Proceedings of the 10th International Conference on Economic Management and Green Development
DOI: 10.54254/2754-1169/2026.38260

# *The Economics of Scarcity and Ecosystem Lock-in: NVIDIA's Premium Pricing, Market Power and Consumer Willingness to Pay in the AI Accelerator Market*

**Henan Wang**

*Canadian International School of Guangzhou, Guangzhou, China*
*evsae@icloud.com*

***Abstract.*** This study analyzes NVIDIA's high pricing model in the AI chip market by combining the concepts of product differentiation, market power, and consumer behavior. It suggests that NVIDIA has established very high switching costs with the CUDA ecosystem, making the price elasticity of demand very low, if not approaching zero, allowing them to charge a price well above the marginal cost. While this approach is based on using excess profits to invest in technological progress, it has also brought about demand-side issues such as distortions in the allocation of resources resulting from the "computing capacity tax" and consumer surplus reductions due to hardware-and-software bundling. At the industry level, it has exacerbated the excessive concentration of the market and created systemic risks. To address this, this study recommends steps to optimize these measures, like setting up interoperable technical standards, moving towards transparent tiered-pricing regulations, and enhancing the resilience of the supply chain. These measures are designed to maintain a balance between the corporate incentives for innovation, market fairness, and efficiency, and to offer a theoretical reference for the healthy development of the AI industry ecosystem and inclusive growth in the global digital economy.

***Keywords:*** Ecosystem Lock-in, AI Accelerator Market, Switching Costs, Compute Tax, Contestable Markets

## 1. Introduction

### 1.1. Research background

With the rise of generative AI, the computing power landscape in the world has changed. As the foundation of the large-scale AI models, GPUs have become the "new oil" of the digital economy era. In high-performance computing, NVIDIA is the one that is well cultivated and has almost gained a monopoly. In the fourth quarter of 2023, NVIDIA accounted for over 80% of the global discrete graphics card market share and over 90% of the market share for data center artificial intelligence acceleration chips, data sources indicate. This power has enabled it to charge $30,000 to $40,000 for the unit price of its most popular product, the H100, during the period of 2023 and 2024. Although manufacturing costs are high, NVIDIA's gross profit margin has always been very

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2754-1169/2026.38260

high, above 70%, which is much higher than that of traditional semiconductor manufacturers. It has followed a new pricing principle away from the traditional "marginal cost plus". The high-end pricing strategy has had a significant impact on the survival costs of downstream cloud vendors and AI startups, and the global technology industry chain layout. Chips export controls have added to supply chain tensions in a time of global economic inflation and geopolitical competition.

Hence, examining how NVIDIA has been able to sustain its premium price in a complicated setting and analyzing consumers' willingness to pay is valuable for understanding the distribution mechanism of the digital economy, delineating the limits of market power, and ensuring a sustainable development of the AI industry.

## 1.2. Literature review

There are many studies in academia concerning corporate pricing strategies. To understand the brand premium of NVIDIA, Sam states that there are three main factors to premium pricing: higher quality, unique brand image, and scarcity [1]. According to the industrial organization theory of Ji et al., a market power firm often adopts the product differentiation strategy to reduce the influence of the substitute product and make its price high by increasing the switching costs [2]. Guo et al. focus on the semiconductor industry and explain the concept of "architectural lock-in," which is the result of making a technological standard (e.g., CUDA) into a common language, making the consumers less price sensitive [3]. Lampinen's behavioral economics research also shows that people are likely to be irrational towards "brand worship" and "performance anxiety" when they are buying complex technological products, which justifies their preference towards paying a high price for certain brands [4]. But the majority of current research focuses on pricing mechanisms from one point of view. Few consider the interplay of technological ecosystems, market dynamics, and consumer behavior to systematically explore how premium pricing has reached the levels seen in the AI chip industry. Previous research largely deals with pricing models for traditional manufacturing or consumer goods, or with the evolution of the semiconductor technology only, and either on the effects of supply-demand imbalance on prices or on the external limitations on prices by antitrust policy. In contrast, little has been focused on how internally NVIDIA has created an ecosystem moat by combining hardware and software, and pulled the consumers' willingness to pay to the brink with the shortage of AI computing resources.

Furthermore, there are relatively underdeveloped micro-economic studies of consumers' sense of "being forced to take" higher prices and quantitative studies of consumers' switching costs. The disintegration between theory and reality, namely, no research into the interactions of technology and markets, and psychology, is the main gap that is addressed by this study.

## 1.3. Analytical framework

This study adopts a research approach of "phenomenon description, theoretical analysis, problem diagnosis, and policy recommendations" to provide a detailed description of NVIDIA's current business operations and pricing practices to serve as an empirical foundation. Second, it adopts the product differentiation theory to explore, on the basis of the CUDA ecosystem, a technological barrier that the firm constructs to achieve competitive advantages. Third, it uses consumer behavior theory to analyze the factors that lead to less price elasticity of demand, namely, high switching costs and network externalities. Finally, it outlines the potential risks of the premium pricing strategy and offers suggestions to improve from the angles of industrial policy and corporate strategy, aiming to provide a theoretical reference for the healthy and balanced development of the AI industry.

DOI: 10.54254/2754-1169/2026.38260

## 2. Case profile: pricing power in the age of accelerated computing

Founded in 1993 and headquartered in Santa Clara, California, NVIDIA established a strong brand presence in the gaming market through its GeForce series of graphics cards in its early years [5]. The launch of the CUDA parallel computing platform in 2006 marked the company's strategic transformation from a hardware manufacturer into a provider of integrated hardware-and-software solutions, laying the foundation for its dominance in the AI era. NVIDIA has now developed a product portfolio centered on data centers. The H100 chip, based on the Hopper architecture, integrates more than 80 billion transistors and delivers 1,979 TFLOPS of FP8 computing performance, while the next-generation GB200, based on the Blackwell architecture, further overcomes computing bottlenecks through NVLink interconnect technology. Because of constraints on TSMC's CoWoS packaging capacity, demand for NVIDIA's high-end chips has consistently exceeded supply. Between 2023 and 2024, the delivery lead time for the H100 once exceeded 36 weeks, while its spot-market price carried a premium of nearly 100% over the official list price.

In terms of pricing strategy, NVIDIA has abandoned the traditional "per-core pricing" model in favor of "system-level value-based pricing." For example, a DGX H100 server is priced at approximately USD 370,000. The company also implements market segmentation through customized chip variants and software subscription services. By dealing directly with cloud-computing giants such as Microsoft and Google, NVIDIA generated more than USD 47 billion in data-center revenue in fiscal year 2024, effectively gaining control over the lifeblood of global AI computing capacity. As summarized in Table 1, NVIDIA's pricing power is reflected in its financial performance and supply-demand imbalance.

Table 1. NVIDIA data center business performance and market metrics (FY2023–FY2024)

| Metric | Value | Note |
|---|---|---|
| Global AI Accelerator Share | >90% | Training segment, 2023 |
| Discrete GPU Share | ~80% | Desktop, Q4 2023 |
| H100 DGX System ASP | ~$370,000 | List price |
| H100 Spot Price Premium | +80% to +100% | vs. MSRP, 2023 |
| H100 Delivery Lead Time | ≤36 weeks | Peak demand |
| Data Center Revenue (FY24) | $47.5B | YoY +217% |
| Gross Margin (GAAP) | 73–76% | FY24 average |
| R&D Intensity | ~20% | Of the total revenue |

## 3. Economic analysis of the premium pricing strategy

While NVIDIA's premium pricing model has been hugely successful and commercial, the company's justification for premium pricing, which transcends cost considerations, has given rise to several structural microeconomic tensions. This section examines in detail the positive market signals arising from this strategy and the issues that it can mask, using the theories of product differentiation, market power, and consumer behavior.

DOI: 10.54254/2754-1169/2026.38260

## 3.1. Efficiency rationales and positive externalities

### 3.1.1. The innovation–profit feedback loop

On the supply side, NVIDIA's high pricing has created a lot of positive incentives. As per the microeconomic theory of production, in the presence of a high value of marginal profits, firms will have the funds to invest in high-risk R&D. NVIDIA has reinvested in R&D the massive profits it has made from its data-center business, spending over 20% of the total revenue on R&D in fiscal year 2024. This virtuous cycle of "premium pricing"–"high R&D investments"–"increasing technological lead" has further entrenched the company's edge in process technology, including 4N process optimization, and architectural design, including the Transformer Engine, while also setting a standard for the computing performance of the entire AI sector [6]. The high-priced products offer customers access to the latest and most powerful hardware, ensuring efficient and stable training of large models and objectively supporting the rapid development of Generative AI technology.

### 3.1.2. Network effects and ecosystem standardization

NVIDIA's whole system, the hardware and the software, around CUDA is strongly networked. The consumers' high willingness to pay is partially driven by "ecosystem stickiness": when developers widely use the technology to build programs, the number and variety of software libraries become enormous, thus further decreasing the cost of programming new CUDA programs for consumers [7]. While NVIDIA's pricing is high, its standardized CUDA platform lowers the hidden costs that enterprises incur when adapting software and recruiting talent for the NVIDIA platform. This is a market standard that, with NVIDIA dominating the market, has actually effectively decreased technological fragmentation in the early years of AI advancement and boosted collaboration productivity within the sector.

## 3.2. Structural distortions and welfare costs

Despite the positive impacts discussed above, NVIDIA's strategy of using its market power to sustain extreme premium pricing has created more serious problems on the demand side and within the industrial structure. These problems are primarily reflected in the following three areas:

### 3.2.1. Quasi-inelastic demand and the "compute tax"

In a perfectly competitive market, the price should be found to reflect the supply/demand of the market and should direct the allocation of resources. But NVIDIA has constructed very high switching costs with the CUDA ecosystem, leading to the price elasticity of demand for chips approaching zero. No matter how high the prices are (several times higher than the production cost), hyperscale cloud service providers and AI startups are still compelled to buy NVIDIA products; they will lose the existing software ecosystem, and their development efficiency will be lost as well [8]. This "compelled willingness to pay" has led to overblown resource allocation distortions – downstream businesses spend significant capital expenditure (CapEx) on buying computing power, at the expense of investments in data cleaning, algorithm optimization, and talent development. Macroeconomically, this amounts to a significant "computing capacity tax" on the AI innovation chain and will affect the democratization and diversified development of AI technology.

DOI: 10.54254/2754-1169/2026.38260

### 3.2.2. Monopoly pricing, bundling and consumer surplus erosion

NVIDIA's "system-level pricing," which involves "bundling" of hardware and software as a strategy, is a standard monopoly pricing strategy. The higher the index according to Lerner, the higher the degree of market power of the firm and the more its prices will deviate from MC. NVIDIA's gross margin is currently above 70%, meaning that they are charging more for their products than they need to earn a normal rate of return. This policy greatly reduces consumer surplus. The steep entry barrier for such small and medium-sized enterprises and research institutions prevents them from playing a part in state-of-the-art AI research and development, resulting in unfair market opportunities [9]. Furthermore, NVIDIA's "tiered pricing" and "special-market variant" policies—disguising themselves as performance throttling—actually allow NVIDIA to keep prices artificially down, while formally avoiding the risk of geopolitical policies, as the H20 does. This not only damages the interests of consumers but also causes an inefficient use of the computing resources of the world.

### 3.2.3. Concentration risk, X-inefficiency and speculative misallocation

According to Industrial Organization theory, when a market is monopolistic, the monopoly firm may experience "X-inefficiency" and a falling incentive for innovation. With more than 90 percent market share in the AI training chip business, NVIDIA may not be pressing for innovation or focusing on enhancing the energy efficiency of its products if they don't have to. More seriously, this is a very concentrated market structure, which means that it poses high systemic risks [1]. If the world starts relying too heavily on the chips and closed ecosystem of one supplier, a chip supply-chain issue, a technological weakness, or a strategic move at NVIDIA could make the entire AI industry vulnerable to being paralyzed. In the meantime, the price of computing capacity has made it necessary for some AI firms to abandon technological development and resort to "computing capacity arbitrage," or marketing of speculative ideas. This speculative landscape isn't conducive to the healthy and sustainable growth of the AI sector. Additionally, NVIDIA's price-setting power allows it to profit-maximize using price discrimination, ultimately driving up the prices of digital transformation across the globe and creating a digital divide.

To sum up, the issues surrounding NVIDIA's high pricing are not just business-related but also come down to industry-level issues related to fair competition and the innovation ecosystem. The issues are urgent and must be resolved through effective market regulation and optimization.

## 4. Policy and strategic recommendations

### 4.1. Interoperability standards to dissolve ecosystem lock-in

The main problem with the resource allocation distortions is the price elasticity of demand, which is extremely low, so the key is to break the extremely high switching costs in the CUDA ecosystem. Regulatory authorities should encourage and/or mandate that dominant computing capacity providers adhere to the interoperability principle. Specifically, the interconnection mechanisms that are applied in the telecommunications industry could be encouraged in the machine-learning sector, for example, by disseminating the adoption of open-source formats such as ONNX (Open Neural Network Exchange). Introduce policy subsidies or tax incentives to promote cross-platform compatible code library development by software developers to decrease downstream enterprises' reliance on a sole proprietary ecosystem. However, when the switching costs become much smaller,

DOI: 10.54254/2754-1169/2026.38260

the demand in the AI chip market will become price elastic, and NVIDIA will have to charge closer to the marginal cost. This would help to reduce the crowding-out effect of the "computing capacity tax" on innovation resources and shift market resources towards core value-creating activities like optimizing algorithms and improving data quality.

### 4.2. Transparent value disclosure and antitrust safeguards

To overcome the loss of consumer surplus and market welfare, actions should be taken from the corporate self-regulation and external supervision points of view. NVIDIA needs to be more open about cost and value at the corporate level by improving its "system-level value-based pricing" strategy. The suggestion is to offer basic research institutions and the education sector chips at a significant discount in an "academic-edition" tier of prices. This will help the company to meet its social responsibilities and help reduce the digital divide. As for the regulatory aspects, the performance throttling strategies imposed on NVIDIA's "special-market variant" chips should be carefully scrutinized as an "exploitative abuse" under the Anti-Monopoly Law. Regulators could implement a "fair transaction price" assessment model by looking at the ratio of performance to price of similar products in competition, and stop the enterprises from setting an unreasonable difference in price by artificially lowering the performance of the product. NVIDIA's hardware and software should also be carefully examined to ensure consumers have the freedom to choose their operating system and that NVIDIA's bundling does not impede competition and consumer choice, if it reduces their surplus.

### 4.3. Supply-side contestability and compute infrastructure resilience

The answer lies in a diverse set of market suppliers and in increasing the resilience of supply chains to mitigate the innovation-stifling effect of over-concentrated market power and the systemic risk that is generated.

Firstly, downstream cloud service providers (CSPs) should take a multi-sourcing approach. Even if competing products are marginally less capable, providers should set aside some of their purchases for alternatives like the MI300 series chips from AMD, Intel Gaudi accelerators, or, in fact, in-house ASIC chips, so as to have leverage against NVIDIA's pricing. Second, governments should view computing infrastructure as a public good and allocate more funding to research and development of open-source architecture for the instruction set – RISC-V and homegrown AI chips. A fund for breakthroughs in computing technologies should be created to help reduce barriers to entry into the computing industry. Finally, a global reserve of computing capacity and an emergency response mechanism, like strategic petroleum reserves, should be established to lessen the risk from supply disruptions from a single supplier. These would effectively make a contestable market a reality, forcing the monopolist to keep investing in innovation, to not allow X-inefficiency, and to ensure the long-term stability and healthy development of the global AI industry ecosystem.

## 5. Conclusion

### 5.1. Summary of findings

This study examines the rationale underlying NVIDIA's premium pricing in the AI chip market and reveals that its primary foundation lies in the high switching costs created by the CUDA ecosystem. These switching costs have caused the price elasticity of demand to approach zero, enabling

DOI: 10.54254/2754-1169/2026.38260

NVIDIA to sustain price premiums far above marginal cost. Although this strategy uses excess profits to drive technological advancement and has accelerated the development of the AI industry in the short term, it has also produced significant adverse effects. The substantial "computing capacity tax" distorts resource allocation, hardware-and-software bundling erodes consumer surplus, and excessive market concentration creates systemic risks and suppresses innovation. In response, this study proposes measures such as establishing interoperable technical standards, implementing transparent tiered-pricing regulations, and strengthening supply-chain resilience. These measures are intended to reduce ecosystem lock-in, introduce contestable market mechanisms, and balance corporate incentives for innovation with market fairness and efficiency.

### 5.2. Practical implications

The practical significance of this study lies in providing regulators with a microeconomic theoretical basis for identifying and regulating monopolistic conduct in digital markets. It also offers guidance to upstream and downstream enterprises in the AI industry chain when formulating cost-control and risk-management strategies. By demonstrating the necessity of ecosystem openness and diversified competition, this study can help transform AI computing capacity from a "capital-intensive luxury" into an "inclusive form of infrastructure", thereby promoting inclusive growth in the global digital economy.

### 5.3. Limitations and avenues

This study is, however, based mainly on secondary data for its qualitative analysis and was not able to get into details about the cost structure of NVIDIA or detailed data on the willingness of consumers to pay. This constraint compromises, to a certain extent, the accuracy of the conclusions that can be drawn. Future research may be performed to gather primary data that will be based on questionnaires and in-depth interviews with the corporate procurement decision makers. Econometric models could then be used to generate more exact quantitative estimates of the price elasticity of demand for the computing capacity and switching costs, lending additional support and further development for the conclusions of this study.

## References

[1] Sam, R. (2025). How Sirripa addresses the irrelevance of the P/E ratio in sectoral comparison and reveals hidden market rationality in equity valuation—the cases of Palantir, Nvidia, and Micron Technology. *Journal of Mathematical Finance, 15*(4), 8. https://doi.org/10.4236/jmf.2025.154033.

[2] Ji, H., Jiang, H., Ma, W., Johnson, D. S., Myers, R. M., & Wong, W. H. (2008). An integrated software system for analyzing chip-chip and chip-seq data. *Nature Biotechnology, 26*(11), 1293-1300. https://doi.org/10.1038/nbt.1505.

[3] Guo, H., Shen, Z., Chen, Y., & Dong, M. (2025). Analyzing the impact of government R&D subsidy and digital transformation on supply chain risk dynamics management and firm performance in China's chip industry. *International Journal of Production Economics, 281*(0). https://doi.org/10.1016/j.ijpe.2025.109524.

[4] Lampinen, M. (2024). NVIDIA's Blackwell unleashes the AI revolution. *Automotive World Magazine* (TN.52). https://www.automotiveworld.com/articles/nvidias-blackwell-unleashes-the-ai-revolution/

[5] Alqubaisi, S. N., Alqadi, A. M., Ameri, S. A. A., Alzaabi, N. A., & Nobanee, H. (2025). NVIDIA and Amid: Financial analysis amid the AI revolution and geopolitical challenges. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.5073807.

[6] Lu, L.-T. (2026). Nvidia's competitive strategy in the AI era: A case study of competitive advantage and strategic choice. *International Journal of Management and Commerce, 8*(2), 57–68. https://doi.org/10.66856/ijmc.2026.8.2.8049

DOI: 10.54254/2754-1169/2026.38260

[7] Philippidis, A. (2024). GTC 2024: Nvidia highlights AI' revolution' in drug discovery, genomics. *GEN Edge, 6*(1), 244–248. https://doi.org/10.1089/genedge.6.01.047.
[8] Wang, Y. (2025). NVIDIA's comprehensive market analysis: Technologies and competitors. *Finance & Economics, 1*(11). https://doi.org/10.61173/ja8wsk61.
[9] Xu, K. (2025). NVIDIA's research and development investment: Impact on financial performance and market valuation. *Advances in Economics, Management and Political Sciences, 148*(1), 109–117. https://doi.org/10.54254/2754-1169/2024.ld19178.