for specialized hardware, including networking equipment, high-speed memory, and cooling systems.

Chip Shortage Fears Can Spur Investments and Partnerships

Developers and engineers are rushing to bring AI-first products to market amid a shortage of AI processing hardware. Orders for new GPUs have a six-month backlog, and prices for Nvidia's A1000 and H1000 lineup of GPUs have shot up significantly, with chips being sold in secondary marketplaces at hefty premiums.[13,14]

Anticipating a supply crunch and overdependence on chip suppliers, hyperscalers have been working on developing their own AI processing hardware, and this in-house capability is proving critical in handling generative AI demand. For example, Alphabet runs its Bard AI system on natively designed chips called the Tensor Processing Units (TPUs), which are also available for outside customers through the Google Cloud service.[15] Microsoft has also been working on its own AI chip alongside a low-cost ARM based CPU.[16] Private market ventures also want to capitalize on the looming scarcity for AI compute. In the past five years, over $6 billion in outside capital has been invested in AI chip ventures.[17]

Partnerships are another rising trend to address the chip shortage. Microsoft entered a multi-billion-dollar deal with GPU-as-a-service provider Core Weave, which buys and then supplies GPU power as a service to the broader market.[18] Amazon Web Services partnered with Nvidia to allow predictable and stable access to Nvidia's Hopper GPUs to Amazon's cloud customers. Dedicated instances will combine Nvidia's hardware with AWS networking and scalability solutions to deliver up to 20 exaflops of compute performance for developers.[19]

AI is still in its early innovators and experimental phase, so most demand for hardware comes from technologically advanced companies at this point. But we believe the entire semiconductor value chain, including foundries, chip designers, and semi equipment suppliers can benefit in the near term as AI spreads. Multi-trillion-dollar markets such as advertising, e-commerce, digital media and entertainment, online services, communications, and productivity are likely to ramp up spending on turnkey hardware setups as well. We also believe the industry will likely be prone to deal activity, with large accelerator vendors looking to acquire adjacent component companies to solidify research and development and product innovation.

Conclusion: There's No AI Without Specialized AI Hardware

The AI boom will likely spur a data center upgrade cycle that favors a new computing stack with the GPU at its core. The rapid proliferation of LLMs will likely result in exponential demand for AI processing and accelerated spending on specialized chips, which could open a hundred-billion dollar plus market for GPUs in the near future.[20] Meanwhile, large cloud hyperscalers will likely continue to invest in R&D to build and deploy chips of their own, in a bid to reduce dependence on large chip providers and reduce costs. AI chip demand may be lumpy, but we believe that the semiconductor value chain is well positioned to capture this opportunity and create a potential investment alternative as AI penetrates new markets.

Footnotes

1. Global X Forecast with information derived from StratView Research. (2023, Jan). Artificial Intelligence Chips Market Projected to Reach $127.7 Billion by 2028.
2. Nvidia Corporation Inc. (2023, May). NVIDIA Quarterly Revenue Trend Q1 2023 Earnings.
3. Ina Fried. Axios. (2023, April 27). AI boosts enthusiasm in Microsoft, but cloud is what's driving sales.
4. Alex Blake. Digital Trends. (2023, May 23). Nvidia GPUs see massive price hike and huge demand from AI.
5. Silver Linings. (2023, January 23). AWS, Google Cloud, Microsoft continue their capex climb.
6. FS Community. (n.d.). Deep Comparison Between Server CPU and GPU. FS Blog.
7. Ibid.
8. Global X Estimates with estimates derived from Statista Intelligence. (2023). AI Chip Market Report.
9. Fierce Electronics. (2022, January 20). Nvidia has an 80% share of AI processors, Omdia says.
10. Global X Estimates
11. Ibid.
12. Gartner. (2023, April 6). Gartner forecasts worldwide IT spending to grow 5% in 2023.
13. Digital Trends. (2023, May 12). Nvidia AI GPUs See Massive Price Hike and Huge Demand from AI.
14. Kan, M. PC Mag. (2022, September 21). Nvidia CEO: High Prices for GPUs Are Here to Stay.
15. Reuters. (2023, April 5). Google Says Its AI Supercomputer Is Faster, Greener Than Nvidia.
16. Zoe Schiffer. The Verge. (2023, April 18). Microsoft reportedly working on its own AI chips that may rival Nvidia's.
17. Greenberg, A. (2023, May 3). Chipmakers Are in Hot Demand as Venture Capitalists Bet Big on AI.
18. CNBC. (2023, June 1). Microsoft Inks Deal With CoreWeave to Meet OpenAI Cloud Demand.
19. NVIDIA Newsroom. (2023, April 20). AWS and NVIDIA Collaborate on Next-Generation Infrastructure for Training Large Machine Learning Models and Building Generative AI Applications.
20. Gartner. (2023, April 6). Gartner forecasts worldwide IT spending to grow 5% in 2023.