Dean&Francis

ISSN 2959-6130

Nvidia in the AI Chip Market Competition: Based on SWOT Model and Financial Analysis

Fangyi Shao[1, *]

[1]School of Business, Macau University of Science and Technology, Macau, China, 999078

*Corresponding author: sfy041110@outlook.com

Abstract:

The semiconductor industry is the core of the digital information age and the technological linchpin of global connectivity. Its technological innovation and application expansion represent the cutting edge of contemporary scientific and technological advancement. In recent years, with the rise and proliferation of artificial intelligence technologies, particularly generative AI, the global semiconductor technology industry is undergoing unprecedented transformation, giving rise to numerous industry giants and emerging players. Among these industry leaders, NVIDIA has leveraged its years of technological accumulation and strategic positioning in the GPU field to gain significant competitive advantages. It has emerged as a core participant in the new wave of AI, surging to become the world's most valuable company by market capitalization. Its development trajectory and market performance have drawn widespread societal attention. This article focuses on NVIDIA's data center business, analyzing its development trajectory and the characteristics of the AI chip industry. It employs a SWOT model to examine NVIDIA's internal strengths and external environment, concluding with a summary of NVIDIA's performance and financial status, providing an objective overview of its market position and current development. Through this analysis, NVIDIA is expected to maintain its market position by sustaining R&D investment, deepening its software ecosystem, optimizing supply chain management, and proactively addressing geopolitical challenges. This strategic approach positions the company to sustain high-growth momentum in the increasingly competitive AI chip market.

Keywords: NVIDIA; AI chip; SWOT analysis; financial analysis

1. Introduction

The semiconductor industry is the pivot of today's digital revolution, powering the cutting-edge technologies that make our world hyper-connected. Increasing demands for smart, rapid, and efficient computing continue to drive semiconductor innovation to new frontiers of possibility. In the next decade, the world semiconductor market will see explosive growth driven by disruptive technologies like artificial intelligence (AI), autonomous cars, 5G networks, and the Internet of Things (IoT). Of these, AI semiconductors, or AI chips, are proving to be a game-changer, offering peak processing and power efficiency for demanding machine learning and deep learning applications [1]. Currently, artificial intelligence technologies, particularly generative AI represented by large language models, are spearheading a new wave of innovation, reshaping the global landscape of technological competition and industrial development pathways. The meteoric ascent of large language models—epitomized by OpenAI's GPT lineage—is catapulting AI from narrow, task-specific expertise toward versatile, broadly general intelligence, igniting global technological rivalry and sweeping industrial upheaval [2]. Graphics processing units (GPUs) excel at handling computationally intensive tasks involving large-scale data sets due to their massive parallel processing capabilities, and their importance continues to grow. The GPU-accelerated applications are used in computational physics, computational chemistry, life sciences, medical imagining, mathematics (linear algebra, differential equations solvers), computer vision, signal and image processing and many more [3]. As the undisputed leader in the GPU sector, NVIDIA has benefited from the latest wave of artificial intelligence, experiencing rapid expansion in company scale. In July 2025, NVIDIA surpassed a market capitalization of $4 trillion, becoming the world's most valuable company. Its development journey and strategic approach hold significant reference value for technology innovation enterprises. This paper focuses on NVIDIA's data center business, employing research methods such as SWOT analysis to examine its internal and external environments. It discusses the impact of strategic choices on business data and provides a comprehensive analysis of the company's current status.

2. Overview and Development History of NVIDIA Corporation

2.1 Company Profile

NVIDIA is a leading company in the technology and semiconductor industries, synonymous with innovation and investment potential. This American technology firm specializes in visual computing technology and is widely recognized as the inventor of the graphics processing unit (GPU). The GPU is a high-performance processor that has revolutionized graphics processing across various platforms, from workstations to mobile devices [4]. After over three decades of technological accumulation, NVIDIA has emerged as the undisputed leader in AI chips amid the AI wave. Centered around its GPU technology, the company has built a full-stack computing solution encompassing hardware, software, and ecosystem support. Its core businesses span data centers, gaming and AI PCs, professional visualization, automotive, and robotics. NVIDIA's products are widely deployed in large language model training, intelligent driving, and robotics applications. NVIDIA is emerging as a central player in the new wave of AI innovation.

2.1.1 The early development process of the company

NVIDIA Corporation was founded in January 1993. Its initial core business focused on the design of graphics chips and chipsets, specializing in fabless IC semiconductor design. Headquartered in Santa Clara, California, USA, the company operates globally. In May 1995, NVIDIA launched its first multimedia accelerator for game consoles, the NV1. This integrated chip combined graphics, sound, and controller driver functions, earning favor from Sega of Japan. However, technical compatibility issues with the mainstream Windows Direct3D graphics programming interface caused significant setbacks in its commercial rollout. In 1997, NVIDIA introduced the RIVA 128 2D processor, which received an unprecedented market response. In 1998, it established a strategic partnership with TSMC, adopting a fabless model focused on chip design and sales. In 1999, NVIDIA released the world's first fully functional GPU, the GeForce 256. That same year, NVIDIA went public on the NASDAQ stock exchange with a market capitalization of $230 million. In 2000, NVIDIA acquired the former graphics card giant 3DFX. In 2006, NVIDIA introduced the CUDA standard, a milestone that provided developers with a direct programming environment, ushering in the era of GPU general-purpose computing (GPGPU). Through years of promotion and sustained investment, CUDA and the GPU general-purpose computing capabilities it represents gained widespread market recognition. This established NVIDIA's proprietary software ecosystem, cementing its position as the global leader in programmable graphics processing technology. In 2018, NVIDIA introduced RTX technology, integrating ray tracing, artificial intelligence, and other innovations to achieve deep integration of photorealistic rendering, AI capabilities, and precise physics simulation.

3. Industry Overview

At present, the AI chip market is undergoing rapid development and transformation, with the global market size expanding at a fast pace and holding significant potential for further growth. In 2024, the global AI chip market reached approximately $90.2 billion, with projections indicating it could hit $400 billion by 2027, drawing intense scrutiny from capital markets. Characterized by high market barriers and rapid technological iteration, this article examines these industry traits through the lens of Moore's Law in semiconductors and the capital expenditure strategies of major industry players.

3.1 High Market Threshold

AI tasks impose different performance requirements on chips compared to general-purpose chips. General-purpose chips, such as central processing units (CPUs), have strong sequential operation capability, but they cannot provide sufficient performance for techniques like DNNs that require intensive parallel computation and high-bandwidth memory. Specialized AI chips can be up to thousands of times faster than CPUs for training and inference of DNNs [5]. On the other hand, manufacturing AI chips requires continuous breakthroughs from advanced process technologies to advanced packaging techniques. Simultaneously, as chip scale rapidly expands, the complexity of AI chip design and R&D investment are rising sharply. As the global leader in advanced chip manufacturing processes, TSMC has already launched mass production of its 3nm process and is expected to commence mass production of its 2nm process by the end of 2025, securing a core competitive advantage.

The software ecosystem forms the moat of the AI chip industry. After years of accumulation, NVIDIA's CUDA platform has built a vast developer ecosystem with a large user base, creating extremely high switching costs and network effects that make it difficult for latecomers to challenge.

In terms of industrial chain development, synergies within the AI chip industry chain are increasingly evident across the entire value chain—from raw material supply, chip design, manufacturing, and packaging/testing to final product applications [6]. The upstream and downstream segments of the AI chip industry are largely controlled by industry giants, resulting in high market concentration. For instance, companies like TSMC and Samsung Electronics dominate advanced manufacturing processes, while NVIDIA and AMD hold nearly the entire market share for discrete GPUs. Internet giants such as Microsoft and Google, leveraging their capital reserves and network effects, serve as the primary forces driving the training and development of large AI language models.

In terms of talent demand, the AI chip industry faces unprecedented demand for highly skilled professionals. An increasing number of universities are establishing AI-related majors and courses, forging deep partnerships with enterprises to set up joint laboratories and internship bases, and implementing comprehensive talent development programs that bridge theory and practice.

3.2 High Iteration Speed

In 1975, Gordon Moore, one of the co-founders of Intel Corporation, made a prediction that the number of transistors integrated onto a chip would double every two years. This widely circulated prophecy became known as Moore's Law. It turned out that Moore was a visionary thinker who correctly foresaw that semiconductor technology would advance at an astonishing pace. Although the details of how we shrink transistors have changed considerably over the past few years, many of Moore's predictions about the rapid development of integrated circuits have come true [7]. However, in recent years, with the explosive growth of chip computing power, Moore's Law appears to no longer align with market developments. In 2023, NVIDIA Chief Scientist Dally stated that over the past decade, GPU AI inference performance has increased by a factor of 1,000—far exceeding Moore's Law projections.

Faced with such rapid performance improvements and iteration speeds, chip companies must maintain high R&D investment ratios to preserve their hardware performance advantages. Looking at the 2024 R&D spending of major global semiconductor firms, Intel leads with $16.546

billion, while NVIDIA and AMD invested $12.5 billion and $6.4 billion, respectively. Sustained large-scale R&D investments will continue to widen the technological gap between industry giants and other competitors, further accentuating the monopolistic competitive landscape within the sector.

4. SWOT Analysis

4.1 Strength

4.1.1 Hardware Architecture

First, NVIDIA continuously updates its GPU hardware architecture, providing a robust hardware foundation to meet the computational demands of the AI era. From the initial Tesla and Fermi architectures to the Hopper and Blackwell architectures of the AI era, NVIDIA has consistently maintained its leading position in GPU hardware architecture. Take the Blackwell architecture released by NVIDIA in 2024 as an example. It introduces a disruptive design innovation in GPU chip architecture—transitioning from single-die to "dual-die" configurations—significantly breaking through the performance limitations imposed by single-chip GPU computing power. From a yield perspective, while maintaining consistent overall performance, the low manufacturing yield of single-chip designs leads to higher production costs. Consequently, the Blackwell architecture offers significantly improved economic efficiency compared to previous generations.

4.1.2 Software Ecosystem

The network effects generated by its massive user base constitute one of NVIDIA's core competitive advantages. Since the introduction of CUDA in 2006, after over a decade of accumulation, the CUDA developer community had surpassed 5.9 million members by the end of fiscal year 2025, wielding extensive and profound influence within the developer ecosystem. By leveraging CUDA, developers can harness the powerful parallel computing capabilities of NVIDIA GPUs to accelerate a wide range of computationally intensive tasks, including scientific computing, data analysis, and artificial intelligence [8]. The vast number of programs and projects developed using CUDA not only makes it the optimal choice for newcomers entering the field but also imposes significant migration costs on existing developers.

4.2 Weakness

NVIDIA's fabless model results in relatively weak control over upstream supply. This highly centralized supply chain structure leaves the company vulnerable to disruptions such as sudden supply cuts from core manufacturers. For instance, NVIDIA is deeply reliant on TSMC for advanced chip manufacturing processes and CoWoS packaging. Data from 2024 indicates that NVIDIA is TSMC's second-largest customer, contributing 11% of its revenue. Analysts predict that by 2026, NVIDIA's demand for CoWoS wafers will reach 595,000 units, accounting for 60% of global demand [9]. So far, TSMC's semiconductor fabrication plants (usually called "fabs") have been centered in Taiwan, where technology and human resources are accumulated. Behind the superior technological capabilities of TSMC is the presence of a large number of young, professional, and hard-working engineers and a group of experienced managers at TSMC and a concentration of material and equipment suppliers. All of them are located in Taiwan [10]. This centralized production layout ensures reliability in chip manufacturing, facilitates quality control, and improves product yield rates. However, it also means that if affected by unforeseen events such as natural disasters, TSMC's production capacity would be significantly reduced, dealing a heavy blow to NVIDIA's supply chain.

4.3 Opportunity

The diversity of compute-intensive applications in modern cloud data centers has driven the explosion of GPU-accelerated cloud computing. Such applications include AI deep learning (DL) training and inference, data analytics, scientific computing, genomics, edge video analytics and 5G services, graphics rendering, and cloud gaming [11]. The AI market remains a burgeoning blue ocean, with widespread optimism about its growth potential. It serves as the core growth engine for NVIDIA's performance. According to a Cargoson report, the global artificial intelligence market is projected to reach $244 billion by 2025 and is expected to grow to $827 billion by 2030, with a compound annual growth rate (CAGR) of 27.2%. Among these, generative AI stands as one of the fastest-growing segments, with its market size already reaching $33.9 billion in 2024. It is anticipated to account for 33% of all AI software spending by 2027 [12]. The continuous expansion of market share has mitigated the intensifying market competition to some extent, providing NVIDIA with am-

ISSN 2959-6130

ple room for performance growth.

4.4 Threat

Under the threat of international geopolitical tensions, NVIDIA's procurement and sales operations face adverse impacts. For NVIDIA, the Chinese market ranks as its fourth-largest globally. According to its fiscal year 2025 financial report, revenue from China operations reached approximately $17 billion, accounting for roughly 13% of its total revenue. Jensen Huang previously stated that China's AI market will grow to $50 billion within the next two to three years, and if NVIDIA were excluded, "it would be a tremendous loss." Discussing future strategy, he emphasized that NVIDIA will continue to increase its investment in the Chinese market rather than maintain the status quo, because "competition is fierce, and rivals are investing heavily " [13]. The significance of the Chinese market within the global AI landscape is evident. In August 2025, although the U.S. government reinstated export licenses for H20 chips, it required NVIDIA to remit 15% of sales revenue as tax, which undoubtedly squeezed NVIDIA's profit margins in the Chinese market. Furthermore, in July 2025, China's Cyberspace Administration of China (CAC) held discussions with NVIDIA regarding the security protocols for H20 chips. Concerns over "vulnerability backdoor security risks" garnered widespread market attention, prompting some enterprises to adopt a cautious approach toward purchasing NVIDIA chips. Overall, considering China's massive and continuously expanding AI market share, coupled with high policy uncertainty, the stability of NVIDIA's performance faces a severe test.

On another hand, competition is one of the biggest risks that NVIDIA may have to face in the near future. It is known that NVIDIA operates in a highly competitive environment, as AMD and Intel are its major competitors, and both of them are investing heavily in AI hardware [14]. Vertical business expansion by downstream customers may reshape the competitive landscape of the AI chip market. Driven by considerations such as supply security and computational power shortages, downstream internet companies are developing their own chips, such as Google's Trainium chip and Microsoft's Azure Maia AI accelerator. These initiatives may further intensify competition in the AI chip market and potentially erode Nvidia's market share.

5. Company Performance

NVIDIA's total revenue for the second quarter of fiscal year 2026 reached $46.7 billion, setting another all-time high. This represents a 6% year-over-year increase and a 56% quarter-over-quarter growth, exceeding guidance expectations. Considering factors such as geopolitical influences, this performance is undoubtedly impressive. By business segment, Data Center revenue reached $41.1 billion, significantly outpacing Gaming ($4.3 billion), Professional Visualization ($601 million), and Automotive and Machinery ($586 million), accounting for 88% of total revenue [15]. Data center business revenue grew from $14.5 billion in the third quarter of fiscal year 2024 to $41.1 billion in the third quarter of fiscal year 2026, achieving a compound annual growth rate (CAGR) of 68%. Data center business revenue surged from $14.5 billion in Q3 FY2024 to $41.1 billion in Q3 FY2026, achieving a 68% CAGR [16], serving as the core growth driver of the company's performance. Given that NVIDIA's data center business accounts for nearly 90% of its revenue structure, it is reasonable to use overall financial data to reflect the financial performance of its data center operations.

In terms of profitability metrics, NVIDIA has maintained its consistently strong earnings power. For the second quarter of fiscal year 2026, NVIDIA reported operating revenue and net income of $30.2 billion and $25.8 billion, respectively, representing year-over-year increases of 51% and 52%—growth rates largely in line with revenue expansion. Reflected in gross margin, the synchronized growth in revenue and profit demonstrates stable profitability. NVIDIA's second-quarter gross margin stood at 72.4%, maintaining a high level significantly above other traditional chip manufacturers (AMD's gross margin was 40%, and Intel's was 29.7% during the same period). This highlights NVIDIA's superior pricing power and cost control capabilities relative to its competitors.

From a cash flow perspective, NVIDIA's total cash, cash equivalents, and marketable securities reached $56.8 billion in the second quarter of fiscal year 2026, representing a 6% year-over-year increase and a 63% quarter-over-quarter rise—a growth rate largely in line with revenue expansion. As of the end of Q2, NVIDIA held total cash of $56.8 billion and total debt of $8.5 billion, demonstrating the high sustainability and resilience of its business model. This position also supports its next phase of production scale expansion and R&D investment. In summary, NVIDIA has demonstrated strong financial performance, with its data center business serving as the core driver of growth. This segment generated nearly 90% of total revenue and grew faster than the company as a whole, becoming the primary engine of expansion. Against direct rivals, NVIDIA retains formidable competitive edges: sustained, exceptional profitability, robust pricing power across its products and services, and rigorous cost control. Finally, NVIDIA's overall financial position remains robust, with ample cash and other short-term assets providing solid financial backing to navigate market uncertainties and support further investments.

6. Conclusion

This paper combines NVIDIA's development history and the characteristics of the AI chip industry, employing the SWOT model and financial analysis methods. Focusing on the data center business, it provides a systematic analysis of NVIDIA's current status in the AI chip market. Analysis reveals that through early strategic investments in GPU hardware/software and GPGPU technologies, NVIDIA established and progressively expanded its competitive edge in the GPU market, building a robust industry moat. In recent years, capitalizing on the explosive growth of the AI industry, NVIDIA has achieved significant expansion in company scale and influence, surging to become the world's highest-valued company. However, behind this prosperous development lie certain hidden risks. NVIDIA's fabless model, focused on chip design, has generated substantial economic benefits. Yet this approach has created a highly concentrated supply chain, exposing the company to risks such as supplier production cuts. Geopolitical risks, including Sino-US technological competition, and shifts in market dynamics also pose external threats, potentially leading to under-performance relative to expectations. From a performance and financial standpoint, NVIDIA has delivered impressive results, with its data center business serving as the core growth engine driving rapid expansion. Concurrently, its robust financial position ensures NVIDIA's capacity for future R&D investment and market expansion, playing a significant positive role in maintaining competitive advantages and enhancing resilience against market risks. Looking ahead, the burgeoning blue ocean of the AI chip market will open up vast growth opportunities for NVIDIA. By maintaining its technological leadership in chip design, deepening its developer software ecosystem, and optimizing supply chain management through diversified approaches while proactively addressing external threats like geopolitical tensions, the author believes NVIDIA is poised for a new phase of growth. It is expected to sustain its dominant position amid intensifying competition in the AI chip market, offering valuable insights for other high-tech enterprises through its developmental trajectory and strategic initiatives.

References

[1] Talati, D. (2021). Silicon minds: The rise of AI-powered chips. International journal of Science and Research Archive 1, (02):097-108.

[2] Wei, W., Zeng, Z., & Liu, L. (2025). From the DeepSeek breakthrough to see the innovation paradigm, challenges and responses of China's artificial intelligence industry. Economic Review, (06), 102-114.

[3] Mišić, M. J., Đurđević, Đ. M., & Tomašević, M. V. (2012, May). Evolution and trends in GPU computing. In 2012 Proceedings of the 35th International Convention MIPRO. IEEE, pp. 289-294.

[4] Zhang, J. W. (2012). Research on the development strategy of NVIDIA Corporation. Master's thesis, Lanzhou University.

[5] Pang, G. (2022). The AI chip race. IEEE Intelligent Systems, 37(2), 111-112.

[6] Evans, Gao, H. L., & Zhai, T. (2025). Analysis of the current status and competitive landscape of AI chip technology. Robot Industry, (03), 33–40.

[7] Li, W., & McKenzie, J. (2023). Moore's Law may be touching physical and economic boundaries. World Science, (12), 51–55.

[8] Pang, W., Wang, J., & Weng, C. (2024). GPGPU and CUDA unified memory research status survey. Computer Engineering, 50(12), 1–15.

[9] Eastmoney. (2025, August 6). U.S. stock market pre-trading: Three major index futures edge up, Li Auto surges over 6%. East money Finance. From https://stock.10jqka.com.cn/usstock/20250806/c670170301.shtml

[10] Momoko, K. (2022). Taiwan's TSMC as a focal point of US-China high-tech conflict. Asia-Pacific Review, 29(1), 5-12.

[11] Choquette, J., Gandhi, W., Giroux, O., Stam, N., & Krashinsky, R. (2021). Nvidia a100 tensor core gpu: Performance and innovation. IEEE Micro, 41(2), 29-35.

[12] Cargoson. (2024, July 10). How big is the AI market? Statistics and facts. Cargoson Blog. From https://www.cargoson.com/en/blog/how-big-is-the-ai-market-statistics

ISSN 2959-6130

[13] Dou, S. P. (2025, July 21). Starting from Jensen Huang's visit to China: What does the Chinese market mean? Shanghai Securities News · China Securities Net. https://news.10jqka.com.cn/20250721/c669779333.shtml

[14] Kalera, G., Wane, D., & Esteves Dietzold, B. (2025). NVIDIA and the future of AI infrastructure: A financial and strategic analysis of a high-growth tech stock. Available at SSRN 5327765.

[15] NVIDIA. (2025, August 27). NVIDIA announces financial results for second quarter fiscal 2026. NVIDIA Newsroom. From https://nvidianews.nvidia.com/news/nvidia-announces-financial-results-for-second-quarter-fiscal-2026

[16] NVIDIA. (2025). Revenue by market quarterly trend (Q2 fiscal 2026). NVIDIA Investor Relations. From https://s201.q4cdn.com/141608511/files/doc_financials/2026/Q226/Rev_by_Mkt_Qtrly_Trend_Q226.pdf