DOI:10.1145/3631537

Michael Cusumano

Technology Strategy and Management

NVIDIA at the Center of the Generative AI Ecosystem—For Now

Assessing the ascent of NVIDIA.

GENERATIVE AI HAS attracted worldwide attention as a foundational technology with almost unlimited applications (see my previous column, "Generative AI as a New Innovation Platform," *Communications*, October 2023). Goldman Sachs estimates applications of this technology could raise global GNP by $7 trillion (7%) during the next decade.[9] At the center of the new ecosystem is NVIDIA, whose high-end graphical processing units (GPUs) account for approximately 80% of the market for GPUs that power generative AI software.[5,20]

Established in 1993, NVIDIA's founders, led by Jen-Hsun (Jensen) Huang, initially saw a need for powerful specialized chips that could take over graphics processing from PC or workstation central processing units (CPUs), a market dominated by Intel. The company went public in 1999 and exceeded $1 billion in revenue in 2002. In its most recent quarter, NVIDIA reported sales of $13.5 billion (double the prior year) and net profits of $6.2 billion. It is now the world's most valuable semiconductor company, with a market cap surpassing $1 trillion, compared to $159 billion for AMD and $154 billion for Intel, its top competitors. This column explores several questions behind NVIDIA's extraordinary history.

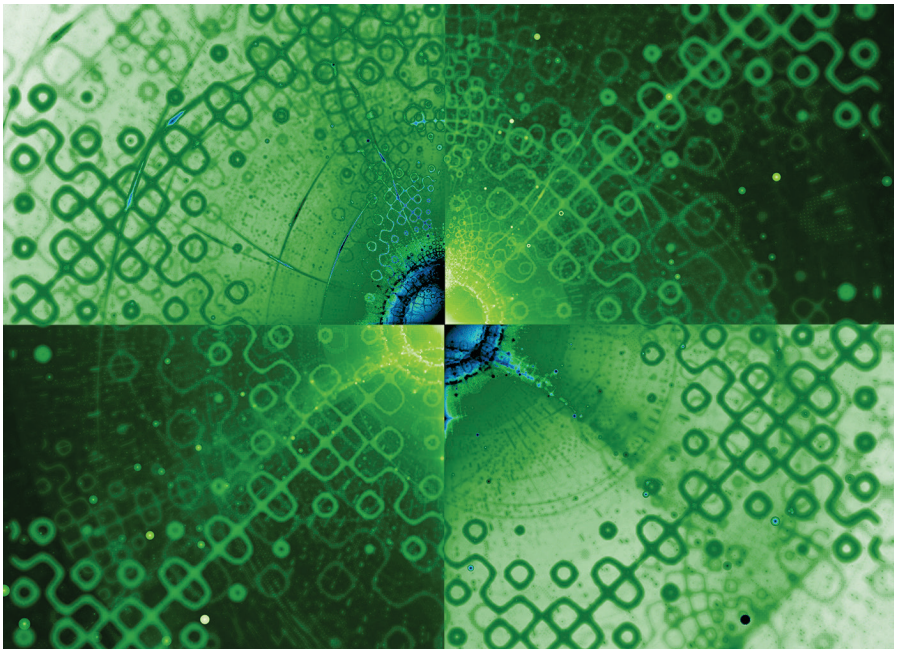

IMAGE BY ANDRIJ BORYS ASSOCIATES, USING SHUTTERSTOCK

Why Has NVIDIA Dominated the GPU Market?

First, NVIDIA early on introduced architectural innovations that made its GPUs the hardware of choice, initially for gaming and then for many other applications. The key product introduction dates to 2006, with the G80 Tesla series GPU. NVIDIA switched from arrays of a few specialized compute cores (sub-processors) that could perform complex tasks independently, as in a CPU, to an array of many more simple cores running twice as fast or faster. Each core could handle a few pixels on a graphics display or perform many specific tasks in parallel. This new design was 100% faster than NVIDIA's previous generation. Its next Fermi microarchitecture, released in 2010, was eight times faster, with many more compute cores.[31]

Second, also in 2006, NVIDIA introduced a new programming model

Demand for NVIDIA GPUs during the past few years has exceeded supply.

and language for its GPUs with a free software development kit (SDK) called CUDA, for Compute Unified Device Architecture. CUDA started as an extension of C/C++ to support fast parallel processing by directly accessing instruction sets in the GPU hardware.[8] An abstraction layer isolated the software from other underlying hardware, enabling CUDA to run on different PCs, workstations, and servers—as long as they incorporated NVIDIA GPUs as graphics cards or part of the server stack. Although not a direct comparison of device speeds, according to NVIDIA data from 2006–2008, programs written using CUDA with its next GeForce 8 series GPUs were 100 to 400 times faster than programs running on the general-purpose Intel Xeon CPUs.[31]

Third, we keep finding new ways to deploy GPUs as accelerators, and NVIDIA has facilitated this expansion of use cases with industry-specific versions of CUDA.[16] We now see NVIDIA GPUs not only in gaming, artificial intelligence and machine learning (AI/ML), and generative AI software, but also in cryptocurrency mining, virtual reality applications, self-driving vehicles, robotics, and datacenter cloud services. In 2023, gaming was still the company's largest single source of revenue (18%), though datacenters accounted for half of revenues and were on pace to reach 85% by 2024.[13]

Why Are GPUs so Useful for AI/ML and Generative AI Applications?

CPUs typically have dozens or at most a few hundred compute cores that can perform complex tasks; GPUs have many thousands of simpler compute cores that operate in parallel. The GPU architecture is perfectly suited to the huge number of matrix multiplication tasks and logic layers that lie at the heart of neural networks.

Back to 2006: Researchers in France used NVIDIA graphics cards to train their neural networks.[4] More famous work later occurred at the University of Toronto during 2011–2012 ("AlexNet").[31] NVIDIA closely followed these developments and invested heavily in software tools and libraries for building deep-learning applications, such as cuDNN (CUDA Deep Neural Network), released in 2014.[22]

In 2016, NVIDIA introduced its Pascal microarchitecture, targeting the high-performance computing market and datacenters hosting ML/AI and other compute-intensive applications. Now, NVIDIA was able to sell rack severs costing tens of thousands of dollars, not just PC graphics cards. NVIDIA priced its top-end DGX-1 server at $129,000 and even marketed this as an "AI supercomputer in a box." To stimulate the applications ecosystem, NVIDIA donated several servers to universities as well as to OpenAI, then organized as a non-profit research laboratory.[31] OpenAI would go on to partner with Microsoft in 2019 and introduce ChatGPT in November 2022. Overall, since 2017, when Google's work on language transformers grabbed the attention of the AI/ML community, NVIDIA has invested aggressively in optimizing its GPUs and CUDA software for LLMs and inference engines.[3]

In 2022, NVIDIA released its latest Hopper microarchitecture (named for programming pioneer Grace Murray Hopper). The new GH200 systems include more CPU-like capabilities as well as thousands of compute cores and staggering amounts of memory, all meant to "supercharge" generative AI applications.[11]

What Has Kept NVIDIA's Sales and Profits so High?

Demand for NVIDIA GPUs during the past several years has exceeded supply, leading to high GPU prices and profits, even though recent shipments have slowed.[14] U.S. government restrictions on exports of advanced technology also may reduce future revenues, especially since China accounts for 20% to 25% of NVIDIA's datacenter sales.[23] Nonetheless, as of late 2023, NVIDIA claimed an installed base of more than 500 million GPUs, with thousands of CUDA-based applications.[15] The company's H100 processors, introduced in 2022, cost approximately $40,000 each and are essential purchases for datacenters, which represent a trillion-dollar market.[5,20]

Network effects between NVIDIA's GPU platform and third-party applications also create a kind of flywheel, fueling demand. The growing installed base of NVIDIA hardware, particularly in datacenters such as Amazon Web Services (AWS), Microsoft Azure, and Google Cloud, enables more developers to build more CUDA-based applications. CUDA software only runs on NVIDIA GPUs (unless there is a virtual machine emulation layer, which degrades performance). Rising application usage, such as for training LLMs and running inference engines built with CUDA software, requires more NVIDIA hardware. The positive feedback loops resemble what Intel and Microsoft achieved with Intel x86 microprocessors and free Microsoft SDKs paired with PCs running DOS and then Windows.[34] In this case, NVIDIA dominates both the hardware and software sides of the platform.

What Competition Does NVIDIA Face in Hardware?

AMD has targeted NVIDIA's H100 with its MI300X line, specifically designed for generative AI. AMD GPUs may already be slightly ahead of NVIDIA in terms of price-performance, but only for gaming.[29] Intel in 2019 acquired Israel's Habana Labs for $2 billion and then in 2022 introduced the Gaudi2

Network effects between NVIDIA's GPU platform and third-party applications also create a kind of flywheel, fueling demand.

chip, which targets NVIDIA's H100 as well. Intel's product line does particularly well in inference processing.[6,12] Several startups, led by SambaNova and Cerebras, also have raised billions of dollars to design new generations of GPU platforms.[2]

Cloud service providers have been building their own systems to reduce their GPU purchases. Google introduced its famous Tensor Processing Units (TPUs) for in-house use in 2016 and then third-party use in 2018. These have relatively limited software compared to CUDA and require Google Cloud.[28] However, Google TPUs and its JAX AI library, introduced in 2018, reportedly outperform NVIDIA systems in some applications.[3,10] AWS introduced its Trainium machine-learning accelerator in 2020, optimized for deep-learning training, with some software support.[19] Microsoft intended to release a custom AI chip for its datacenters in late 2023.[30] Meta/Facebook also has an in-house GPU and supercomputer effort under way.[32]

What Competition Does NVIDIA Face in Software?

Software is the "moat" that keeps users from switching away from NVIDIA hardware, with some 250 CUDA libraries widely used by GPU programmers.[7] Still, NVIDIA has vulnerable spots. Some programmers complain CUDA is proprietary and not open source (cannot access and modify the source code) as well as difficult to use if you are not familiar with C/C++. NVIDIA has recently introduced support for more popular languages, including Python (PyCuda).[17,21] Of course, programmers can use other languages and avoid CUDA entirely, though they would have to recreate all the CUDA drivers, libraries, and other tools, and they lose direct access to the NVIDIA GPU instruction sets.

A major weakness with AMD and Intel has been their limited GPU software support.[33] As a competitive move, both companies have made drivers and libraries open source. Cooperation with the open source community should help AMD and Intel evolve their software assets faster, but this will still take years.[26]

Other open source frameworks exist for GPU programming, such as OpenCL, introduced in 2009 and based on C.[24] New languages include OpenAI's Triton, introduced in 2021 and based on Python.[27] Triton seems to work especially well with PyTorch 2.0, an open source machine-learning library used to train deep-learning models, originally developed at Meta/Facebook.[25] Triton still requires a CUDA compiler, but it avoids CUDA propriety libraries in favor of open source alternatives. Future versions should run on Intel, AMD, and other GPU hardware.[18]

NVIDIA is at the center of the generative AI ecosystem and is likely to remain there for several years.

Conclusion

NVIDIA is at the center of the generative AI ecosystem and is likely to remain there for several years. However, competitors (for example, AMD and Intel) and users (for example, datacenters and the open source community) are actively developing or exploring alternatives. If NVIDIA GPUs remain scarce and expensive, users will find substitutes or ways around NVIDIA's proprietary software. Datacenters also may turn to cheaper hardware, including CPUs, to host less-demanding generative AI software, such as smaller, focused LLMs and inference engines dedicated to specific tasks.[35] C

References

1. Anirudh, V. Forget ChatGPT vs Bard, the real battle is GPUs vs TPUs. *Analyticsindiamag.com* (Feb. 8, 2023).
2. Bradshaw, T. Start-ups seek to challenge NVIDIA's dominance over AI chip market. *Financial Times*. (July 21, 2023).
3. Bradshaw, T. and Waters, R. How NVIDIA created the chip powering the generative AI boom. *Financial Times*. (May 26, 2023).).
4. Chellapilla, K. et al. High performance convolutional neural networks for document processing. In *Proceedings of the 10th Intern. Workshop on Frontiers in Handwriting Recognition*. (Oct. 2006) Université de Rennes 1, La Baule France.
5. Fitch, A. NVIDIA sales surge as AI boom lifts earnings. *Wall Street J.* (Aug. 23, 2023).
6. Freund, K. Intel Gaudii2 looked to be a credible alternative to NVIDIA. Until..." *Forbes.Com* (Sept. 11, 2023).
7. Gallagher, D. How NVIDIA got huge—and almost invincible. *Wall Street J.* (Oct. 6, 2023).
8. Geeks for Geeks. Introduction to CUDA programming. *Geeksforgeeks.org* (Mar. 14, 2023).
9. Goldman Sachs Research. Generative AI could raise global GNP by 7%. *Goldmansachs.com* (Apr. 15, 2023).
10. Hampton, J. Google claims its TPU v4 outperforms NVIDIA A100. *Datanami* (Apr. 5, 2023).
11. Moore, R. NVIDIA unveils enhanced AI chip configuration to supercharge generative AI applications. *Linkedin.com/pulse* (Aug. 9, 2023).
12. Moore, S.K. NVIDIA still on top in machine learning; Intel chasing. *IEEE Spectrum* (Sept. 18, 2023).
13. Morgan, T. NVIDIA: There's a new kid in datacenter town. *Nextplatform.com*, (Aug. 24, 2023).
14. Mujtaba, H. GPU Shipments Continued to Decline in Q1 2023: NVIDIA at 84%, AMD at 12%, Intel at 4% Market Share. *Wccftech.com* (June 8, 2023).
15. NVIDIA Corporation. About CUDA; https://developer.NVIDIA.com/about-cuda
16. NVIDIA Corporation. Solution Areas; https://developer.NVIDIA.com/solutions-and-industries
17. NVIDIA Corporation. CUDA Python; https://developer.NVIDIA.com/cuda-python
18. Patel, D. How NVIDIA's CUDA monopoly in machine learning is breaking—OpenAI Triton and PyTorch 2.0 *SemiAnalysis.com*, January 16, 2023.
19. Rand, C. A first look at AWS trainium. *Towardsdatascience.com* (Nov. 28, 2022).
20. Reuters. Explainer: Why are NVIDIA's shares soaring and what is its role in the AI boom? *Reuters.com* (June 14, 2023).
21. Saturn Cloud. Python GPU programming—A guide for data scientists. *Saturncloud.io/blog*, (June 13, 2023).
22. Serrano, J. NVIDIA Introduces cuDNN, a CUDA-based library for deep neural networks. *InfoQ* (Sept. 29, 2014).
23. Shah, A. Fearing China, U.S. blocks the sale of NVIDIA GPUs to the Middle East. *Hpcwire.com* (Sept. 1, 2023).
24. Smistad, E. Getting started with OpenCL and GPU computing. *Eriksmistad.no* (June 21, 2010).
25. Solegaonkar, V. Introduction to PyTorch. *Towarddatascience.com*. (Sept. 20, 2019).
26. Tiernan, R. Chip industry is going to need a lot more software to catch NVIDIA's lead in AI. *ZDNET* (Oct. 21, 2020).
27. Tiernan, R. OpenAI proposes open-source Triton language as an alternative to NVIDIA's CUDA. *ZDNET* (July 28, 2021).
28. Tovar, J. GPUs vs TPUs: A comprehensive comparison for neural network workloads. *Linkedin.com/pulse* (Apr. 1, 2023).
29. Walton, J. AMD vs NVIDIA: Who makes the best GPUs? *Tomshardware.com* (June 16, 2022).
30. Warren, T. Microsoft reportedly working on its own AI chips that may rival NVIDIA's. *The Verge* (Apr. 18, 2023).
31. Watkins, M. et al. NVIDIA: Wining the deep-learning leadership battle. *IMD Case 980* (Apr. 15, 2019).
32. Wiggers, K. Meta bets big on AI with custom chips—and a supercomputer. *TechCrunch* (May 18, 2023).
33. Wodecki, B. AMD vs. NVIDIA: The battle for GPU supremacy begins. *Aibusiness.com*. (June 21, 2022).
34. Yoffie, D.B. and Cusumano, M.A. *Strategy Rules: Six Timeless Lessons from Bill Gates, Andy Grove, and Steve Jobs*. Harper Business, New York, 2015.
35. Yoffie, D. et al. AI21 labs in 2023: Strategy for generative AI. Harvard Business School Case # 9-724-383 (Sept. 13, 2023).

Michael A. Cusumano (cusumano@mit.edu) is a professor and Deputy Dean at the Massachusetts Institute of Technology Sloan School of Management, Cambridge, MA, USA, coauthor of *The Business of Platforms* (2019), and a member of the MIT Center for Quantum Engineering (https://cqe.mit.edu/).

The author thanks Vivek Farias, Aidan Mattrick, Rama Ramakrishnan, Sarah von Bargen, and David Yoffie for their many helpful comments.

For information on NVIDIA's GPUs for confidential computing, see p. 60.