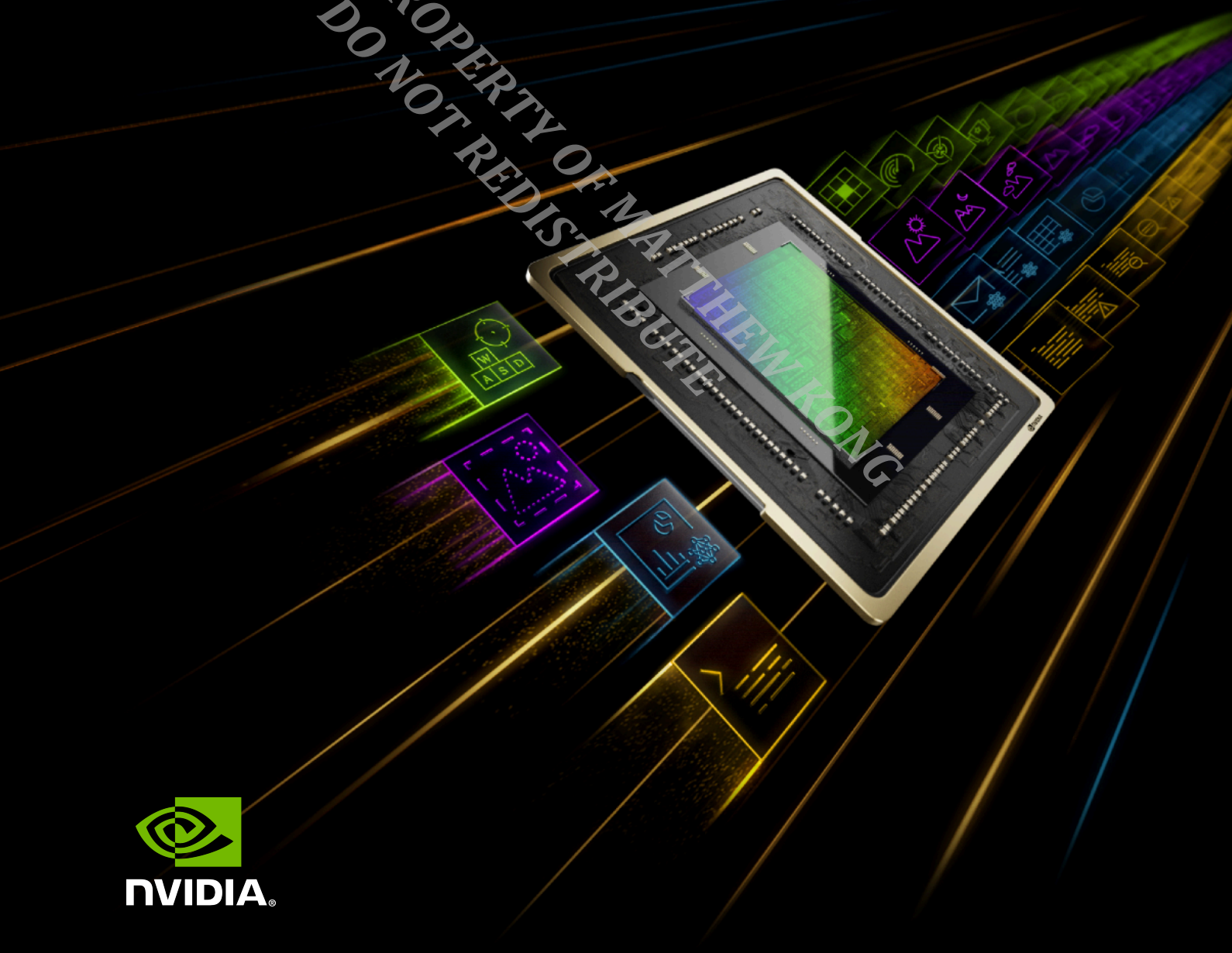
IB Business Management Extended Essay

Maintaining NVIDIA's Leadership in The AI Accelerator Market

Research question: "How and to what extent can NVIDIA maintain its leadership in the AI accelerator market?"

Word count: 3992



NVIDIA

(NVIDIA, 2006)

(NVIDIA, 2023)

Table of Contents

Introduction

About NVIDIA

Founded in 1993, NVIDIA is an American multinational corporation that develops Graphics Processing Units (GPUs)—computer hardware that accelerates graphics-related calculations in most smartphones, laptops, and desktops (Awati et al., 2023). NVIDIA is one of the world's leading GPU manufacturers (JPR, 2023). With the increase in popularity of AI through services like ChatGPT—which are very computationally intensive—demand for NVIDIA's AI accelerators has caused revenue to skyrocket in recent years (Hollister, 2023).

NVIDIA's superior software strengthens their position in this market. Most prominently, CUDA—software that enables AI developers to interact with GPU hardware—has two decades of active support, allowing it to become the most widely used and supported GPU development platform. CUDA only supports NVIDIA GPUs, forcing developers into exclusively purchasing NVIDIA products. This USP[1] facilitated their dominance in the AI accelerator space, effectuating an estimated 90% market share (Brock et al., 2024) and a revenue of $14.5B in Q3 2023 (NVIDIA, 2023).

Fig. 1: *An image of an Nvidia GPU. (Nelius, 2023)*

While NVIDIA has dominated the market, AMD's new MI300X accelerators offer both lower prices and higher performance. Furthermore, ROCm—AMD's competitor to CUDA—is gaining ground despite its immaturity, steadily disrupting NVIDIA's market standing. Thus, NVIDIA's continued dominance is uncertain—and therefore, this extended essay aims to answer the research question, "How and to what extent can NVIDIA maintain its leadership in the AI accelerator market?"

[1]USP: Unique Selling Point

Methodology: Sources

This essay uses solely secondary research. To achieve a comprehensive analysis, several academic sources were used, including research publications. Additionally, news reports and blog posts were used, though differing viewpoints were examined to reduce the impact of bias. Even still, the topic's recency limited the selection of sources. Some articles, for instance, were published mere days before examination because of new developments within the industry. Furthermore, although this paper is primarily qualitative, data gathered through examination of quantitative sources provides a basis for discussion.

Methodology: Tools & Theory

A VRIO analysis assessed NVIDIA's market position by evaluating their resources (value, rarity, imitability, and organization) to determine their ability to maintain leadership. The analysis detailed internal factors that affected their market position, which was corroborated by a Porter's Five Forces analysis—a tool that considers an industry's competitive landscape—which focused on external factors, like competitive rivals. Throughout these two analyses, certain shortcomings in NVIDIA's strategies were identified. Then, to articulate how NVIDIA can maintain their leadership, strategies to remedy these difficulties were proposed and evaluated qualitatively with information from prior analyses. Furthermore, descriptive statistics were used to complement qualitative analyses with relevant quantitative data.

This essay focuses mainly on content from the marketing unit of the IB Business Management course syllabus, with discussion of the competitive environment and demands of customers. However, it also discusses aspects of the operations unit, as supplier information is particularly relevant to NVIDIA's production issues.

NVIDIA's Current Market Leadership

VRIO Analysis

To examine the reasons for NVIDIA's AI accelerator success, a VRIO analysis—which is a framework assessing an organization's internal competitive advantages—was conducted. Through this analysis, the value, rarity, imitability and organization of NVIDIA's resources was evaluated, thus determining the extent to which NVIDIA can maintain their market leadership in this space.

Value

Value involves the degree to which a customer benefits from purchasing an organization's products (Barney, 1991). NVIDIA's AI accelerators add substantial value to their customer's operations, as AI models like ChatGPT often require hardware as powerful and advanced as NVIDIA's AI accelerators (Brock et al., 2024).

Additionally, NVIDIA's software provides astronomical value to AI developers. CUDA is proprietary, which reduces compatibility requirements and thus simplifies software development with GPUs. This simplicity is highly regarded by developers because it reduces the amount of time they need to invest in creating AI models (Harvard, 2020).

Furthermore, NVIDIA can optimize CUDA for their own hardware since it's proprietary, resulting in better performance—saving their customers both time and money. Thus, during its nearly two-decade existence, many developers learnt CUDA, leading to increased industry adoption (Harvard, 2020).

Rarity

Rarity describes the degree to which a resource is available to competitors in an industry (Barney, 1991). NVIDIA's hardware is moderately rare; despite high costs, components used to manufacture their products are available to other companies. In particular, production of their GPU dies[2] is outsourced to TSMC—an external chip manufacturer—with which NVIDIA doesn't have exclusivity agreements (Teer et al., 2022). As such, competitors can access the same resources. However, NVIDIA's chip designs are closed-source and safeguarded by intellectual property (IP) protection, so competitors must develop their own designs—making their hardware a rarer resource.

Imitability

Imitability describes the degree to which an organization's strengths can be copied by competitors (Barney, 1991). Generally, many resources involved in NVIDIA's AI accelerators are very difficult to imitate, which is a major reason for their success.

[2]GPU dies: the core component of a GPU, often costing the most to manufacture.

The main difficulty that competitors experience in imitating NVIDIA's success is the extensive capital required for GPU development. For instance, NVIDIA spent $7.3B on R&D[3] in 2023, which comprised a staggering 27% of their total revenue—a trend that's remained relatively constant for over a decade (See Table 1 & Fig. 2). Few companies within the same industry can replicate NVIDIA's same R&D expenditure, especially those that are more diversified and thus cannot solely develop GPUs—making imitation substantially more difficult. In comparison, AMD—NVIDIA's largest GPU competitor—spent only $5B in R&D in 2022, split across a more diversified product development portfolio than NVIDIA (Macrotrends, 2023).

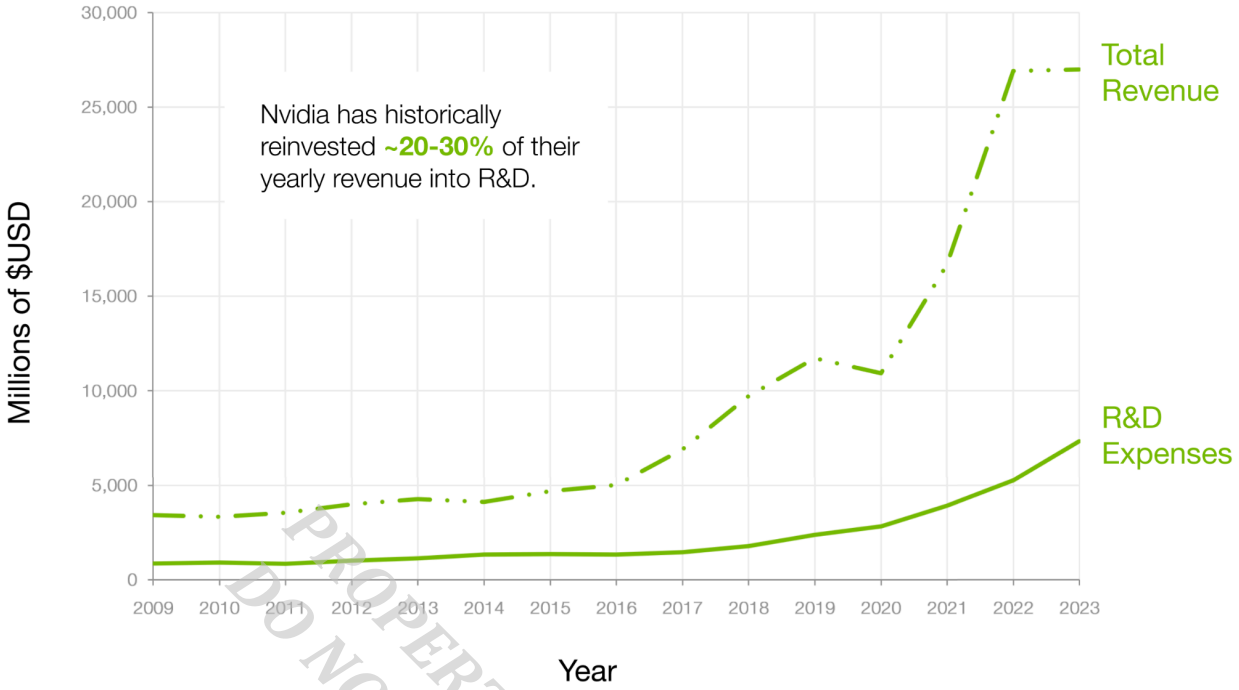
Table 1: *NVIDIA's R&D spending, from 2009 to 2023.*

Year	R&D Expenses ($M USD) [4]	Total Revenue ($M USD) [4]	R&D Expenses as percentage of total revenue
2023	7,339	26,974	27.21%
2022	5,268	26,914	19.57%
2021	3,924	16,675	23.53%
2020	2,829	10,918	25.91%
2019	2,376	11,716	20.28%
2018	1,797	9,714	18.50%
2017	1,463	6,910	21.17%
2016	1,331	5,010	26.57%
2015	1,360	4,682	29.05%
2014	1,336	4,130	32.35%
2013	1,147	4,280	26.80%
2012	1,003	3,998	25.09%
2011	849	3,543	23.96%
2010	909	3,326	27.33%
2009	856	3,425	24.99%

[3]R&D: Research and development.

[4]NVIDIA's R&D expenses and total revenue sourced from *Macrotrends* (Macrotrends, 2023).

Fig. 2: Nvidia's Yearly R&D Costs and Total Revenue Over Time



Additionally, NVIDIA's extensive software success is difficult to copy, as it requires control over both GPU hardware and software, as well as active use in industry development. (Tibazarwa, 2021) No other companies have the same software "moat" as NVIDIA among AI developers. CUDA is proprietary and has been supported for nearly two decades, facilitating performance optimizations and improved ease of use, leading to its estimated 90% market share (Brock et al., 2024). Current competitors are less mature, and thus unlikely to replace CUDA's USPs.

Additionally, incentivizing developers to move platforms is difficult, as they incur immense switching costs. And since most AI developers are already familiar with CUDA, incentives would need to be enticing for such a switch to be worthwhile—which would be difficult given previously mentioned factors.

However, replicating CUDA's offerings is not impossible. AMD has recently been developing tools that port CUDA applications to non-NVIDIA GPUs through their ROCm framework without significant developer intervention, effectively nullifying their CUDA USP (Larabel, 2024). This is one of the biggest threats to NVIDIA's AI accelerator sales, as it enables developers to use AI accelerators from AMD without incurring switching costs. However, this feature is immature and thus sparsely supported, so it is unlikely to pose a significant threat currently (Khan, 2023).

Organization

Organization describes the effectiveness with which a company's structure can capture value from a competitive advantage (Jurevicius, 2023). NVIDIA is generally well positioned in this aspect, as they tightly control their entire product stack,

enabling them to ensure quality and compatibility between products (Raynovich, 2023). This allows them to charge a premium; for instance, their flagship H100 accelerator has an extraordinarily high estimated gross profit margin (GPM) of ~87% (see Table 2).

Table 2: *Calculations NVIDIA's H100 gross profit, based on estimated sales and costs.*

	Lower range	Upper range
Retail price ($USD) (Norem, 2023)	25000	30000
Direct cost per H100 sold ($USD) (Norem, 2023)	$3,320.00	
Quantity sold (Norem, 2023)	550,000	
Cost of Goods Sold (estimated, $M USD)	$1,826	
Gross Profit (estimated, $M USD)	$11,924	$14,674
GPM (estimated)	86.72%	88.93%

However, NVIDIA lacks full control over the manufacturing processes of their AI accelerators. Though they design their chips in-house, their manufacturing is outsourced to specialized chip manufacturers like TSMC—and thus they are bound by the capabilities of these external companies. This reliance cannot be feasibly avoided, as chip manufacturing is highly capital-intensive (Blanchard et al., 2023). Additionally, these companies often experience long lead times, and so NVIDIA has to forecast future demand. With little ability to adjust orders based on current needs (Nussey et al., 2023). In essence, NVIDIA can only fulfill the predicted demand for AI accelerators, leaving potential customers unserved if projections are too low.

Evaluation

Overall, NVIDIA exhibits a very strong competitive advantage in the AI accelerator market; their accelerators provide immense value to NVIDIA's customers, are produced from relatively rare resources, are very difficult to imitate, and are supported by a strong organizational structure.

That said, the VRIO analysis also revealed several faults in NVIDIA's AI dominance, suggesting that they'll experience difficulty maintaining future leadership; particularly, their CUDA "moat" may be replicated, at least to an extent, by AMD's ROCm. Though imitation itself won't be sufficient to gain substantial market share—as high switching costs will likely prevent many AI developers from moving over—it would likely establish AMD as a viable competitor to NVIDIA. Furthermore, NVIDIA doesn't produce AI accelerators in-house, so they're bound by external companies like TSMC, which may inhibit their ability to fulfill demand. As such, customers will likely turn to NVIDIA's competitors for AI accelerators if demand greatly exceeds their own supply, further weakening their dominance.

The VRIO analysis on its own is not sufficient to fully detail the extent to which NVIDIA can maintain its market leadership though, as it mostly focuses on aspects of their products. A more detailed investigation into the external market, therefore, will be conducted through a Porter's Five Forces analysis.

A Novel Competitive Landscape

Porter's Five Forces

Although NVIDIA's AI accelerators exhibit many competitive advantages, the rapidly evolving AI accelerator market may hinder NVIDIA's market dominance. As such, a Porter's Five Forces analysis—which examines an industry's competitive landscape—will be conducted (Porter, 1979). In it, the AI accelerator industry's threat of new entrants, supplier power, customer power, threat of substitutes, and competitive rivals will be considered.

Threat of New Entrants

The threat of new entrants is the ability for other companies outside of the market to become competitive rivals—and in the AI accelerator industry it is low, as factors mentioned in the VRIO analysis make the production of similar products highly difficult. One notable new entrant is Intel, who recently unveiled the Gaudi3—their own AI accelerator competitor (Leswing, 2023; Porter, 1979). They are one of few companies in the world that—like NVIDIA—have pre-established non-AI GPU manufacturing processes and IP, which uniquely positions them to compete in this industry (Leswing, 2023). However, they have much less experience in GPU development than NVIDIA (~4 years vs. over two decades respectively), so they are unlikely to pose a significant threat in the near future (Peddie, 2020; Britannica, 2024).

Besides Intel though, no other companies are likely to enter the AI accelerator segment, as they'd likely have to invest considerably more capital—and thus the threat of new entrants into this industry is very low.

Supplier Power

Supplier power is the ability of suppliers of a product's resources to influence aspects of the product, such as price and quality—and it is very high within the AI accelerator industry (Porter, 1979). Though several semiconductor manufacturers produce chips advanced enough for use in AI accelerators, they each use their own proprietary processes, so switching between them within a single product generation is nearly impossible. As such, NVIDIA is bound by the manufacturer for which they initially designed their chips, which for their recent accelerators, is TSMC. This reliance is problematic, as TSMC is currently unable to fulfill its customer's orders due to manufacturing capacity constraints; Mark Liu, TSMC's chairman, stated in late 2023 that they "cannot fulfill 100% of [their] customers' needs, but [they] try to support about 80%" (Ting-Fang, 2023). As a result, orders of accelerators have reportedly had a lead time of up to a year—which, given the rapidly evolving pace of the AI industry, would be detrimental to prospective customers' operations (Shilov, 2023). Therefore, many have turned to NVIDIA's competitors to satisfy their needs, weakening NVIDIA's foothold over the market.

Additionally, while TSMC is interested in their success, NVIDIA only contributed 6.3% of TSMC's total revenue in 2022—and so TSMC may prioritize orders from higher-paying customers, such as Apple or Qualcomm (Table 3 and Figure 3). For instance, in 2023, Apple reportedly bought every one of TSMC's most advanced chips (Cunningham, 2023). This prioritization is likely a short-term issue, as TSMC is responding to increases in AI accelerator demand by rapidly increasing their output through new manufacturing plants. Until they come online, however, NVIDIA's output is greatly hindered, which incentivizes customers to switch to competing solutions, thereby diminishing their market leadership.

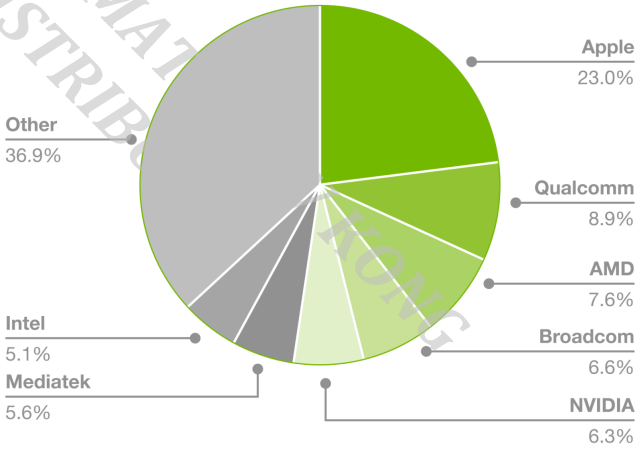
Customer Power

Customer power, which is the ability for customers to negotiate lower prices or higher quality products, is moderately low within the AI accelerator industry (Porter, 1979). Although there's a fairly high number of customers, who each purchase in large quantities, customers have little bargaining power because they rely on NVIDIA's GPUs to operate their AI models; they cannot feasibly be replaced by alternative technologies like Central Processing Units (CPUs), as AI accelerators are several orders of magnitude more performant for AI-related workloads (Suchard et al., 2010).

Table 3: *The proportion of TSMC's total revenue that each of its customers contributes.*[5]

Customer	% of TSMC's Total Revenue
Apple	23.00%
Qualcomm	8.90%
AMD	7.60%
Broadcom	6.60%
NVIDIA	6.30%
MediaTek	5.60%
Intel	5.10%
Other	36.90%

Fig. 3: Proportion of TSMC's Total Revenue by Customer (2022) [5]



Note: The "Other" category consists of customers that each comprise **<5%** of TSMC's total revenue.

[5]Data sourced from *Sravan's Substack* using data published by TSMC (Kundojjala, 2023)

Additionally—as discussed in the VRIO analysis—switching costs are very high, so customers with CUDA-based AI-models are essentially required to continue purchasing NVIDIA's products to operate them. Furthermore, customers within this industry are separate businesses that each generate substantial revenue from their AI ventures, so they are not particularly price conscious. For instance, OpenAI—one of NVIDIA's main customers in the AI accelerator segment—generated $1.6B in revenue from their AI services in 2023, which is considerably higher than their fixed equipment costs (Joseph et al., 2023). These costs are therefore likely not significant factors in purchase decisions, especially since accelerators are needed to run their models, and therefore generate revenue.

Threat of Substitutes

The threat of substitutes, which is the ability for new, different technologies to replicate the needs fulfilled by another product, is a weak force in the AI accelerator industry (Porter, 1979). Potential replacements do exist, but none are likely to successfully replace GPUs within the AI segment. For instance, CPUs can also perform AI-related workloads; however, they are considerably less effective at them than GPUs, and are thus irrelevant as substitutes (Suchard et al., 2010).

Google's Tensor Processing Units (TPUs) are perhaps more appropriate, as they are specialized for neural network workloads, like AI-model training (Google, 2024). Unlike NVIDIA's accelerators though, they are not sold to customers outright; instead, they are only available through Google's own cloud subscription service (Google, 2024; Deepgram, 2023). This complicates AI-development with TPUs, as it forces developers to build on Google's proprietary hardware that is fully managed by Google—ballooning costs due to other mandatory Google services required. As such, Google's TPUs are unlikely to successfully substitute NVIDIA's AI accelerators at a large scale (Deepgram, 2023).

Competitive Rivals

Although NVIDIA has few competitors in the AI accelerator sector, competition is nonetheless moderately high. Namely, AMD poses a moderate threat to NVIDIA, as their AI accelerators exhibit several competitive advantages. For instance, their latest MI300X accelerators are less than half the price of NVIDIA's, while—at least on paper—being 10-80% faster. With that said, the AI industry's dependence on NVIDIA's software ecosystem may interfere with AMD's ability to compete, as CUDA-based AI models may not function on their accelerators, thus negating any of their potential value or performance propositions. However—as described in the imitability section—AMD is developing ROCm to emulate CUDA, allowing for CUDA-based software to run on non-NVIDIA hardware. Nevertheless, it is less stable, more difficult to work with, and less performant than CUDA, which may deter developers from AMD's platform (Khan, 2023). AMD has improved ROCm quite significantly in these areas, but it still remains uncertain if their software will successfully overcome NVIDIA's USPs.

Considering NVIDIA's current position though, AMD may not *have to* develop a superior software

package to weaken their leadership; NVIDIA simply cannot fulfill enough orders of accelerators to satisfy current demand, which forces potential customers into seeking alternatives (Dobberstein, 2023). As discussed in the Supplier Power section, this is largely due to TSMC's supply issues. And while AMD also sources chips from TSMC, they use a more mature and standard process instead, which TSMC manufactures in higher quantities (Moore, 2023; Schor, 2021). As such, AMD doesn't face the same production issues—and this has resulted in high demand from several large customers, such as Meta and OpenAI (Leswing, 2023).

NVIDIA's production issues are likely short-term though, as TSMC is building new manufacturing facilities to keep up with AI demand; Mark Liu, TSMC's chairman, stated in late 2023 that these constraints "should be alleviated in one and a half years," when they complete their expansions (Ting-Fang, 2023). Even still, this lead time may be too long for potential customers, as it massively increases their working capital cycles, which may jeopardize their ability to operate. For instance, Peter Marrs—Dell's Asia Pacific and Japan chief—indicated that NVIDIA's customers simply "can't wait a year" for their GPUs (Dobberstein, 2023). As these customers switch to NVIDIA's competitors, the AI industry's dependence on CUDA will steadily decrease, thus greatly diminishing NVIDIA's greatest USP and therefore their market leadership.

Evaluation

Though NVIDIA is currently a market leader in the AI accelerator segment, some external factors in their current operations may hinder their ability to sustain this position. Particularly, AMD's latest accelerators offer better value, which has supplanted them as a competitive rival and enticed numerous customers into switching. Furthermore, AMD is steadily improving their ROCm framework, thus weakening NVIDIA's software USP. Above all though, the greatest threat to NVIDIA's dominance is their immense dependence on their suppliers; TSMC's production issues have prevented NVIDIA from fulfilling orders, leading to a year-long lead time—and thus enticing customers to purchase competing offerings.

On the other hand, the threat of new entrants, customer power, and the threat of substitutes in the AI accelerator industry are all comparatively low, which will help NVIDIA stay competitive; however, given the relative strength of the other external forces, they will likely struggle to maintain their leadership position if they continue with their current strategies.

Adapting to Competition

Pricing Strategies

One major selling point of competitors' products is their superior value; AMD sells their similarly-performing accelerators at much lower prices. However, accelerator costs pale in comparison to the revenue generated by NVIDIA's biggest customers, and thus AI accelerator prices are generally not of major concern. Nevertheless, lower prices may attract businesses with less capital, like AI startups. Though AMD may not source much revenue from these low-volume customers, they'll still spur increased adoption among developers, steadily weakening NVIDIA's market dominance.

To deter these businesses from switching, NVIDIA could price their own accelerators below those of AMD, which—with an estimated GPM of nearly 90% (Table 1)—likely wouldn't greatly harm their profitability. And since NVIDIA is a price leader in this industry, competitors would have to decrease their own prices in response, inhibiting their ability to fund development of future products and making them less competitive. However, they already have a majority market share, so lowered prices would likely just result in proportionally lower overall revenue—especially since customers gained from this strategy are presumably highly price-sensitive. Simply decreasing prices across their entire product portfolio is thus inadvisable.

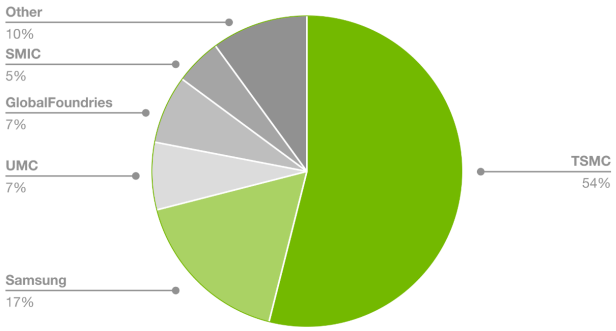
Instead, NVIDIA could sell lower-tier accelerators at a loss, while maintaining the prices of their flagship products. This would fulfill the needs of smaller, more price-sensitive businesses, while also granting more prolific customers access to the high-performance GPUs they require, at the same prices they would otherwise be paying. Since NVIDIA wouldn't rely on loss-leaders for revenue, they could more easily price competitors out of the market—especially considering the difficulty in replicating USPs like CUDA.

However, NVIDIA's production issues have caused unacceptably long lead times, pushing customers to alternative products—an issue that lower prices do not solve. According to projections by TSMC, these issues are short-term because they'll be alleviated in less than two years. However, as previously discussed, many customers simply cannot afford to wait, causing NVIDIA to lose valuable market share to AMD.

Alternative Suppliers

NVIDIA's production issues impact their market standing the most; their current output cannot meet the needs of the rapidly growing AI segment, facilitating losses of market share to competitors. Seeking alternative suppliers, then, may alleviate this issue. However, doing so within a single product generation is highly difficult, so this change is effectively permanent and long-term.

Fig. 4: Global Market Share of Semiconductor Manufacturers (2023) [6]



Note: Of the manufacturers represented, only **TSMC** and **Samsung** produce chips advanced enough for AI-accelerators.[20] **Intel**, which isn't shown, does as well.

Table 4: *The global market share of each semiconductor manufacturer in 2023.*[6]

Supplier	Market share	Produces Advanced chips
TSMC	54%	Yes
Samsung	17%	Yes
UMC	7%	No
GlobalFoundries	7%	No
SMIC	5%	No
Other	10%	No

Besides TSMC, there aren't many viable semiconductor manufacturers—and of the few that exist, none have the same manufacturing capacity. As depicted in Figure 4, the only other company with significant market share that can produce advanced chips is Samsung. Since they have a lower market share (and thus likely lower output), a full switch likely wouldn't alleviate production issues.

Instead, NVIDIA could segment their AI accelerator portfolio into two designs: one built with TSMC, and one with Samsung. This diversification would substantially increase development costs, but also improve output. In the past, NVIDIA has successfully employed this multi-supplier strategy between different product segments, with increased revenue from selling to different markets offsetting high development costs (NVIDIA, 2020; Techpowerup, 2020). The demand for AI accelerators, therefore, may produce a similar result, while also maintaining their market share and stabilizing production through supply chain redundancies.

The aforementioned strategy is very costly though, so NVIDIA could instead build accelerators using multiple TSMC nodes[7] instead of one, which would reduce development costs because they wouldn't have to adapt chip designs to different proprietary nodes. Furthermore, it would increase output, as they are manufactured at different facilities, and are thus largely independent from each other. However, different nodes could perform differently, which would affect AI accelerator quality and thus deter customers.

Overall, no single strategy would effectively secure market dominance. However, a combination may prove effective. For instance, segmenting accelerator offerings by TSMC's processes, *and* selling accelerators produced from lower-performing processes at lower prices, may negate supply issues while satisfying value-oriented customers. NVIDIA could also source chips from multiple suppliers, but this would be highly costly and thus less favourable. Nevertheless, the proposed strategies are likely needed for maintaining NVIDIA's leadership

[7]Data sourced from *Visual Capitalist* (Bhutada, 2021)

[7]Nodes are specific chip sizes that a semiconductor manufacturing company produces (Gray, 2014).

Conclusion

NVIDIA is currently dominant in the AI accelerator industry. As detailed in the VRIO analysis, their products are highly sought after because they offer immense value for their customers, who need them to generate revenue from their own AI ventures. Furthermore, NVIDIA's accelerators are both rare and difficult to imitate, mainly resulting from their proprietary and high-quality CUDA software, which allow developers to interface with their GPUs. They're also well positioned organizationally to take advantage of these factors, as they control nearly their entire product stack—promoting quality and justifying their high GPM.

However, as the Porter's Five Forces analysis revealed, many external forces may inhibit this dominance. Particularly, AMD's latest accelerators are poised to be both cheaper and more performant than NVIDIA's best, with improvements to their ROCm software weakening NVIDIA's CUDA USP. Since AI developers would still incur immense switching costs in adopting AMD's offerings, these factors alone likely won't significantly impact NVIDIA's dominance. However, the high power of suppliers is detrimental, especially since TSMC's capacity issues have resulted in year-long lead times for NVIDIA's accelerators. *For both these reasons, they will likely lose market share to competitors.*

As such, strategies to combat these threats were proposed. Namely, lowering prices could negate AMD's higher value USP. This could be effective in the long term as ROCm becomes a more viable competitor to CUDA; supply issues are more pertinent in the short term though, as they prevent NVIDIA from fulfilling customers' orders. Switching suppliers outright, however, would be inefficacious, as TSMC has the highest manufacturing capacity of any supplier. Instead, NVIDIA could use multiple suppliers, though this would be expensive as they'd have to develop new chip designs. Alternatively, they could build accelerators on different nodes from the same manufacturer, which would be cheaper and increase output. Ultimately though, NVIDIA should adopt a combination of price and supplier strategies, as this would combat both increasing competition and supplier struggles.

There are limitations in the research method. In particular, the topic's recency made finding sources with balanced opinions difficult. For example, when new competing products are released, the media tends to exaggerate their capabilities (Scopelliti, 2011). This was pertinent with AMD's MI300X, which was touted to be a serious threat to NVIDIA's H100—which, while partially true, ignores many of its disadvantages. As such, several sources were examined to procure a balanced analysis. Furthermore, some statistics used are estimations, rather than official published data. This was problematic because it may have skewed discussions of financial information; however, descriptive statistics were used only to support qualitative analyses, so its impact was likely minimal.

Overall though, this essay effectively answered the question "How and to what extent can NVIDIA maintain its leadership in the AI accelerator market?"—a topic that, given the rapid growth of the segment, is crucial in ensuring their success.

Bibliography

Suchard, M. A., Wang, Q., Chan, C., Frelinger, J., Cron, A., & West, M. (2010). Understanding GPU Programming for Statistical Computation: Studies in Massively Parallel Massive Mixtures. *Journal of Computational and Graphical Statistics*, *19*(2), 419–438. http://www.jstor.org/stable/25703576

Scopelliti, I. (2011). The role of concept exaggeration in promoting new products.

Adjustment to Q422 JPR dGPU Market Report. (2023, March 6). Jon Peddie Research. https://www.jonpeddie.com/news/adjustment-to-q422-jpr-dgpu-market-report/

AMD Quietly Funded A Drop-In CUDA Implementation Built On ROCm: It's Now Open-Source. (n.d.). Www.phoronix.com. Retrieved March 2, 2024, from https://www.phoronix.com/review/radeon-cuda-zluda

AMD Research and Development Expenses 2006-2020 | AMD. (n.d.). Www.macrotrends.net. https://www.macrotrends.net/stocks/charts/AMD/amd/research-development-expenses

Awati, R. (2023, November). *What is a GPU?* SearchVirtualDesktop. https://www.techtarget.com/searchvirtualdesktop/definition/GPU-graphics-processing-unit

Barney, J. (1991). *Firm Resources and Sustained Competitive Advantage*. https://josephmahoney.web.illinois.edu/BA545_Fall%202022/Barney%20(1991).pdf

Brock, C. (n.d.). *Nvidia Stock: Is It Still A Top AI Stock In 2024?* Forbes. Retrieved March 2, 2024, from https://www.forbes.com/sites/investor-hub/article/nvidia-stock-still-top-ai-stock/?sh=43e9fdb6a326

Cunningham, A. (2023, August 7). *Report: Apple buys every 3 nm chip that TSMC can make for next-gen iPhones and Macs*. Ars Technica. https://arstechnica.com/gadgets/2023/08/report-apple-is-saving-billions-on-chips-thanks-to-unique-deal-with-tsmc/

Dobberstein, L. (2023, December 7). *Dell APJ chief says the industry won't wait for Nvidia H100*. Www.theregister.com. https://www.theregister.com/2023/12/07/dell_apj_president_says_industry/

Get Next-Level AI On RTX GPUs. (n.d.). NVIDIA. https://www.nvidia.com/en-us/ai-on-rtx/

Gray, I., Acquaviva, A., & Audsley, N. (2014). *What is Technology Node | IGI Global*. Www.igi-Global.com. https://www.igi-global.com/dictionary/technology-node/44459

Hollister, S. (2023, August 23). *Nvidia just made $6 billion in pure profit over the AI boom*. The Verge. https://www.theverge.com/2023/8/23/22608145/nvidia-q2-2024-profit-revenue-ai-chips

Introduction to Cloud TPU. (n.d.). Google Cloud. https://cloud.google.com/tpu/docs/intro-to-tpu

Jon Peddie. (2020, November 26). *Famous Graphics Chips: Intel's GPU History | IEEE Computer Society*. https://www.computer.org/publications/tech-news/chasing-pixels/intels-gpu-history

Joseph, J. (2023, December 30). *OpenAI annualized revenue tops $1.6 billion - The Information*. Reuters. https://www.reuters.com/technology/openai-an

nualized-revenue-tops-16-billion-information-2023-12-30/

Jurevicius, O. (2023). *VRIO Framework Explained - SM Insight*. Strategic Management Insight. https://strategicmanagementinsight.com/tools/vrio/

Kundojjala, S. (2023, May 25). *TSMC's top-10/20/30/40 customers; Who spends how much on TSMC? Apple revenue; Top-10 customer dynamics*. Sravan's Substack. https://exploresemis.substack.com/p/tsmcs-top-10203040-customers-who

Leswing, K. (2023a, December 6). *Meta and Microsoft say they will buy AMD's new AI chip as an alternative to Nvidia's*. CNBC. https://www.cnbc.com/2023/12/06/meta-and-microsoft-to-buy-amds-new-ai-chip-as-alternative-to-nvidia.html

Leswing, K. (2023b, December 14). *Intel unveils new AI chip to compete with Nvidia and AMD*. CNBC. https://www.cnbc.com/2023/12/14/intel-unveils-gaudi3-ai-chip-to-compete-with-nvidia-and-amd.html

Li, J., Aitken, W., Bhambhoria, R., & Zhu, X. (2023, July). Prefix Propagation: Parameter-Efficient Tuning for Long Sequences. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 1408–1419). doi:10.18653/v1/2023.acl-short.120

Nelius, J., & 2023. (2024, February 7). Nvidia's RTX 4090 is a phenomenal graphics card, but you probably shouldn't buy it. Reviewed. https://reviewed.usatoday.com/gaming/content/nvidia-geforce-rtx-4090-review

Moore, S. (2023, December 6). *AMD's Next GPU Is a 3D-Integrated Superchip - IEEE Spectrum*. Spectrum.ieee.org. https://spectrum.ieee.org/amd-mi300

Nussey, S., Potkin, F., & Sterling, T. (2023, September 15). Exclusive: TSMC tells vendors to delay chip equipment deliveries, sources say. *Reuters*. https://www.reuters.com/technology/tsmc-tells-vendors-delay-chip-equipment-deliveries-sources-2023-09-15/

NVIDIA GA100 GPU Specs. (2024, March 2). TechPowerUp. https://www.techpowerup.com/gpu-specs/nvidia-ga100.g931

NVIDIA Research and Development Expenses 2006-2019 | NVDA. (2019). Macrotrends.net. https://www.macrotrends.net/stocks/charts/NVDA/nvidia/research-development-expenses

NVIDIA. (2020) Updated with NVIDIA RTX A6000 and NVIDIA A40 Information V2.0 NVIDIA AMPERE GA102 GPU ARCHITECTURE Second-Generation RTX. (n.d.). https://www.nvidia.com/content/PDF/nvidia-ampere-ga-102-gpu-architecture-whitepaper-v2.pdf

NVIDIA's Winning Platform Strategy with CUDA. (2020, March 22). Digital Innovation and Transformation. https://d3.harvard.edu/platform-digit/submission/NVIDIAs-winning-platform-strategy-with-cuda/

Porter, M. E. (1979). *How Competitive Forces Shape Strategy*. Harvard Business Review. https://hbr.org/1979/03/how-competitive-forces-shape-strategy

Raynovich, R. S. (2023, August 24). *The Untold Story Behind Nvidia's Earnings: Full-Stack AI*

Dominance. Forbes. https://www.forbes.com/sites/rscottraynovich/2023/08/24/the-untold-story-behind-nvidias-earnings-full-stack-ai-dominance/?sh=74c8058d5246

Reuters. (2023, June 29). TSMC sending more workers to speed up building of new Arizona plant. *Reuters*. https://www.reuters.com/technology/tsmc-sending-more-workers-speed-up-building-new-arizona-plant-2023-06-29/

Schor, D. (2021, October 26). *TSMC Extends Its 5nm Family With A New Enhanced-Performance N4P Node*. WikiChip Fuse. https://fuse.wikichip.org/news/6439/tsmc-extends-its-5nm-family-with-a-new-enhanced-performance-n4p-node/

Shilov, A. (2023, November 28). *Nvidia sold half a million H100 AI GPUs in Q3 thanks to Meta, Facebook — lead times stretch up to 52 weeks: Report*. Tom's Hardware. https://www.tomshardware.com/tech-industry/nvidia-ai-and-hpc-gpu-sales-reportedly-approached-half-a-million-units-in-q3-thanks-to-meta-facebook

Teer, J., Bertolini, M., Ritoe, J. A., Heyster, S., Sweijs, T., de Wijk, R., Rademaker, M., Vlaskamp, M., Patrahau, I., Thompson, J., Kim, S., Minicozzi, R., Meszaros, A., Cisco, G., & Gorecki, M. (2022). Fragile balance: the semiconductor and critical raw material ecosystem. In *Reaching breaking point: The semiconductor and critical raw material ecosystem at a time of great power rivalry* (pp. 7–26). Hague Centre for Strategic Studies. http://www.jstor.org/stable/resrep44057.6

Tibazarwa, A. (2021). Strategic Integration for Hardware and Software Convergence Complexity. IEEE Engineering Management Review, 49(3), 92-102.

Tensor Processing Unit (TPU). (n.d.). Deepgram. https://deepgram.com/ai-glossary/tensor-processing-unit-tpu

The Editors of Encyclopedia Britannica. (2018). NVIDIA Corporation | global corporation. In *Encyclopædia Britannica*. https://www.britannica.com/topic/NVIDIA-Corporation

TING-FANG, C. (2020, November 26). *TSMC sees AI chip output constraints lasting 1.5 years*. Nikkei Asia. https://asia.nikkei.com/Business/Tech/Semiconductors/TSMC-sees-AI-chip-output-constraints-lasting-1.5-years