Harvard Business School

9-725-360
SEPTEMBER 5, 2024

DAVID B. YOFFIE
SARAH VON BARGEN

Nvidia, Inc. in 2024 and the Future of AI

In 2024, Nvidia was on top of the world. On June 18th, Jensen Huang, co-founder and CEO of Nvidia, could brag that he had built the world's most valuable company, worth $3.4 trillion. While Microsoft and Apple quickly recaptured the value crown, some analysts forecasted that Nvidia was so strongly positioned that it might become the world's first $4 trillion company. Indeed, Huang announced that revenues for Q2 2024 jumped more than 250% year over year. (See **Exhibits 1a-1d**) Yet the semiconductor business was notoriously cyclical. Leadership in global semiconductors changed many times over its 60-year history, moving from Texas Instruments to the Japanese, Intel, Samsung, TSMC, and in 2024 Nvidia. Huang's wanted to break that cycle to ensure that Nvidia would stay on top.

Huang asked his team to explore two pressing issues in the fall of 2024. First, Nvidia's spectacular performance was built on generative AI, but the GenAI market was rapidly evolving. In 2024, Nvidia Graphic Processor Units (GPUs) were the best solution for training large language models (LLMs) and inference (providing the answer to a query). Since demand for Nvidia GPUs far outstripped supply in 2024, Nvidia had a huge backlog and needed to decide who should receive its cutting-edge systems. In addition, Nvidia charged outsized premiums, delivering unprecedented gross margins for the semiconductor business. Yet competition was beginning to emerge: AMD was offering GPUs at lower prices that were becoming more competitive with Nvidia, and many other companies had announced specialized chips that could replace Nvidia for some applications. The Department of Justice also launched an anti-trust investigation in August,[1] sending subpoenas to Nvidia in early September.[2] Under these circumstances, who should get GPUs? And was monopoly pricing the best strategy?

Second, Huang had lingering questions about where Nvidia should play in the value chain. On the supply side, Nvidia was deeply dependent on a single supplier, Taiwan Semiconductor Manufacturing Company (TSMC). (See **Exhibit 2)** On the customer side, Nvidia faced highly concentrated customers for the data center. Nvidia announced its intention to invest in data centers and compete with those customers. While the largest cloud providers, such as Microsoft and Amazon, made substantial profits renting Nvidia GPUs, they were also unhappy that an important supplier was becoming a direct competitor. The largest "hyperscalers" (the common name for the biggest cloud service providers) were designing their own chips to replace Nvidia. Nvidia was also going head-to-head with its largest OEM customers, such as Dell and HP. This left two obvious questions: Should Huang continue to move forward with vertical integration and go head-to-head with its biggest customers, or stay a component supplier? And were there options to hedge its risk with TSMC?

Professor David B. Yoffie and Research Associate Sarah von Bargen prepared this case. This case was developed from published sources. Funding for the development of this case was provided by Harvard Business School and not by the company. HBS cases are developed solely as the basis for class discussion. Cases are not intended to serve as endorsements, sources of primary data, or illustrations of effective or ineffective management.

Nvidia's Origins

The driving force behind Nvidia was Jensen Huang, born Jen-Hsun Huang in 1963 in Taiwan. At five years old, his family moved to Thailand. A few years later, his parents sent him and his older brother to live with relatives in the United States, seeking better educational opportunities. Huang's parents later migrated to the U.S. and settled in Oregon where Huang earned his undergraduate degree in electrical engineering from Oregon State University in 1984 and his master's degree in electrical engineering from Stanford University in 1992.[3]

Huang's first job was as a dishwasher at a Denny's restaurant. Later, Huang worked as a director at LSI Logic and a microprocessor designer at AMD, before co-founding Nvidia. Huang believed these formative experiences shaped his work ethic and leadership style. In a 2024 speech to Stanford GSB students, Huang emphasized the importance of struggling in life, stating "Unfortunately, resiliency matters in success. I don't know how to teach it, but I hope suffering happens to you."[4]

At Nvidia, Huang believed that "the flattest organization is the most empowering one" and fostered a culture of shared responsibility and collaboration. Huang had an unconventional management style, which included avoiding one-on-one meetings and up to 60 direct reports. According to Huang, this approach fostered open communication, encouraged collective problem-solving, and allowed for swift decision-making. Huang encouraged leaders to email him their "top five things," which he read each morning. He also promoted transparency by allowing employees of all levels to join meetings, ensuring equal access to information.[5] Speaking about Nvidia's structure, Huang stated:

> The architecture of the company (to me) is a computer with a computing stack, with people managing different parts of the system. Who reports to whom, your title is not related to anywhere you are in the stack. It just happens to be who is the best at running that module on that function on that layer is in-charge. That person is the pilot in command…[6]

Under Huang's leadership, Nvidia's revenue grew from $159 million in FY 2000 to $60.9 billion in FY 2024.[7] Analysts forecasted Nvidia sales to be over $100 billion in calendar year 2024.[8] The company's workforce also expanded from around 250 employees at the time of its IPO in 1999, to 29,600 by early 2024.[9] Nvidia's valuation reflected similar growth: in May 2023, it reached $1 trillion; within 9 months, Nvidia's valuation doubled.[10] On June 5th, 2024, Nvidia's market cap hit $3 trillion, and two weeks later, it briefly surpassed Microsoft to become the world's most valuable company.[11] While Nvidia's stock proved volatile (1.7 beta)[12], Huang emerged as the 11th richest person in the world.[13]

Early Days: 1993-1999

Huang founded Nvidia in 1993 with Chris Malachowski, an engineer at Sun Microsystems, and Curtis Priem, a graphic chip designer at IBM and Sun. The trio started the company at Denny's restaurant in California, planning the company over a series of Grand Slam breakfasts. Initially, their strategy was to produce the gaming console inside of PCs. Within a few years, they decided to focus instead on improving graphics-based processing, specifically for gaming applications. Huang noted that "video games were simultaneously one of the most computationally challenging problems and would have incredibly high sales volume. Those two conditions don't happen very often. 'Video games' was our killer app."[14]

The first few years at Nvidia were especially challenging. Between 1993-1997, Nvidia's only significant customer was Sega, the video game company. In its early days, Nvidia relied on a quadrilateral architecture to develop its game chips to fit with Sega's architecture. However, when

Microsoft announced changes to its preferred chip architecture, Sega followed. This posed a serious problem for Nvidia: it was only halfway through production on its Sega chips, which suddenly became unusable. This left Nvidia in a cash crunch, forcing it to develop a new chip in 1996.[15]

At the time, chip development and production generally required 18 to 24 months. Since Nvidia had 9 months of cash left, it needed to develop a new chip in record time. By shortening the production time to less than 6 months using software prototypes, Nvidia built a new chip, called RIVA 128, which became one of the largest graphics chips ever created. It was also the first fully accelerated 3D graphics pipeline for a computer, essential at a time when video game graphics were transitioning from flat 2D pictures to more realistic, 3D images. The chip was so powerful that Huang was able to convince developers to use the chip for game development, focusing on the chip's strong performance and 3D rendering. Nvidia sold one million units of RIVA 128 within 4 months.[16]

In 1999, Nvidia released an add-in graphics card called GeForce 256, marketed as the world's first GPU. GeForce 256 brought advanced 3D graphics capabilities to consumer PCs and established Nvidia's reputation in the gaming industry. Nvidia went public on January 22, 1999. By the end of 1999, the company had grown to 1,000 employees and $159 million in revenue.[17]

Nvidia: 2000-2020

Nvidia built on its early success by focusing on GPUs for gaming applications. Unlike CPUs (central processing units), which handled computing tasks sequentially for the computer's operating system, GPUs utilized parallel processing, which performed multiple computations at the same time. Parallel processing was a key component for increasing the computing power of these chips, making it essential for graphics in video games, which required constant real-time rendering of high-quality images. As time progressed, users realized that GPUs could be utilized for other applications.

Nvidia relied on Moore's Law to drive performance in the 1990s and early 2000s. Moore's Law, named after Intel's co-founder Gordon Moore, was based on his prediction that the number of transistors on an integrated circuit would double roughly every 18 to 24 months. The industry largely tracked Moore's Law since Moore's article was published in 1965. For Nvidia's GPUs, Moore's Law enabled rapid improvements in hardware acceleration and rendering capabilities. These features allowed game developers to create immersive experiences optimized for Nvidia's GPUs. However, Nvidia remained a small player: With FY 2001 revenue of $735 million, it was only 3% of industry leader Intel's $26 billion in revenues.[18] Indeed, Huang was terrified of Intel at the time. He later reflected, "I don't go anywhere near Intel…Whenever they come near us, I pick up my chips and run."[19]

To compete with the giants of the semiconductor industry, Nvidia pursued a different strategy. Intel and Samsung, the world's largest semiconductor companies, were vertically integrated, designing as well as manufacturing their chips. By contrast, Nvidia built its business around outsourcing its manufacturing mostly to TSMC. Nvidia designed its GPUs, while TSMC would fabricate the chips, and a variety of third parties would assemble the chips, put them into packages, and test the chips. Contract manufacturers were then employed to put GPUs onto graphics cards, which could be inserted into a gaming console, PC, or server. Finally, Nvidia was responsible for sales. The company relied on a direct sales force that sold to OEMs like Dell and HP, and data center customers including Amazon and Microsoft. They also sold GPUs to retailers and direct-to-consumers.

Semiconductor fabrication was one of the most capital-intensive businesses in the world. In 2024, the cost of a single state-of-the-art-fab could be as much as $25 billion.[20] By outsourcing, Nvidia escaped the capital investment required by TSMC, Intel, Samsung, and others, but it also had less direct control

over its supply chain. During normal times, it would take 8-12 weeks from order to delivery of a GPU. During periods of high demand for manufacturing at TSMC, it could take four months or more.[21] From the late 1980s until roughly 2018, Intel had been the leader in semiconductor manufacturing. But poor execution at Intel and steady progress at TSMC turned TSMC into the most advanced fabricator of integrated circuits. TSMC manufactured over 90% of the world's leading edge logic chips, including 7nm, 5nm, and 3nm nodes in 2024.[22] TSMC's position was so strong that it could demand that its largest customers pre-pay billions of dollars to reserve capacity before it was built.[23]

In the early days, when GPUs were almost exclusively used for video games, Nvidia sold its graphics cards for an average of $300-$500.[24] This compared to less than $50 for Intel's low-end graphic chips that dominated the market for non-gaming PC graphics. Nvidia's challenge was that GPUs were one of the most expensive chips to manufacture. Since each chip had a large number of computing cores, the chips were big and often required the most advanced technology to deliver performance and keep them from overheating. To improve performance over time, Nvidia added more cores, more memory, etc., which meant the chips got bigger and costs went up. By the early 2020s, gaming GPUs were selling for as much as $1500.[25] GPUs for training AI were much bigger, much more expensive to manufacture, and priced differently. While a high-end CPU might have a small number of complex cores and ~60 billion transistors in 2024,[26] Nvidia's data center GPUs had over 16,000 cores and ~80 billion transistors.[27] TSMC charged Nvidia roughly $900 per chip,[28] and in 2024, Nvidia was selling those high-end GPUs (H100) for $25,000-$40,000 (see **Exhibits 3 and 4**).[29]

High performance came with high power requirements. An Nvidia GPU could consume up to 1,000 watts or more per chip, and most servers had multiple GPUs.[30] According to one analysis, "power is the ultimate constraint."[31] Another analyst pointed out that Nvidia GPUs sold in 2023 consumed more power than 1.3 million homes, and by 2026, the incremental power requirements for data centers were roughly equivalent to the entire country of Japan.[32] Mark Zuckerberg warned that when the GPU "drought" ended, energy issues could severely constrain growth. He believed that data centers might start pushing "500 megawatts or even pushing a gigawatt," and worried that the lack of growth in energy production would impact AI availability.[33] In a dramatic example of the growing concern about energy, Microsoft announced in September 2024 a $1.6 billion deal to re-open the nuclear plant at 3 Mile Island in Pennsylvania to provide power for AI (and its GPUs).[34] (see **Exhibits 5a and 5b**)

With higher costs and higher performance, Nvidia's challenge was to find customers willing to pay a premium. A common problem in the semiconductor industry was that hardware improvements were often available before the applications that could utilize a chip's processing power. Huang's answer was CUDA, or Compute Unified Device Architecture. Introduced in 2006 as a programming language similar to C++, it soon became a platform and programming model that enabled developers to create new capabilities for GPUs. CUDA was proprietary and locked to Nvidia's architecture: development of new software on CUDA only worked on Nvidia's GPUs. Another distinguishing feature of CUDA was that it morphed from a programming language into a platform. Nvidia allowed third-party developers to write programs for the system and share these programs with other developers. This open platform approach fostered a developer community around CUDA, allowing developers to utilize the platform's SDKs (software development kits) for free. As Huang later explained in 2022:

> We've been advancing CUDA…for 15 years… We optimize across the full stack iterating between GPU, acceleration libraries, systems, and applications continuously all while expanding the reach of our platform by adding new application domains that we accelerate… We have over 150 SDKs that serve industries from gaming and design to life and earth sciences, quantum computing, AI, cybersecurity, 5G, and robotics.[35]

CUDA, however, was not an immediate success. Despite heavy spending, demand for CUDA lagged. When downloads of CUDA declined from 2010-2013, and Nvidia's stock was depressed, Nvidia's board began to worry about corporate raiders.[36] The emergence of machine learning models of AI saved CUDA. Princeton professor Fei-Fei Li (who later moved to Stanford) and her team had developed an image-classifying database known as ImageNet. By 2009, the database consisted of 3.2 million images and growing. To prompt other researchers to use the database, Professor Li started the annual ImageNet Large Scale Visual Recognition Challenge in 2010, where programmers competed to develop software for accurately classifying the most images. In 2012, the record for most accurate classification was broken by AlexNet, a deep learning program developed by three academics. To power AlexNet, the students utilized parallel computing implemented in CUDA on Nvidia GPUs.

Scientists in fields including biosciences, computing, and chemistry began to realize that Nvidia's GPUs could complete calculations in a fraction of the time of standard scientific computers. A research scientist at Nvidia also recognized the potential of AI for GPU sales. Soon after the researcher published a paper on neural networks in 2013, he met with Huang. He detailed why deep learning would be essential to AI, and why Nvidia should place its bets on the technology. After hearing the pitch, Huang agreed, and decided to make a big bet on AI applications, even though there were limited revenue opportunities for AI at the time. Huang "sent out an e-mail on Friday evening saying everything is going to deep learning, and that we were no longer a graphics company…By Monday morning, we were an A.I. company. Literally, it was that fast," remembered Greg Estes, an Nvidia VP.[37] Similar to Kevin Costner in the movie Field of Dreams, who built a baseball field in the middle of Iowa and hoped that customers would come, Huang was betting that AI would be a big revenue driver. Unfortunately for Huang, though, customers were slow to implement deep learning models at scale from 2012 to 2020. While Nvidia had tried to diversify its revenue sources, including chips for automotive systems and cloud gaming, non-gaming-related revenues remained low. Fortunately for Huang, new customers found an unexpected use for GPUs: crypto mining.

Cryptocurrency mining The boom in Bitcoin brought explosive demand for GPUs. Cryptocurrency mining involved solving complex mathematical problems to validate transactions and create new units of the cryptocurrency, in a field where fractions of a second mattered. GPUs, with their parallel processing capabilities, were highly efficient at performing the calculations required for mining. By 2017, crypto mining farms were growing around the world, requiring significant computing power. To solve this, miners scooped up Nvidia and AMD's GPUs by the millions. This caused a shortage of gaming GPUs.[38] In response, Nvidia released a new line of GPUs in 2017 specifically designed for crypto mining. This allowed them to preserve the gaming market without sacrificing the crypto space.[39]

The success of GPUs in crypto mining highlighted a challenge that Huang wanted to address. While a single GPU would generally be used in a gaming console or PC, networking hundreds or even thousands of GPUs were needed to deliver high-performance compute for data centers. This led Nvidia to buy Mellanox in 2019 for almost $7 billion. Mellanox was an Israeli company that focused on developing chips for high-bandwidth networking. Fast communications between GPUs and other components in an AI cluster were critical for training and inference. Moreover, Mellanox was deeply embedded in data centers, giving Nvidia new routes into its new customers.

Crypto, however, turned out to be a volatile space. When miners experienced a "crypto winter" in the 4th quarter of 2018,[40] Nvidia's GPU revenue from mining halved. After recovering from the 2018/2019 winter, Nvidia's crypto segment dropped again in Q4 2022. However, this time crypto's losses were quickly offset by the launch of ChatGPT.

The Emergence of Artificial Intelligence

ChatGPT Elon Musk, Sam Altman, Ilya Sutskever, Greg Brockman, and John Schulman founded OpenAI as a non-profit lab in 2015. Transitioned to a for-profit company in 2019, OpenAI attracted top AI talent from Google and other large tech companies. Six months after becoming a for-profit company, Microsoft invested $1 billion into the business. OpenAI developed generative AI chatbots called GPTs, or Generative Pre-trained Transformers,[41] which were trained on extremely large datasets. (see **Exhibit 6**) After 7 years, OpenAI debuted ChatGPT in November 2022. This 3rd generation of GPT was trained on 175 billion parameters, compared to 120 million for GPT-1.[42] Within months, ChatGPT exploded in popularity, becoming the fastest consumer application in history to reach 100 million users.[43]

Given the size of their training datasets, OpenAI required massive computing power to train their models. OpenAI revealed that training ChatGPT 4, with roughly 1.7 trillion parameters, cost $100 million and required 25,000 Nvidia GPUs.[44] At the time, Nvidia's GPUs were the only chips powerful enough to train foundation models, and most developers relied on CUDA. As a result, demand for Nvidia's most powerful GPUs skyrocketed.[45] With Nvidia's GPUs in short supply, prices soared.[46] GenAI suddenly revolutionized Nvidia's business.

Nvidia's Business in 2024

Once a niche provider of graphic chips for video games, Nvidia had morphed into a diversified semiconductor, systems, and software company in 2024. While almost 90% of its revenues came from data centers ($30 billion in Q2 2024), Nvidia continued to lead in gaming GPUs ($2.9 billion in Q2 2024), while it was investing in chips for automotive applications ($346 million in Q2) and visualization chips for high-end workstations ($454 million in Q2).[47]

For high performance graphics, Nvidia's customers continued to be gamers and gaming enthusiasts. Nvidia's primary product in 2024 was its GeForce GPUs for PCs, and its Tegra GPUs for game consoles. GeForce GPUs were generally viewed as having higher performance with higher prices compared to AMD's Radeon GPUs. Intel held about two-thirds of the graphics market for PCs, while Nvidia GPUs captured about 18% and AMD 15% (primarily for gaming PCs).[48]

In the data center market, Nvidia's largest customers for its most advanced GPUs (H100s) were Microsoft and Meta (150,000 GPUs each in 2023), followed by AWS, Google Cloud, and Oracle (50,000 each, see **Exhibit 7**).[49] Analysts estimated that Nvidia sold roughly 3.7 million GPUs in 2023, and captured 98% market share in the data center.[50] Nearly half of Nvidia's revenues came from only four large data center customers.[51] Historically, Nvidia as well as AMD introduced new products on an 18-24 month cadence, but both companies were moving to annual upgrades. Nvidia planned to deliver a new chip designed for large LLMs called Blackwell in early 2025, followed by more advanced architectures in 2026 and 2027.[52] Blackwell promised 2.5X the performance for only 25% more cost. Yet maintaining these timelines would be challenging.[53] In August 2024, Huang confirmed that Blackwell would be late due to poor yields at TSMC.[54] To drive these improvements, Nvidia spent heavily on R&D. In 2023, it spent $8.6 billion on research and development, a 17% increase over 2022. However, as revenues surged, R&D fell to only 14% of sales (compared to 20% for competitor AMD).

Recognizing the inherent cyclical nature of semiconductors, Huang wanted to expand Nvidia's footprint from just selling GPUs and related networking to complete systems and cloud services for AI. Huang's biggest bet on this strategy was its attempt to purchase ARM in 2020, which licensed CPU technology for most cellphones and increasingly PCs and servers.[55] Huang's objective was "to boost ARM's R&D development and to strengthen its ecosystem by adding in Nvidia's graphics and AI IP to

ARM's portfolio."[56] However, British anti-trust authorities and the FTC worried that the $40 billion acquisition "would give one of the largest chip companies control over the computing technology and designs that rival firms rely on to develop competing chips."[57] In February 2022, Huang abandoned the ARM acquisition. Nonetheless, Nvidia introduced an ARM-based CPU for high performance computing and AI (called Grace) in 2021 and continued to support ARM's technology for the data center.[58]

Nvidia's next move was to build complete systems for data centers called DGX H100. The DGX 100 was a 375-pound box that could cost up to a half-million dollars. It was one of the highest performance systems for accelerating AI workloads in 2024. With up to 8 H100 GPUs, Nvidia incorporated its proprietary networking technology, which allowed these boxes to be connected and scaled. Nvidia hoped to sell a half million of these devices by the end of 2024.[59] Yet many customers viewed these boxes as premium priced, only appropriate for the most demanding AI workloads, and expensive to maintain. In addition, this strategy put Nvidia into direct competition with its OEM customers, such as Dell and HP.

In another move to get closer to the end customer, Nvidia started a server rental business in 2023 called DGX Cloud. Through a complicated partnership with some of their largest customers including Google, Microsoft, and Amazon, Nvidia rented "servers with its chips located in those companies' data centers. It then turned around and rented those servers to customers, including ServiceNow and Amgen."[60] This provided Nvidia with another revenue source in addition to direct hardware sales. Nvidia's main selling point was "promising better performance" on DGX Cloud by promising preferred access to its GPUs.[61]

A critical element of Nvidia's strategy was that virtually all of Nvidia's hardware products were dependent on TSMC in 2024. Access to the most advanced nodes allowed TSMC customers to deliver more transistors per square millimeter than its competitors (which generally meant better performance). In 2024, Apple and Nvidia were reputed to be TSMC's largest customers, which usually translated into preferred access to its leading-edge process technology. While TSMC was building fabs in Arizona and Japan, the vast majority of TSMC's leading-edge capacity remained in Taiwan. One potential challenge for Nvidia was that TSMC usually required 18-24 months to build a new fab, but Huang wanted Nvidia to operate on an annual cadence for new products.

Demand for GPUs

The sustainability of Nvidia's competitive advantage depended heavily on future demand for AI. Some analysts predicted generative AI could be a $1 trillion market opportunity for chip makers over the next decade, with Nvidia poised to capture most of that growth.[62] Other analysts predicted a less optimistic outlook but still forecasted a robust 30%+ CAGR in GPU revenues and volumes.[63] Finally, a few independent observers remained concerned about the slow adoption of Gen AI in the enterprise market, which could dramatically slow the torrent growth for GPUs.[64]

Some of the complexity of predicting demand for GPUs was related to the different customer segments. The largest GPU segment in volume was video game consoles and PCs. Roughly 65-70% of GPU units went to gaming hardware and cloud gaming services. While gaming boomed during the pandemic, there was some evidence that growth in consumer gaming might be slowing as TikTok and YouTube captured more of gamers' time with short videos.[65]

The second largest unit volume segment and largest revenue segment in 2024 was data centers, which was 25-30% of volume, but closer to 80% of revenues for Nvidia. Other segments, such as

automotive and healthcare, remained relatively small in 2024. The biggest data centers were buying virtually every high-end GPU they could find. Demand seemed insatiable. Larry Ellison, the chairman of Oracle reported, "I went to dinner with Elon Musk, Jensen Huang, and I. I would describe the dinner as me and Elon begging Jensen for GPUs. Please take our money; no, take more of it. You're not taking enough of it…"[66] But semiconductors had frequently experienced boom and bust cycles. Historically, when markets for semiconductors such as CPUs or memory boomed, customers chronically over-ordered, built excessive inventory, and later canceled orders. In addition, some analysts worried that the build-out of data centers may be overbuilt. Similar to the over-building in telecom in 2000, when companies such as WorldCom and Enron invested in dark fiber, hoping for future growth, some argued that the insane amount of investment going into GPUs and data centers would not be sustained.[67] Since there was no "killer app" for generative AI yet, and the cost of a query was as much as seven times the price of a Google search, some observers feared there was a "whiff of irrational exuberance" for Nvidia.[68] But enthusiasm for generative AI led many analysts to believe that this cycle was different. These analysts believed that enterprise customers investing heavily in GenAI in 2023-24 would quickly see huge benefits and continue buying cloud services.

A related question about AI demand for GPUs was the emergence of smaller LLMs, which could be "run on the edge" (e.g., end-user devices) as well as smaller data centers. While OpenAI, Anthropic, and others were building larger and larger LLMs, with 2 trillion or more parameters, companies including Meta, Apple, Alphabet, and numerous startups were also deploying smaller LLMs, which would not require as much processing power to train and run. Smaller models were also faster and lower cost to train, and lower cost to maintain.

One of the most talked about LLMs in 2024 was Llama from Meta. Unlike many of its competitors, Llama was open source and much smaller (8B, 70B, and 405B parameters in mid-2024).[69] Apple also announced its own AI model called Apple Intelligence, which used a small LLM that would run on the iPhone with Apple's ARM chips. Apple also used Google's TPU (described below) to train its AI, not Nvidia GPUs.[70] Similarly, in the summer of 2024, Microsoft announced a new "AI PC" that promised longer battery life and sophisticated AI processing without a GPU. It relied on a Qualcomm ARM CPU and a Qualcomm accelerator chip for inference.

One of the biggest technical question marks was demand for inference vs. training. GPUs were essential for training most large AI models. There were no good alternatives in 2024. But once a model was trained, there were several solutions for providing inference. While GPUs could deliver more computational power for demanding inference tasks, using GPUs for inference consumed more power, and they were more costly to buy and operate. Competitors that designed CPUs, including Intel, AMD, and Qualcomm, were pitching data centers to switch from GPUs to CPUs and special accelerator chips designed specifically for AI inference workloads. The total cost of ownership (TCO) for CPUs with accelerators were potentially much lower than the TCO of high-end Nvidia GPUs. But Huang frequently argued that Nvidia's superior performance generated superior returns for customers.

Another big unknown about demand for American companies, including Nvidia and AMD, was restrictions on sales of GPUs to China. The U.S. imposed export controls on advanced semiconductors, including GPUs. China comprised 14% of Nvidia's data center sales in FY 2024, down from 19% the year before.[71] Additionally, some of Nvidia's GPUs, including those specifically developed for China, could not be sent without an export license. As a result, Nvidia was expecting further declines in Chinese sales. Nvidia also stated it was planning to develop chips that could be sent to China while still complying with the trade restrictions.[72] Whether the U.S. President in 2025 would allow such sales was unknown. Finally, there was evidence that China had circumvented the U.S. chip export ban by

using shell companies and shipping via Hong Kong.[73] If U.S. policy cracked down on these shipments, American GPU sales to China might be severely impacted.

Competition

Inevitably, the combination of high market share, forecasted growth, and customer concerns over supplier concentration would induce new competitors to enter the fray. The challenge for would-be competitors was that Nvidia GPUs led in performance, networking, and programming with CUDA.

Direct Competitors: AMD and Intel

AMD Nvidia's most direct rival was Advanced Micro Devices (AMD). In 2014, Lisa Su (Huang's first cousin once removed) took charge as CEO. A PhD in electrical engineering from MIT, she turned AMD into a leading semiconductor player after decades of the company playing second fiddle to Intel. By 2024, AMD had captured roughly 35% of the CPU market for PCs[74] and 24% of CPUs in the data center.[75] While CPUs were AMD's core business, it entered the GPU market in 2006 when it purchased ATI for $5.4 billion.[76] At the time, ATI had roughly a quarter of the GPU market.[77] While Nvidia remained the performance leader, AMD gained market share in gaming by capturing the PlayStation business for Sony, and Xbox for Microsoft.

In an effort to match (or surpass) Nvidia after the launch of ChatGPT, AMD invested aggressively in higher performance GPUs. However, AMD's best GPU (MI250X) delivered roughly 380 teraflops,[78] while Nvidia's H100 could deliver up to 1,000 teraflops.[79] One independent analyst also found that Nvidia's next generation GPU, Blackwell (H200), ran 44% faster than AMD's best GPUs. But some analysts predicted that AMD would eventually reduce the performance gap. AMD also priced its GPU significantly lower ($8,000-10,000 vs. $25,000-40,000).[80]

In August 2024, AMD made its biggest move when it bought ZT Systems for $4.9 billion.[81] ZT designed data center equipment for cloud computing companies, and sold servers, racks, and other infrastructure. AMD was hoping it would allow them to compete with Nvidia's DGX100. While the DGX100 allowed very limited customization, ZT systems could build more customized systems for customers. Lisa Su stated emphatically that "there is no one size that is going to fit all…openness will allow the ecosystem to innovate."[82]

AMD was gaining some traction, raising its expected revenue from data center GPUs from $4.0 to $4.5 billion in 2024.[83] Yet AMD struggled with the same supply problem as Nvidia. While Nvidia was expecting several million GPUs from TSMC in 2024, AMD would only get a few hundred thousand.

Intel The volume leader in graphics chips remained Intel in 2024. As the leading CPU company, Intel started to sell low-end graphics for PCs in the late 1990s. Over time, Intel embedded low-performance graphics into the chipsets it sold for PCs, and it captured between 70-80% of that market.[84] Around 2006, Intel leadership recognized the potential market for high-end graphics, and Pat Gelsinger (Intel's CEO in 2024) was assigned to build the chip. After four years and hundreds of millions of dollars, the project was canceled (and Gelsinger left the company for a decade). Intel continued to invest in graphics, including purchasing several AI chip companies that offered specialized accelerators for training and inference. Its biggest purchase was Habana Labs for $2 billion in 2019. Habana's Gaudi2 processor was positioned to compete with Nvidia's H100, at roughly half the price point ($8,000-10,000). Like AMD, this made Gaudi2 a cost-effective alternative. However, Gaudi2 did not yet match the H100's performance, its next-generation chip to compete with Blackwell was not expected until late 2025, and its ecosystem support and software were not as mature as CUDA.[85]

Customer-Competitors: Alphabet and Amazon

Major tech giants including Alphabet, Amazon, and Microsoft viewed developing their own AI chips in-house as a way to reduce costs and eliminate shortages. This led to a race among hyperscalers to design AI chips that could replace or supplement Nvidia.

Google's TPU In December 2023, alongside the debut of Google LLM Gemini, Alphabet announced the release of their newest chip called Google Cloud TPU v5p, the 5th generation of their TPU line. TPUs, or tensor processing units, were "Google's custom-developed application-specific integrated circuits (ASICs) used to accelerate machine learning workloads."[86] Unlike GPUs, which were designed for parallel processing and were modified for AI processing, Google specifically designed TPUs for neural networks, including training and inference.[87] This efficiency in operations meant that Google Cloud's TPUs consumed less energy than Nvidia's chips: Nvidia's Tesla V100 and A100 utilized between 250-400 watts per card, while Google Cloud TPU v3 and v4 used 120-150 and 200-250 watts per chip, respectively.[88] It also saved Google a significant amount of money: while the company reportedly purchased Nvidia chips for around $15,000 each, Google spent between $2,000-3,000 for each TPU it designed in-house. As one analyst stated, "When [Google] encountered a vendor that held them over a barrel, they reacted very strongly."[89]

Amazon Amazon started developing chips in 2015 through its acquisition of the Israeli chip-design firm Annapurna Labs. Through this acquisition, Amazon released three chips; Graviton, Trainium, and Inferentia. Graviton was a CPU that Amazon claimed had "40% better price performance than comparable x86 chips."[90] By 2024, Amazon had released the 4th generation of Graviton, which had significant improvements in computing performance and memory over the previous three generations.[91] Graviton4 chips were also "based on Arm architecture and consume less energy than chips from Intel or AMD."[92] According to Amazon, more than 50,000 AWS customers were already utilizing Graviton chips by late 2023.[93] Trainium was used by large AI companies such as Anthropic for training LLMs (a company which received a $4 billion investment from Amazon)[94], and Inferentia was utilized for AI inference workloads by customers including Airbnb and ByteDance. Amazon also used Inferentia for powering Alexa, Amazon Ads, and Rufus, its new shopping bot.[95] According to an Amazon VP, Trainium and Inferentia offered "up to 40%, 50% in some cases of improved price (and) performance – so it should be half as expensive as running that same model with Nvidia."[96] With AWS controlling more than one-third of the cloud computing market, efficiency gains of 40-50% across their three AI chips meant real financial benefits.

In November 2023, Amazon released the second generation of its AI chip Trainium2, which was 4x faster for training than the first-generation chip, while improving energy efficiency by up to 200%.[97] In July 2024, Amazon announced that it was focused on developing a new series of AI chips that were less expensive and faster than Nvidia's GPUs. Its goal was to lessen its reliance on Nvidia and decrease the "Nvidia tax", or the premium paid for the company's chips. Andy Jassy commented that he had "heard loud and clear from customers that they relish better price performance."[98]

Startups Numerous startups were building alternatives to Nvidia GPUs, receiving millions in funding to develop chips powerful enough for AI and data centers. One of the best-known startups was Cerebras Systems, which developed one singular chip (rather than Nvidia's multi-chip approach) for GPU capabilities, central processing, and memory. According to CEO Andrew Feldman, this single chip model was better for training LLMs. "We use a giant chip, they [Nvidia] use a lot of little chips. They've got challenges of moving data around, we don't."[99] At their most recent funding round in 2021, Cerebras Systems was valued at $4 billion. In August 2024, Cerebras released Cerebras Inference, their fastest AI product which claimed to be 20X faster than Nvidia and one-fifth the cost per token.[100]

However, Cerebras was not as versatile as Nvidia; while Nvidia's chips could be used across many applications, Cerebras' chips could not.[101] Additionally, CUDA was much more developed and well-known than the software required to program with Cerebras' chips.

Venture capitalists were jumping on the AI bandwagon as well, investing $6 billion into AI semiconductor companies in 2023 alone.[102] A startup known as D-Matrix was founded in 2019 and received $100 million of funding in September 2023. They announced they were releasing "a semiconductor card for servers" in late 2024 which would "reduce the cost and latency of running AI models." In July 2024, another startup named Groq received $300 million in additional funding from a VC group led by Blackrock, bringing the startup's valuation to $2.2 billion. Founded in 2016 by Jonathan Ross, a Google veteran who helped develop the first set of TPUs, Groq received additional attention in early February 2024 when it posted videos showing that its chips could power LLMs "in a fraction of a second."[103] This speed was due to Groq's chip architecture, which was designed specifically for LLMs. However, this speed also meant limited versatility:

> Groq's chips are so fast because they're specialized for the architecture underlying LLMs. However, there's a tradeoff between speed and flexibility: one reason why Nvidia's chips are so popular… is because they can be more easily used for different types of AI models.[104]

Finally, there was a cost issue; as CEO Ross explained, Groq had pivoted to cloud computing services for developers, since it was cost-prohibitive for other customers.[105]

Jim Keller, the CEO of Tenstorrent (another start-up competitor to Nvidia), believed that new, smaller, more efficient AI models could alter the landscape for data centers. "There could be a model published next month that takes computation [costs] down ten times," according to Keller. This would leave the data center world with massive overcapacity. Having worked at AMD and Intel, Keller noted that "semiconductors are cyclical. Every time there's an upcycle, sooner or later, somebody builds too much capacity. And then when there's a crash, everyone says 'you should have seen that coming." He thought many companies would go bankrupt because "There's already a gap between what people are paying to train models versus what they're charging [for AI]…they're already over their skis."[106] Indeed, even OpenAI seemed likely to lose $5 billion in 2024.[107]

Competition for CUDA: To reduce dependence on Nvidia, customers searched for alternatives to CUDA. One of the most important customer concerns was code developed with CUDA could not be transferred to non-Nvidia GPUs or accelerators. As an AWS vice president explained, "Nvidia's got great chips, and more importantly, they have an incredible ecosystem," making it "very, very challenging" to get customers to utilize other chips.[108] In 2021, OpenAI released an open-source language called Triton, which was similar to Python and allowed "researchers with no CUDA experience to write highly efficient GPU code."[109] Triton also allowed developers to run code on many types of chips, not just Nvidia ones. By 2024, Microsoft, Meta, and Google were investing in Triton to diversify away from their exclusive dependence on CUDA, since Triton could be utilized on Nvidia GPUs as well as other GPU brands. As Intel's CTO Greg Lavender explained, "Essentially it breaks the CUDA lock-in."[110] Jim Keller (CEO of Tenstorrent), argued that CUDA "is a swamp, not a moat," because the software was too complicated.[111] CUDA, he argued, could have a similar fate to UNIX, which ultimately lost out to easier-to-use open-source alternatives, like Linux.[112]

AMD also worked on developing an alternative to CUDA, called ROCm. Introduced in 2016,[113] AMD described this as an "open software stack that includes programming models, tools, compilers, libraries, and runtimes for AI and HPC solution development on AMD GPUs." However, ROCm never

gained the popularity of CUDA, and by 2024, AMD was developing other approaches for building out its software.[114]

Jensen Huang's decisions in late 2024

While Nvidia's short-term prospects seemed unassailable, the semiconductor industry was a business of constant innovation and new challenges. The immediate challenge was how to allocate and price in a world of shortages and exploding demand. The longer-term challenges related to how far up or down the value chain should Nvidia play. With exploding revenues and cashflow, all options were on the table.

Product allocation might seem easy: let the market decide. Run an auction that would likely produce very high prices, high margins, and a highly concentrated customer base. The problem was that Huang knew that it was important to balance Nvidia's relationship with strategic customers, such as large cloud providers, with OEM partners, such as Dell and HP, and spread GPUs across all key customer segments, such as Data Centers vs. gaming. In addition, Nvidia was committed to supporting a broad ecosystem, which meant it wanted researchers and software developers to have access to leading-edge products. At the same time, Huang had to worry about customers hoarding and over-purchasing GPUs and then canceling; or customers buying GPUs and scalping them. Finally, Huang didn't want GPUs concentrated in one geography.

Instead of an auction, a second approach was to use Nvidia's virtual monopoly power to price discriminate and charge different customers different prices, based on likely willingness to pay. Andrew Feldman, CEO of Cerebras, argued that this was happening in 2023-4, and "They're just extorting their customers, and nobody will say it out loud."[115] Another competitor argued that Nvidia was charging a reverse premium: While the largest customers usually got volume discounts, Nvidia could charge a higher price to customers who wanted more allocation.[116] Nvidia could also allocate "strategically," intentionally favoring some customers over others. For example, Nvidia gave CoreWeave, a small, private venture-backed cloud computing company that specialized in GPU-intensive AI compute, a significant allocation (40k GPUs – See **Exhibit 7**). Alternatively, it could create a queue: customers who bought low-end, lower margin GPUs and other Nvidia products could be given higher priority and more allocation. In other words, Huang had many alternatives. Yet he had to make choices that could pass muster with anti-trust authorities investigating the company. He insisted publicly that Nvidia allocated "fairly," but what was "fair"?[117]

The longer-term questions of where to play in the value chain were a never-ending struggle for a semiconductor company like Nvidia. On the one hand, it was deeply dependent on a single supplier (TSMC) and increasingly vulnerable to highly concentrated and sophisticated hyperscalers. Down the road, it could be squeezed on both sides. On the supplier side, Nvidia had a close relationship with TSMC, but TSMC had a history of raising prices in periods of high demand; moreover, TSMC's dependence on production in Taiwan created geopolitical risks. The most likely alternative to TSMC was Intel, which was behind TSMC in process technology in 2024. Intel had committed tens of billions of dollars to build out a foundry business that would be competitive with TSMC, but Intel was not expected to match TSMC's foundry capabilities for several years. Intel was also a competitor to Nvidia. Huang publicly explained that: "In the event that we have to shift from one fab to another, we have the ability to do it. We won't be able to get the same level of performance or cost, but we will be able to provide supply."[118]

On the customer side, the biggest cloud providers - AWS, Google, and Microsoft - were developing GPU alternatives. Huang was confident that Nvidia could out-innovate the cloud providers, but

would that be enough to retain the business? Should Nvidia push aggressively forward to build a direct connection with AI customers? Would competition with cloud providers accelerate their move to alternative products? Or should Nvidia stay the course, compete with AWS, Google and Microsoft, and use its performance advantage to demonstrate to the final customer that Nvidia GPUs were the best solution? Finally, by offering data center customers complete systems, such as the DGX H100, Nvidia was going head-to-head with Dell, HP, Lenovo and other OEMs who built servers. Since Nvidia chip prices were very high, most OEMs struggled to make good margins on Nvidia-based servers.[119] As Nvidia moved into servers, OEMs could see the writing on the wall: If Nvidia sells servers in volume directly to data centers, what role will they play? If competitors to Nvidia emerge at the chip level, will OEMs flee? When Intel tried to do something similar seven years earlier, its largest OEM customers defected to AMD, forcing Intel to backtrack.[120]

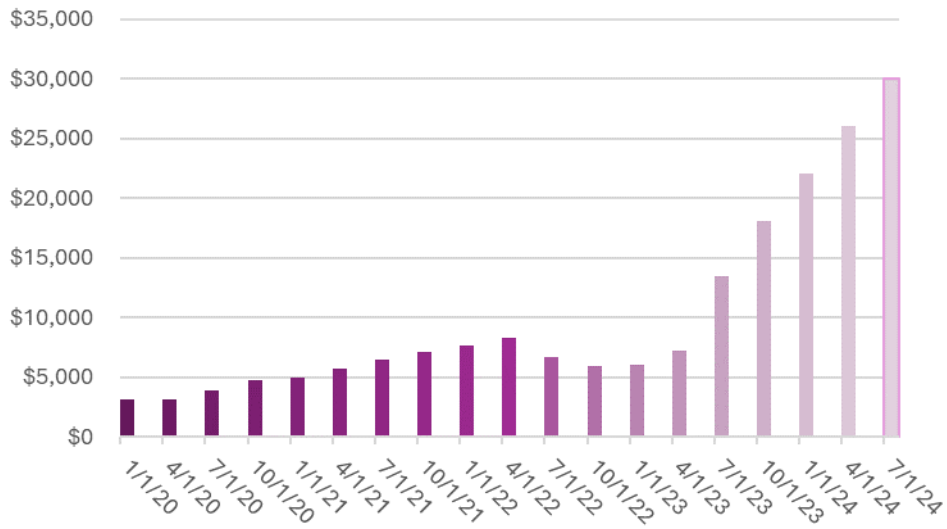
Despite these longer-term strategic challenges in late 2024, every firm in the tech world looked with awe on Jensen Huang and Nvidia. Its astounding growth in revenues and profits since the launch of ChatGPT was the envy of the world.

Exhibit 1a Nvidia Financial Summary (in $ millions) by Fiscal Year (Feb – Jan)

	2010	2015	2020	2021	2022	2023	2024
Revenue	3,326	4,682	10,918	16,675	26,914	26,974	60,922
COGS	2,138	2,083	4,150	6,118	9,439	11,618	16,621
R&D	818	1,360	2,829	3,924	5,268	7,339	8,675
SG&A	329	480	1,093	1,912	2,166	2,440	2,654
Operating income (loss)	41	759	2,846	4,721	10,041	5,577	32,972
Net income	68	631	2,796	4,332	9,752	4,368	29,760
EBITDA	238	979	3,227	5,819	11,215	7,121	34,480
Cash and cash equivalent	447	497	10,896	847	1,990	3,389	7,280
Total assets	3,586	7,201	17,315	28,791	44,187	41,182	65,728
Total liabilities	921	2,783	5,111	11,898	17,575	19,081	22,750
Total shareholders' equity	2,665	4,418	12,204	16,893	26,612	22,101	42,978
Gross margin	36%	56%	62%	63%	65%	57%	73%
Return on equity	(3%)	14%	26%	30%	45%	18%	92%
Data Center Revenue	-	-	2,983	6,696	10,613	15,005	47,525
Gaming Revenue	-	-	5,518	7,759	12,462	9,067	10,447

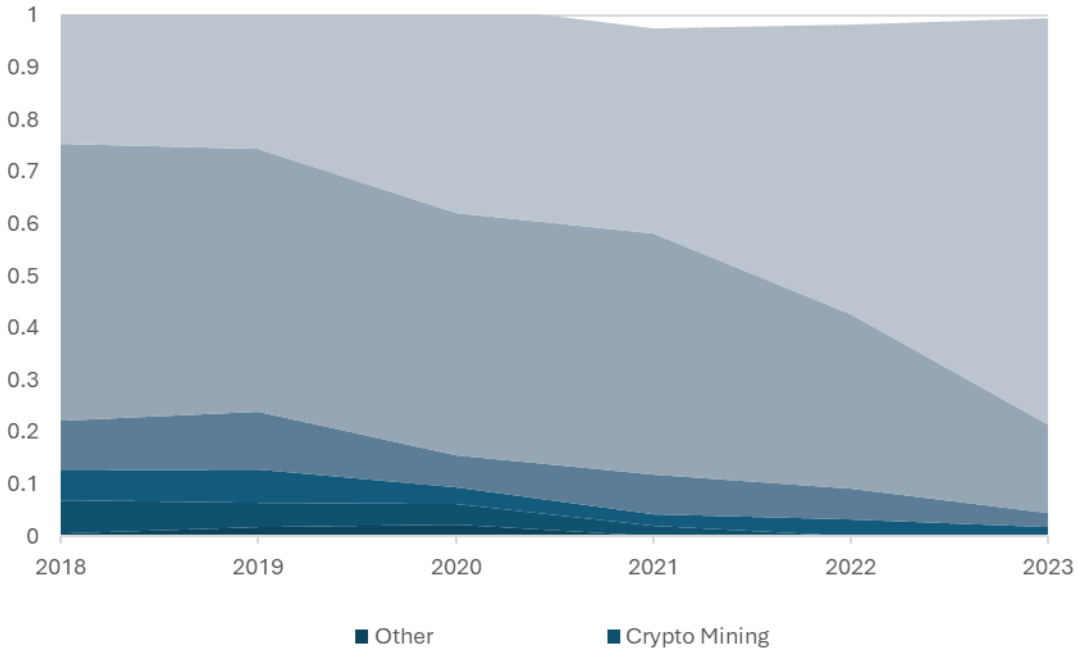
Source: Created by casewriter using data from Nvidia's 10-Ks and Capital IQ, accessed August 14, 2024.

Exhibit 1b Nvidia Quarterly Revenue by Calendar Quarters, 2020-2024



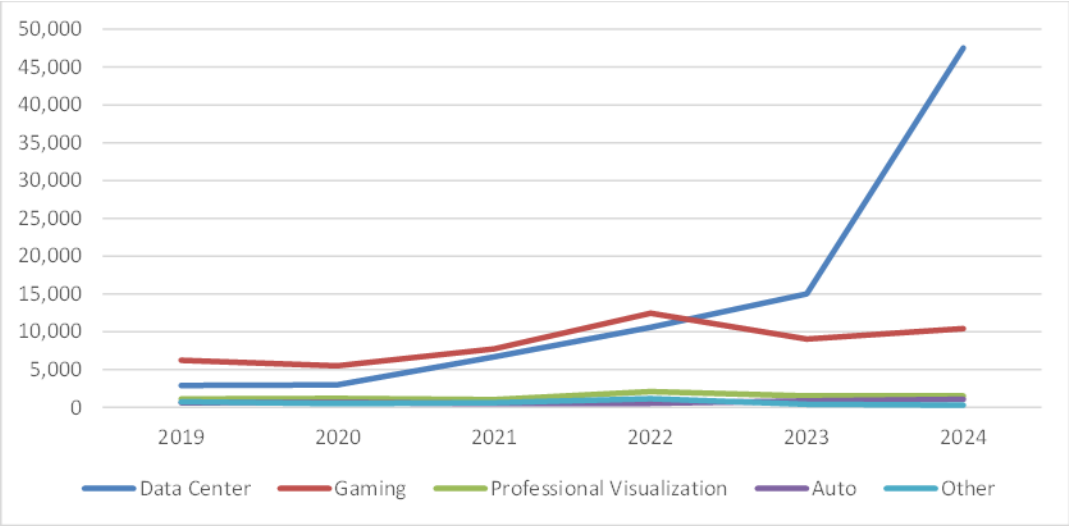
Source: Created by casewriter using data from Statista, "Nvidia revenue worldwide from fiscal year 2021 to 2025." https://www.statista.com/statistics/1424787/nvidia-market-revenue-by-quarter/ , accessed September 10, 2024.

Exhibit 1c Percentage of Nvidia's Revenue by Product Line, 2018-2023 (by Calendar Year)



Source: Created by casewriter using data from Stratechery, "Nvidia on the Mountaintop," https://stratechery.com/2023/nvidia-on-the-mountaintop/, accessed September 22, 2024.

Exhibit 1d Nvidia's Global Revenue by Market Segment in $M USD, FY 2019-2024



Source: Created by casewriter using data from Statista, "Nvidia Specialized market revenue worldwide from fiscal year 2019 to 2025, by quarter," https://www-statista-com.ezp-prod1.hul.harvard.edu/statistics/1120484/nvidia-quarterly-revenue-by-specialized-market/, accessed September 7, 2024.

Exhibit 2 TSMC, Intel, and AMD Financial Summary (in $ millions) by Year

TSMC Financial Summary (in $ millions) by Fiscal Year (same as Calendar Year)

	2010	2015	2020	2021	2022	2023
Revenue	14,388	25,580	47,672	57,281	73,692	70,557
COGS	7,287	13,135	22,359	27,708	29,802	32,202
R&D	1,019	1,988	3,897	4,501	5,314	5,952
SG&A	623	695	1,266	1,605	2,065	2,333
Operating income (loss)	5,459	9,759	20,169	23,464	36,525	30,075
Net income	5,542	9,184	18,181	21,526	33,089	27,368
EBITDA	8,394	16,456	31,751	38,457	50,502	47,160
Cash and cash equivalent	5,072	17,064	23,500	38,429	43,710	47,830
Total assets	24,656	50,262	98,267	134,432	161,609	108,572
Total liabilities	4,809	14,024	32,921	56,102	65,242	66,881
Total shareholders' equity	19,847	36,239	65,346	78,329	96,367	113,691
Gross margin	49%	49%	53%	52%	60%	54%
Return on equity	30	27%	30%	30%	40%	26%

Source: Compiled by casewriter using data from company documents, accessed August 14, 2024.

Exhibit 2 Intel Financial Summary (in $ millions) by Fiscal Year (same as Calendar Year) (continued)

	2010	2015	2020	2021	2022	2023
Revenue	43,623	55,355	77,867	79,024	63,054	54,228
COGS	15,132	20,676	34,255	35,209	35,949	32,517
R&D	6,576	12,128	12,556	15,190	17,528	16,046
SG&A	6,309	7,930	6,180	6,543	7,002	5,634
Operating income (loss)	15,588	14,356	22,082	22,082	2,575	31
Net income	11,464	11,420	20,899	19,868	8,014	1,689
EBITDA	20,226	23,067	36,115	33,874	15,610	9,633
Cash and cash equivalent	5,498	15,308	5,865	4,827	11,144	7,079
Total assets	63,186	101,459	153,091	168,406	182,103	191,572
Total liabilities	13,756	40,374	72,053	73,015	78,817	81,607
Total shareholders' equity	49,430	61,085	81,038	95,391	103,286	109,965
Gross margin	65%	63%	56%	55%	43%	40%
Return on equity	25%	20%	26%	23%	8%	2%

Source: Complied by casewriter using data from company documents, accessed August 14, 2024.

Exhibit 2 AMD Financial Summary (in $ millions) by Fiscal Year (same as Calendar Year) (continued)

	2010	2015	2020	2021	2022	2023
Revenue	6,494	3,991	9,763	16,434	23,601	22,680
COGS	3,632	2,911	5,416	8,505	11,550	11,278
R&D	1,405	947	1,983	2,845	5,005	5,872
SG&A	934	482	995	1,448	2,336	2,352
Operating income (loss)	462	(352)	1,369	3,648	1,264	401
Net income	471	(660)	2,490	3,162	1,320	854
EBITDA	845	(185)	1,681	4,055	4,700	3,854
Cash and cash equivalent	447	497	10,896	847	1,990	3,389
Total assets	3,586	7,201	17,315	28,791	44,187	41,182
Total liabilities	921	2,783	5,111	11,898	17,575	19,081
Total shareholders' equity	2,665	4,418	12,204	16,893	26,612	22,101
Gross margin	36%	56%	62%	63%	65%	57%
Return on equity	(3%)	14%	26%	30%	45%	18%

Source: Compiled by casewriter using data from company documents, accessed August 14, 2024.

Exhibit 3 Cost of GPUs (in USD), and their Selling Prices

Brand	Model	Estimated Cost	Selling Price
Nvidia	H100	$3,100	$25,000-40,000
Nvidia	B200 (Blackwell series)	>$6,000	$30,000-40,000
AMD	MI300X	n/a	$10,000-15,000

Source: Compiled by the casewriter using data from Raymond James and Tom's Hardware "Nvidia's Jensen Huang says Blackwell GPU to cost $30,000-$40,000, later clarifies that pricing will vary as they won't sell just the chip," (https://www.tomshardware.com/pc-components/gpus/nvidias-jensen-huang-says-blackwell-gpu-to-cost-dollar30000-dollar40000-later-clarifies-that-pricing-will-vary-as-they-wont-sell-just-the-chip, https://www.tomshardware.com/tech-industry/artificial-intelligence/nvidias-h100-ai-gpus-cost-up-to-four-times-more-than-amds-competing-mi300x-amds-chips-cost-dollar10-to-dollar15k-apiece-nvidias-h100-has-peaked-beyond-dollar40000), and TechRadar, "AMD pulverizes Nvidia's RTX 4090 in popular Geekbench OpenCL benchmark – but you will need a small mortgage to buy AMD's fastest GPU ever produced." (https://www.techradar.com/pro/amd-pulverizes-nvidias-rtx-4090-in-popular-geekbench-opencl-benchmark-but-you-will-need-a-small-mortgage-to-buy-amds-fastest-gpu-ever-produced), accessed September 5, 2024.

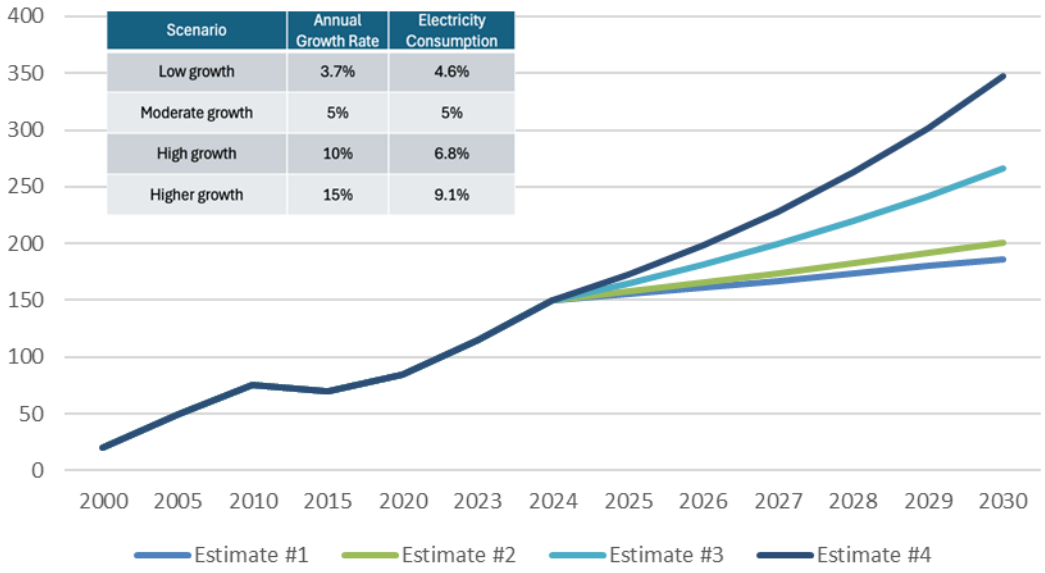
Exhibit 4 TSMC Pricing for their Wafers

Wafer size	Price	Used in
3nm	$20,000	Nvidia's H200 and B100

Source: Created by casewriter using data from PCgamer, "TSMC may increase wafer pricing by 10% for 2025: Report," https://www.tomshardware.com/tech-industry/tsmc-may-increase-wafer-pricing-by-10-for-2025-report, accessed September 5, 2024.

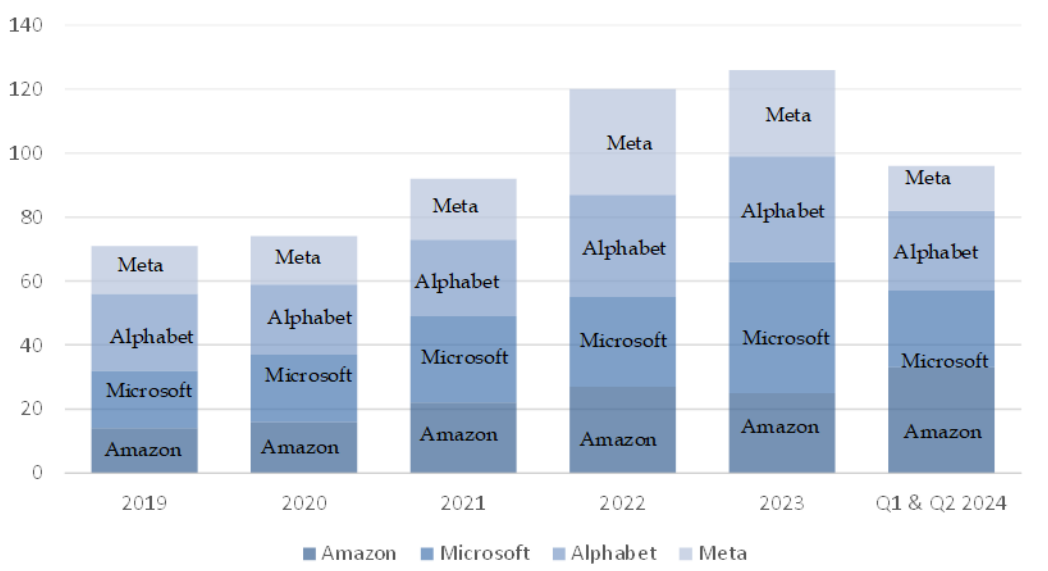
Note**:** If TSMC was able to attain high yields of close to 90%, a 3nm wafer would produce roughly 100-120 B100 GPUs (Nvidia's new Blackwell GPUs). If yields were low, which was the reported to be the case in 2024, the number of chips would be substantially less.

Exhibit 5a Current and Forecasted Commissioned Power for U.S. Data Centers, 2000-2030



Source: Created by casewriter using data from EPRI, "Powering Intelligence: Analyzing Artificial Intelligence and Data Center Energy Consumption." (https://www.epri.com/research/products/000000003002028905)

Exhibit 5b CapEx on Data Center Power by Major Company, 2019-2023 & 2024 Q1-Q2



Source: Compiled by casewriter using data from WSJ "The AI Spending Spree, in Charts," and GeekWire, "Capital Spending soars in the cloud as Microsoft, Google, and others bet big on AI demand." https://www.wsj.com/tech/ai/artificial-intelligence-investing-charts-7b8e1a97?mod=Searchresults_pos2&page=1, https://www.geekwire.com/2024/capex-and-the-cloud-microsoft-google-and-other-tech-giants-are-betting-big-on-ai-demand/, accessed September 11, 2024.

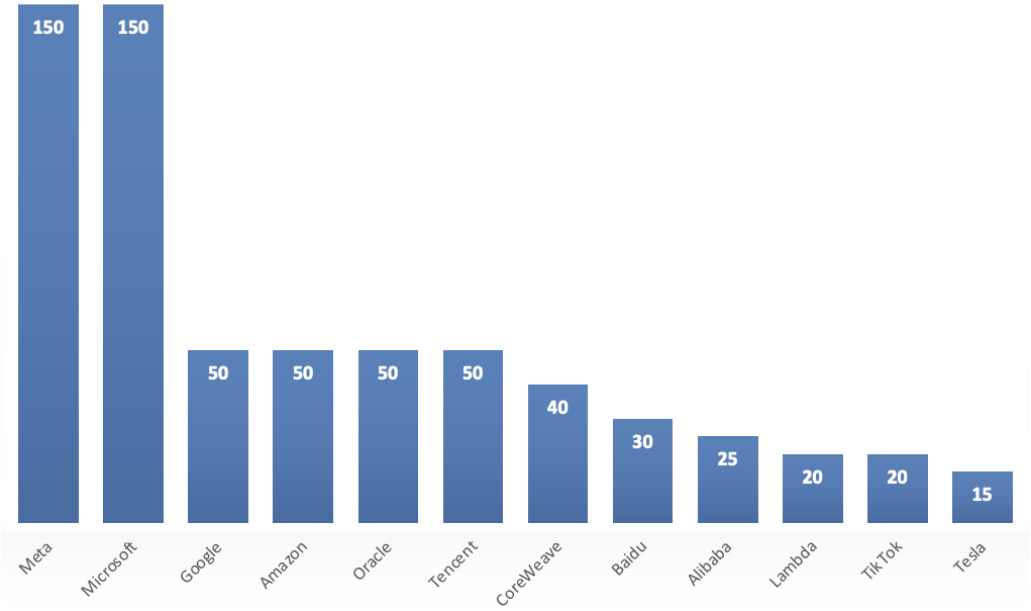
Exhibit 6 LLMs and their Parameters

Large Language Model	Parameters
Phi-1.5	1.3B
Phi-2	2.7B
Llama2	7B, 13B, or 70B
Llama 3	8B, 70B, and 400B
BloombergGPT	50B
Claude2	130B
GPT-3	175B
GPT-4 "32k"	1.76T

Source: Compiled by casewriter using data from Kelvin Legal, "Understanding Large Language Models – Parameters," https://kelvin.legal/understanding-large-language-models-what-are-paramters/, accessed September 5, 2024.

Note: The parameters for OpenAI's GPT are estimates.

Exhibit 7 Shipment estimates of Nvidia H100 GPUs worldwide in 2023, by customer



Source: Created by casewriter using data from Statista, "Estimated shipments of Nvidia H100 graphics processing units (GPUs) worldwide in 2023, by customer," https://www.statista.com/statistics/1446564/nvidia-h100-gpu-shipments-by-customer/, accessed August 12, 2024.

Endnotes

[1] https://www.reuters.com/technology/nvidia-faces-us-doj-probe-over-complaints-rivals-information-reports-2024-08-02/, accessed August 26, 2024.

[2] Ian King and Leah Nylen, "Nvidia Gets DOJ Subpoena in Escalating Antitrust Probe." *Bloomberg*, September 3, 2024, https://www.bloomberg.com/news/articles/2024-09-03/nvidia-gets-doj-subpoena-in-escalating-antitrust-investigation, accessed September 9, 2024.

[3] "Jensen Huang". *Stanford Engineering*, https://engineering.stanford.edu/about/heroes/2018-heroes/jensen-huang, accessed May 10, 2024.

[4] Jensen Huang, Founder and CEO of NVIDIA." *Stanford Graduate School of Business*, YouTube, March 5, 2024, https://www.youtube.com/watch?v=lXLBTBBil2U, accessed June 5, 2024.

[5] Jyoti Mann, "Jensen Huang's 14-hour days and workaholic lifestyle helped him turn Nvidia into a $3 trillion company." *Business Insider*, June 7, 2024, https://www.businessinsider.com/jensen-huang-nvidia-routine-management-leadership-style-ceo-founder-2024-4 , accessed June 9, 2024.

[6] "NVIDIA CEO Jensen Huang." *Acquired Interviews*, October 15, 2023, https://www.acquired.fm/episodes/jensen-huang, accessed June 10, 2024.

[7] Nvidia 10-Ks, 2000 and 2024.

[8] https://fxopen.com/blog/en/analytical-nvidia-stock-forecast-for-2024-2025-2030-and-beyond/, accessed September 16, 2024.

[9] Dean Takahashi, "Shares of Nvidia Surge 64% After Initial Public Offering." *Wall Street Journal*, Jan 25, 1999, accessed Aug 26, 2024.

[10] Dominic Kesterton, "The Big Number: $3.34 Trillion." *The New York Times*, June 21, 2024, https://www.nytimes.com/2024/06/21/business/3-34-trillion-nvidia-market-value.html, accessed July 10, 2024

[11] Dominic Kesterton, "The Big Number: $3.34 Trillion." *The New York Times*, June 21, 2024, https://www.nytimes.com/2024/06/21/business/3-34-trillion-nvidia-market-value.html, accessed July 10, 2024.

[12] "NVIDIA Corporation (NVDA)." *Yahoo Finance*, August 26, 2024, https://finance.yahoo.com/quote/NVDA/, accessed August 26, 2024.

[13] "Jensen Huang." *Bloomberg Billionaires Index*, https://www.bloomberg.com/billionaires/profiles/jenhsun-huang/, accessed August 26, 2024.

[14] "The rise of Jensen Huang, the Nvidia CEO who was born in Taiwan, raised in Kentucky, and is now one of the richest men on earth." *Fortune*, Feb 22, 2204, https://fortune.com/2024/02/22/who-is-jensen-huang-nvidia-net-worth-biography-success/, accessed June 12, 2024.

[15] "Nvidia Part 1: The GPU Company (1993-2006)." *Acquired*, Season 10, Episode 5, https://www.acquired.fm/episodes/nvidia-the-gpu-company-1993-2006 , accessed June 10, 2024.

[16] "Nvidia Part 1: The GPU Company (1993-2006)." *Acquired*, Season 10, Episode 5, https://www.acquired.fm/episodes/nvidia-the-gpu-company-1993-2006 , accessed June 10, 2024.

[17] "Nvidia Part 1: The GPU Company (1993-2006)." *Acquired*, Season 10, Episode 5, https://www.acquired.fm/episodes/nvidia-the-gpu-company-1993-2006 , accessed June 10, 2024.

[18] Intel 2001 10-K and Nvidia 2001 10-K.

[19] Stephen Witt, "How Jensen Huang's Nvidia is Powering the A.I. Revolution." *The New Yorker*, November 27, 2023, https://www.newyorker.com/magazine/2023/12/04/how-jensen-huangs-nvidia-is-powering-the-ai-revolution#:~:text=For%20decades%2C%20the%20major%20manufacturer,up%20my%20chips%20and%20run.%E2%80%9D, accessed August 14, 2024.

[20] Alexandra Alper, Stephen Nellis, and Heekyong Yang, "Exclusive: Samsung's new Texas chip plant cost rises above $25 billion." *Reuters*, March 15, 2023, https://www.reuters.com/technology/samsungs-new-texas-chip-plant-cost-rises-above-25-billion-sources-2023-03-15/, accessed August 26, 2024.

[21] "Wait times for Nvidia's AI GPUs ease to three to four months, suggesting peak in near-term growth – the wait list for an H100 was previously eleven months: UBS." *Tom's Hardware*, Feb 16, 2024, https://www.tomshardware.com/tech-industry/artificial-intelligence/wait-times-for-nvidias-ai-gpus-eases-to-three-to-four-months-suggesting-peak-in-near-term-growth-the-wait-list-for-an-h100-was-previously-eleven-months-ubs, accessed August 26, 2024.

[22] Eric Cheung and Will Ripley, "Everyone wants the latest chips. That;s causing a huge headache for the world's biggest supplier." *CNN*, March 22, 2024, https://www.cnn.com/2024/03/22/tech/taiwan-tsmc-talent-shortage-training-center-intl-hnk/index.html#:~:text=Sometimes%20called%20the%20most%20important,smartphones%20to%20artificial%20intelligence%20applications, accessed August 29, 2024.

[23] Anton Shilov, "TSMC Gets Billions in Pre-Payments for Fab Capacity." *Tom's Hardware*, November 17, 2021, https://www.tomshardware.com/news/tsmc-collects-huge-prepayments, accessed August 29, 2024.

[24] "Graphic card price inflation over the years." *Reddit*, https://www.reddit.com/r/pcmasterrace/comments/856hzf/graphic_card_price_inflation_over_the_years/. Accessed August 29, 2024.

[25] "NVIDIA GeForce RTX 4080 Founders Edition Graphics Card 16GB GDDR6X- Titanium and Black." https://www.newegg.com/nvidia-founders-edition-video-cards-900-1g136-2560-000-geforce-rtx-4080-16gb-gddr6x/p/1FT-0004-00844?item=9SIB9ZZJN44231&nm_mc=knc-googleadwords&cm_mmc=knc-googleadwords-_-video%20card%20-%20nvidia-_-nvidia-_-9SIB9ZZJN44231&utm_source=google&utm_medium=organic+shopping&utm_campaign=knc-googleadwords-_-video%20card%20-%20nvidia-_-nvidia-_-9SIB9ZZJN44231&source=region&srsltid=AfmBOop8uA_lQSH2fsHM67bTwMkSqIbC-RYH8-_VLthAsPawhyHA7sLeOGQ, accessed August 12, 2024.

[26] "How many transistors are in your Intel CPU, and what is its frequency (GHz)? How do you calculate it?" https://www.quora.com/How-many-transistors-are-in-your-Intel-CPU-and-what-is-its-frequency-GHz-How-do-you-calculate-it, accessed August 29, 2024.

[27] Janakiram MSV, "10 Interesting Facts about Nvidia Hopper H100 GPU." *Forbes*, March 29, 2022, https://www.forbes.com/sites/janakirammsv/2022/03/27/10-interesting-facts-about-nvidia-hopper-h100-gpu/, accessed August 10, 2024.

[28] TSMC used a 5 nm process to make Nvidia H100 chips; Wall Street analyst Robert Castellano calculated that TSMC earned $13,400 per 5nm wafer, and each wafer has 86 H100 chips. Each H100 chip earned 155 US dollars, and the superposition package generated 722.85 US dollars. So when TSMC is the Nvidia OEM H100, it can earn an average of nearly 900 US dollars per chip. https://www.moomoo.com/news/flash/16377502/analyst-on-average-each-chip-in-tsmc-s-oem-h100?level=1&data_ticket=1724692496338262

[29] Tae Kim, "Mark Zuckerberg Says Meta will Own Billions Worth of Nvidia H100 GPUs by Year End." *Barron's*, Jan 19, 2024, https://www.barrons.com/articles/meta-stock-price-nvidia-zuckerberg-b0632fed, accessed Aug 28, 2024; "DELL DGP4C NVIDIA H100 PCIe Tensor Core GPU 80GB Memory Interface 5120 Bit HBM2E Memory Bandwidth 2TB/s PCI-E 5.0 x16 128GB/s Graphics Processing Unit." *ServerSupply*, https://www.serversupply.com/GPU/EXPANSION%20MODULE/APPLICATION%20ACCELERATOR/NVIDIA/900-21010-0100-030_396177.htm?gad_source=1&gclid=Cj0KCQjwz7C2BhDkARIsAA_SZKa7p0F74dHLJr3X804mw9bsECSAiA2wXp9ehHVNwyfrq5SlNkIdMogaAmhDEALw_wcB, accessed August 10, 2024.

[30] Diana Goovaerts and Matt Hamblen, "Could GPU power levels break the data center ecosystem?" *Fierce Network*, April 30, 2024, https://www.fierce-network.com/cloud/could-gpu-power-levels-break-data-center-ecosystem, accessed August 19, 2024.

[31] Ibid.

[32] Jowi Morales, "A single modern AI GPU consumes up to 3.7 MWh of power per year – GPUs sold last year alone consumed more power than 1.3 million homes." *Tom's Hardware*, June 14, 2024, https://www.tomshardware.com/desktops/servers/a-single-modern-ai-gpu-consumes-up-to-37-mwh-of-power-per-year-gpus-sold-last-year-alone-consume-more-power-than-13-million-households, accessed August 19, 2024. https://iea.blob.core.windows.net/assets/6b2fd954-2017-408e-bf08-952fdd62118a/Electricity2024-Analysisandforecastto2026.pdf, accessed September 26, 2024.

[33] Jowi Morales, "A single modern AI GPU consumes up to 3.7 MWh of power per year – GPUs sold last year alone consumed more power than 1.3 million homes." *Tom's Hardware*, June 14, 2024, https://www.tomshardware.com/desktops/servers/a-single-modern-ai-gpu-consumes-up-to-37-mwh-of-power-per-year-gpus-sold-last-year-alone-consume-more-power-than-13-million-households, accessed August 19, 2024.

[34] https://www.washingtonpost.com/business/2024/09/20/microsoft-three-mile-island-nuclear-constellation/ accessed Sept 18, 2024.

[35] "Nvidia Part 2: Nvidia Part II: The Machine Learning Company (2006-2022)." *Acquired*, Season 10, Episode 6, https://www.acquired.fm/episodes/nvidia-the-machine-learning-company-2006-2022 , accessed June 12, 2024.

[36] Stephen Witt, "How Jensen Huang's Nvidia is Powering the A.I. Revolution." *The New Yorker*, November 27, 2023, https://www.newyorker.com/magazine/2023/12/04/how-jensen-huangs-nvidia-is-powering-the-ai-revolution, accessed July 26, 2024.

[37] Stephen Witt, "How Jensen Huang's Nvidia is Powering the A.I. Revolution." *The New Yorker*, November 27, 2023, https://www.newyorker.com/magazine/2023/12/04/how-jensen-huangs-nvidia-is-powering-the-ai-revolution, accessed July 26, 2024.

[38] Stephen O'Neal, "Crypto Winter Survivor: Inside Nvidia's Difficult Relationship With Mining." *Cointelegraph*, Feb 19, 2019, https://cointelegraph.com/news/crypto-winter-survivor-inside-nvidias-difficult-relationship-with-mining, accessed June 10, 2024.

[39] "Nvidia Part 2: Nvidia Part II: The Machine Learning Company (2006-2022)." *Acquired*, Season 10, Episode 6,

[40] Prashant Jha, "History of Crypto: Crypto winter and Ethereum landmarks." *Cointelegraph*, April 5, 2024, https://cointelegraph.com/news/history-crypto-ethereum-winter-landmarks, accessed July 10, 2024.

[41] "What is GPT?" *AWS*, https://aws.amazon.com/what-is/gpt/, accessed July 10, 2024.

[42] Nefi Alarcon, "OpenAI Presents GPT-3, a 175 Billion Parameters Language Model." *Nvidia*, July 7, 2020, https://developer.nvidia.com/blog/openai-presents-gpt-3-a-175-billion-parameters-language-model/, accessed July 21, 2024.

[43] Krystal Hu, "ChatGPT sets record for fastest-growing user base – analyst note." *Reuters*, February 2, 2023, https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/, accessed July 22, 2024.

[44] Tor Björn Minde, "Generative AI does not run on thin air!" *RISE*, August 10, 2023, https://www.ri.se/en/news/blog/generative-ai-does-not-run-on-thin-air#:~:text=The%20Cost%20of%20Training%20GPT%2D4&text=OpenAI%20has%20revealed%20that%20it,of%20energy%20usage%20during%20training, accessed August 29, 2024.

[45] Ben Gilbert and David Rosenthal, "Nvidia Part II: The Machine Learning Company (2006-2022)." *Acquired*, March 27, 2022, Season 10, Episode 6, https://www.acquired.fm/episodes/nvidia-the-machine-learning-company-2006-2022, accessed June 12, 2024.

[46] Hassan Mujtaba, "NVIDIA AI GPU Demand Blows Up, Chip prices Increase By 40% & Stock Shortages Expected Till December." *Wccftech*, May 22, 2023, https://wccftech.com/nvidia-ai-gpu-demand-blows-up-chip-prices-increase-40-percent-stock-shortages-till-december/, accessed June 20, 2024.

[47] "NVIDIA Announces Financial Results for Second Quarter Fiscal 2025." *Nvidia Newsroom*, August 28, 2024, https://nvidianews.nvidia.com/news/nvidia-announces-financial-results-for-second-quarter-fiscal-2025, accessed August 29, 2024.

[48] Thomas Alsop, "PC graphics processing unit (GPU) vendor shipment share worldwide from 2009 to 2023, by quarter." *Statista*, April 18, 2024, https://www.statista.com/statistics/754557/worldwide-gpu-shipments-market-share-by-vendor/#:~:text=As%20of%20the%20fourth%20quarter,market%20share%20of%2018%20percent, accessed July 29, 2024

[49] "Nvidia GPU Shipments by Customer." *Reddit*, https://www.reddit.com/r/singularity/comments/1890o9y/nvidia_gpu_shipments_by_customer/#lightbox; accessed July 29, 2024.

[50] "Nvidia Shipped 3.76 Million Data-center GPUs in 2023 According to Study." *HPCwire*, June 10, 2024, https://www.hpcwire.com/2024/06/10/nvidia-shipped-3-76-million-data-center-gpus-in-2023-according-to-study/, accessed June 12, 2024.

[51] https://fortune.com/2024/08/29/nvidia-jensen-huang-ai-customers/, accessed 9/10/2024.

[52] "Nvidia Shipped 3.76 Million Data-center GPUs in 2023 According to Study." *HPCwire*, June 10, 2024, https://www.hpcwire.com/2024/06/10/nvidia-shipped-3-76-million-data-center-gpus-in-2023-according-to-study/, accessed July 29, 2024.

[53] https://www.sequoiacap.com/article/ais-600b-question/, accessed September 18, 2024.

[54] Anissa Gardizy, "Nvidia Reports 122% Revenue Growth, Confirms Chip Delay." *The Information*, August 29, 2024, https://www.theinformation.com/briefings/nvidia-reports-122-revenue-growth-confirms-chip-delay?utm_campaign=%5BREBRAND%5D+%5BTI-AM%5D+Th&utm_content=1095&utm_medium=email&utm_source=cio&utm_term=124&rc=ewwd92, accessed August 29, 2024.

[55] Kim Lyons, "Nvidia is acquiring Arm for $40 billion." *The Verge*, Sep 13, 2020, https://www.theverge.com/2020/9/13/21435507/nvidia-acquiring-arm-40-billion-chips-ai-deal, accessed July 24, 2024.

[56] "What's Next After Nvidia Ends Quest To Acquire Arm From Softbank." *Forbes*, Feb 10, 2022, https://www.forbes.com/sites/tiriasresearch/2022/02/09/whats-next-after-nvidia-ends-quest-to-acquire-arm-from-softbank/, accessed July 24, 2024.

[57] "FTC Sues to Block $40 Billion Semiconductor Chip Merger." *Federal Trade Commission*, Dec 2, 2021, https://www.ftc.gov/news-events/news/press-releases/2021/12/ftc-sues-block-40-billion-semiconductor-chip-merger, accessed July 24, 2024.

[58] "NVIDIA Grace CPU". *Nvidia Cloud & Data Center*, https://www.nvidia.com/en-us/data-center/grace-cpu/, accessed July 24, 2024.

[59] Stephen Witt, "How Jensen Huang's Nvidia is Powering the A.I. Revolution." *The New Yorker*, November 27, 2023, https://www.newyorker.com/magazine/2023/12/04/how-jensen-huangs-nvidia-is-powering-the-ai-revolution, accessed July 26, 2024.

[60] Anissa Gardizy, "Nvidia's Cloud Spending Soars, as It Looks Beyond Chips." *The Information*, May 28, 2024, https://www.theinformation.com/articles/nvidias-cloud-spending-soars-as-it-looks-beyond-chips, accessed August 17, 2024.

[61] Anissa Gardizy, "Nvidia's Cloud Spending Soars, as It Looks Beyond Chips." *The Information*, May 28, 2024, https://www.theinformation.com/articles/nvidias-cloud-spending-soars-as-it-looks-beyond-chips, accessed August 17, 2024.

[62] "Generative AI to Become a $1.3 Trillion Market by 2032, Research Finds." *Bloomberg Law*, June 1, 2023, https://www.bloomberg.com/company/press/generative-ai-to-become-a-1-3-trillion-market-by-2032-research-finds/, accessed August 10, 2024.

[63] "Graphic Processing Unit (GPU) Market Will Grow at CAGR of 33.8% by 2032." *Precedence Research* , August 30, 2024, https://www.precedenceresearch.com/press-release/graphic-processing-unit-market#:~:text=The%20global%20graphic%20processing%20unit,33.8%25%20between%202023%20and%202032., accessed August 30, 2024.

[64] https://www.sequoiacap.com/article/ais-600b-question/, accessed September 24, 2024.

[65] Salvador Rodriguex, Sarah E. Needleman, and Sebastian Herrera, "Amazon Paid Almost $1 Billion for Twitch in 2014. It's Still Losing Money." *The Wall Street Journal* , July 29, 2024, *https*://www.wsj.com/tech/twitch-amazon-video-games-investment-9020db87?mod=tech_trendingnow_article_pos1, accessed July 29, 2024.

[66]https://fortune.com/2024/09/16/larry-ellison-elon-musk-begged-nvidias-jensen-huang-more-gpus-fancy-sushi-dinner/, accessed September 18, 2024.

[67] Kevin Maney, "Is Nvidia Remaking the Movie: 'Corning 2000'?" *Kevin Maney Substack* , March 21, 2024, *https*://kevinmaney.substack.com/p/is-nvidia-remaking-the-movie-corning, accessed July 29, 2024.

[68] "What could stop the Nvidia frenzy?" *The Economist*, Aug 26, 2024, https://www.economist.com/business/2024/08/26/what-could-stop-the-nvidia-frenzy, accessed August 29, 2024.

[69] "Meet Llama 3.1." https://llama.meta.com/, accessed September 10, 2024.

[70] Max A. Cherney, "Apple used Google's chips to train two AI models, research paper shows." *Reuters*, July 30, 2024, https://www.reuters.com/technology/apple-says-it-uses-no-nvidia-gpus-train-its-ai-models-2024-07-29/?mc_cid=b04a390009&mc_eid=001d905869, accessed August 19, 2024.

[71] Nvidia 10-K, FY 2024.

[72] Fanny Potkin, "Exclusive: Nvidia preparing version of new flagship AI chip for Chinese market." *Reuters*, July 22, 2024, https://www.reuters.com/technology/nvidia-preparing-version-new-flaghip-ai-chip-chinese-market-sources-say-2024-07-22/, accessed July 24, 2024.

[73] https://www.wsj.com/tech/the-underground-network-sneaking-nvidia-chips-into-china-f733aaa6, accessed July 29, 2027.

[74] Thomas Alsop, "Distribution of Intel and AMD x86 computer central processing units (CPUs) worldwide from 2012 to 2024, by quarter." *Statista*, July 19, 2024, https://www-statista-com.ezp-prod1.hul.harvard.edu/statistics/735904/worldwide-x86-intel-amd-market-share/, accessed August 28, 2024.

[75] Anton Shilov, "AMD records its highest server market share in decades — Intel fights back in client PCs". *Tom's Hardware*, Aug 12, 2024, https://www.tomshardware.com/pc-components/cpus/amd-records-its-highest-server-market-share-in-decades-but-intel-fights-back-in-client-pcs, accessed August 28, 2024.

[76] "AMD to Buy ATI for $5.4 Billion." *CBS News*, July 24, 2006, https://www.cbsnews.com/news/amd-to-buy-ati-for-54-billion/, accessed August 28, 2024.

[77] "ATI May See More High-End Market Share Gains." *Forbes*, Feb 1, 2006, https://www.forbes.com/2006/02/01/ati-nvidia-0201markets05.html, accessed Aug 28, 2004.

[78] "AMD Instinct™ MI200 Series Accelerators". *Boston Ltd*, March 22, 2022, https://www.boston.co.uk/blog/2022/03/22/amd-instinct-mi200-accelerators.aspx, accessed Aug 28, 2024.

[79] "NVIDIA H100 Tensor Core GPU Architecture." *Nvidia*, 2022, https://www.advancedclustering.com/wp-content/uploads/2022/03/gtc22-whitepaper-hopper.pdf, accessed Aug 28, 2024.

[80] "Stacking Up AMD MI200 versus Nvidia A100 Compute Engines." *The Next Platform*, December 6, 2021, https://www.nextplatform.com/2021/12/06/stacking-up-amd-mi200-versus-nvidia-a100-compute-engines/, accessed Aug 28, 2024; Tae Kim, "Mark Zuckerberg Says Meta will Own Billions Worth of Nvidia H100 GPUs by Year End." *Barron's*, Jan 19, 2024, https://www.barrons.com/articles/meta-stock-price-nvidia-zuckerberg-b0632fed, accessed Aug 28, 2024.

[81] Asa Fitch, "AMD Buys AI Equipment Maker for Nearly $5 Billion, Escalating Battle with Nvidia." *The Wall Street* Journal, August 19, 2024, https://www.wsj.com/tech/ai/amd-buys-ai-equipment-maker-for-nearly-5-billion-escalating-battle-with-nvidia-e1106142, accessed August 19, 2024.

[82] "An Interview with AMD CEO Lisa Su About Solving Hard Problems," June 6, 2024, Stratechery (stratechery.com/2024).

[83] Tae Kim, "AMD Stock Jumps After Earnings. The Chip Maker Raised Its AI Chip Guidance Again." *Barron's*, July 31, 2024, https://www.barrons.com/articles/amd-earnings-stock-price-ai-chips-f3ca0324, accessed Aug 28, 2024.

[84] Thomas Alsop, "Distribution of Intel and AMD x86 computer central processing units (CPUs) worldwide from 2012 to 2024, by quarter." *Statista*, July 19, 2024, https://www-statista-com.ezp-prod1.hul.harvard.edu/statistics/735904/worldwide-x86-intel-amd-market-share/, accessed August 28, 2024.

[85] https://spectrum.ieee.org/nvidia-ai, accessed September 24, 2024.

[86] "Introduction to Cloud TPU." *Google Cloud*, https://cloud.google.com/tpu/docs/intro-to-tpu, accessed August 17, 2024.

[87] Kurtis Pykes, "Understanding TPUs and GPUs in AI: A Comprehensive Guide." *DataCamp*, May 2024, https://www.datacamp.com/blog/tpu-vs-gpu-ai, accessed August 17, 2024.

[88] Kurtis Pykes, "Understanding TPUs and GPUs in AI: A Comprehensive Guide." *DataCamp*, May 2024, https://www.datacamp.com/blog/tpu-vs-gpu-ai, accessed August 17, 2024.

[89] Cade Metz, Karen Weise, and Mike Isaac, "Nvidia's Big Tech Rivals Put Their Own A.I. Chips on the Table." *New York Times*, Jan 29, 2024, https://www.nytimes.com/2024/01/29/technology/ai-chips-nvidia-amazon-google-microsoft-meta.html, accessed August 17, 2024.

[90] Eric J. Savitz, "The Cloud Giants Are Taking On Nvidia in AI Chips. Here's Why – And How." *Barron's*, April 19, 2024, https://www.barrons.com/articles/ai-stocks-chips-nvidia-amazon-microsoft-google, accessed August 17, 2024.

[91] Kyle Wiggers, "Amazon unveils new chips for training and running AI models." *TechCrunch*, November 28, 2023, https://techcrunch.com/2023/11/28/amazon-unveils-new-chips-for-training-and-running-ai-models/, accessed August 17, 2024.

[92] Jordan Novet, "Amazon announces new AI chip as it deepens Nvidia relationship." *CNBC*, November 28, 2023, https://www.cnbc.com/2023/11/28/amazon-reveals-trainium2-ai-chip-while-deepening-nvidia-relationship.html, accessed August 18, 2024.

[93] Jordan Novet, "Amazon announces new AI chip as it deepens Nvidia relationship." *CNBC*, November 28, 2023, https://www.cnbc.com/2023/11/28/amazon-reveals-trainium2-ai-chip-while-deepening-nvidia-relationship.html, accessed August 18, 2024.

[94] Cade Metz, Karen Weise, and Mike Isaac, "Nvidia's Big Tech Rivals Put Their Own A.I. Chips on the Table." *New York Times*, Jan 29, 2024, https://www.nytimes.com/2024/01/29/technology/ai-chips-nvidia-amazon-google-microsoft-meta.html, accessed August 17, 2024.

[95] Eric J. Savitz, "The Cloud Giants Are Taking On Nvidia in AI Chips. Here's Why – And How." *Barron's*, April 19, 2024, https://www.barrons.com/articles/ai-stocks-chips-nvidia-amazon-microsoft-google, accessed August 17, 2024.

[96] Max A. Cherney, "Amazon racing to develop AI chips cheaper, faster than Nvidia's executives say." *Reuters,* July 25, 2024, https://www.reuters.com/technology/artificial-intelligence/amazon-racing-develop-ai-chips-cheaper-faster-than-nvidias-executives-say-2024-07-25/, accessed August 17, 2024.

[97] "AWS Unveils Next Generation AWS-Designed Chips." *Amazon*, November 28, 2023, https://press.aboutamazon.com/2023/11/aws-unveils-next-generation-aws-designed-chips, accessed August 17, 2024.

[98] Christiaan Hetzner, "Tech stocks take a pounding as hedge fund Elliott warns AI trades like Nvidia are in 'bubble land'". *Fortune*, August 2, 2024, https://fortune.com/2024/08/02/ai-nvidia-intel-arm-stock-bubble-elliott/, accessed August 17, 2024.

[99] Kif Leswing, "Nvidia dominates the AI chip market, but there's more competition than ever." *CNBC*, June 2, 2024, https://www.cnbc.com/2024/06/02/nvidia-dominates-the-ai-chip-market-but-theres-rising-competition-.html#:~:text=Nvidia%20has%20generated%20about%20%2480,better%20chip%20for%20particular%20tasks., accessed June 20, 2024.

[100] "Introducing Cerebras Inference: AI at Instant Speed." *Cerebras*, August 27, 2024, https://cerebras.ai/blog/introducing-cerebras-inference-ai-at-instant-speed, accessed September 10, 2024.

[101] "Can Cerebras Take on Nvidia?" *Brownridge Research*, May 31, 2024, https://www.brownridge.com/can-cerebras-take-on-nvidia/, accessed Sept 24, 2024.

[102] Kif Leswing, "Nvidia dominates the AI chip market, but there's more competition than ever." *CNBC*, June 2, 2024, https://www.cnbc.com/2024/06/02/nvidia-dominates-the-ai-chip-market-but-theres-rising-competition-.html#:~:text=Nvidia%20has%20generated%20about%20%2480,better%20chip%20for%20particular%20tasks., accessed June 20, 2024.

[103] Natasha Mascarenhas and Stephanie Palazzolo, "Groq, an Nvidia Rival, Nears $2.2 Billion-Valuation BlackRock Deal; Musk's Mystery Company Revealed." *The Information*, July 10, 2024, https://www.theinformation.com/articles/groq-an-nvidia-rival-nears-2-2-billion-valuation-blackrock-deal-musks-mystery-company-revealed, accessed July 18, 2024.

[104] Natasha Mascarenhas and Stephanie Palazzolo, "Groq, an Nvidia Rival, Nears $2.2 Billion-Valuation BlackRock Deal; Musk's Mystery Company Revealed." *The Information*, July 10, 2024, https://www.theinformation.com/articles/groq-an-nvidia-rival-nears-2-2-billion-valuation-blackrock-deal-musks-mystery-company-revealed, accessed July 18, 2024.

[105] Natasha Mascarenhas and Stephanie Palazzolo, "Groq, an Nvidia Rival, Nears $2.2 Billion-Valuation BlackRock Deal; Musk's Mystery Company Revealed." *The Information*, July 10, 2024, https://www.theinformation.com/articles/groq-an-nvidia-rival-nears-2-2-billion-valuation-blackrock-deal-musks-mystery-company-revealed, accessed July 18, 2024.

[106] Stephanie Palazzolo and Amir Efrati, "Nvidia's Aggressive Sales Tactics will Backfire, says Rival." *The Information*, Aug 21, 2024, https://www.theinformation.com/articles/nvidias-aggressive-sales-tactics-will-backfire-says-rival, accessed Aug 28, 2024.

[107] https://www.theinformation.com/articles/why-openai-could-lose-5-billion-this-year?utm_campaign=%5BREBRAND%5D+%5BTI-AM%5D+Th&utm_content=1095&utm_medium=email&utm_source=cio&utm_term=124&rc=ewwd92, accessed August 29, 2024.

[108] Cade Metz, Karen Weise, and Mike Isaac, "Nvidia's Big Tech Rivals Put Their Own A.I. Chips on the Table." *New York Times*, Jan 29, 2024, https://www.nytimes.com/2024/01/29/technology/ai-chips-nvidia-amazon-google-microsoft-meta.html, accessed August 17, 2024.

[109] "Introducing Triton: Open-source GPU programming for neural networks." *OpenAI*, July 28, 2021, https://openai.com/index/triton/, accessed August 20, 2024.

[110] Tim Bradshaw, "Nvidia's rivals take aim at its software dominance." *Financial Times*, May 21, 2024, https://www.ft.com/content/, accessed July 13, 2024.

[111]https://www.theinformation.com/articles/nvidias-aggressive-sales-tactics-will-backfire-says-rival?utm_campaign=article_email&utm_content=article-13473&utm_medium=email&utm_source=sg&rc=ewwd92, accessed August 26, 2024.

[112] Ibid.

[113] "AMD Releases New Version of ROCm, the Most Versatile Open Source Platform for GPU Computing." *AMD*, https://www.amd.com/en/newsroom/press-releases/2016-11-14-amd-releases-new-version-of-rocm-the-most-versati.html, accessed Sept. 17, 2024.

[114] Matthew S. Smith, "Challengers are Coming for Nvidia's Crown." *IEEE Spectrum*, Sept 16, 2024, https://spectrum.ieee.org/nvidia-ai, accessed Sept 16, 2024.

[115] Stephen Witt, "How Jensen Huang's Nvidia is Powering the A.I. Revolution." *The New Yorker*, November 27, 2023, https://www.newyorker.com/magazine/2023/12/04/how-jensen-huangs-nvidia-is-powering-the-ai-revolution#:~:text=For%20decades%2C%20the%20major%20manufacturer,up%20my%20chips%20and%20run.%E2%80%9D, accessed July 26, 2024, p. 19.

[116] Interview with an AI competitor, May 15, 2024.

[117] Kylie Robinson, "Customer demand for Nvidia chips is so far above supply that CEO Jensen Huang had to discuss how 'fairly' the company decides who can buy them." *Fortune*, Feb 21, 2024, https://fortune.com/2024/02/21/nvidia-earnings-ceo-jensen-huang-gpu-demand-supply-allocate-fairly/, accessed Aug 19, 2024.

[118] https://www.thestreet.com/technology/nvidias-jensen-huang-addressed-three-big-questions-about-ai-future, accessed September 16, 2024.

[119] Interview with former Intel executive, September 25, 2024.

[120] Interview with former Intel executive, September 25, 2024.