A Work Project, presented as part of the requirements for the Award of a Master's degree in Management from Nova School of Business and Economics.

NVIDIA's Bet on Artificial Intelligence

The Impact on Healthcare

Daniela da Silva Ferreira Fernandes (55425)

Work Project carried out under the supervision of:

Professor Luís Almeida Costa

09/01/2024

ABSTRACT

Artificial Intelligence is a transformative technology that has the potential to revolutionize various industries and drive unprecedented advancements. This case study aims to assess the pivotal role NVIDIA plays in the AI landscape, exploring in depth the strategic choices the company has undertaken, and how it will need to adjust to the rapidly evolving tech environment. The analysis will encompass NVIDIA's competitive position in its data center, gaming, professional visualization, and automotive markets, as well as the healthcare industry. Furthermore, a comprehensive analysis on NVIDIA's Corporate Social Responsibility initiatives will be conducted.

KEYWORDS:

NVIDIA, Strategy, Artificial Intelligence, AI Chips, GPUs.

This work used infrastructure and resources funded by Fundação para a Ciência e a Tecnologia (UID/ECO/00124/2013, UID/ECO/00124/2019 and Social Sciences DataLab, Project 22209), POR Lisboa (LISBOA-01-0145-FEDER-007722 and Social Sciences DataLab, Project 22209) and POR Norte (Social Sciences DataLab, Project 22209).

TABLE OF CONTENTS

Group part

CASE STUDY

The release of ChatGPT on November 30, 2022, marked the beginning of a new era of AI (Appendix 1). The human-like chatbot showcased the abilities of generative AI. ChatGPT reached one million customers in only five days, making companies need to rethink productivity, products, and business models.

ChatGPT is a Large Language Model (LLM) trained on 570 GB of text data, comprising 380 billion words. After the user inputs a text, or 'prompt,' the trained data is leveraged to repeatedly predict the next word to generate an answer. However, LLMs need a large amount of processing power to train and generate output. This processing mostly happens in data centers, so AI software like ChatGPT can be accessed over the cloud. One company especially profited from the rise of generative AI and the high demand for specialized hardware for data centers.

NVIDIA recognized the trend early and invested heavily in advanced chips and software that can perform the complex tasks required by AI. The technology company has built itself a reputation in the AI industry after leveraging powerful chips that origin in the gaming industry. While focusing on building a comprehensive ecosystem, the launch of ChatGPT transformed NVIDIA's business. Since then, NVIDIA Corporation's (NVDA) stock has risen by 155%, reaching a market capitalization of $1.1 trillion[1] (Appendix 49).

THE DATA CENTER AI CHIP MARKET

Status Quo & Progression of the AI Chip Market

From 2017 until 2022, the global adoption of AI among businesses has increased by 250%, with 50% of companies introducing AI in at least one business function (Appendix 2). The AI

[1] All monetary Appendices presented in this document are denominated in U.S. dollars (USD) unless otherwise specified.

ecosystem can be subdivided into two layers: software and hardware. AI software involves three main categories: applications, technologies and models, and programming languages and platforms. AI software runs on specialized high-performance hardware chips built from semiconductor material (Appendix 3). The worldwide market size for semiconductors reached $574.1 billion in sales in 2022, owing to its indispensable role in electronic devices (Appendix 4). Within the semiconductor market, the global integrated circuits (IC) market stood at a total of $474.4 billion in 2022 and is forecasted to grow 14% in the next year (Awati 2021) (Appendix 5). ICs are sets of electronic circuits on a thin substrate of semiconductor material. Underlying the IC market is the AI chip market, valued at $28.0 billion globally in 2023 and forecasted to reach $165.0 billion by 2030 (Appendix 6).

In 2022, a country-specific analysis determined the regions with the highest share in the AI chip market. In the North American region, the United States has dominated the AI chip market (Precedence Research 2023). The Asian-Pacific region is forecasted to be the fastest-growing market, with China leading in projected growth rates through 2032. This is mainly due to a rise in investments in AI security solutions. The significant growth potential in the Asian-Pacific region is primarily attributable to high chip design advancements, which are projected to gain market share from the North American region in the upcoming years. Africa is the dominant player in the LAMEA region due to investments flowing foremost towards research and development (R&D) (Appendix 7).

The Multifaceted World of AI Chips

The AI chip market can be categorized into various segments, including chip type, industry vertical, application, and processing type (Appendix 9). There are four main AI chip types: Central Processing Units (CPUs), Graphics Processing Units (GPUs), Field-Programmable Gate Arrays (FPGAs), and Application-Specific Integrated Circuits (ASICs). The trade-off between flexibility and efficiency is a key consideration when it comes to AI chips. Currently,

a chip that can perform various tasks is not as efficient, and vice versa. AI is based on large sets of data. Therefore, it is important for AI chips to process data efficiently. CPUs are general-purpose chips and are not optimized for parallel processing. This means they cannot efficiently process multiple data streams simultaneously, leading to limited performance for AI tasks. GPUs were originally designed for rendering graphics in video games and other graphics-intensive applications. They excel at parallel computation and are capable of a wide range of tasks beyond graphics, including scientific simulations and AI applications. Developers can write their own software for CPUs and GPUs. FPGAs offer high customization for AI applications but may involve complex development and relatively high costs. Users can reprogram FPGAs to tailor them to their applications. ASICs are specialized chips designed for a specific AI task with a fixed architecture and function. They offer high efficiency and energy savings at the cost of flexibility (Appendix 3). In 2022, in terms of revenue, the GPU segment dominated the AI chip market. However, concerning CAGR, the ASIC segment was forecasted to obtain the highest growth rate in the forecast period between 2023 and 2032 (Appendix 7).

The AI chip market can be divided into various categories regarding the industry vertical. Segments in this category include media and advertising, financial services, information technology and telecom, retail, healthcare, automotive and transportation, and others. In 2022, the healthcare segment dominated the market and was projected to acquire the highest market share until 2032 (Appendix 8). This increase is mainly attributable to a rise in demand for investigating and managing patient data (Precedence Research 2023).

By application, the AI chip market can be divided into the following segments: language models, robotic, computer vision, network security, and others. The application segment is currently and expected to be dominated by the robotic segment, which offers wide applications for AI. AI-based industrial robots allow for improved security, precision, and overall efficiency concerning manufacturing (Appendix 7).

Lastly, the AI chip market can be characterized by processing type: edge and cloud. Typical edge devices for AI chips include smartphones, security cameras, autonomous vehicles, or on-premise data centers (Appendix 10). Data centers are where companies store and process vast amounts of data, which are integral for AI operations as AI-based applications need to analyze and learn from big datasets. While edge AI chips come with relatively high costs and power constraints, they allow for many use cases. In 2022, the edge segment generated over 75% of revenue share and was forecasted to retain the highest market share through 2029 (Precedence Research 2023). Meanwhile, the cloud segment is expected to observe notable growth in the upcoming years caused by the rapid adoption of cloud services within AI. Using AI chips in the cloud instead of in edge data centers presents numerous benefits, including improved flexibility and scalability, as cloud resources can easily be adjusted to match AI computing needs (Appendix 11).

Dynamics in the AI Chip Market

Driving Trends in the Market Landscape

The accelerated digitalization of the global economy, mainly attributable to mandatory remote work and online education caused by the COVID-19 pandemic, has rapidly increased society's demand for advanced technologies and automation (Appendix 12). The digitized economy is forecasted to grow 40% annually, reaching 163 trillion gigabytes by 2025. This trend will be accompanied by a drastic increase in big data availability, fostering potential for AI systems (Appendix 13).

AI-based applications require enormous computing capabilities to process data. Since 1959, the computing power of chips doubled every two years, in what became known as "Moore's Law". However, computational power has doubled every six months during the past ten years, remarkably exceeding Moore's Law (Appendix 14). These advancements in performance can

mainly be attributed to new and improved algorithmic techniques and the availability of large datasets.

Due to the rise of importance of AI, LLM-based chatbots have been rapidly and widely adopted in numerous industries, such as e-commerce, retail, or healthcare. The wide applications for AI chatbots enabled companies to improve corporate processes and systems. For example, chatbots allowed organizations like Amazon to meet marketing objectives, improve their sales process, and establish further customer engagement and brand loyalty (Amazon Web Services n.d.). In 2023, AI systems were estimated to generate more than $8.0 billion per year in savings and contribute to $112.0 billion in retail sales (Maximize Market Research 2023).

Organizations are focusing more on sustainability as concerns regarding carbon footprint are becoming progressively predominant in society. Aligning with ESG goals is crucial for maintaining customer engagement. An increasing number of organizations are using AI to address sustainability efforts by improving efficiency in business processes and daily operations (Appendix 15).

Constraints & Restrictions in the Market Landscape

Conversely, the surge in demand for AI chips is associated with various challenges that restrict market growth. The absence of expertise and skilled workforce is a crucial restriction in the growth of the AI chip market. This scarcity is due to the novelty and complexity of AI chips, with various high-growth technology industries competing for these professionals (Appendix 16).

The world's most important advanced technology, including AI chips, is predominantly manufactured in a single facility in Taiwan belonging to the Taiwan Semiconductor Manufacturing Company (TSMC) (Appendix 17). The high concentration of manufacturing and the sourcing of raw materials from regions with political instability, including China,

Taiwan, and Ukraine, amplifies the ongoing chip shortage and increases wait times (Appendix 21). Moreover, ongoing conflicts are impacting the global supply chain, such as tensions between China and Taiwan and the Russia-Ukraine war. The commercial war between the United States and China is escalating, with the United States imposing export controls that restrict transactions with China.

Additionally, as concerns over the sustainability of the industry are rising, the AI chip market is expected to decarbonize its value chain. Intel's 700-acre manufacturing plant in the United States produced 15,000 tons of waste in the first quarter of 2021, with 60% being hazardous (Belton 2021). Regarding the production of electronic devices, the manufacturing of AI chips is considered responsible for the highest carbon output. As demand for AI chips continues to rise, production is increased, which conflicts with the environmental objectives and international climate goals to reach net zero carbon emissions by 2050 (Ng 2023).

Growing computational requirements for data centers increases the pressure on existing infrastructure. Larger and, therefore, more complex data sets are leading to significantly higher power consumption. The training of a single AI model emits as much energy as five times the lifetime emissions of a typical car, including manufacturing and fuel (Appendix 18). While GPUs are powerful enough for current AI applications, the bottleneck lies in the infrastructure. Data centers are limited in the number of chips, like GPUs, they can connect as they would require more energy than the grid could supply (Barber 2023).

The adoption of AI in our daily lives has created an extensive fear of redundancy and inequality. Recent advancements have affected numerous industries, automating processes and systems. The fear of job losses has increased in the past and roughly 41% of surveyed CEOs agreed to the statement that AI will displace more jobs than it will create in the long run (Appendix 19).

Group part

Due to the widespread adoption of AI and the consequential social challenges arising legal and regulatory norms have been developing rapidly, especially concerning data protection and intellectual property (IP) (Kemp 2021). With the increased applicability of technology, organizations face ethical concerns related to fairness, transparency, security, and safety of their customers. In 2022, 49% of companies in Germany stated that one of the biggest obstacles to adopting AI in their operations is uncertainty due to legal hurdles. Moreover, 33% mentioned a lack of trust in AI as another obstacle to employing AI (Appendix 20). The European AI Act was introduced in 2021, in which distinctive regulations apply for different levels of risk, providing the first comprehensive AI law (European Parliament 2023).

In the coming years, the focus will lie on enhancing productivity across various sectors of the economy, with AI playing a central role in achieving this goal. In a recent study, AI integration in conversational assistance has led to an increase of 14% in issues solved per hour (McKendrick 2023). In this context, security and trust in AI chips and data center reliability will become increasingly important, driven by the need for greater automation (Hilson 2023). Hardware design is expected to lay the foundations for upcoming AI developments, exhibiting companies' importance in protecting their intangible assets. The software industry fundamentally focuses on selling IP rights rather than protecting tangible assets (Moriggi 2018). In addition to having a strong patent portfolio, it is extremely important to update patents according to new innovations and not rely on existing patents in the long run, especially in such a fast-paced market like AI.

COMPETITION AND SUBSTITUTES

Legacy Chip Companies

Despite its constraints, the industry is evolving and growing rapidly, in which numerous companies have gained a foothold. Legacy chip companies NVIDIA, Intel, AMD, IBM, and

Huawei are actively working on new generations of AI chips. At the same time start-ups gained traction, not only focusing on semiconductors but also on quantum computing. Cloud providers started or have been active in developing chips to support their service offerings. However, especially NVIDIA, Intel, and AMD have already shown significant traction, generating billions in revenue.

Regarding data center AI chips, Intel was leading in terms of revenue in Q2 2021. The data center branch accounted for 31% of Intel's $63.1 billion revenue in the fiscal year 2022 (Intel 2021). In this quarter, the company sold $5.7 billion worth of chips. However, sold chips consist mainly of CPUs that are not recommended for LLM workloads. Sales decreased by 18% from 2021 to 2022, followed by a 15% drop from 2022 to 2023, resulting in data center AI chip sales of $4.0 billion in Q2 2023 (Appendix 23). Intel planned to bring an AI-focused GPU called "Ponte Vecchio" to the market. It suffered years of delays and was eventually removed from the website's product line (Appendix 1). Intel announced a strategy shift during a supercomputer conference in Germany in May 2022 and is planning to introduce the "Falcon Shores" chips in 2025 to compete against AMD and NVIDIA (Appendix 1). As an established company in the semiconductor market, it has the resources to establish a strong ecosystem for hardware and software, including the Intel® AI Analytics Toolkit. Intel builds its chips in its own factories (Intel n.d.). In 2019, Intel made an impactful move by purchasing Habana Labs, an Israeli AI chip startup focused on ASICs, for $2.0 billion (Appendix 1).

AMD publicly showed its first interest in AI in 2016 when launching purpose-built chips and applications. Further chips focused on data centers, namely “MI100” and “MI250”, were launched in 2020 and 2021. The “MI250” reached 80% of the processing power of NVIDIA’s then market-leading chip. AMD has worked extensively on its platform "ROCm" to match the capabilities of NVIDIA's platform. Using third party software, developers showedhat the same LLMs can be run on both platforms without code changes (Intel n.d.). In Q2 2021, data center

chip sales accounted for $0.8 billion. The revenues grew 83% from 2021 to 2022 and fell 11% the following year. In Q2 2023, AI data center revenues reached $1.3 billion (Appendix 23). The market is anticipating the announced aggressive launch of AMD's new "MI300" data center AI GPU in the fourth quarter of 2023. AMD expects it to fill some of the supply-demand gap effectively and to exceed the $6.0 billion data center revenue of 2022 in the fiscal year of 2023 (Intel n.d.). Citi analysts noted: *"The scenario is reminiscent of Intel (NASDAQ:INTC) vs AMD ten years ago when Intel had better performance and ecosystem and over 90% share"* (Shane 2023). For production, AMD heavily relies on TSMC. In 2022, AMD acquired industry-leading FPGA designer Xilinx for $50.0 billion, expanding its expertise in AI (Appendix 1).

IBM invested $53.0 million into researching alternative approaches to traditional digital hardware in 2008 via their Defense Department's research section. In 2014, they launched their brain-like “TrueNorth” AI chip. Moreover, IBM showcased a prototype analog AI chip in September 2023, which demonstrated an estimated 14 times higher energy efficiency, particularly in natural-language AI tasks, when compared to digital counterparts (Appendix 1). The AI and data platform for enterprises, namely “IBM Watsonx”, is expected to leverage the in-house developed chips.

Huawei unveiled its first AI chip, "Ascend 910", in October 2019 (Appendix 1). Further down the road, they partnered with the well-known Chinese AI company HKUST Xunfei to produce their own GPUs. The founder of Xunfei, Liu Qingfeng, announced their plans to compete with NVIDIA during the Chinese Entrepreneurs Forum 2023. NVIDIA recognized them as aotentially strong competitor, especially in the Chinese market (Neuro 2023). In September 2023, Huawei announced its companywide strategy shift towards AI (Appendix 1).

Group part

AI Chip Start-Ups

The AI chip market is undergoing a significant transformation with the emergence of start-ups that have the potential to disrupt the industry. The global average total deal value of venture capital (VC) investments in semiconductors averaged $1.3 billion in the span of 2017 to 2020. Since then, VC investment has seen exponential growth, reaching a peak of $7.8 billion in 2021, according to Deloitte (Appendix 24). However, startup deals and investments are continuing to fall throughout 2022 and 2023, as United States market data shows. The number of deals dropped from 23 to 4 in the first three quarters of each year, with total investments of $1.8 billion and $881.4 million, respectively (Reichow 2023). Companies like SambaNova, Graphcore, and Tenstorrent have raised over $3.0 billion from VC in the past decade (AI Startups 2023).

SambaNova Systems, founded in 2017, introduced the “SN40L” chip optimized for intensive AI workloads, which they believe will address the growing model sizes more effectively compared to legacy architectures. Having secured over $1.3 billion in funding, they also lease their platform to businesses, promoting a user-friendly AI-as-a-service model (Appendix 1).

Cerebras Systems, founded in 2015, launched its own chip in April 2021 and secured funding of $750.0 million at a $4.0 billion post-money valuation. The "Wafer Scale Engine 2" aims to condense the work of hundreds of GPUs into a single chip. With its efforts, Celebras has attracted pharmaceutical giants like AstraZeneca and GlaxoSmithKline, potentially accelerating drug research (Appendix 1).

Meanwhile, the British contender Graphcore, established in 2016, unveiled the “IPU-POD256” as their flagship AI chip. With nearly $700.0 million in funding, it formed partnerships with major storage firms and prestigious global research institutions.

Group part

Tenstorrent is developing a technology that emphasizes high throughput and low power consumption, fundamental requirements for generative AI models. It not only plans to produce chips in collaboration with Samsung but also provide IP for data centers. The Canadian startup raised $234.5 million with a $1.0 billion valuation (Appendix 1).

Cloud Providers

Cloud providers are a key customer base for AI chips (Appendix 22). Google and Amazon started to develop their own chips early, while others like Microsoft and Alibaba started after the high demand and respective chip shortages became apparent.

Google has been designing and deploying ASICs since 2016. While Google's chips power applications such as Translate, Photos, Search, and more, they are exceptionally well suited for AI, also powering Google's Bard (Leswing 2023). While Google has a competitive advantage due to its mature software and hardware, rather than selling ASICs independently, it makes this technology available externally via Google Cloud. Further, Google refuses to document hardware externally and discloses chips well after they are deployed.

Amazon Web Services (AWS) started to build its own specialized hardware in 2013. In 2015, Amazon moved into the AI chip market with its acquisition of an Israeli startup called Annapurna Labs for an estimated $350.0-400.0 million. AWS rolled out their custom chips Inferentia in 2019 for inference workloads, which predict new output, and Trainium for training workloads, which refers to understanding patterns in data, in 2020. With these chips, AWS aims to enable its customers to leverage AI fully. Still, AWS invested $100.0 million in a generative AI innovation center powered by NVIDIA chips. Amazon announced their own LLM in April 2023 called "Titan" together with the service "Bedrock" to support developers leveraging generative AI (Appendix 1).

Microsoft primarily relies on NVIDIA GPUs to power LLMs used within its cloud service Azure and for Copilot. However, recent articles report that it has been secretly developing its own chips since 2019. Also, the partly Microsoft-owned company OpenAI is exploring building its own chips, according to internal discussions in October 2023. Microsoft's own AI chip, "Maia 200", a GPU, was launched in late November 2023. At the same time, it was announced that the latest AI chips from NVIDIA and AMD will be accessible through Azure. Azure offers an end-to-end AI architecture containing a set of tools that support users in developing, deploying, and managing AI applications (Appendix 1).

Alibaba, a popular cloud computing provider in China, has announced its own AI chip, the "Hanguang 800", to be leveraged within its services with no intention to sell it as a standalone product. The cloud computing division shows the highest growth within Alibaba. Two of China's largest tech companies, Huawei and Alibaba, have joined the semiconductor industry, showcasing how crucial semiconductors are for China's "Made in China 2025" plan (Appendix 1).

From Traditional to Quantum Computing

Quantum computing is a new way of computing and requires innovative specialized hardware fundamentally different from traditional chips, such as GPUs. Quantum computers are able to process large amounts of data at once and solve problems much faster than regular computers. A high level of parallel processing power is fundamental to AI. The technology and materials required for quantum chips are more distinct, complex,nd still in the developmental phase (McKinsey & Company 2023). While significant developments have been in recent years, including Google showcasing a quantum computer that solves a specific task in seconds that regular supercomputers could solve in thousands of years, the technology is still not mature (Appendix 1). Further hardware is only of value if software and applications are available for

it. Researchers were able to run quantum computing software on traditional GPUs to support the development of hardware applications that is currently unavailable (Rahul 2023).

In 2022, startup investors allocated a record amount of $2.4 billion to quantum technology start-ups. While investments grew only by 1% from 2021 to 2022, 68% of quantum computing investments from 2001 to 2022 were made in the last two years. Major technology companies like Google, Microsoft, NVIDIA, IBM, and Intel are also increasingly investing in quantum computing (Appendix 25).

THE AI CHIP VALUE CHAIN

All IC chips are made from semiconductor materials and are categorized in three primary categories: Logic, Memory and Digital, Analog and Other (DAO). AI chips are recognized as advanced logic chips. All IC chips undergo the same value chain stages, which will be analyzed below (Appendix 26).

The Semiconductor Value Chain

The need for technical expertise and scalability has led to a highly specialized global supply chain characterized by both high R&D and capital expenditure (CAPEX) investments (Appendix 27). The value chain consists of four stages: pre-competitive research, chip design, wafer fabrication and assembly, testing, and packaging (ATP) (Appendix 28). A specialized network of software design tools and core IP suppliers, materials, and equipment supports these stages. Each activity requires varying R&D and CAPEX, contributing to different proportions of the total value.

Countries perform different roles according to their comparative advantages, resulting in six major participants: the United States, South Korea, Taiwan, China, Japan, and Europe, ordered by added value (Appendix 29).

Group part

In the semiconductor industry, it is possible to distinguish between four types of companies, depending on their level of integration and business model: Integrated Device Manufacturers (IDMs), Fabless Design firms, Foundries, and Outsourced Assembly and Test companies (OSATs). IDMs, like Intel and Samsung Electronics, vertically integrate across multiple parts of the value chain, performing design, manufacturing, assembly, testing, and packaging activities in-house. Fabless Design firms, such as NVIDIA and AMD, focus on design while outsourcing wafer fabrication and ATP. Foundries are responsible for the wafer fabrication of Fabless firms and IDMs, as most IDMs lack sufficient in-house manufacturing capacity. OSATs offer contract services for ATP to both IDMs and Fabless companies (Appendix 30 and Appendix 31).

Pre-Competitive Research

Pre-competitive research involves fundamental scientific and engineering investigations conducted globally by scientists from various sectors, including industry, universities, government-sponsored national labs, and research institutes. This type of research is distinct from proprietary R&D, as its outcomes are frequently disseminated through publications and shared broadly within the scientific community.

Chip Design

Chip Design is the outcome of R&D efforts, IP, and the expertise of a highly trained workforce, encompassing both hardware design and software development work. Various companies engage in chip design, primarily falling into four categories: Fabless companies, which are the most common business model for the design of advanced logic chips (Appendix 29); IDM; Original Equipment Manufacturers (OEMs) and Electronic Design Automation (EDA)/IP Providers. IDMs are the most common business model, followed by Fabless companies and OEMs (Appendix 32). OEMs use semiconductors as input for other products and some have

begun to design their own chips. Consequently, they have a growing presence in chip design and increasingly participate in the same product and talent markets that fabless companies and IDMs tap into for their needs. Finally, EDA companies are trusted intermediaries between design companies and foundries. They provide software tools for chip design and ensure seamless compatibility with foundries' manufacturing processes. This phase accounts for 56% of total added value and is the most R&D-intensive phase, contributing 53% of overall R&D spending (Appendix 26). The United States holds a dominant position as the global leader in chip design. In 2020, 46% of total design companies operated within its borders (Appendix 33). In fact, seven out of the top ten design companies, ranked by annual revenue, are based in the United States (Appendix 34). However, a decline is evident due to the growing influence of China and South Korea, which are expected to reduce the United States' market share to 36% by 2030.

Equipment and Materials

Equipment contributes 12% of the value added within the value chain (Appendix 26). In terms of geographic distribution, the United States holds the majority at 42%, followed by Japan at 27%, and Europe at 21% (Appendix 29). The manufacturing of semiconductors depends on over 50 types of advanced equipment, and lithography tools, essential for manufacturing advanced logic chips, make up a substantial portion of the CAPEX for fabrication entities (Appendix 35). There are three main equipment providers worldwide: Applied Materials, ASML, and Tokyo Electron ranked by revenue (Appendix 36). Currently, ASML is the exclusive global supplier of high-performance lithographic machines tailored for such chips (Tarasov 2022). Materials contribute to a total added value of 5% (Appendix 26). Notably, key material suppliers are highly concentrated in China and East Asia, constituting 73% of total suppliers (Appendix 29). The process involves up to 300 different inputs, with silicon crucial in wafer fabrication and organic substrates essential for ATP (Appendix 37). Silicon and noble

gases are the primary materials used in wafer fabrication. China is the largest global silicon supplier, contributing over 68% of the world’s production, while Ukraine holds the distinction of being the primary supplier of noble gases, providing approximately 70% and 40% of the global supply of neon and krypton, respectively (Statista 2023). However, in the aftermath of Russia's invasion of Ukraine, China has emerged as a substitute supplier of noble gases.

Manufacturing: Wafer Fabrication and ATP

Wafer fabrication contributes to a total added value of 19% and is the most CAPEX-intensive phase, constituting 64% of overall CAPEX spending (Appendix 26). Approximately 75% of global semiconductor manufacturing capacity is concentrated in China and East Asia (Appendix 29). Such an outcome is attributed to robust government incentives coupled with access to a skilled workforce and robust infrastructure. However, it's noteworthy that the exclusive production capacity for advanced logic chips is currently situated in Taiwan and South Korea (Appendix 38). Specifically, the exclusive capability to manufacture advanced logic chips is held by just three companies: TSMC, Samsung, and Intel. The industry's focus on these entities is anticipated to strengthen in the future, fueled by the growing demand for these types of chips. Moreover, TSMC is the global foundry leader, contributing approximately 58% of global foundry revenue, followed by Samsung and Global Foundries, with estimated shares of around 11% and 7%, respectively (Appendix 17). ATP contributes to a total added value of 6% (Appendix 26). Furthermore, mainland China and Taiwan collectively represent 57% of the global capacity for ATP (Appendix 29). The concentration of manufacturing in China and East Asia is attributed to lower labor costs in the region, coupled with high technical expertise.

Group part

Global Semiconductor Shortage

Since late 2020, a global semiconductor shortage has led to product cancellations, extended lead times, and higher prices due to demand surpassing production capacity, rooted in manufacturing constraints and a skilled worker shortage (Wassen Mohammad 2022). The COVID-19 pandemic worsened supply conditions, creating a widespread imbalance with anticipated demand spikes, especially for healthcare and remote work products, along with fluctuations in chip demand for sectors like automotive and consumer electronics. Simultaneously, global lockdowns forced chip manufacturing plants to close. Geopolitical tensions, such as Japan's 2019 export restrictions and the United States-China trade war, are marked by export controls limiting transactions with China.

In response, foundries utilization rates exceeded the 80% standard, maintaining high production levels until 2022 in an attempt to meet increased demand (Appendix 39). While the global chip shortage has eased, short-term declines in sales are anticipated to persist through 2023 due to macroeconomic challenges and market cyclicality.

Governments are addressing concerns about future shortages and geopolitical tensions related to manufacturing by regulating and incentivizing domestic semiconductor production. The United States, European Union, China, South Korea, Taiwan, Japan, and others have committed to boosting their manufacturing capacity through comprehensive subsidies and tax incentives. Both TSMC and Samsung have forged agreements with the United States government to establish foundry plants in the country. However, if regional supply chains aimed for total self-sufficiency, it would require around an additional $1.0 trillion in initial investments to match current levels of semiconductor consumption. This would also result in an overall semiconductor price increase ranging from 35% to 65% (Appendix 40).

Group part

Marketing and Sales

Companies' Ecosystem

AI chip companies provide not just cutting-edge software and hardware but also entry into comprehensive ecosystems. These ecosystems include software tools, developer support, and AI platforms, attracting developers and businesses. Recognizing the importance of software development in AI, many companies focus marketing efforts on developers, providing resources, documentation, and support to encourage the creation of AI applications on their platforms.

Strategic partnerships with key players in sectors like cloud service providers, automakers, and data center operators enhance the attractiveness and functionality of these ecosystems. NVIDIA stands out for having one of the most developed ecosystems.

Distribution Channels

In the semiconductor industry, distribution channels are diverse, comprising both direct and indirect sales. Indirect sales are made through independent distributors and each company's partner network, which distribute to OEMs, electronic manufacturing service providers, and companies across various industries. The distribution strategy employed by semiconductor companies varies. According to the Semiconductor Industry Association, sales generated through distributors range from 25% to 85% of total revenues. On average, distributors handle 50% of the semiconductor industry's revenues (Semiconductor Industry Association 2021).

Group part

NVIDIA CORPORATION

Founding and Initial Steps

In 1993, NVIDIA was founded by Jensen Huang, who at the time was working as a microprocessor designer at AMD, along with Chris Malachowsky and Curtis Prime, who were working as staff engineers and chief designers at a company called Sun Microsystems.

During its early-stages, NVIDIA's vision was to bring 3D graphics to the gaming and multimedia markets. Two years after its founding, NVIDIA introduced its first-ever product, the 'NV1', a multimedia accelerator capable of handling both regular 2D and 3D graphics. By 1999 the company had established its presence in the market with the launch of its first GPU, the GeForce 256. The GPU was a groundbreaking innovation that redefined modern computer graphics and revolutionized parallel computing (Michael Justin Allen Sexton 2018).

NVIDIA Today

For over three decades, scientists, researchers, developers, and creators have utilized NVIDIA's technologies to achieve remarkable accomplishments. Today, more than 40,000 companies use NVIDIA's AI technologies (NVIDIA 2023). As of now, the company has operations across four continents: Asia, Europe, North America, and South America, and more than 50 offices worldwide. With its headquarters in Santa Clara, NVIDIA presents itself as a global leader in high-end computing and AI, being one of the world's largest graphics processing and chip manufacturing companies. The company's success is built on innovation and high-quality products. NVIDIA boasts a global workforce of 26,196 employees across 35 countries, retaining them with a low turnover rate of 5%. Among these employees, 19,532 were involved in R&D, while 6,664 were engaged in sales, marketing, operations, and administrative roles. This distribution is reflected in their operating expenses: $7.3 billion (66%) is allocated to R&D

Group part

and $2.4 billion (22%) to administrative expenses, out of $11.1 billion in 2022[2] (NVIDIA 2023).

Additionally, employee referrals significantly influenced their recruitment strategy, accounting for over 37% of new hires in 2022 (Appendix 41). The Employee Stock Purchase Plan, which allows employees to purchase NVIDIA's stock at a discount, delivered more than $300 million in value to its employees (NVIDIA 2023). During 2022, Nvidia supported its workforce by giving constant feedback on their performance and offering a library of technical content accessible on-demand, mentorship schemes, career coaching services, and pulse surveys to capture employee feedback directly (NVIDIA 2023).

Notably, NVIDIA launched over 160 products this year, with expenditures of $11,6 million dedicated to the acquisition of semiconductors needed for these new offerings (NVIDIA 2023).

NVIDIA's financial results are structured into two main segments: Compute & Networking and Graphics[3]. In the Compute & Networking segment, NVIDIA reported revenues of $15.1 billion in 2022. Meanwhile, the Graphics segment generated $11.9 billion in revenues (NVIDIA 2023). Within these two segments, NVIDIA generates revenues from product sales, hardware and systems, licensing, and cloud service.

---

[2] Note: In this document, references to specific years about NVIDIA's financial data pertain to NVIDIA's Fiscal Year that spans from February of that year to January of the following year. For example, '2022' refers to Fiscal Year 2023, which covers the period from February 2022 to January 2023. This convention is applied consistently to all fiscal year references for NVIDIA throughout this case study.

[3] The Compute & Networking segment includes Data Center accelerated computing platform; networking; automotive
AI Cockpit, autonomous driving development agreements, and autonomous vehicle solutions; electric vehicle computing platforms; Jetson for robotics and other embedded platforms; NVIDIA AI Enterprise and other software; and cryptocurrency mining processors, or CMP.
The Graphics segment includes GeForce GPUs for gaming and PCs, the GeForce NOW game streaming service and
related infrastructure, and solutions for gaming platforms; Quadro/NVIDIA RTX GPUs for enterprise workstation graphics; virtual GPU, or vGPU, software for cloud-based visual and virtual computing; automotive platforms for infotainment systems; and Omniverse Enterprise software for building and operating metaverse and 3D internet applications.

Group part

Central to NVIDIA's production strategy is their fabless business model. Since 1998, it has outsourced its chip manufacturing to the foundry TSMC and later diversified its supply to Samsung Electronics. NVIDIA designs its chips and focuses on technological innovation while foundries manufacture the physical product. For the manufacturing equipment, TSMC relies on the company ASML, which provides advanced machinery, such as its $200 million extreme ultraviolet lithography machine, needed to make cutting-edge chips (Kharpal, Two of the world's most critical chip firms rally after Nvidia's 26% share price surge 2023). Recently, TSMC has invested in two new Arizona manufacturing plants.VIDIA has already marked its commitment to leverage the new production capacity, shifting part of its production to the United States (S. H. Lee 2022).

At its core, NVIDIA specializes in four large markets in which its computing platforms can provide great acceleration for applications: Gaming, Automotive, Professional Visualization, and Data Center. Starting in their early days with a focus on gaming, NVIDIA had become a leader in the market with its high-performance GPUs that have revolutionized gaming graphics and performance. In the automotive sector, NVIDIA is pivotal for self-driving technologies and advanced driver assistance systems using AI-based technologies. NVIDIA serves the professional visualization market with GPUs like the Quadro series, crucial for digital modeling and 3D simulations. In data centers, NVIDIA's GPUs play a key role in managing large-scale data workloads and complex computations, boosting the performance of cloud computing and AI applications. Currently, the company is also investing in hybrid systems that integrate both quantum computing and traditional computing, leveraging their GPU capabilities. Its CUDA Quantum platform provides a bridging technology that facilitates the integration of different AI chips (NVIDIA 2023). Recently the company also announced its DGX Quantum, the world's first GPU-accelerated quantum computing system, that caters to researchers and practitioners to facilitate the integration of quantum in their applications (NVIDIA 2023). Furthermore,

Group part

NVIDIA is engaging in Corporate Social Responsibility (CSR) activities to reduce its carbon footprint. The company's GPUs are already 20x more energy efficient than traditional CPUs for certain AI and HPC workloads (NVIDIA 2023).

At the end of 2022, NVIDIA registered $27.0 billion in revenues with a 57% gross margin and net income of $4.4 billion (Appendix 42). The Data Center market contributed $15.0 billion, accounting for 55% of total revenues; Automotive added $9.1 billion, making up 34%; Professional Visualization generated $1.5 billion, representing 6%; the remaining $455 million came from OEM & Other (Appendix 43 and Appendix 44). 31% of total revenues were generated in the United States with $8.3 billion. Taiwan followed with $6.9 billion, accounting for 26% of the total revenue. China (including Hong Kong) generated $5.8 billion, representing 21%. Revenue from other countries totaled $5.9 billion, making up the remaining 22% of the overall revenues (NVIDIA 2023).

To deliver its products and services to the end user, NVIDIA sells through various direct and indirect channels. This includes direct sales through its online store, dedicated sales teams for enterprise clients, and OEM partnerships for broader market reach. Additionally, NVIDIA utilizes a global network of retailers, subscription-based cloud services like DGX Cloud and participates in events and conferences while also engaging in government contracts and licensing its IP. In 2022, NVIDIA's main customers were enterprises and cloud computing providers leveraging data center GPUs for AI and deep learning, gamers using GeForce GPUs, and designers, academic institutions, OEMs, and the automotive industry. Individually, NVIDIA's customers do not contribute to more than 10% of the company's overall revenues (NVIDIA 2023). However, NVIDIA maintains relationships with a few large customers, mainly large technology companies, that span across its various sectors and are key to NVIDIA's business in its different target markets (NVIDIA 2023).

Group part

NVIDIA's Growth from Gaming to AI Giant

In 2006, NVIDIA released a software toolkit called CUDA that would eventually propel it to the center of the AI boom. It is a parallel computing platform and programming model that helps developers effectively utilize the parallel processing capabilities of NVIDIA's GPUs. CUDA provides the ability to divide complex tasks into thousands of smaller ones, which the GPU can then process in parallel, leading to much faster processing. The company designed the software to work only with NVIDIA's GPUs. Today, the platform boasts a community of four million developers and has been downloaded 40 million times (Shu and Liao 2023). After the launch of CUDA, NVIDIA's strategy was to target software developers, for which NVIDIA conducted university courses and training webinars, besides offering the platform for free. From 2006 to 2007, NVIDIA's advertising and promotion expenses increased by $4.2 million (NVIDIA 2006). This rise in expenditure was primarily to highlight the benefits of CUDA, emphasizing its ability to allow developers to use familiar programming languages for high-level computing tasks, thereby removing the need for specialized learning. Bryan Catanzaro, the vice president of applied deep learning research at NVIDIA and one of the only employees working on AI when he joined in 2008, said, "For ten years, Wall Street asked NVIDIA, 'Why are you making this investment? No one's using it.' And they valued it at $0 in our market cap." (Tarasov2023). A significant turning point was in 2012 when a team from the University of Toronto won the ImageNet computer vision contest with their deep learning model called AlexNet. It was powered by CUDA-enabled NVIDIA GPUs and achieved accuracy in recognizing images with an error rate of only 15% compared to 25% from the previous years (Tilley 2016). It validated NVIDIA's long-term investment in CUDA, demonstrating the immense potential of GPUs in accelerating deep learning and AI tasks.

From then on, NVIDIA's strategy shifted. The recognition of the immense potential of GPUs led NVIDIA to explore new markets where such capabilities could be transformative. It began

diversifying the application of its GPU beyond gaming, specifically targeting data center markets. To develop data center solutions, the company invested heavily in R&D by capitalizing on the solid profit and constant cash flow its gaming business generated. During the years 2012, 2013, and 2014, they incurred increasing R&D expenses of $1.2 billion, $1.3 billion, and $1.4 billion, respectively (Appendix 42). By then, 6,658 full-time employees were engaged only in R&D activities, accounting for 72 % of all employees (NVIDIA 2015).

From 2013 to 2015, the number of companies NVIDIA engaged with on deep learning grew nearly 35 times to over 3,400 partners from industries such as manufacturing, healthcare, energy, life sciences, automotive, financial services, and entertainment (NVIDIA 2016).

In 2016, NVIDIA introduced the first AI Data Center GPU, the ‘Tesla P100’, which was designed for deep learning and AI tasks. In the same year, NVIDIA took eight of these GPUs and put them into the DGX-1, a rectangular container NVIDIA called "the world's first AI supercomputer in a box.” (McHugh 2016). NVIDIA’s CEO Jensen Huang personally delivered the first unit to Elon Musk and his non-profit OpenAI. He gifted the system one week before Intel unveiled its first processor designed specifically for AI workloads at its big annual conference in San Francisco. NVIDIA's data center revenue experienced a substantial increase from $0.3 billion in 2015 to $1.9 billion in 2017, marked by a notable 133% growth from 2015 to 2017, contributing to an overall three-year CAGR of 85% (Appendix 44).

Since then, NVIDIA has developed more GPUs for data center applications by constantly redesigning and improving their chip architecture. Launching the ‘A100 Tensor Core GPU’ in 2020 represented a monumental leap, achieving an 11x increase in higher computational performance compared to the ‘P100 GPU’. Introducing the ‘H100 Tensor Core GPU’ in 2022, NVIDIA achieved up to nine times faster training times for complex AI models, showcasing H100’s superior processing capabilities compared to the previous model (Appendix 46). With

the announcement of the NVIDIA ‘H200 Tensor Core GPU’, available for sale in 2024, this path of innovation will continue. NVIDIA claims it will double the inference or AI prediction performance of the H100, particularly for LLMs (Appendix 47).

These GPUs are not intended for consumer-grade hardware but are built to handle the extensive computational demands of large-scale AI training and inference tasks in on-premise data centers and cloud-based data centers. NVIDIA’s GPUs began gaining prominence in cloud computing as major cloud service providers like AWS, Microsoft Azure, and Google Cloud Platform integrate them into their cloud infrastructures.

As the demand for AI applications increased, companies and government agencies worldwide needed to hire engineers, developers, researchers, and data scientists with AI expertise. However, the demand for AI-trained developers surpassed the number of available professionals in the field. To help bridge that gap, NVIDIA established the Deep Learning Institute to train developers with hands-on courses in fundamental and advanced AI topics in 2016. By 2020, the institute had trained over 250,000 people worldwide (Appendix 1).

To further enhance NVIDIA’s high computing power and AI capabilities, it acquired Mellanox Technologies in March 2019 for $6.9 billion (NVIDIA 2019). In a “Mad Money” interview, Jensen Huang highlighted the reasoning behind the acquisition “We’re combining the leaders of AI computing and high-speed networking and data processing into one company.” “With that, hopefully we could, …, accelerate the innovation and create amazing things for data centers going forward,” he said (Clifford 2020). From a strategic point of view, the acquisition also enabled NVIDIA to merge the two company's human resources (HR) talents. Following this acquisition, NVIDIA experienced significant financial growth. NVIDIA reported record revenues of $3.9 billion for the second quarter of 2020, a 50% increase from $2.6 billion a year earlier (Appendix 48). Mellanox contributed approximately 14% of the company's total revenue

Group part
(Infotechlead 2020). The data center revenue, which included Mellanox, saw a remarkable 167% increase to $1.75 billion in that quarter (Newman 2020).

The company has also launched a free program called NVIDIA Inception, designed to provide access to technology, software support, expertise, market exposure, and capital to early-stage companies that are employing AI. With 7,000 technology start-ups in 2020, the program now supports over 15,000 (NVIDIA 2023). It is one of the largest AI startup ecosystems in the world.

The release of ChatGPT from OpenAI in 2022, with a brain composed of more than 20,000 NVIDIA GPUs, caused many companies to look for ways to add similar generative AI capabilities to their software. Demand for NVIDIA's GPUs strengthened as a result (Appendix 1). In the same year, the data center revenues surpassed the gaming revenues for the first time. They rose 41% to a record of $15.0 billion, making up 55% of total revenues. Gaming experienced a decrease of 27% to 9.07 billion, contributing 34% to total revenues (Appendix 43).

NVIDIA's Position in AI, Deep Learning and Data Centers

NVIDIA has a market capitalization of $1.1 trillion as of November 2023, making it the sixth most valuable company in the world. In contrast, back in 2021, NVIDIA was ranked as the 24th most valuable company. NVIDIA competitor AMD is currently in position 53 with a market cap of $195.0 billion, followed by Intel in position 59 with a market cap of $183.9 billion (Appendix 49). By the end of February 2023, NVIDIA's share price had nearly jumped more than 60% since January (Appendix 1). NVIDIA also controls 87% of the market for GPUs, followed by AMD with 10% and Intel with 3% (Appendix 50). The company received prestigious awards at the NeuIP2022 conference for its groundbreaking work in generative AI models and achieved new records in the latest MLPerf benchmark in 2023, which is a set of standardized tests measuring the performance of AI hardware (Appendix 1). One of NVIDIA's

supercomputers was able to train a GPT-3 model three times faster than the previous record. NVIDIA’s data center revenues have quadrupled over the last two years. In the third quarter of 2023, they reported a record of $14.5 billion, up 41% from the previous quarter and 27% more than one year before (NVIDIA 2023).

Consequently, NVIDIA's data center processor revenue grew significantly, exceeding 30% in a decade (Value 2023). 50% of data center revenue came from cloud service providers, while the remaining 50% came from consumer internet entities and large companies (Novet 2023). NVIDIA powers over 70% of the supercomputers on the Global TOP500 list, including 23 of the top 30 systems on the Green500 list (Appendix 1).

The company has sold more than half a million of its H100 GPUs in just three months of 2023. Their AI technologies were mainly used for public cloud services such as ChatGPT or embedded in applications and used by clients who want to create their own LLM. The price of the GPU H100 ranges from $20,000 to $40,000. It approximately costs NVIDIA around $3,320 to manufacture one, resulting in a gross profit of up to 1000% (Schreiner 2023). Jensen Huang has publicly declared that falling GPU prices are “a story of the past” (White 2023). Financial consulting firm Raymond James has validated NVIDIA's high pricing strategy. The justification for these prices is rooted in the high cost of innovation and development, which increases as complexity is added to AI technologies. Glassdoor reported an average salary of around $202,000 annually for NVIDIA's electronics hardware engineers (Glassdor 2023). Developing advanced chips like the H100 involves thousands of hours of work by numerous specialized professionals, contributing to NVIDIA's substantial R&D expenses. In 2022 NVIDIA invested 27% of its revenue in R&D, compared to AMD's 21%, Intel's 17%, and IBM's 11% (Appendix 51). NVIDIA's market position is supported by its patent portfolio that protects NVIDIA’s AI innovations. The company has a total of 11,214 patents, out of which 6,973 have been granted, and more than 74% are active. Most of these patents are filed in the United States, China, and

Germany. Intel holds comparatively few patents, with 741, followed by AMD, with 185 patents. (Appendix 52).

Over the years, NVIDIA has developed a platform strategy that creates a thriving ecosystem. A key part of this strategy is the NVIDIA AI Platform, which is anchored in the CUDA software and relies on NVIDIA's GPUs. This strategy is not only about the software or the hardware, but it is also about the community that uses and develops on the platform. NVIDIA provides the tools and the foundation — CUDA and the GPUs. Then, developers come in and start building applications that are specifically designed to run on NVIDIA's hardware. As they create more and more of these applications, the platform itself becomes more powerful, which attracts a wider range of users. These users are businesses and technology enthusiasts who use these applications for innovative AI projects. As more people start using the platform, more feedback and demand are created, which in turn encourages developers to build more tailored applications.

Nevertheless, NVIDIA's ecosystem extends beyond this platform. NVIDIA created a developer program that gives developers access to over 150 free Software Development Kits, technical training, webinars, expert help, and networking opportunities (NVIDIA 2023). On average, 39,000 developers sign up for monthly membership. The program currently has more than two million members. At the beginning of 2023, NVIDIA organized the GPU Technology Conference. It brought together experts, developers, and industry leaders and featured over 1,000 sessions covering a broad range of AI topics (see Appendix 53 for NVIDIA’s Support Services and Community Engagement). In February 2023, NVIDIA announced that it was partnering with the Italian oil and gas giant Eni, which involved the expansion of Eni's Green Data Center just outside Milan with the addition of the HPC4 supercomputer. This supercomputer included 3,200 NVIDIA Tesla GPU accelerators. Two years before, NVIDIA collaborated with the University of Florida to build the world's fastest AI supercomputer.

Group part

NVIDIA's Partner Network, including OEMs and Universities, currently has over 1,500 members worldwide (see Appendix 54 for NVIDIA's Partner Types). Along the way, NVIDIA created an ecosystem encompassing hardware, software solutions, a comprehensive range of services, community support, and a robust partner network.

GLOBAL CHALLENGES AND OPPORTUNITIES

The case study explores NVIDIA and the AI chip market, uncovering key elements that are crucial for understanding the future of this rapidly evolving industry.

While it is evident that AI already made a significant impact, it is essential to consider the regulatory and social challenges that may arise due to the increased adoption of AI. How far will the adoption of AI go, and how can short-lived trends be differentiated from disruptive technologies?

Another critical element is the importance of the AI value chain. With concerns around the security and reliability of the value chain, companies need to have a clear strategy for ensuring the capacity and efficiency of their supply chain. How will those strategies foster or hinder the industry's growth?

NVIDIA has already made a significant impact in the AI chip market. However, as different chip types and new technologies are being explored and numerous competitors are entering the market, NVIDIA will have to plan its next steps strategically. Considering its positioning in the market, how well is NVIDIA equipped to overcome internal and external challenges?

Group part

TEACHING NOTE

CASE SUMMARY

The case study examines the strategic choice made by NVIDIA Corporation to invest heavily in the development of chips and software optimized for AI applications, a move that significantly influenced its market position. This decision was particularly pivotal after the introduction of ChatGPT on November 30, 2022, which marked a new era in AI. ChatGPT, an LLM trained on a vast amount of text data, showcases the immense processing power required for such advanced AI systems, primarily operated in data centers and accessible over the cloud.

NVIDIA's ability to anticipate the emerging trend of generative AI and the subsequent high demand for specialized data center hardware positioned the company to benefit greatly from this technological shift. The effect of this strategic move was reflected in NVIDIA's financial performance, as its stock value increased by 155% and reached a market capitalization of $1.1 trillion. The case study delves into the analysis of NVIDIA's strategic decision to capitalize on the AI revolution, focusing on the innovations and challenges that come with such a transformative shift in the technology sector.

TEACHING OBJECTIVES & APPROACH

Target Audience & Teaching Objectives

This case study is tailored for MBA and master’s courses on strategy, as well as executive education seminars. The objective of this case discussion is to conduct a strategic analysis of NVIDIA’s background, evaluate its strategic decisions and operations, and understarnd how the company has attained its current market position. The resulting analysis will help students in assessing NVIDIA's competitive advantage and formulating strategies for its long-term sustainability. The following guiding questions were established as guidelines for achieving the teaching objectives (Figure 1):

Group part

- **Strategic analysis**: To reach the teaching objective, we will use analytical frameworks addressing various aspects of NVIDIA's internal ecosystem and external environment. The evaluation involves PESTEL, Porter's Five Forces, the Business Model Canvas, the Size-Uniqueness Model[4], and the Ansoff Matrix. To complete the analysis, students must integrate essential insights and adapt them into actionable decisions.
- **Strategic decision-making**: The case involves the elaboration of precise business strategies. The formulated strategies should explore long-term visions to set a direction for NVIDIA to achieve sustainable growth and prosper future success.

Figure 1: Guiding Questions

Section	Questions
Opening Question	What is Nvidia's market position and why?
External Analysis	**Q 1.1** What are the main trends shaping the future of AI chips?
	Q 1.2 How do you evaluate the attractiveness of the industry?
Internal Analysis	**Q 2.1** How do you evaluate Nvidia's competitive positioning?
	Q 2.2 How do you characterize Nvidia's existing business model?
	Q 2.3 How do you evaluate the sustainability of Nvidia's competitive advantage?
Strategic Decision-Making	**Q 3.1** What fundamental challenges does Nvidia face and what strategic recommendations would you make?
Closing Question	How do you evaluate Nvidia's chances of long-term success?

Teaching Approach

The teaching method includes a period dedicated to self-study, where students are expected to read and prepare the case before the class. Furthermore, students take part in a 90-minute in-class discussion, resulting in an in-depth assessment of the case. The case lays the foundation for class discussion by providing crucial information about NVIDIA and the AI chip market. This teaching note is intended as a roadmap for instructors (Figure 2). Instructors can explore

[4] Framework introduced by Professor Luís Almeida Costa in his strategy classes.

certain sections deeply or address them at a broader level, depending on academic requirements and desires. The teaching note also involves dynamic features such as teamwork exercises.

To prepare for the class, we suggest exploring recent developments in NVIDIA. It would be helpful to review the Size – Uniqueness Framework before the class, as this is not a standardized strategy framework. The structure of the framework can be found in Appendix 55. To enhance students' preparation, we advise presenting the first two guiding questions regarding NVIDIA's external environment alongside the case. During the in-class discussion, it is vital to briefly cover the results of the external analysis before touching on NVIDIA's internal analysis. To wrap up the case, students should synthesize learnings from the case and theoretical expertise to explain how their presented solutions help NVIDIA set a strategy for growth and improve its chances of success.

The teaching note is structured into three sections, enclosed by the opening and closing questions. The guiding questions help instructors in evaluating students' evolving perspectives concerning NVIDIA's chances of success. Ideally, the following agenda should be presented before introducing the opening question:

- **External analysis**: Determine present and future trends and analyze market attractiveness.
- **Internal analysis**: Establish NVIDIA's competitive positioning and business model.
- **Strategic decision-making**: Evaluate NVIDIA's strategic position and assess value-creating opportunities for NVIDIA's growth.

Group part

Figure 2: Teaching Plan

Part	Plan	Duration	
Preparation for Case Discussion	Students are given the case approximately a week before the in-class discussion for self-study purposes.	**60 min**	
	Students are provided with the first two guiding questions based on the external analysis (Q 1.1 & Q 1.2) to prepare in advance of the class.		
In-Class Case Discussion	The instructor demonstrates the agenda and presents the opening question.	**5 min**	**90 min**
	The instructor and the students briefly review the outcomes of the external analysis done prior to the class.	**10 min**	
	The instructor guides the students through the internal analysis by posing the respective guiding questions (Q 2.1 & Q 2.2). Additionally, students form groups to formulate a draft of the Business Model Canvas.	**30 min**	
	The instructor introduces the Size – Uniqueness Framework and facilitates dialogue regarding the analysis of sustainable competitive advantage (Q 2.3).	**20 min**	
	Following a brief discussion of the results, the students synthesize learnings and evaluate feasible and attainable strategic actions Nvidia can take to grow its business using the Ansoff Matrix (Q 3.1). Conclude with the closing question.	**25 min**	

CASE ANALYSIS

External Analysis

The first two guiding questions assess the industry using PESTEL (Figure 3) and Porter's Five Forces (Figure 4) and analyze the dynamics of competition. These frameworks allow for an extensive evaluation of various environmental factors, present and future, crucial for informative decision-making.

Considering the applied tools are standard frameworks, students should be familiar with the structure. Therefore, guiding questions Q 1.1 and Q 1.2 are meant to motivate self-study and research as preparation for the in-class discussion. Before starting the discussion, we suggest presenting a description of the history of the AI industry, with a specific focus on AI chips. A brief timeline of the history of AI can be found in Appendix 1. To retain emphasis on main teaching objectives, students are encouraged to share key take-aways only.

Q 1.1 What are the main trends shaping the future of AI chips?

The AI chip market has experienced a drastic increase in relevance over the past years, especially concerning rapid advancements in AI, 5G, autonomous driving, data centers, or ML. We have used a PESTEL analysis to examine the macro-environmental factors impacting the

AI chip market by focusing on political, economic, social, technological, environmental, and legal aspects.

Political: The high dependency on Taiwan through centralized manufacturing of AI chips in a single facility introduces significant political risks mainly attributable to Taiwan's delicate political situation concerning China. Additionally, the Russia-Ukraine war has impacted the global supply chain. Considering Ukraine was not able to keep up with the production of critical chip components, China has become the main supplier, creating an intensified overreliance on China. The ongoing commercial war between the United States and China may have extensive consequences for the global supply chain and has the potential to intensify geopolitical tensions. This is particularly crucial as China is one of the main players in the semiconductor value chain and is the second largest end-consumer, responsible for 24% of total semiconductor consumption (Appendix 40).

Economic: One reason for the growth of the AI chip market is the availability of big data, which flourished through the progression of the digitized economy, amounting to a growth of 40% per year. On the other hand, growing demand presents new restrictions to the global supply chain. This may intensify the ongoing chip shortage as manufacturing cannot keep up with the increase in demand due to limited manufacturing capacity. Moreover, material suppliers are concentrated in high-risk regions, leading to a higher uncertainty about the availability of raw materials and price volatility. This impacts final costs overall. Manufacturing suppliers are concentrated within a small number of companies, which means that any regional events can potentially escalate into more substantial and far-reaching problems for the industry. With the surge in demand for AI chips, a shortage of skilled professionals has emerged. Higher wages may be the result as organizations try to attract scarce talent, increasing operational costs. Furthermore, the lack of a skilled workforce may hinder production and innovation in the AI industry, deterring the overall growth of the market.

Social: The global acceptance of AI has risen by 250% since 2017 as it has emerged as an impacting technology rather than a short-term trend. Since the introduction of AI chatbots, numerous industries have adapted chatbots within their processes and systems. As AI chatbots have been widely applicable and are expected to be increasingly utilized in society, the use of AI chips is developing. Rising concerns about social acceptance of AI-based processes are a consequence of increased fear with regard to redundancy rates for both blue- and white-collar workers. The trend of automating processes and systems can impact consumer behavior and employee morale, leading to a feeling of inequality.

Technological: The progression of new and improved algorithmic techniques has allowed exponential growth of computational power. Such growth led to significant improvements in chip performance. Notably, surpassing Moore's Law presents significant potential for growth. Despite the growth opportunities, AI-based hardware applications, such as GPUs, pose power and infrastructure issues. Due to the intensity of power consumption of GPUs often exceeding the capacity of existing data center infrastructure, the full potential of data centers persists to stay unexploited and limits their scalability. This poses unique challenges for companies specializing in GPUs. With AI-powered decisions being utilized progressively in day-to-day activities, hardware trust is becoming increasingly important. Therefore, companies must ensure that both the software and underlying hardware are reliable and secure. To enable seamless AI integration, innovative hardware, and well-defined strategies are essential for future innovation.

Environmental: Due to the production of AI chips, especially GPUs, being responsible for the highest carbon output, concerns over industry sustainability are rising. The decarbonization of the AI chip market is expected to restrict future growth potential. While demand for AI chips continues to rise, production is being boosted to keep up with industry pressure, making it naturally more challenging for the market to persist with environmental objectives and climate

goals. However, organizations are increasingly turning to AI-based systems as a method to reduce their carbon footprint, tackling ESG challenges. The employment of AI chips allows for the optimization of processes, increased energy efficiency, and reduced resource consumption by enabling data-driven decisions. AI chips and their implementation are seen as vital for long-term sustainable progression and to meet organizational requirements without compromising operations.

Legal: The increase in AI-powered decisions is leading to higher complexity regarding regulations, security, and data privacy concerns. Due to the rapidly changing regulatory environment, companies with AI-based systems must consistently assess internal processes against compliance requirements. The EU AI Act provides a risk-based regulatory framework for AI, impacting the way organizations can develop and implement AI solutions and providing increased security for consumers. Moreover, companies must prioritize data protection measures and guarantee reliable handling of sensitive information to sustain public trust and uphold legal responsibilities. Many AI-based companies have built their business model around patent licensing, proving the importance of IP strategy and protection.

Figure 3: PESTEL Analysis

Political	Economic	Social	Technological	Environmental	Legal
• **Centralized manufacturing** of AI chips in **Taiwan** increases dependency, especially concerning due to the delicate **political** situation between **China** and **Taiwan** • The **Russia-Ukraine** war has led to an intensified **overreliance** on **China** as they became the new **main supplier** of chip components • The **United States-China** trade war may have extensive consequences for the global supply chain, due to China's role in the value chain, posing risks of **unpredictable disruptions**	• The **digitized economy** has driven the availability of **big data,** critical for AI progression and adoption • **Manufacturing** cannot keep up with demand, intensifying ongoing chip shortage • **Concentration** of suppliers within **high-risk** regions increases uncertainty about **availability of raw materials** and **price volatility** • **Complexity** of AI poses a **lack of expertise** and **skilled workforce,** potentially leading to higher **operational costs** and **bottlenecks** in production or innovation	• **Acceptance** of AI rose drastically since 2017, emerging as an **impacting technology** • Growing demand of **AI chatbots** to meet and improve marketing objectives, sales processes, etc. • The trend of **automating** processes and systems has created a **fear** of increased **redundancy rates,** potentially harming **social acceptance** and impacting consumer behavior and employee morale	• New and improved **algorithmic techniques** accelerate advancements in **computing power**, outperforming Moore's Law • **Power and infrastructure issues** of AI-based **hardware** applications arise, keeping the potential of **data centers** unexploited • With a shift from software-centric to holistic, **hardware trust** is becoming progressively more important, putting future focus on **innovative hardware** and **well-defined strategies**	• The **manufacturing** of **AI chips** is responsible for the **highest carbon output** with regards to production of electronic devices. • Due to increasing concerns over **industry sustainability** and the increasing **demand for production**, the **decarbonization** of the AI chip market is becoming naturally tougher • Organizatios are employing AI solutions to **tackle ESG challenges**, in which AI chips are seen as vital for **long-term sustainable progression**	• **Security and regulations** are becoming increasingly complex, forcing companies to **consistently** assess processes against **compliance requirements** • The **EU AI Act** has given businesses a **framework** for providing **higher security** for consumers • To sustain **public trust**, **data protection measures,** guaranteeing **reliable handling** of sensitive information, have to be developed • Market leaders have built their business model around **patent licensing**, proving the importance of **IP strategy** and **protection**

Group part

Source (adapted): Alanzi, Salem. “Pestle Analysis Introduction.” ResearchGate. Salford, England: University of Salford, July 11, 2022. https://www.researchgate.net/publication/327871826_Pestle_Analysis_Introduction.

Q 1.2 How do you evaluate the attractiveness of the industry?

Industry Structure on Competition

Porter’s Five Forces is widely used to analyze how industry structure affects the intensity of competition. Not only does it cover direct and indirect competition, but it also extends to the bargaining power of industry participants and the rivalry between companies. The framework makes it possible to assess the strategic position of a business and its potential for profitability.

Threat of New Entrants: Moderate

The case reveals that multiple companies are entering the AI chip market, fine-tuning their products, either developing in-house or using strategic acquisitions. This influx suggests that barriers to entry might not be overly restrictive. The emergence of start-ups like SambaNova, Graphcore, and Tenstorrent further underscores this threat. The substantial VC investments in the semiconductor industry also hint at a conducive environment for new players. However, slowing VC and PE investments, caused by NVIDIA's market dominance, might decrease the threat of new entrants in the future.

Challenges like R&D costs, technical know-how, scalability, and establishing market credibility can act as barriers, potentially moderating the threat level, especially with fewer investors willing to commit capital over a long period.

While investing in AI and AI chips will likely always attract some investors, it is yet to be proven that new entrants can enter the market successfully and gain market share. Further, the ecosystem (software, developer community) and network effects will be restrictive for new entrants not building on an existing ecosystem.

Threat of Substitutes: High

Group part

Cloud providers could be a main customer for AI chip companies, and they are starting to develop their own chips. When a potential customer decides to build a product in-house instead of purchasing the hardware, that in-house solution is considered a substitute for AI chip companies. In economic and business terms, a substitute is a product or service that a consumer can use in place of another product or service. In this scenario, the in-house solution is fulfilling the need that the company's product would have served. Further, cloud provider's AI services are a substitute to edge offerings. Therefore, cloud providers building their own chips will not only reduce the revenue from this customer but can affect the edge data center revenue of other customers, as cloud providers offer higher flexibility and scalability.

Quantum computing is also highlighted, suggesting a potential long-term substitute for traditional AI chip functionalities. However, currently, quantum hardware is not mature enough and commonly used AI chips can only support the development of quantum application.

Bargaining Power of Buyers: Moderate to Low

The case mentions long waiting times for AI chips, even at high prices, suggesting strong demand and thus giving AI chip companies within the industry more bargaining power. Further, buyers tend to be less price sensitive compared to consumer products as AI hardware is bought less frequently, while high quality and longevity are essential for the buyer's operations. However, with more companies entering the market and providing diverse product options, buyers might gain more negotiating power over time. This is especially true for large customers building multiple data centers. The bargaining power of big technology companies like Amazon or Microsoft is expected to increase significantly as their cloud revenue constitutes a substantial part of the market demand.

Bargaining Power of Suppliers: Moderate

Some companies, such as Intel, produce their own chips, while others, like AMD, heavily rely on suppliers such as TSMC. The option to produce in-house potentially reduces supplier power.

As the industry grows and diversifies, the dynamics between manufacturers and suppliers will influence this force. Production is highly complex, and factories are designed and built for specific chips. The machinery and training of employees may not be relevant if the supplier decides to produce different AI chips, leading to high sunk costs. However, depending on the severity of the frequent product iteration of AI chips new training and adjustment of machines may become necessary either way. The emergence of the chip shortage in 2020 underscored the limited availability of expertise and manufacturing facilities, granting suppliers a degree of bargaining power. Moreover, as production capacity is negotiated long in advance, the recent rise of AI will give suppliers more options for future customers and thus more bargaining power. For the suppliers it is essential to strengthen business with customers established in the industry that can drive revenue in the following decade, hence market leaders for AI chips are especially well positioned.

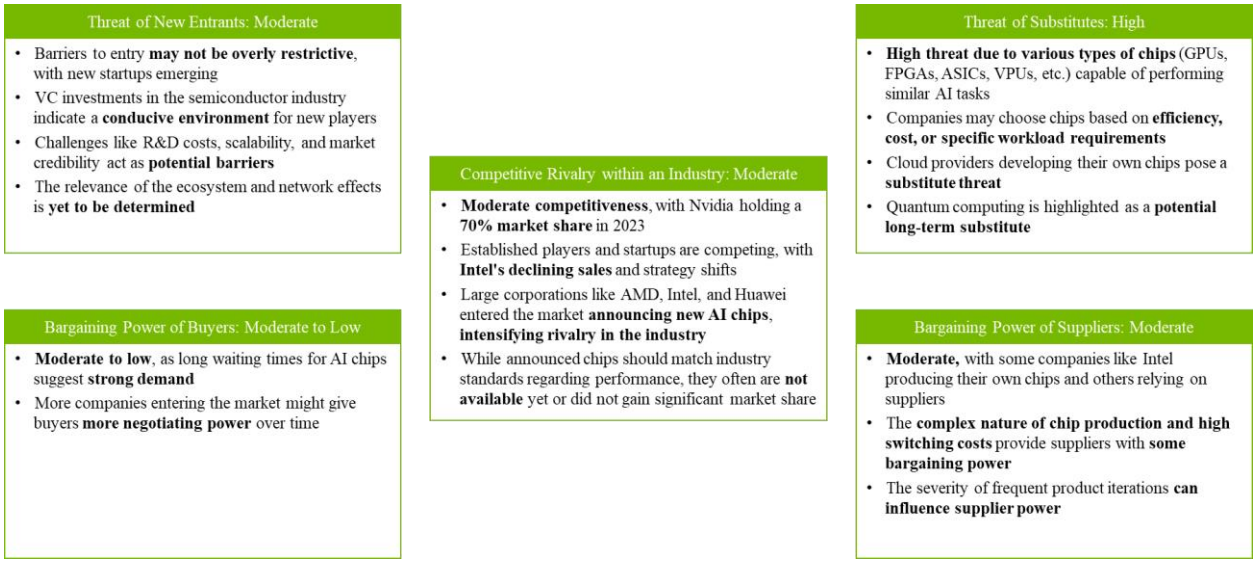
Competitive Rivalry within an Industry: Moderate

In the year 2023 NVIDIA's market share was around 70% and most chips from competitors are not ready or proven yet. Nevertheless, the AI chip market appears to be competitive. Established players like Intel, AMD, and IBM are mentioned, but there's also a note of start-ups trying to disrupt the industry. Intel's declining sales and shifts in its strategy further highlight the competitive pressures. The main rival for all companies within the AI chip industry is NVIDIA, which set the status quo for hardware and software. However, high prices and low availability lead to customers anticipating new chip releases from companies besides NVIDIA.

In conclusion, Porter’s Five Forces analysis reveals a highly dynamic landscape concerning the AI chip industry. The threat of new entrants is moderate, as VC investments reduce barriers such as high R&D costs and market credibility. The threat of substitutes is high, considering diverse chip types are competing for dominance, and the intention of in-house developments is rising. Buyer power is moderate to low but has the potential to increase with an increase in

competition and influence of large technology companies. Supplier power remains moderate due to established strategic relationships and production complexities. Competitive rivalry is also moderate, as there is currently a clear market leader. However, other established firms and start-ups are challenging its dominance. How these forces interact will influence the future of the AI chip industry.

Figure 4: Porter's Five Forces



Source (adapted): Bhaskar, Mahesh P, Snehal Pawar, and Yogesh Hole. “Porter’s Five Forces Model: Gives You A Competitive Advantage.” International Journal of Physics 11, no. 04 (January 2019).

Dynamics of Competition

Competition is affected by the structure of the industry, but it is not fully determined by it. To gain a better understanding of competition, it is necessary to examine how companies interact with each other.

Out of 17 data center AI chips, most are still under development, and only a few have been deployed following the release of ChatGPT. The rapid development of new chips to match the competition showcases the industry's fast pace. For the companies that already have chips available for sale, the emphasis always lies on performance and efficiency, not price. The

market can still be described as uncontested. Business-to-business (B2B) prices are not standardized and not fully disclosed. However, price estimates suggest high markups for AI chips, indicating the focus on innovation. As the market is relatively new and companies need to offset their high R&D costs, low-cost AI chip alternatives are currently unavailable. However, the value added for the customer and their potential efficiency gains, for example, in drug research, outweigh the high acquisition costs.

The adoption of GPUs in AI data centers created new demand due to their superior processing power, enabling new applications. While NVIDIA and AMD are succeeding according to market share and analyst reports, Intel is losing market share mainly due to selling CPU-based AI chips. Customers seem to value the flexibility that GPUs offer as they are applicable to a wide range of systems. However, different chip types, such as FPGAs and ASICs, are also being developed, which are well suited for AI workload and superior in terms of efficiency. Companies might choose one over the other based on efficiency, cost, or specific workload requirements. In the short-term, the flexibility of GPUs is appreciated while LLMs still develop, but companies might switch to ASICs in the long run when the purpose and perfect size of the LLMs are determined to reduce operating cost or overcome infrastructure restraint due to the high amount of electricity that is needed.

Working together and establishing business relationships with important suppliers, such as TSMC, is essential for fabless AI chip companies, as the demand currently outweighs the supply. Customers in need of AI solutions will prefer paying a premium to avoid long waiting times if cost savings from their AI solution outweigh acquisition costs. Partnerships with large customers, such as cloud providers, are commonly used to generate traction for the brand and secure market share. For example, NVIDIA and AMD are closely cooperating with Microsoft for their newest generation of chips. However, as a few cloud providers are generating large revenue streams, AI chip companies face customer concentration risk. The pricing power of

cloud providers will increase as soon as multiple AI chips are available. Further, cloud providers are working on their own chips to reduce dependency and increase their bargaining power.

Start-ups are adopting innovative approaches to develop specialized hardware solutions to cater to the needs of the market. Challenges start-ups face include the need for significant investment in R&D, the ability to scale efficiently, and the need to establish credibility in the market. Start-ups, on average, become profitable after three to four years. However, for AI chip start-ups, time-to-market and path to profitability are expected to be significantly longer. By then, start-ups will face stiff competition from established players who have already built a brand and gained market share.

In conclusion, AI chip companies compete not only on innovation to gain market share but also for partnerships with large customers and suppliers. The creation of new demand and the value of innovation in a still uncontested market indicates that companies are pursuing a Blue Ocean Strategy (Kim and Mauborgne 2004). The redefined market boundaries and altered competitive landscape can lead to the decline of traditional players like Intel, who fail to adapt. Meanwhile, innovative companies can gain profits and grow rapidly. However, if market leaders fail to continue to innovate, the attracted competitors can create a contested “Red Ocean”, and market participants will start to compete on price and efficiency.

Network Effects

The dynamics of competition can vary significantly in markets with network effects. Thus, it is essential to analyze them to gain a comprehensive understanding of what companies are competing for (Zhu and Iansiti 2019).

AI chip hardware from different manufacturers is often not compatible with each other and, therefore, creates an ecosystem lock-in. However, the value of this hardware does not grow with an increasing number of data centers employing it besides brand recognition. To leverage

network effects, many AI chip companies run specific software on their chips that is only compatible with their own hardware. Network effects play a crucial role in reinforcing the competitive advantage of companies that recognized the trend for AI early.

Direct network effects are a phenomenon where the value of a product or service increases as more people use it. A larger developer community can, therefore, attract more developers. Companies can benefit from this by growing developer communities around their hardware and software. This is highly relevant as the documentation and software available for chips can increase the ease of use, making the ecosystem more attractive.

Cross-side indirect network effects happen when more developers and software are available for the chips, which consequently attracts more hardware customers. Better software and chip documentation can decrease implementation costs and operational expenses. AI chip companies can, therefore, steer the revenue of their hardware by offering better software that, again, will attract more developers. The strength of network effects can dramatically shape value creation and capture increasing barriers to entry.

Network structure refers to the difference between building multiple small local networks or a single large global network. The structure influences the ability of each company to sustain the scale of its community. AI chips and developers operate on a global scale, as the company's software is accessible anywhere. It is worth noting that the code, in general and for AI chips, is written in English. Therefore, developer communities act as one global network cluster. Local clusters, in which each country would have its own developer community, could increase the vulnerability to business challenges. As the networks would be smaller, a new developer community would need fewer developers to become as attractive as existing communities. Therefore, local network structures would decrease the difficulty of reaching critical mass, the minimum of developers needed to operate a community in each market.

Disintermediation, the risk of customers bypassing the platform, increases as third-party software that can connect to software from different chip manufacturers becomes available. While one does not exit the ecosystem in a typical way, it would bypass the need for the developer community of a specific company, reducing switching costs. Disintermediation is mainly reduced by innovation and increasing the value of the hardware product itself.

Multi-homing, or using multiple software platforms at once, is inconvenient as chips are often not compatible with each other, but it can become a factor for large data center providers that provide services to customers that need more processing power than a single data center can provide.

Network bridging occurs when companies from different business areas have a developer ecosystem that can be transferred. For example, Intel already has developers for other chip types that might be relevant for AI chips. The same occurs for hardware that applies to multiple solutions, for example, data centers and gaming devices.

In conclusion, lock-ins can be created by locking in one or both sides of the market. Customers that only use software from one AI chip company employ data centers to use these chips. Data centers can create a similar lock-in by providing only chips from one brand. However, this can be surpassed if third-party software that is operable on both chips is used. Therefore, creating a large developer community is extremely important for AI chip companies as a lock-in focused on software is more robust. Existing network effects can reduce investments in other companies and thus reduce competitive pressures for the competing platform.

Internal Analysis

In this part of the teaching note, we look at NVIDIA's competitive positioning and business model, concluding on whether the company has a sustainable competitive advantage. The

internal analysis aims to convey to students in what way the company's competitive positioning and business model can set it apart from its competitors.

In Q 2.1, we will examine NVIDIA’s competitive positioning. We will introduce its generic strategy to gain a competitive advantage, and then we will delve into its value proposition, explaining how NVIDIA has been trying to create value for clients in a unique way. We will conclude with an analysis of its ecosystem. Evaluating a company's competitive positioning is imperative as it offers insights into its market standing, facilitating the identification of profitability and comparative positioning against competitors, thereby revealing whether the company holds a competitive advantage.

To answer Q 2.2, we will use the Business Model Canvas (Figure 5). We chose this template as it provides an in-depth deconstruction of the business model and enables us to understand the specific actions the company has been taking to achieve a competitive advantage. Before starting the analysis, we suggest recapping the nine basic Business Model Canvas building blocks.

To assess NVIDIA's sustainable competitive advantages, Q 2.3 analyzes the company’s strategic positioning and business model drivers through the lens of the Size-Uniqueness Framework (Figure 6).

Q 2.1 How do you evaluate NVIDIA’s competitive positioning?

In general, firms can pursue two primary strategies to gain a competitive advantage within their industries: cost leadership or differentiation. Cost leadership involves becoming the lowest-cost producer, while differentiation focuses on creating unique and valued products or services, allowing the company to charge premium prices (Porter 1985). In NVIDIA’s case, the company follows a differentiation strategy, setting itself apart in the competitive landscape by prioritizing and addressing specific needs highly valued by customers in the AI chip market. These needs

include high performance and energy efficiency, strong data storage capabilities, computational power, as well as reliability and security.

NVIDIA’s commitment to innovation is at the core of its value proposition. This dedication has empowered the company to deliver cutting-edge solutions that meet the rising demand for high-performance technologies. The company's advanced GPUs play a critical role for customers requiring powerful computational capabilities, such as gamers, graphic designers, and professionals in data-heavy industries. These GPUs are renowned for their speed and reliability, offering superior processing power for complex tasks. In the field of AI and deep learning, NVIDIA leadership provides immersive value. The company offers solutions that enable organizations to integrate AI into their businesses quickly and seamlessly. NVIDIA relies heavily on a strong pool of talented human capital to cater these solutions to the market. Furthermore, the company's commitment to sustainability and corporate responsibility aligns seamlessly with the values of a modern customer base.

NVIDIA's ultimate value creation comes from the synergy of its innovative solutions, encompassing both hardware and software offerings, fortified by a diverse range of services, including cloud solutions. All these offerings culminate in establishing a robust ecosystem where customers are equipped with all the tools for effective application performance. This ecosystem stands out as the most developed among its competitors.

NVIDIA's ecosystem is a network of connected partners, components, and technologies working together to support the company's products and services. The following five key elements best describe NVIDIA's ecosystem:

1. Hardware: Cutting-edge graphic cards for various applications in the gaming, data center, visualization, and automotive markets.

2. Software and Services: Comprehensive software solutions like the CUDA computing platform and cloud services that enhance and complement NVIDIA’s hardware.

3. Community and Support: Customer service, educational programs, and forums in which NVIDIA's collaborative community of users and developers can support and communicate with each other.

4. Partner Network: Strategic alliances with manufacturers, OEMs, software developers, academic institutions, and other stakeholders, helping to integrate NVIDIA's products into a wide range of solutions to reach a broader customer base.

5. AI and High-Performance Computing (HPC) Applications: Specialized AI and HPC platforms that can solve the computational needs of these modern technological fields, like the NVIDIA AI Platform for enterprises.

Q 2.2 How do you characterize NVIDIA's existing business model?

The Business Model Canvas is a widespread strategic management template used to describe, document, and challenge a business model. It provides a structured way to visualize and analyze nine key components of a business and how they interact. The Value Proposition component has been omitted in this section, as it was already addressed in Q 2.1.

Key Activities: One of NVIDIA's core activities is R&D, enabling it to innovate and offer cutting-edge solutions to meet its customer's fast-changing needs and maintain its leading position in the AI chip market. Notably, NVIDIA’s significant investment in R&D surpasses that of its competitors and has enabled the company to anticipate market trends and capitalize on them. By following a fabless business model, NVIDIA can focus on designing and engineering high-performance and energy-efficient GPUs, translating the R&D breakthroughs into market-ready products. Besides that, NVIDIA invests in updating and creating software to

complement its hardware, like CUDA, ensuring a seamless integration of its products into user ecosystems.

Among NVIDIA’s strategic activities, mergers and acquisitions play a pivotal role, as demonstrated by its strategic acquisition of Mellanox and initiatives like the Inception Program. By supporting start-ups with the Inception Program, NVIDIA contributes to developing cutting-edge technologies and solutions in AI and creating a robust ecosystem of companies that might use or contribute to NVIDIA's own technologies and products. Forming strategic partnerships with key stakeholders allows the company to continuously leverage its partners' capabilities to deliver high-quality chip design.

NVIDIA's technical support and customer service use multiple communication channels to ensure customer satisfaction and retention, thereby gaining crucial insights into customer needs and improving products through the Community Feedback Forum.

Furthermore, NVIDIA's sales and marketing activities are critical in promoting products and services to potential customers. This is done via various marketing channels and direct sales to consumers through its informative website and NVIDIA’s representatives.

Key Partners: NVIDIA has built key strategic partnerships across various industries to maintain its market reach and enhance product functionality. The company has formed alliances with Original Equipment Manufacturers (OEM) and top-tier PC and laptop manufacturers like Dell and Lenovo to incorporate NVIDIA products, technologies, and solutions directly into their end-user devices. NVIDIA also works closely with independent software vendors (ISVs) that use NVIDIA's GPUs to run their software. NVIDIA ensures that ISV software takes full advantage of NVIDIA’s graphics cards' capabilities, leading to enhanced performance and visual quality. To reinforce NVIDIA's presence in the cloud computing area, the company teams with major cloud service providers like Google Cloud, Microsoft Azure, and Amazon

Group part

Web. NVIDIA collaborates with both emerging and established companies in the AI universe to promote the adoption of its unique AI platforms. Global academic and research institutions partner with NVIDIA to advance AI and computer graphics research, enriching NVIDIA's R&D activities. Moreover, NVIDIA, at the forefront of AI, is partnering with the United States government to help tackle challenges in healthcare and sustainability to cybersecurity and connectivity with its innovative technologies.

Key Resources: NVIDIA's key resources are fundamental for its innovative edge and leading market position. One critical resource is the company's human capital, which is vital for continuous innovation and product development, especially its team of highly skilled engineers. NVIDIA's human capital is remarkably retained, with a low turnover of 5% in 2022. NVIDIA establishes and maintains connections with universities and technical organizations to attract its human capital and actively participates in industry conferences. Employee referrals are also an important instrument responsible for over 37% of new hires in 2023. To retain its human capital, the company ensures a dynamic learning environment with on-the-job training, coaching, and continuous feedback. Additionally, NVIDIA generates customized learning journeys for its staff to address developmental wants and needs, continuously enhancing its product range to guarantee employees are offered the latest applications and programs. NVIDIA's compensation program stimulates future investment, such as employee equity, as well as rewarding performance.

Another key resource that the company established is its active engagement with the software developer community, encouraging the development of applications optimized for NVIDIA's platforms and software. Engineering and marketing teams collaborate closely with important software developers to discuss and advocate platforms, address product requirements, and solve technical issues. NVIDIA's developer program ensures early access to products, motivating the creation of software tools. Additionally, NVIDIA's Deep Learning Institute focuses on training

developers and empowering them to construct AI applications that bargain NVIDIA’s platform capabilities.

A robust technological infrastructure supports NVIDIA's hardware and software solutions, such as the CUDA platform and AI systems.

Another crucial resource is the company's intellectual property, particularly patents in advanced GPU designs and AI applications. It protects NVIDIA's proprietary technology, making it more challenging for competitors to recreate.

NVIDIA has generated a brand image exhibiting high performance and quality. This reputation has served as an impactful asset, strengthening market leadership and new product acceptance. The company’s extensive laboratories and research center facilities assist NVIDIA in staying ahead of the technology curve. A worldwide distribution network ensures product availability across various markets. Lastly, NVIDIA's solid financial position enables the company to invest in R&D, acquisitions, and other critical projects for its success.

Customer Segments: NVIDIA's customer segments are diverse, expanding its operations across gaming, data centers, professional visualization, and automotive markets. Gamers are one of the main customers, relying on NVIDIA's GeForce GPUs for high-end gaming experiences. With the advancements in AI technology, enterprise clients seeking robust AI and data center solutions have become increasingly important for NVIDIA. The company supplies them with advanced computing products designed for their needs. Another customer segment is the automotive industry, which includes manufacturers and tech firms that develop autonomous vehicles and rely on NVIDIA's Drive platform. Designers and content creators from the media and entertainment sector use NVIDIA's products to speed up graphic rendering and other visualization tasks. Scientific and academic institutions utilize NVIDIA's GPUs for complex data analysis across climate studies or bioinformatics fields. Finally, NVIDIA serves

customers, including OEMs and cloud computing providers, who incorporate NVIDIA’s advanced GPU technologies to elevate their respective services and applications. These diverse customer segments share their need for cutting-edge technology to perform high-performance computing demands.

Customer Relationship: NVIDIA ensures a strong ongoing customer relationship by offering a range of services that suit the different needs of its diverse customer groups. For individual consumers, NVIDIA offers extensive online support, including live chat, FAQs, and forums accessible through their comprehensive online platform. Enterprise clients receive more customized personal support from account managers and specialists. This direct support is critical for businesses that rely on NVIDIA's technology for their operations. Additionally, NVIDIA offers training and educational programs such as the NVIDIA Deep Learning Institute, fostering a relationship with developers and students by empowering them with the skills needed to use NVIDIA's technology effectively. Furthermore, NVIDIA strengthens its user base by nurturing active community platforms where gamers and professional developers can exchange ideas, share experiences, and support each other.

Channels: NVIDIA delivers its products to the end user through various integrated channels. Customers can make purchases directly from NVIDIA's official online store. Further, NVIDIA has a sales team with NVIDIA representatives who work directly with enterprise customers. Through OEM partnerships, NVIDIA reaches a broader market without direct selling. Having a global network of retailers and distributors enables NVIDIA to reach casual consumers by making their products available in physical and online stores. A direct-to-consumer channel represents NVIDIA's cloud-based services like the DGX Cloud. It operates on a subscription model, providing a continuous touchpoint for customers and an ongoing revenue stream. Lastly, NVIDIA hosts and participates in events and conferences to showcase new technologies and updates of existing features and engage with partners and customers.

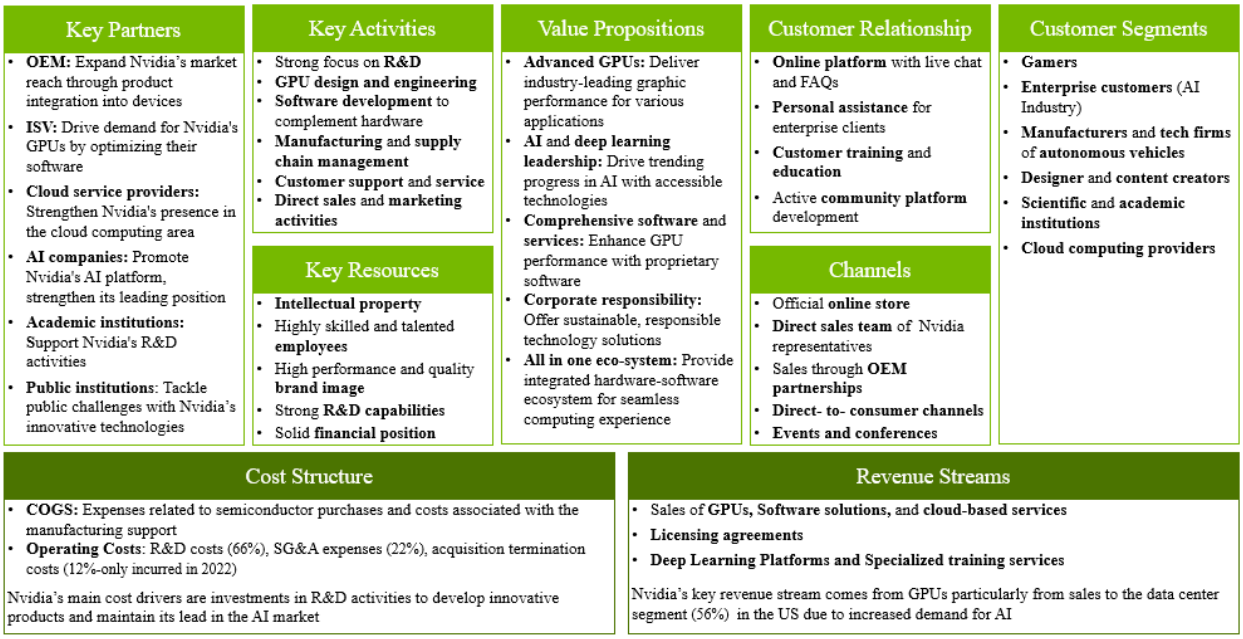
Group part

Cost Structure: NVIDIA's cost structure can be categorized into cost of goods sold (COGS) and operating costs. NVIDIA's COGS primarily encompasses expenses related to the semiconductor purchase from subcontractors and costs associated with manufacturing support. These contain direct labor and overheads, inventory and warranty provisions, and shipping costs. NVIDIA's operating costs mainly include R&D as well as sales, general, and administrative (SG&A) expenses. With 66% of the total operating costs, R&D expenses are the biggest cost driver.

Revenue Streams: NVIDIA reports its financial results in two primary segments: Compute & Networking and Graphics. Within these two segments, NVIDIA generates revenues from several primary sources, reflecting the diverse markets and industries it serves. The main revenue stream is selling AI Chips, mainly GPUs designed for gaming, professional visualization, data centers, automotive markets, and OEMs. NVIDIA also generated revenues through software solutions like the AI Platform along with cloud-based services. Furthermore, the company earns income from licensing agreements with other companies that use NVIDIA technology patents. NVIDIA also capitalizes on its deep learning and AI expertise by offering a deep learning platform and specialized training services. A key driver of its rapid revenue growth comes from data center solutions, which comprise 56% of the total revenues, including high-performance GPUs and networking equipment used for machine artificial intelligence. Looking at the geographical distribution, NVIDIA generates 31% of its income in the United States, followed by 26% in Taiwan and 21% in China.

Group part

Figure 5: NVIDIA's Business Model Canvas



Source (adapted): Osterwalder, Alexander, and Yves Pigneur. 2010. Business model generation: a handbook for visionaries,game changers, and challengers. Vol.1. Rio de Janeiro: John Wiley & Sons.

Q 2.3 How do you evaluate the sustainability of NVIDIA's competitive advantage?

The Size-Uniqueness Framework is a tool to analyze the sustainability of competitive advantage introduced by Professor Luís Almeida Costa in his strategy classes. It identifies two fundamental sources of sustainable competitive advantages: size and uniqueness.

When considering size-driven advantages, we encounter two distinct scenarios. The first is the situation of unprofitable imitation. The dominant market position held by a company, or a small group of companies, dissuades potential imitators from replicating key factors. In this scenario, imitators recognize that engaging in imitation means confronting the established competitors, who, because of their size, have both the ability and motivation to respond aggressively. The established firms' capacity for aggressive reactions arises from a variety of elements, including

economies of scale and scope, brand and geographic proliferation, or buying switching costs[5]. Additionally, size can provide an even higher level of protection and put in motion a "success breeds success" process, influenced by factors such as network and word-of-mouth effects.

Turning to uniqueness-driven advantages, sustainable competitive advantages can arise from two sources. The first is exclusivity, where imitation is not permitted. This occurs due to ownership of patents, copyrights, and licenses or exclusive access to inputs and distribution channels. The second source is causal ambiguity, uncertainty, and social complexity. Here, imitation is difficult or time-consuming for competitors. This arises from organizational advantages, distinctive capabilities, external architectures like platforms, networks, partnerships, or a strong brand and reputation. Below, we will apply this framework to NVIDIA and conclude on its sustainable competitive advantages:

Size – Unprofitable Imitation

Economies of Scale: Focus on the cost advantages achieved through increased production levels of a particular product. In NVIDIA's industry, scale is crucial due to the substantial fixed costs involving significant R&D and CAPEX. Leveraging its size, NVIDIA effectively harnesses economies of scale with a high production volume, leading market share, and strategic operational efficiency. Moreover, with global market access, NVIDIA leverages a broader customer base to scale operations and profits strategically reinvested in R&D, leading to a virtuous cycle of innovation and efficiency. NVIDIA's scale also enhances bargaining power with suppliers and potential partners, which is especially crucial given the industry's concentration among a select few manufacturing suppliers. However, the industry's limited manufacturing capacity potentially disrupts these conventional dynamics.

[5] In the case of NVIDIA, we will focus solely on developing elements related to economies of scale and scope, as other factors are not present.

Economies of Scope: Arise when a business gains cost advantages by producing a variety of goods or services rather than specializing in the production of a single product. NVIDIA benefits from economies of scope due to its diverse product portfolio and expertise spanning various areas within the technology sector. Its broad portfolio allows the sharing of R&D efforts, achieving efficiencies as innovations developed for one product can be adapted or applied to another, leading to cost savings. This is particularly significant given NVIDIA's focus on GPU architecture, which is leveraged across various applications. Synergies extend to marketing efforts, particularly those directed at developers and customers.

Buyer Switching Costs: The concept refers to customers' costs when switching to a different provider. NVIDIA leverages its interconnected system of software and hardware offerings to increase the switching costs for its customers. NVIDIA’s CUDA platform and software is only compatible with their own AI chips. If applications are tailored to NVIDIA’s architecture and the customer switches hardware, not only rewriting or adapting the software might be necessary, but the developers might also need additional training for the new software. Further, AI chips in an installation are interconnected. Therefore, customers would need to switch AI chip provider for the whole installation. While options for third-party solutions that can bypass those obstacles may increase, for now, NVIDIA successfully created hardware and software lock-ins as current developers are already familiar with the CUDA platform.

Size – "Success Breeds Success"

Direct Network Effects: The earlier introduced concepts apply especially to digital and technology markets. NVIDIA leverages network effects by focusing on and growing its software platform, CUDA. Due to direct network effects, the value of the CUDA platform increases as more people use it. This can be explained by the fact that with more users, the code for AI applications is better documented, and developers receive more feedback, which

increases the ease of use. Offered for free, today, more than four million developers are using CUDA.

Cross-side indirect network effects: Not only did NVIDIA grow a large software community, but it was also capable of leveraging it using cross-side indirect network effects. As the CUDA software can only run on NVIDIA's GPUs, the growing developer community positively affects hardware sales. If the developers of a customer are already familiar with the CUDA platform, they are likely to buy NVIDIA chips as it is the only way to use the CUDA software. By leveraging network effects and exploiting the high switching costs caused by the incompatibility of AI chips, NVIDIA was able to establish a sustainable competitive advantage.

Uniqueness – Exclusivity

Patents, Copyrights, and Licenses: Innovations, such as GPU Architecture and AI platforms, have been key drivers for establishing and leveraging NVIDIA's performance. Notably, NVIDIA holds more patents than its main competitors, Intel and AMD. However, relying solely on patent protection does not make market entry impossible for competitors. Despite the elevated imitation costs linked to patents, research revealed that 60% of patented innovations were imitated within four years. Moreover, about half of the surveyed innovations led firms to believe that patents had only delayed imitators' entry by a few months or less (Mansfield 1981). Consequently, the conclusion arises that patent protection does not confer a sustainable competitive advantage for NVIDIA.

Exclusive access to inputs and distribution channels: NVIDIA, along with its competitors, relies on a limited number of suppliers for raw materials and manufacturing processes concentrated in politically sensitive regions like Taiwan. Distribution channels in the industry are generally similar. The adoption of a fabless business model in a highly specialized and

globalized value chain further heightens NVIDIA's dependence on manufacturing suppliers, posing a significant weakness in NVIDIA's strategic positioning.

Uniqueness – Causal Ambiguity, Uncertainty, and Social Complexity

Internal Capabilities: NVIDIA excels in an innovation-driven industry, showcasing impressive flexibility. Such ability is a testament to its organizational advantage and distinctive capabilities. NVIDIA’s consistent success lies in its remarkable capacity to anticipate market trends and capitalize on them, benefiting from a first-mover advantage. This foresight has led the company through transformative phases involving strategic restructuring and the delivery of innovative products. Initially, by recognizing the future trajectory of computing and identifying the immense demand for graphics improvement, particularly propelled by the evolving PC gaming industry. Consequently, the company launched its GPUs that revolutionized industry standards by enhancing processing capabilities. On a second moment, NVIDIA recognized the suitability of its GPUs for the computational requirements of AI applications and strategically shifted its focus to the rapidly expanding field of AI. Leveraging its GPU architecture, the company started developing cutting-edge AI platforms that are relevant to its CUDA program.

Such achievements are the result of strong HR and R&D efforts. NVIDIA spends more on R&D than its competitors and most of its personnel is dedicated to this area. The company's effective HR policies contribute to attracting and retaining top talent, although the challenge of a skilled workforce shortage is anticipated to grow. NVIDIA’s R&D efforts are exponentiated by its fabless business model, which enables a more efficient allocation of resources towards chip design. In contrast to a company engaged in in-house production, NVIDIA does not face logistical challenges and costs when modifying its manufacturing processes to accommodate rapid changes in design.

Group part

The software-centric nature of NVIDIA further underscores its adaptability, as software development is not bound by physical limitations, making it easier to update. One last key factor contributing to NVIDIA's agility in the market is its GPU-focused approach since GPUs allow for multiple applications across different industries. However, they are less efficient than other AI chips, such as FPGAs and ASICs. This raises questions about the trajectory of these newer chips in the future. This is particularly pertinent given that chip efficiency is intricately tied to sustainability, a pressing concern in today's landscape. The greater the efficiency of a chip, the higher its computing capabilities, all while consuming the same amount of power. Subsequently, efforts around quantum computing are also being made, but NVIDIA has already positioned itself to capitalize on it here.

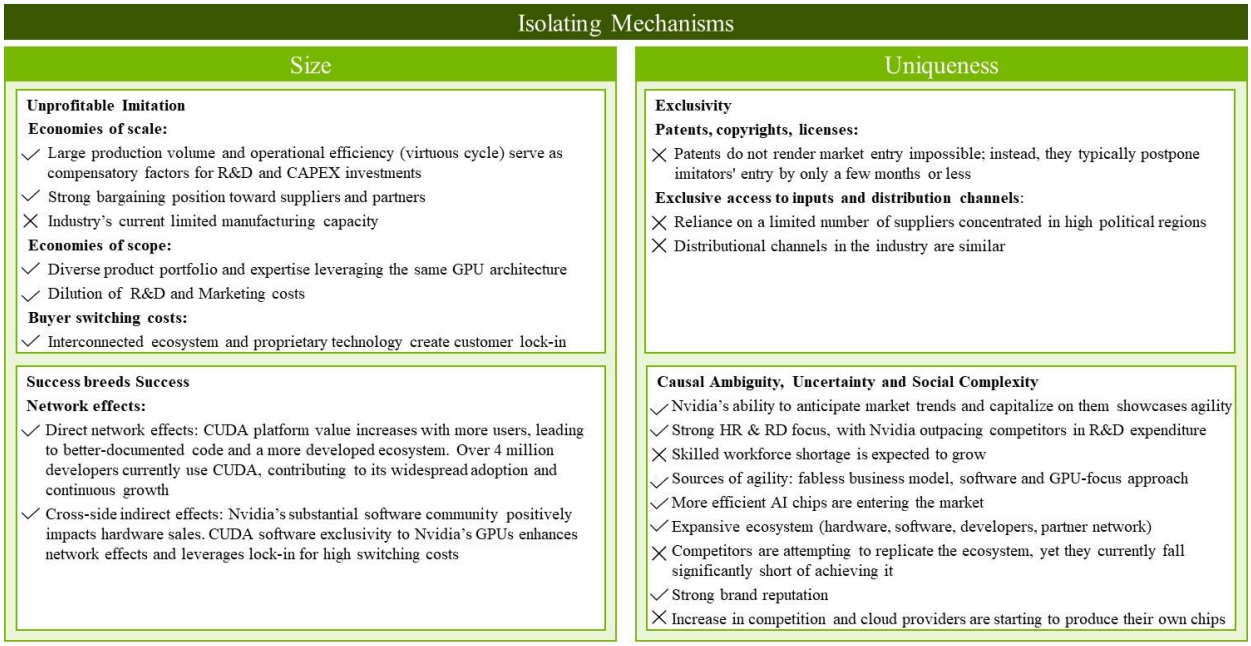
External Capabilities: NVIDIA has managed to build a robust ecosystem by investing in the development of software tools that seamlessly integrate and complement its hardware offerings. This synergy between hardware and software creates a virtuous cycle of efficiency that is fortified by its active engagement with a growing network of strategic partners and developers. Competitors are developing their ecosystems, thus fostering a growing availability of software compatible with multiple platforms. However, such efforts are still far from the dimension of NVIDIA's growing ecosystem that will continue to benefit from network effects.

Brand and reputation: NVIDIA's proven history performance and favorable consumer feedback allowed the company to attain a strong brand reputation and position itself as an industry leader. Famous for its hardware and software innovations, NVIDIA is closely associated with AI excellence. However, competition is increasing, especially from companies such as AMD and Huawei. The latter's growth is particularly relevant given their capability to serve the expanding Chinese market, while NVIDIA is facing challenges to do so due to the ongoing commercial war between the United States and China. Competitors, such as AMD, are

strategically entering the FPGA market. Lastly, cloud providers are starting their own chip production, indicating a potential shift in the industry's structure.

In summary, NVIDIA's sustainable competitive advantages are rooted in its size-driven benefits, encompassing economies of scale and scope, as well as network effects. Additionally, the company derives distinct advantages from its organizational agility, the development of a robust ecosystem and a strong brand reputation. Nonetheless, challenges persist, particularly in terms of securing inputs, specifically in manufacturing. The evolving competitive landscape poses additional hurdles with new competitors entering the scene, accentuated by the rapid pace of the industry, necessitating continuous and dedicated efforts in innovation.

Figure 6: Size – Uniqueness Framework



Source (adapted): Almeida Costa, Luís. "Size – Uniqueness Framework." Framework introduced in Strategy Courses.

Strategic Decision-Making

Building upon our previous analysis of NVIDIA, where we examined the external and internal environment, this section first aims to identify the main challenges faced by the company.

Consequently, our goal is to identify new value-creating opportunities by leveraging on the firm's unique resources and capabilities and previously identified market trends. To present the value creation opportunities, we will apply the Ansoff Matrix, a framework that allows for the evaluation of both risks and opportunities associated with each identified strategy.

Q 3.1 What fundamental challenges does NVIDIA face and what strategic recommendations would you make?

Considering NVIDIA's existing resources and capabilities and the outcomes of the external analysis conducted, the company faces different challenges. The AI chip industry operates at an accelerated innovation pace and needs substantial operational costs, particularly in R&D. Moreover, the AI chip value chain is susceptible to various disruptions, and NVIDIA is particularly vulnerable to them, given its fabless business model. Key disruptions arise from the dependence on a limited number of material suppliers and manufacturers, mainly concentrated in regions with high political risk, with particular emphasis on the dependence on TSMC and Taiwan. As a consequence, the industry has shown limited manufacturing capacity coupled with an increasing shortage in human capital which led to a chip semiconductor shortage since late 2020.

The company is exposed to considerable risk in the form of potential new United States regulations that could further constrain commercial relations with China. This risk is particularly pronounced as China constitutes a significant portion of NVIDIA's total revenue, namely 21% in 2022.

Despite having no single customer contributing more than 10% of revenues, NVIDIA depends on a few large customers. One of its important customer segments, cloud service providers, is starting to produce its own chips. This will likely reshape traditional market dynamics. Finally,

the industry faces challenges related to shifting to more environmental and sustainable practices.

Following the Ansoff Matrix, a strategic planning tool that identifies four basic growth opportunities for businesses, namely market penetration, market development, product development or diversification, the following strategies for value creation emerge as viable options for NVIDIA to overcome its primary challenges (Ansoff 1957):

Market penetration: NVIDIA should further strengthen its partnership with the United States government, addressing challenges ranging from healthcare and sustainability to cybersecurity. This strategic alliance would not only allow for a reliable revenue stream but would also shield the company from potential disruptions arising from future United States regulations impacting commercial relations with China. Additionally, the widespread accessibility of generative AI has transformed the landscape, enabling every company to leverage its versatile capabilities. NVIDIA's commitment to utilizing generative AI to aid companies in constructing their own LLMs is a noteworthy initiative. To propel this momentum further, NVIDIA should strategically capitalize on this opportunity, persist in these collaborative efforts, and actively seek to attract more companies.

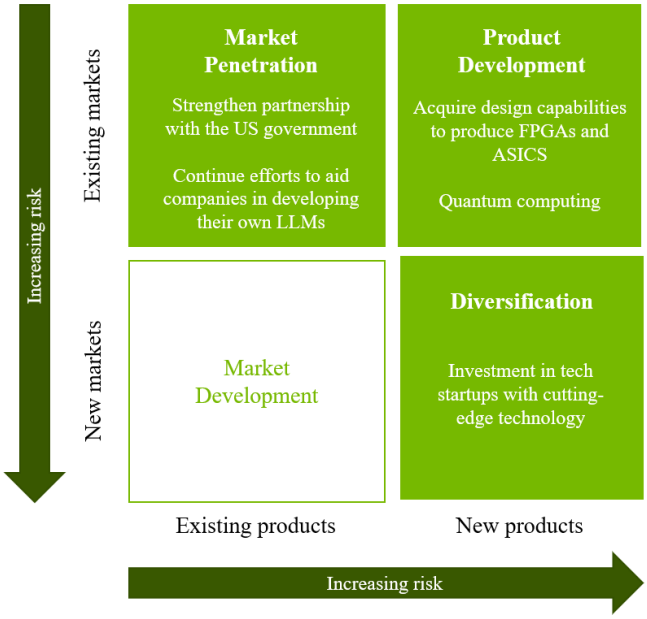
Product Development: NVIDIA, primarily centered on GPUs, faces a concern regarding the efficiency of its chips, particularly in comparison to FPGAs and ASICs. A recommended strategy for NVIDIA is to acquire chip design capabilities tailored for those chips, a path already taken by competitors AMD and Intel. Notably, concerning CAGR, the ASIC segment was forecasted to obtain the highest growth rate between 2023 and 2032. This strategic move would be driven by the superior efficiency of these chips in contrast to GPUs, although they cater to more specific use cases and are less flexible. This becomes especially pertinent considering the connection between efficiency and sustainability, as previously highlighted. Lastly, by opting

for an acquisition strategy over in-house development, NVIDIA stands to rapidly attain the required chip design capabilities.

Furthermore, NVIDIA should continue pursuing its focus on quantum computing, an emerging technology with potential as a long-term substitute for AI chip functionalities. Quantum computers are able to process large amounts of data at once and solve problems much faster than regular computers. However, being still at a nascent stage, it is uncertain when enterprises will achieve significant breakthroughs in the field, adding complexity to the landscape. This technology's effective operation will also need pairing with powerful digital computers. Despite these challenges, NVIDIA has proactively entered this domain with the introduction of NVIDIA CUDA Quantum, a hybrid quantum-classical computing platform that facilitates the integration and programming of different AI chips in a unified system. Subsequently, the company launched DGX Quantum, the world's first GPU-accelerated quantum computing system. This system combines NVIDIA's accelerated computing platforms with a quantum control platform developed with one of its partners. NVIDIA has established connections with quantum hardware providers and forged partnerships with companies in this field. By following this strategy, NVIDIA will also be able to develop deeper relationships within its ecosystem.

Diversification: Although NVIDIA continues its R&D efforts, it is important to remain vigilant about emerging technologies developed by competitors. Innovations from tech start-ups could potentially render NVIDIA's products obsolete. NVIDIA should monitor the venture capital AI landscape and prepare corporate vehicles to quickly invest in tech start-ups and capitalize on technological breakthroughs. Crucial to this end is to continue investing in its NVIDIA Inception program, which is targeted at start-ups.

Figure 7: The Ansoff Matrix



Source (adapted): Ansoff, Harry Igor. 1957. "Strategies for Diversification." *Harvard Business Review* 35 (5): 113-124.

Manufacturing approach

NVIDIA follows a fabless business model, focusing on chip design within the intricate value chain of AI chips. However, this choice has rendered the company more vulnerable to supply chain conditions. Amidst heightened commercial tensions between the United States and China, coupled with the aftermath of the semiconductor chip shortage, NVIDIA is compelled to reassess its business model. Delays, increased product prices, and the inability to fulfill all customer orders demand proactive measures. Consequently, maintaining the current fabless business model does not seem a viable option, especially given the current political scene where governments worldwide are advocating for localized supply chains and offering incentive packages to do so. In light of these considerations, NVIDIA contemplates two options:

Bring manufacturing in-house: Opting for in-house manufacturing would reduce vulnerability to supply chain disruptions and diminish dependence on TSMC and its bargaining

power. However, the escalating complexity of the AI chip market necessitates specialization more than ever. Pursuing this approach would entail a significant and prolonged investment, resulting in a scenario where NVIDIA would have to allocate resources across various activities, prompting substantial restructuring within the company. Key questions arise: would this option be viable? Could NVIDIA match TSMC's efficiency? Is securing the necessary human capital feasible, or would a substantial investment in training be required?

Maintain a fabless business model but diversify suppliers: This strategy would enable NVIDIA to focus on research, product quality, and innovation while mitigating risks associated with concentrated manufacturing. This is in case manufacturing would diverge from Asia-Pacific regions such as China, Taiwan, and South Korea, reducing NVIDIA's dependence on these countries. Moreover, when choosing which supplier, it is important to rely on factors such as trustworthiness and efficiency. TSMC and Samsung pose as viable solutions since they are among the few foundry companies capable of producing advanced chips. However, the fact they are not American companies could still present challenges. One option would be to partner with them in United States territory, incentivizing them to open manufacturing plants in the United States and leveraging recent United States incentives aimed at localizing supply chains. Both TSMC and Samsung are currently building their manufacturing plants in the United States, and NVIDIA has already expressed interest in making use of this. Lastly, it would be advisable to negotiate exclusivity contracts with suppliers to ensure manufacturing capacity. This could involve acquiring specific lots or a percentage of manufacturing capacity, even though such exclusivity contracts would most likely come with a further price increase.

The same strategy should be followed to diversify equipment manufacturing, namely to reduce the dependence on ASML, the exclusive global supplier of high-performance lithographic machines tailored for advanced chips.

Group part

The current business model stands out as the most cost-effective option, with any deviation leading to increased prices and substantial upfront investments, as illustrated in Appendix 40 from the case study. The core debate centers around finding an equilibrium between cost-effectiveness and the security and stability provided by a localized supply chain. Given the global reliance on these chips by end customers, governments, and various entities, all of which entrust them with highly sensitive information, any disruption or cessation in their production would impose unprecedented constraints on the world. The semiconductor chip shortage has brought awareness to this critical dependence. Consequently, there is a belief that customers would be willing to pay a premium for assurance of the reliability and security of their product orders.

The optimal strategy appears to be maintaining NVIDIA's fabless business model while diversifying suppliers. This way, NVIDIA would maintain its focus on chip design and would leverage its manufacturer’s efficiency in the United States without incurring the substantial investment and restructuring required for in-house manufacturing. This is despite the increase in price due to the higher cost of American labor compared to that of Asia-Pacific countries since it is believed final customers will value the security and reliability of services over the price increase. Finally, this strategy aligns with industry demands for sustainability, as localized supply chains can reduce transportation emissions, contributing to more environmentally friendly practices. Considering all factors, the envisioned changes cannot be executed by NVIDIA alone. Collaborative efforts between governments and companies across all value chain stages will be essential to ensure the success of all options. This collaboration is vital to facilitating an efficient transition and guaranteeing the availability of material and human resources needed to implement such changes successfully. Governments play an important role in tackling challenges concerning infrastructure, workforce development and securing a nurturing environment where businesses can thrive.

Group part

To conclude, in order to overcome main challenges, NVIDIA should diversify revenue streams, broaden its customer base and reconsider its manufacturing approach. Furthermore, given the industry's dynamic landscape, NVIDIA must continue its commitments to R&D and in attracting and retaining a skilled workforce, especially as their shortage is forecasted to increase. This proactive strategy will allow for anticipation of market trends and sustained innovation and adaptability in the long run.

CONCLUDING REMARKS & KEY TAKEAWAYS

Demand for AI chips is increasing due to the transformative potential of AI in our daily lives. Nonetheless, the industry faces many challenges. The main challenges include a chip shortage as a result of limited manufacturing capacity, skilled labor shortages, government regulations, and political instability. To overcome these challenges, governments are incentivizing the localization of supply chains and increasing domestic manufacturing capacity.

Environmental sustainability concerns also impact the industry due to increased power consumption, as the development of advanced semiconductors like AI chips results in higher power consumption and increased pollution. Porter's Five Forces analysis reveals moderate to low bargaining power for buyers and suppliers, high threat of substitutes, and moderate competitive rivalry. The dynamics of competition emphasize innovation over price, with network effects influencing the market's structure and sustainability.

Shifting to NVIDIA's perspective, the company has demonstrated a historic ability to anticipate trends and adapt its business and resources, accordingly, leveraging its GPUs for superior computational power and, more recently, developing a robust ecosystem. This integration of hardware, software, community support, partnerships, and AI applications serves as a unique selling proposition, fostering customer loyalty and differentiation in the continually evolving technology space. NVIDIA's entrance into the AI scene marked a significant milestone, being

among the first to enter the AI chip market and currently holding a 70% market share. Achieving such results is attributed to the company's innovative approach, marked by continuous investment in research and development, coupled with strong relationships with the developer community, showcasing NVIDIA's commitment to staying ahead in the dynamic technology landscape. Additionally, NVIDIA benefits from skilled human capital and a strong brand reputation, further solidifying its competitive position in the market. However, NVIDIA is not immune to challenges. The fabless business model exposes the company to vulnerabilities in the value chain, and the overreliance on China raises concerns, especially in the face of potential new United States regulations affecting commercial relations. Additionally, the concentration of revenue from a few major customers poses a risk, and the fast-paced, highly competitive industry demands continual adaptation to sustain its competitive edge.

Looking ahead, NVIDIA stands to benefit from a strategic shift focused on diversifying revenue streams, expanding its customer base, and enhancing manufacturing capacity. Possible value-creation strategies include strengthening partnerships with the United States government, continuing efforts to attract more companies as customers to assist in developing their own LLMs, acquiring design capabilities to produce FPGAs and ASICs, venturing into quantum computing, and investing in technology start-ups pioneering cutting-edge technology. These initiatives not only provide growth opportunities but also foster deeper relationships within the company's ecosystem. To address manufacturing vulnerabilities, NVIDIA should adopt a balanced approach, maintaining its fabless model while supporting the establishment of United States-based manufacturing plants.

NVIDIA'S BET ON ARTIFICIAL INTELLIGENCE: THE IMPACT ON HEALTHCARE

THE AI HEALTHCARE MARKET

Size, Growth, and Segments of the Market

The highest levels of disruption due to AI can be found in industries such as healthcare, financial services, the automotive industry, and education (Statista 2023). The AI in healthcare market size was estimated at USD 11.0 billion in 2021 and is expected to reach USD 188.0 billion in 2030, increasing at a compound annual growth rate of 37 % from 2022 to 2030 (Appendix 56).

Based on components, the AI Healthcare market is divided into software, hardware, and services. Software holds the largest market share (40.5%) and is projected to experience the fastest growth with a CAGR of 40.8% from 2023 to 2031. Technologies includes ML, NLP, computer vision, robotics, and expert systems. ML and NLP account for more than 50% of total technologies. Applications include drug discovery and development, robot-assisted surgery, connected machines, fraud detection, virtual assistants, administrative workflow assistants, diagnosis, dosage error education and cybersecurity. Leading with a 24.6% share, drug discovery and development is expected to show a CAGR of 41.5% from 2023 to 2031 closely followed by robot-assisted surgery. End-users include healthcare providers, pharmaceutical and biotechnology companies, healthcare payers, and patients. The predominant segments are healthcare providers and pharmaceutical and biotechnology companies. Regions covered are North America, Asia Pacific, Europe, Latin America, and the Middle East and Africa. Leading in 2023, North America secured the largest revenue share at 57.7%. However, the Asia Pacific region is expected to have the fastest CAGR of 40.1% from 2024 to 2030 (Appendix 57).

Finally, the femtech market, closely linked with the AI healthcare sector, is projected to reach

$103 billion by 2030 (Boston Consulting Group 2023). This market addresses various women's health needs, including fertility, menopause, and general health conditions affecting women disproportionately or distinctly (Emma Kemble 2022).

Growth Drivers

The AI healthcare market is on a trajectory of remarkable growth, marked by exponential expansion. Key drivers fueling this surge include a shortage of medical staff coupled with workforce burnout, the imperative to cut care expenses, and the drive to enhance overall efficiency (Straits Research 2023) (Boyd 2023). Additionally, the upward trend is accentuated by a rising geriatric population, contributing to a surge in chronic diseases. This demographic shift underscores the urgency for advancements in diagnosing and understanding chronic conditions at their initial stages. Furthermore, the evolving lifestyles spurred by the Covid-19 pandemic have heightened expectations, fostering a demand for more convenient and personalized medical care (Grand View Research 2023). Examples of AI applications in healthcare showcase its potential to completely transform how services are provided:

Improvement in patient care and experience: The rise of AI personal assistants is leading to the creation of a pre-primary care market, by filtering out non-urgent cases, allowing doctors to focus on critical cases (Statista 2023). This process efficiently directs patients still requiring primary care to their relevant medical specialties. An example is Your.MD that uses AI to match patients' symptoms with public data for personalized advice. AI developments are facilitating home-based healthcare, with home testing and monitoring, telemedicine, e-prescriptions, and doctor-free health checks (Matthew Huddle 2023).

Improvement in hospital management: AI facilitates optimal resource utilization by automating tasks such as documentation, claims handling, patient onboarding, and scheduling. Companies like Doximity, Abridge, and DeepScribe are exploring AI applications for these

tasks (Matthew Huddle 2023). AI allows the analysis of historical data and real-time situations to forecast resource needs, enabling adjustments in staffing and resource levels. Automating low-value tasks cuts costs significantly (Deloitte 2019). This is crucial, considering that nearly a quarter of U.S. national health expenditure is allocated to administrative costs (Boyd 2023).

Improvement in technical support in the medical area: AI is crucial in advancing precision, efficiency, and speed in diagnosis, primarily through innovations in the analysis of medical imaging. Given that medical imaging constitutes almost 90% of healthcare data, AI's data processing capabilities result in exponential efficiency, surpassing that of human practitioners (NVIDIA 2023). Other advancements besides enhancing medical imaging include, for example, Beyond Verbal's software capable of examining vocal patterns to detect emotions. Preliminary health-related studies suggest potential connections between specific voice patterns and certain disease processes, paving the way for non-invasive markers in healthcare (Deloitte 2019).

Treatment and monitoring will experience enhancements through the utilization of AI medical devices, including connected and cognitive devices, such as portable, wearable, and implantable technologies. These devices monitor health information, including vital signs, engaging patients and caregivers and autonomously administering therapies (Carlton 2019). An example is AiCure, a company with software designed to verify if patients have taken their medication at the correct times, (Deloitte 2019). Robots have also demonstrated success in patient tratment, being employed in surgical operations, integrated into implants and prosthetics, and assisting physicians and healthcare staff in various tasks (Memarzadeh 2020).

Enhancing the drug development process*:* this process has historically been time- and resource-intensive. Many drugs take over a decade to reach the market, with an average success rate for drug candidates of just 10% (Benemann 2023). AI can streamline various stages, primarily in discovery and development and in clinical trials, accelerating the overall process

and enabling quicker transitions from discovery to market (Matthew Huddle 2023). Drug discovery aims to identify small molecules that selectively modulate the functions of target proteins. Generative AI excels in efficiently screening through extensive research results, establishing connections between data points, and significantly reducing the pool of candidate molecules (Deloitte 2019). Similar to models like ChatGPT generating text, generative AI, leveraging LLMs trained on biomolecular data, can produce blueprints for new molecules and proteins (Benemann 2023). Clinical trials are poised for AI-driven progress in protocol generation, where AI rapidly creates drafts of clinical protocols using inputs from published literature, prior trials, and various medical sources (Licholai 2023). Looking ahead, the use of generative AI in the drug development process has the potential to expedite the availability of therapeutic solutions, even for rare conditions that have historically faced challenges or economic barriers (Matthew Huddle 2023).

Improvements in the area of genomics: Genomic sequencing is the process of determining the genetic makeup of a specific organism or cell type. It is set to become widely available globally within the next decade, offering crucial insights for precision medicine (Memarzadeh 2020), by enabling early diagnostics and personalized treatments. Instead of relying on general population data, doctors will consider an individual's genetics, environment, medical history, and lifestyle to assess the risk of developing certain diseases and to consider medical options. Companies are still in the early stages of genomic research. For example, Flow Health is building a vast knowledge graph of medicine and genomics, using 30 million gigabytes of clinical data from 22 million veterans over 20 years to understand how each gene variant affects an individual's observable traits, such as height, eye color and blood type (Statista 2023).

Advancements in genomics are reshaping drug development. Through detailed patient profiling encompassing demographics, medical history, and genetics, AI aids clinical researchers in efficiently identifying and matching patients with trials. This precision approach accelerates

recruitment and enhances trial data quality. Furthermore, a groundbreaking application in clinical research is the development of digital twins for patients. AI creates virtual replicas of individual patients. These dynamic models predict treatment efficacy and simulate safety outcomes, allowing researchers to optimize strategies and minimize risks through virtual trials (Licholai 2023).

If these efforts succeed, specialized diagnostics and treatments will shift healthcare from disease management to prevention. Organizations will proactively monitor healthy individuals and conduct preventative interventions (Deloitte 2019). Moreover, the upcoming trend is toward establishing a personalized care ecosystem centered around patients, integrating medical and social caregivers (Appendix 58). This ecosystem will aim to enhance the patient experience and streamline their access to care (Carlton 2019).

Growth Constraints

Healthcare organizations are primarily concerned about the cost of technologies (36%), integrating AI into the organization (30%), and implementation issues, including AI risks and data issues (28%) (Kumar Chebrolu 2020). Integration of AI solutions is further hindered by the lack of standardized protocols and interoperability among different healthcare systems (Spherical Insights 2022). Furthermore, .as AI systems take on critical roles in patient care decisions, ethical concerns emerge, particularly related to potential biases in algorithms and accountability for errors. Ethical considerations are heightened in the domains of genomics and preventive care, where issues like genetic modification and discrimination surface, alongside the administration of drugs to healthy individuals based on future disease predictions. Rapid technological changes necessitate the evolution of existing legislative and regulatory frameworks. However, ethical concerns, coupled with a lack of trust in AI technologies due to limited familiarity, may strict legislation, hindering advancements in technology. This is particularly relevant concerning patient data privacy. The need for patient confidentiality and

strengthened cybersecurity has led to a complex regulatory environment surrounding healthcare data (Matthew Huddle 2023). Further complexities in this regulatory landscape may negatively impact achieving an interconnected healthcare ecosystem.

COMPETITION

In the AI healthcare market, competition mirrors that of the AI data center, detailed in the Teaching Note. The AI sector demands substantial investment in R&D and CAPEX for rapid innovation. Recognizing AI's technical nature, companies form strategic partnerships to leverage core competencies and engage in M&A activity. With the focus on innovation, companies seek to differentiate their products. Most of them have developed general software and hardware applicable to various sectors, including healthcare. Others, like IBM and Nvidia, have focused on tailored platforms, software, and services for healthcare. While IBM faced challenges with IBM Watson Health, leading to its eventual sale in 2022, Nvidia succeeded with its healthcare-specific offerings (Weiss 2022) (Benemann 2023). Lastly, various AI chips exist. Notably, AMD, through Xilinx, is actively using FPGAs to advance genomics research (AMD Xilinx 2018).

NVIDIA'S STRATEGIC ADVANCEMENTS IN HEALTHCARE

Nvidia implemented a strategic move that sets it apart from competitors with the creation of the NVIDIA Clara which is a comprehensive healthcare set of platforms, software, and services that addresses enterprises worldwide and is driving AI solutions across different healthcare areas. Clara includes BioNeMo for drug discovery, Holoscan for medical devices, Parabricks for genomics, and MONAI for medical imaging (Appendix 59). Clara has enabled healthcare breakthroughs, such as generating blueprints for two novel proteins with BioNemo, conducting a first-of-its-kind surgery with Holoscan, and deploying MONAI-powered solutions in

radiology departments. Clara already counts with over 100 partners (Benemann 2023) (Appendix 60).

Clara's integrated platforms are revolutionizing smart hospitals by facilitating seamless collaboration among various hospital departments and sensors and enabling swift responses to real-time data. Examples include patient monitoring with intelligent video analytics, advancements in medical imaging applications, support for digital and robotic surgery and integration of telemedicine. Clara optimizes healthcare resources, automates routine tasks for professionals, and enhances global patient-centric care (Niewolny 2022).

Accessible through various cloud providers, Clara creates a mutually beneficial relationship by leveraging the vast user base and extensive datasets of these providers. With thousands of companies globally relying on cloud services, they become potential customers for Clara, enhancing its efficiency through increased usage and a growing network of partners. For cloud providers, offering Clara is attractive, as it expands their cloud offerings, providing a compelling solution for healthcare-related companies. This is particularly relevant in the highly competitive landscape among cloud providers. However, potential risks arise as some cloud providers are developing their AI chips. If these providers were to create their healthcare-specific software, no longer requiring NVIDIA's services, it could pose a significant challenge for NVIDIA.

NVIDIA's Inception program currently endorses over 1800 healthcare start-ups (NVIDIA 2017).

CHALLENGES AND STRATEGIC RECOMMENDATIONS

In the healthcare sector, NVIDIA faces significant challenges, particularly in the intricate ethical, legal, and regulatory landscape, especially regarding future advancements in data privacy. This uncertainty adds complexity to how regulations may impact NVIDIA's

technological progress. Additionally, NVIDIA's competitive landscape includes two key considerations. Firstly, although the company has established itself as a GPU market leader with Clara specifically tailored for healthcare, competitors are bringing to the market more efficient AI chips. Secondly, potential challenges from cloud providers are on the horizon, as they embark on producing their AI chips and may further complicate matters if they decide to develop in-house healthcare software for their cloud services. However, this shift was prompted by the semiconductor chip shortage aftermath and NVIDIA is an industry leader known for its technological advancements and efficiency and has a long history of strategic partnerships with main cloud providers. We believe that as long as NVIDIA continues delivering cutting-edge innovations to the market, cloud providers will lack the incentive to internally produce similar software. Fundamental to prevent this from happening is to ensure that shortages caused by the chip shortage do not recur. Addressing vulnerabilities in the AI chip value chain is imperative, as elaborated in the Teaching Note.

Despite challenges, the application of AI in healthcare presents a considerable growth opportunity for NVIDIA. The company strategically addresses key opportunities through Clara. Drug discovery and development and robot-assisted surgery are the two applications with main growth potential and NVIDIA is actively tackling them through BioNemo and Holoscan. Diagnosis is another pivotal application being addressed with MONAI and Parabricks is addressing the development of genomics which will revolutionize advancements in diagnosis and treatment, leading to the establishment of preventive medicine. In a broader context, Clara is contributing to the establishment of smart hospitals, enhancing overall efficiency and improving patient experience.

To sustain growth and maintain its leadership position, NVIDIA should continue robust R&D efforts, establishing strategic partnerships, and engaging in M&A activities, leveraging initiatives like NVIDIA Inception and NVIDIA Deep Learning Institute. These initiatives

demonstrate its proactive approach to staying ahead of technological advancements, engaging with start-ups for early involvement and ensuring ongoing development and training for its human capital.

NVIDIA should proactively expand its network of partners and augment its capabilities in genomics and cybersecurity, enhancing the range and efficiency of its ecosystem. In cultivating its network of partners, NVIDIA should target specific segments for maximum impact. Firstly, it should engage with drug developers, particularly those from the femtech market, expected to have exponential growth. Additionally, NVIDIA should target drug developers willing to address less common diseases, now feasible with AI technologies accelerating the process and reducing the need for scale. This way, NVIDIA could support a niche-market that would face low competition as there are not many customers for these drugs, the space in the market would be filled by the first companies to join. Lastly, the emphasis should be on fostering collaborations with hospitals in the US, the largest geographical market, laying the groundwork for a global ecosystem among healthcare providers, pharmaceutical and biotechnology companies, healthcare payers and patients as hospitals are the place where patient data is concentrated, serving as a hub center to other participants. Even though China will represent the biggest CAGR, political tensions between US-China advise against increasing commercial relations with China.

Strategies to attract these entities should address their primary challenges, assist in AI integration and tackle implementation issues. NVIDIA should focus on showcasing AI's productivity enhancements for potential customers and aim to build trust in these technologies. Demystifying AI is crucial due to high skepticism and will be crucial for the generalization of these technologies. Customer service and marketing campaigns are crucial for achieving this. Awareness activities should be extended to the public in general. There is a need for a shift in societal attitudes, requiring collaboration with government bodies and educational institutions.

This is important as legislation and regulation reflect societies' mentalities and the population in general is reluctant to change.

Genomics should be the central focus for NVIDIA, representing a crucial avenue to revolutionize diagnosis and drug discovery, and achieve preventive healthcare. NVIDIA should acquire competencies for designing FPGAs and ASICs. This move aligns with industry trends where competitors are already leveraging these technologies. As technology advances, the escalating need for efficiency is evident, prompting NVIDIA to be prepared to deploy more efficient chips than GPUs if necessary. Regarding cybersecurity, the evolving landscape of interconnected systems requires a concerted focus on ensuring data privacy. This emphasis becomes integral not only for regulatory compliance but also to establish NVIDIA as a reliable partner in collaborative ventures.

In conclusion, NVIDIA should aspire to create a fully interconnected healthcare ecosystem, through Clara. Securing a first-mover advantage is crucial to consolidate all data from healthcare providers, pharmaceutical and biotechnology companies, healthcare payers and patients into a centralized accessible point. This consolidation will facilitate the implementation of preventive and individualized treatment plans based on comprehensive data. Pharmacies will oversee drug therapy adherence and personalize patient communications, while clinical-trial project teams will optimize country selection and site activation. Additionally, pharmaceutical companies will optimize drug development needs. This intricate ecosystem will yield a significant lock-in effect that will be difficult to replicate for competitors. The more participants within NVIDIA's healthcare ecosystem, the more appealing it becomes for others to join.

REFERENCES

AI Startups. 2023. *Top 10 startups developing AI hardware.* December 15. https://www.ai-startups.org/top/hardware/.

AI4D. 2023. *Artificial Intelligence For Everyone.* https://www.ai4diversity.org/.

Allyn, Bobby. 2022. *Deepfake video of Zelenskyy could be 'tip of the iceberg' in info war, experts warn.* March 16. https://www.npr.org/2022/03/16/1087062648/deepfake-video-zelenskyy-experts-war-manipulation-ukraine-russia.

Alsop, Thomas. 2023. *Artificial intelligence (AI) chip market revenue from 2022 to 2027.* November 22. https://www.statista.com/statistics/1283358/artificial-intelligence-chip-market-size/.

—. 2023. *Integrated circuit (IC) market revenue worldwide worldwide from 2009 to 2024.* October 17. https://www.statista.com/statistics/519456/forecast-of-worldwide-semiconductor-sales-of-integrated-circuits/.

—. 2023. *Semiconductor market revenue worldwide from 1987 to 2024.* October 17. https://www.statista.com/statistics/266973/global-semiconductor-sales-since-1988/.

Amazon Web Services. n.d. *Amazon Lex.* https://aws.amazon.com/lex/.

AMD Xilinx. 2018. *Mindray High End Medical Imaging Solutions: Powered by Xilinx.* May 21. https://www.xilinx.com/video/corporate/mindray-high-end-medical-imaging-solutions.html.

Ansoff, Harry Igor. 1957. "Strategies for Diversification." *Harvard Business Review* 35 (5): 113-124.

Antonio Varas, Raj Varadarajan, Jimmy Goorich, and Falan Yinug. 2021. *Strengthening The Global Semiconductor Supply Chain in an Uncertain Era.* BCG X SIA.

https://www.semiconductors.org/wp-content/uploads/2021/05/BCG-x-SIA-Strengthening-the-Global-Semiconductor-Value-Chain-April-2021_1.pdf.

Anyoha, Rockwell. 2017. *The History of Artificial Intelligence.* August 28. https://sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence/.

Archibald, Liz. 2023. *How NVIDIA Fuels the AI Revolution With Investments in Game Changers and Market Makers.* December 11. https://blogs.nvidia.com/blog/nvidia-investments/.

Arcuri, Gabrielle Athanasia and Gregory. 2022. *Russia's Invasion of Ukraine Impacts Gas Markets Critical to Chip Production.* March 14. https://www.csis.org/blogs/perspectives-innovation/russias-invasion-ukraine-impacts-gas-markets-critical-chip-production.

Awati, Rahul. 2021. *Integrated Circuit (IC).* September. https://www.techtarget.com/whatis/definition/integrated-circuit-IC.

Barber, Jeff. 2023. *Power Constraints and AI Workloads: The Hidden Challenges of High-Density Data Centers.* June 8. https://www.linkedin.com/pulse/power-constraints-ai-workloads-hidden-challenges-data-jeff-barber/.

Belton, Pádraig. 2021. *The Computer Chip Industry has a Dirty Climate Secret.* September 18. https://www.theguardian.com/environment/2021/sep/18/semiconductor-silicon-chips-carbon-footprint-climate.

Benemann, Kathy. 2023. *100+ Partners Bring NVIDIA Clara AI Healthcare Platform to Enterprises Worldwide.* March 21. https://blogs.nvidia.com/blog/nvidia-clara-ai-healthcare-enterprises/.

Blanchard, Olivier. 2022. *Could Intel, Nvidia and Qualcomm's Radically Different Automotive Strategies Create Opportunities for OEMs But Pain Points For Consumers?* January 18. https://creativestrategies.com/could-intel-nvidia-and-qualcomms-radically-different-automotive-strategies-create-opportunities-for-oems-but-pain-points-for-consumers/.

Boston Consulting Group. 2023. "The Future of Digital Health."

Boyd, Eric. 2023. *Microsoft and Epic expand AI collaboration to accelerate generative AI's impact in healthcare, addressing the industry's most pressing needs.* August 22. https://blogs.microsoft.com/blog/2023/08/22/microsoft-and-epic-expand-ai-collaboration-to-accelerate-generative-ais-impact-in-healthcare-addressing-the-industrys-most-pressing-needs/.

Business Endo. 2023. *16 Nvidia Statistics You Need to See: Revenue, Employee, Net income and More.* https://businessendo.com/NVIDIA-statistics/.

Byford, Lyndal. 2019. *Quantum computer does in 200 seconds what a supercomputer takes 10,000 years to do.* October 24. https://www.scimex.org/newsfeed/quantum-computer-does-in-200-seconds-what-a-supercomputer-takes-10,000-years-to-do#:~:text=Researchers%20at%20Google%20have%20created,outperforms%20the%20fastest%20classical%20supercomputer.

Callaham, John. 2023. *Microsoft is reportedly working on a "Steam of mobile" app store for iOS and Android.* October 13. https://www.neowin.net/news/microsoft-is-reportedly-working-on-a-steam-of-mobile-app-store-for-ios-and-android/.

Carlton, Shubham Singhal and Stephanie. 2019. *The era of exponential improvement in healthcare?* May 14. https://www.mckinsey.com/industries/healthcare/our-insights/the-era-of-exponential-improvement-in-healthcare#/.

Carr, Austin, and Ian King. 2023. *Nvidia's AI chips power CHATGPT-and multibillion-dollar surge.* June 15. https://www.bloomberg.com/news/features/2023-06-15/nvidia-s-ai-chips-power-chatgpt-and-multibillion-dollar-surge.

Chui, Michael, Bryce Hall, Alex Singla, Helen Mayhew, and Alex Sukharevsky. 2022. "McKinsey & Company." *The State of AI in 2022—and a Half Decade in Review.* December 6. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2022-and-a-half-decade-in-review.

Clark, Adam. 2023. *Qualcomm, Mobileye, Other AI Chip Stocks That Can Thrive. It's Not All About the Cloud.* June 7. https://www.barrons.com/articles/qualcomm-mobileye-nvidia-ai-chip-stocks-497a0ab7.

Clifford, Tyler. 2020. *Nvidia completes 'homerun deal' after closing $7 billion acquisition of Mellanox.* April 27. https://www.cnbc.com/2020/04/27/nvidia-ceo-calls-mellanox-acquisition-a-homerun-deal.html#:~:text=Nvidia%20completes%20'homerun%20deal'%20after%20closing%20%247%20billion%20acquisition%20of%20Mellanox,-Published%20Mon%2C%20Apr&text=%E2%80%9CThis%20is%20a%.

Commencis. 2023. *The Key Benefits and Challenges of Moving to the Cloud.* March 15. https://www.commencis.com/thoughts/the-key-benefits-and-challenges-of-moving-to-the-cloud/.

CompaniesMarketCap. n.d. *Largest Companies by Market Cap.* https://companiesmarketcap.com/.

Costa, Alvin Da. 2020. *NVIDIA Partner Program Expands to 1,500 Members, Adds New Benefits.* https://blogs.nvidia.com/blog/npn-expands-1500-members-new-benefits/.

Crumpler, William. 2020. *The Problem of Bias in Facial Recognition.* May 1. https://www.csis.org/blogs/strategic-technologies-blog/problem-bias-facial-recognition.

Darling , Coran, Danny Tobey, Bennett Borden , and Ashley Carr. 2023. *G7 publishes guiding principles and code of conduct for artificial intelligence.* https://www.dlapiper.com/en/insights/publications/ai-outlook/2023/g7-publishes-guiding-principles-and-code-of-conduct-for-artificial-intelligence.

DataGuard. 2023. *30 most important privacy statistics and facts for 2023.* January 18. https://www.dataguard.co.uk/blog/30-most-important-privacy-statistics-and-facts-for-2023.

Davies, Kasia. 2023. *What are the biggest obstacles when it comes to implementing AI?* June 19. https://www.statista.com/statistics/1393469/obstacles-implementing-ai-germany/.

Delgado, Camilo. 2023. *What is AI Accelerated Ray Tracing?* May 25. https://www.pcguide.com/gpu/what-is-ai-accelerated-ray-tracing/.

Deloitte. 2019. "The future of artificial intelligence in health care: How AI will impact patients, clinicians, and the pharmaceutical industry."

—. 2021. *Upping the ante: Venture capital investment in chip companies reaches new highs.* December 1. https://www2.deloitte.com/us/en/insights/industry/technology/technology-media-and-telecom-predictions/2022/semiconductor-investors-venture-capital.html.

Drapkin, Aaron. 2023. *Data Breaches That Have Happened in 2022 and 2023 So Far.* December 12. https://tech.co/news/data-breaches-updated-list.

Emma Kemble, Lucy Pérez, Valentina Sartori, Gila Tolub, and Alice Zheng. 2022. *The dawn of the FemTech revolution.* February 14. https://www.mckinsey.com/industries/healthcare/our-insights/the-dawn-of-the-femtech-revolution.

European Parliament. 2023. *EU AI Act: First Regulation on Artificial Intelligence.* June 14. https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence.

Freund, Karl. 2023. *Intel acquires Habana Labs for $2 Billion.* October 5. https://www.forbes.com/sites/moorinsights/2019/12/16/intel-acquires-habana-labs-for-2b/.

—. 2023. *Microsoft announces Maia AI, ARM CPU, AMD MI300, & New Nvidia for azure.* November 23. https://www.forbes.com/sites/karlfreund/2023/11/16/microsoft-announces-maia-ai-arm-cpu-amd-mi300--new-nvidia-for-azure/.

Fung, Joey. 2023. *Amazon's AI leap with Inferentia and Trainium Chips.* August 16. https://themilsource.com/2023/08/16/amazons-ai-leap-with-inferentia-and-trainium-chips/.

Glassdor. 2023. *Salary Details for an Electronics Hardware Engineer at NVIDIA.* October 29. https://www.glassdoor.com/Salary/NVIDIA-Electronics-Hardware-Engineer-Salaries-E7633_D_KO7,36.htm.

Grand View Research. 2023. *AI In Healthcare Market Size, Share & Trends Analysis Report By Component (Software Solutions, Hardware, Services), By Application (Virtual Assistants, Connected Machines), By Region, And Segment Forecasts, 2024 - 2030.* https://www.grandviewresearch.com/industry-analysis/artificial-intelligence-ai-healthcare-market.

Gupte, Sandeep. 2023. *Redefining Workstations: NVIDIA, Intel Unlock Full Potential of Creativity and Productivity for Professionals.* February 15. Accessed December 2023. https://blogs.nvidia.com/blog/intel-rtx-ada-workstation/.

Hao, Karen. 2019. *Training a single AI model can emit as much carbon as five cars in their lifetimes.* June 6. https://www.technologyreview.com/2019/06/06/239031/training-a-single-ai-model-can-emit-as-much-carbon-as-five-cars-in-their-lifetimes/.

Harding, Tom. 2020. *Osborne Clarke.* September 30. https://www.osborneclarke.com/insights/geforce-now-cloud-based-gaming-challenges-licensing-distribution-models.

Hernandez, Daniela. 2014. *IBM unveils a 'brain-like' chip with 4,000 processor cores.* August 7. https://www.wired.com/2014/08/ibm-unveils-a-brain-like-chip-with-4000-processor-cores/.

Hilson, Gary. 2023. *AI Must Be Secured at the Silicon Level.* January 23. https://www.eetimes.com/ai-must-be-secured-at-the-silicon-level/?_gl=1*e2th8x*_ga*NTQ5MTg2ODUxLjE3MDEwOTY2NDc.*_ga_ZLV02RYCZ8*MTcwMTIwOTc3NS4zLjAuMTcwMTIwOTc3NS42MC4wLjA.&_ga=2.190412395.1915378285.1701096652-549186851.1701096647.

Himanshu, J, and K Vineet. 2023. *Artificial Intelligence Chip Market Research, 2032.* Allied Market Research, Allied Market Research.

IEEE. 2017. *Ethically aligned design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems.* Vol. 2. Toronto: IEEE Canada International Humanitarian Technology Conference (IHTC). https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf.

Infotechlead. 2020. *Nvidia revenue surges 50% due to Mellanox acquisition.* August 21. https://infotechlead.com/networking/NVIDIA-revenue-surges-50-due-to-mellanox-acquisition-62470.

Insights by GreyB. 2023. *NVIDIA Corporation Patents - Key Insights and StatsI.* https://insights.greyb.com/nvidia-corporation-patents/.

Intel. n.d. *AI & Machine Learning Ecosystem Developer Resources.* https://www.intel.com/content/www/us/en/developer/ecosystem/overview.html.

—. 2021. *intc.* December 25. https://www.intc.com/filings-reports/all-sec-filings/content/0000050863-22-000007/intc-20211225.htm.

Jotrin. 2022. *A Brief History of the Development of AI Chips.* January 4. https://www.jotrin.com/technology/details/a-brief-history-of-the-development-of-ai-chips.

Kanungo, Alokya. 2023. *The Green Dilemma: Can AI Fulfil Its Potential Without Harming the Environment?* July 18. https://earth.org/the-green-dilemma-can-ai-fulfil-its-potential-without-harming-the-environment/#:~:text=AI's%20Carbon%20Footprint&text=As%20datasets%20and%20models%20become,gas%20emissions%2C%20aggravating%20climate%20change.

Kemp, Chris. 2021. "Legal Aspects of Artificial Intelligence." *Kemp IT Law* 3.

Khan, Bilal. 2022. *What are Workstations? What are their features and uses?* September 7. https://errorbook.com/knowledgebase/what-are-workstations/.

Kharpal, Arjun. 2019. *Alibaba unveils its first A.I. chip as China pushes for its own semiconductor technology.* September 25. https://www.cnbc.com/2019/09/25/alibaba-unveils-its-first-ai-chip-called-the-hanguang-800.html.

—. 2019. *Huawei launches A.I. chip as it looks to defy us pressure, pitting it against giants like Qualcomm and Nvidia.* August 23. https://www.cnbc.com/2019/08/23/huawei-launches-ai-chip-ascend-910-pitting-it-against-nvidia-qualcomm.html.

—. 2023. *Two of the world's most critical chip firms rally after Nvidia's 26% share price surge.* May 25. https://www.cnbc.com/2023/05/25/tsmc-asml-two-critical-chip-firms-rally-after-nvidias-earnings.html.

Kim, W. Chan, and Renée Mauborgne. 2004. *Blue Ocean Strategy.* October 1. https://hbr.org/2004/10/blue-ocean-strategy.

King, Ian. 2022. *Wait Times for Chips Grow Again in March as Shortages Drag On.* April 5. https://www.bloomberg.com/news/articles/2022-04-05/wait-times-for-chips-grow-again-in-march-as-shortages-drag-on.

Kumar Chebrolu, Dan Ressler, and Hemnabh Varia. 2020. *Smart use of artificial intelligence in health care: seizing opportunities in patient care and business activities.* Deloitte.

LaBerge, Laura, Clayton O'Toole, Jeremy Schneider, and Kate Smaje. 2020. "McKinsey & Company." *How COVID-19 has pushed companies over the technology tipping point—and transformed business forever.* October 5. https://www.mckinsey.com/capabilities/strategy-and-corporate-finance/our-insights/how-covid-19-has-pushed-companies-over-the-technology-tipping-point-and-transformed-business-forever.

Lard, Jeremy. 2023. *Nvidia sells half a million AI chips and bags $14.5 billion in just three months.* November 28. https://www.pcgamer.com/nvidia-sells-half-a-million-ai-chips-in-just-three-months-netting-dollar10-billion-plus/#:~:text=With%20that%20in%20mind%2C%20the,according%20to%20research%20outfit%20Omdia.

Lardinois, Frederic. 2020. *AWS launches Trainium, its New Custom ML training chip.* December 2. https://techcrunch.com/2020/12/01/aws-launches-trainium-its-new-custom-ml-training-chip/.

Lee. 2022. *AI chip startup SiMa.ai launches auto business with former Bosch, Mercedes executive.* November 3. https://www.reuters.com/technology/ai-chip-startup-simaai-launches-auto-business-with-former-bosch-mercedes-2022-11-03/.

Lee, Steve Holland and Jane. 2022. *TSMC triples Arizona chip plant investment, Biden hails project.* December 7. https://www.reuters.com/technology/biden-visit-taiwans-tsmc-chip-plant-arizona-hail-supply-chain-fixes-2022-12-06/.

Leswing, Kif. 2023. *Google reveals its newest A.I. supercomputer, says it beats Nvidia.* April 5. https://www.cnbc.com/2023/04/05/google-reveals-its-newest-ai-supercomputer-claims-it-beats-nvidia-.html.

Licholai, Greg. 2023. "AI In Clinical Research: Now And Beyond." *Forbes.*

Lindgardt, Zhenya, Martin Reeves, George Stalk, and Michael S. Deimler. 2009. *Business Model Innovation: When the Game Gets Tough, Change the Game.* December 1. https://web-assets.bcg.com/img-src/BCG_Business_Model_Innovation_Dec_09_tcm9-121706.pdf.

Loeffler, John. 2023. *Nvidia's AI dominance will be the death knell for its GeForce graphics cards.* August 8. https://www.techradar.com/computing/gpu/nvidias-ai-dominance-will-be-the-death-knell-for-its-geforce-graphics-cards.

Mann, Tobias. 2022. *Intel challenges NVIDIA, AMD with trio of workstation GPUs.* August 9. Accessed December 2023. https://www.theregister.com/2022/08/09/intel_arc_pro/.

Mansfield, Edwin, Mark Schwartz, and Samuel Wagner. 1981. "Imitation Costs and Patents: An Empirical Study." *The Economic Journal* 91 (364): 907-18. https://www.jstor.org/stable/2232499.

Maslej, Nestor, Loredana Fattorini, Katrina Ligett, and John Ecthemendy . 2023. *Artificial Intelligence Index Report.* Stanford University, Stanford: AI Index Steering Committee.

Matthew Huddle, Josh Kellar, Krishna Srikumar, Krishna Deepak, and Daniel Martines. 2023. *Generative AI Will Transform Health Care Sooner Than You Think.* Boston Consulting Group.

Maximize Market Research. 2023. *Artificial Intelligence Chip Market: Global Industry Analysis and Forecast (2023-2029) Trends, Statistics, Dynamics, Segment Analysis.* Maximize Market Research. https://www.maximizemarketresearch.com/market-report/artificial-intelligence-chip-market/185676/.

—. 2023. *Artificial Intelligence in Healthcare Market Size, Growth, Opportunities & Trends: Global Industry Analysis and Forecast (2023-2029).* November. https://www.maximizemarketresearch.com/market-report/global-artificial-intelligence-ai-healthcare-market/21261/.

McHugh, JIm. 2016. *NVIDIA CEO Delivers World's First AI Supercomputer in a Box to OpenAI.* https://outlookseries.com/A0975/Infrastructure/3872_NVIDIA_CEO_AI_Supercomputer_Box_OpenAI.htm.

McKendrick, Joe. 2023. *Yes, AI Increases Productivity, Study Suggests.* April 25. https://www.forbes.com/sites/joemckendrick/2023/04/25/yes-ai-increases-productivity-study-suggests/?sh=163bcbbb12c2.

McKinsey & Company. 2023. *Quantum technology sees record investments, progress on talent gap.* April 24. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/quantum-technology-sees-record-investments-progress-on-talent-gap.

—. 2023b. *What is quantum computing?* May 1. https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-quantum-computing.

Memarzadeh, Adam Bohr and Kaveh. 2020. *The rise of artificial intelligence in healthcare applications.* June 26. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7325854/.

Michael Justin Allen Sexton, Yannick Guerrini & Pierre Dandumont. 2018. *The History of NVIDIA GPUs: NV1 to Turing.* August 25. https://www.tomshardware.com/picturestory/715-history-of-nvidia-gpus.html.

Mikulich, Stephanie. 2023. *Nvidia is leading the AI chip market, but rivals are coming.* October 23. https://finance.yahoo.com/video/why-nvidias-hold-ai-chip-152406471.html.

Mordor Intelligence. 2023. *Industry 4.0 Market Size & Share Analysis - Growth Trends & Forecasts (2023 - 2028).* https://www.mordorintelligence.com/industry-reports/industry-4-0-market.

Moriggi, Andrea. 2018. *The Role of Intellectual Property in the Intelligence Explosion.* January 22. https://www.4ipcouncil.com/research/role-intellectual-property-intelligence-explosion.

Napolitano, Eliyabeth. 2023. *AI eliminated nearly 4,000 jobs in May, report says.* June 2. https://www.cbsnews.com/news/ai-job-losses-artificial-intelligence-challenger-report/.

Neuro, Benzinga. 2023. *NVIDIA CEO Acknowledges Huawei As A Potential Competitor In AI Chipmaking.* December 6. https://www.alsahm.com/news/content/nvidia-ceo-acknowledges-huawei-as-a-potential-competitor-in-ai-chipmaking-2023-12-06.

Newman, Daniel. 2020. *NVIDIA Q2 Delivers Record Revenue: Datacenter Business Up 167%.* August 20. https://convergetechmedia.com/nvidia-q2-delivers-record-revenue-datacenter-business-up-167/.

Ng, Gideon. 2023. *Unveiling the Hidden Environmental Impacts of AI: Strategies to Mitigate this Growing Threat.* June 9. https://kr-asia.com/unveiling-the-hidden-environmental-impacts-of-ai-strategies-to-mitigate-this-growing-threat.

Niewolny, David. 2022. *What is a Smart Hospital?* November 22. https://blogs.nvidia.com/blog/what-is-a-smart-hospital/.

Novet, Jordan. 2023. *Nvidia's revenue triples as AI chip boom continues.* November 21. https://www.cnbc.com/2023/11/21/nvidia-nvda-q3-earnings-report-2024.html.

Nunes, Ashley. 2021. *Automation Doesn't Just Create or Destroy Jobs — It Transforms Them.* November 04. https://hbr.org/2021/11/automation-doesnt-just-create-or-destroy-jobs-it-transforms-them.

NVIDIA. 2023. https://www.nvidia.com/en-us/about-nvidia/careers/life-at-nvidia/.

NVIDIA. 2006. *2006 NVIDIA Annual Report.* California: NVIDIA.

NVIDIA. 2015. *2015 NVIDIA's Anuual Report.* NVIDIA. https://s201.q4cdn.com/141608511/files/doc_financials/annual/2015/2015_NVDA_Annual_Report.pdf.

NVIDIA. 2016. *2016 Nvidia Corporation Anual Review Notice of Annual Meeting Proxy Statement From 10-K.* Santa Clara, California: Nvidia.

—. n.d. *A Timeline of Innovation.* https://www.nvidia.com/en-us/about-nvidia/corporate-timeline/.

—. 2023. *Customer Support.* https://www.nvidia.com/en-us/support/consumer/.

—. 2023. *Enterprise Overview.* December 16. https://docs.omniverse.nvidia.com/enterprise/latest/index.html.

—. 2023. *GeForce NOW.* https://www.nvidia.com/en-eu/geforce-now/.

NVIDIA. 2022. *Human Rights Policy.* California: Legal Department. https://www.nvidia.com/content/dam/en-zz/Solutions/about-us/documents/HumanRightsPolicy.pdf.

—. 2023. *Inception.* https://www.nvidia.com/en-us/startups/.

—. 2023. *Medical Imaging AI Made Easier: NVIDIA Offers MONAI as Hosted Cloud Service.* November 26. https://blogs.nvidia.com/blog/monai-cloud-apis-rsna/.

—. 2021. *NVIDIA A100 Tensor Core GPU.* https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/a100/pdf/nvidia-a100-datasheet-nvidia-us-2188504-web.pdf.

—. 2023. *NVIDIA Announces Financial Results for Third Quarter Fiscal 2024.* https://nvidianews.nvidia.com/news/nvidia-announces-financial-results-for-third-quarter-fiscal-2024#:~:text=NVIDIA%20(NASDAQ%3A%20NVDA)%20today,50%25%20from%20the%20previous%20quarter.

—. 2023. *NVIDIA Announces New System for Accelerated Quantum-Classical Computing.* https://nvidianews.nvidia.com/news/nvidia-announces-new-system-for-accelerated-quantum-classical-computing.

—. n.d. *NVIDIA Annual Reports and Proxies.* https://investor.nvidia.com/financial-info/annual-reports-and-proxies/default.aspx.

—. 2021. *NVIDIA Corporation Annual Review 2021.* https://s201.q4cdn.com/141608511/files/doc_downloads/2021/04/2021-Annual-Review.pdf.

—. 2022. *NVIDIA Corporation Annual Review 2022.* https://s201.q4cdn.com/141608511/files/doc_financials/2022/ar/2022-Annual-Review.pdf.

—. 2023. *NVIDIA Corporation Annual Review 2023.* https://s201.q4cdn.com/141608511/files/doc_financials/2023/ar/2023-Annual-Report-1.pdf.

—. 2023. *NVIDIA Corporation: History.* Accessed 2023. https://www.NVIDIA.com/en-eu/about-NVIDIA/corporate-timeline/.

—. 2023. *NVIDIA Corporation-Financial Info-Quarterly Results.* https://investor.nvidia.com/financial-info/quarterly-results/default.aspx.

—. 2023. *NVIDIA CUDA Quantum.* https://developer.nvidia.com/cuda-quantum.

—. n.d. *NVIDIA DRIVE Partner Ecosystem.* https://www.nvidia.com/en-eu/self-driving-cars/partners/.

—. 2023. *NVIDIA H100 Tensor Core GPU.* https://www.nvidia.com/de-de/data-center/h100/.

—. 2023. *NVIDIA H200 Tensor Core GPU.* https://www.nvidia.com/en-us/data-center/h200/.

—. 2023. *NVIDIA History.* https://www.NVIDIA.com/en-eu/about-NVIDIA/corporate-timeline/.

—. 2023. *NVIDIA Partner Network.* https://www.nvidia.com/en-us/about-nvidia/partners/.

—. 2023. *NVIDIA RTX Technology.* https://www.nvidia.com/en-us/design-visualization/technologies/rtx/.

—. 2019. *Nvidia to Acquire Mellanox for $6.9 Billion.* March 11. https://nvidianews.nvidia.com/news/nvidia-to-acquire-mellanox-for-6-9-billion.

—. 2023. *Platform Overview.* December 16. https://docs.omniverse.nvidia.com/platform/latest/overview.html.

—. 2017. *Power Breakthroughs in Healthcare and Life Sciences.* December 21. https://www.nvidia.com/en-us/industries/healthcare-life-sciences/.

O'Regan, Alaina. 2023. *EnCharge AI reimagines computing to meet needs of cutting-edge AI.* January 27. https://engineering.princeton.edu/news/2023/01/27/encharge-ai-reimagines-computing-meet-needs-cutting-edge-ai.

Osterwalder, Alexander, and Yves Pigneur. 2010. *Business model generation: a handbook for visionaries,game changers, and challengers.* Vol. 1. Rio de Janeiro: John Wiley & Sons.

Oxford Economics. 2023. *Chipping Away: Assessing and Addressing the Labor Market Gap Facing the U.S. Semiconductor Industry.* Exford Economics, Semiconductor Industry Association. https://www.semiconductors.org/chipping-away-assessing-and-addressing-the-labor-market-gap-facing-the-u-s-semiconductor-industry/.

Ozorio, Stacy. 2023. *Next-Level Computing: NVIDIA and AMD Deliver Powerful Workstations to Accelerate AI, Rendering and Simulation.* October 19. https://blogs.nvidia.com/blog/ai-workstations/.

Porter, Michael E. 1985. *The Competitive Advantage: Creating and Sustaining Superior Performance.* New York: Free Press.

Precedence Research. 2023. *Artificial Intelligence (AI) Chip Market.* Precedence Research. https://www.precedenceresearch.com/artificial-intelligence-chip-market.

—. 2023. *Artificial Intelligence in Healthcare Market Size, Report 2022-2030.* February. https://www.precedenceresearch.com/artificial-intelligence-in-healthcare-market.

PwC. 2021. *Global Top 100 Companies by Market Capitalization.* May. https://www.pwc.com/gx/en/audit-services/publications/assets/pwc-global-top-100-companies-2021.pdf.

Rahul, Rao. 2023. *How researchers use Nvidia's gpus to simulate qubits.* October 3. https://spectrum.ieee.org/nvidia-qubit.

Ramiro Palma, Raj Varadarajan, Jimmy Goodrich, Thomas Lopez, and Aniket Patil. 2022. *The Growing Challenge of Semiconductor Design Leadership.* BCG X SIA. https://www.semiconductors.org/wp-content/uploads/2022/11/2022_The-Growing-Challenge-of-Semiconductor-Design-Leadership_FINAL.pdf.

Reddit. 2023. *GPU Marketshare in Q2 2023: NVIDIA 87%, AMD 10% and Intel 3%.* https://www.reddit.com/r/NVIDIA/comments/165tw5b/gpu_marketshare_in_q2_2023_NVIDIA_87_amd_10_and/.

Reichow, Greg. 2023. *Nvidia's dominance in AI chips detering investment in rival start-ups.* December 2. https://www.euronews.com/next/2023/09/11/nvidias-dominance-in-ai-chips-is-detering-investment-in-would-be-rival-semiconductor-start.

Reuters. 2023. *Ai Startup Sambanova launches new chip designed for higher quality AI.* September 19. https://www.reuters.com/technology/ai-startup-sambanova-launches-new-chip-designed-higher-quality-ai-2023-09-19/.

—. 2023. *Exclusive: CHATGPT-owner OpenAI is exploring making its own AI chips.* October 6. https://www.reuters.com/technology/chatgpt-owner-openai-is-exploring-making-its-own-ai-chips-sources-2023-10-06/.

—. 2023. *Intel gives details on future AI chips as it shifts strategy | reuters.* May 22. https://www.reuters.com/technology/intel-gives-details-future-ai-chips-it-shifts-strategy-2023-05-22/.

—. 2023. *Nvidia results show its growing lead in AI chip race.* February 2. https://www.reuters.com/technology/nvidia-results-show-its-growing-lead-ai-chip-race-2023-02-23/#:~:text=The%20key%20to%20the%20company%27s,OpenAI%27s%20wildly%20popular%20ChatGPT%20chatbot.

—. 2023. *Samsung to manufacture chips from AI chip startup Tenstorrent.* October 2. https://www.reuters.com/technology/samsung-manufacture-chips-ai-chip-startup-tenstorrent-2023-10-02/.

Robins, Mark. 2023. *Learning Training and Inference.* November 5. https://www.linkedin.com/pulse/difference-between-deep-learning-training-inference-mark-robins-mdq8c/.

Rocio, Lorenzo, Nicole Voigt, Miki Tsusaka, Matt Krentz, and Katie Abouzahr. 2018. *How Diverse Leadership Teams Boost Innovation.* January 23. https://www.bcg.com/publications/2018/how-diverse-leadership-teams-boost-innovation.

Ross, Jenna. 2023. *Nvidia vs. AMD vs. Intel: Comparing AI Chip Sales.* August 25. https://www.visualcapitalist.com/NVIDIA-vs-amd-vs-intel-comparing-ai-chip-sales/.

Rtology, Eva. 2022. *New AI method allows real-time rendering.* October 31. Accessed December 2023. https://medium.com/mlearning-ai/new-ai-method-allows-real-time-rendering-785ba85fd53e.

Saha, Sudip. 2023. *Visual Computing Market Outlook (2023 to 2033).* July. https://www.futuremarketinsights.com/reports/visual-computing-market.

Salian, Isha. 2023. *Nvidia wins neurips awards for research on Generative AI, generalist AI agents.* April 20. https://blogs.nvidia.com/blog/nvidia-neurips-research/#:~:text=Two%20NVIDIA%20Research%20papers%20%E2%80%94,of%20AI%20and%20machine%20learning.

Scheer, Steven, and Ari Rabinovitch. 2019. *Intel buys Israeli AI startup Habana Labs for $2 billion.* December 16. https://www.reuters.com/article/us-habana-labs-m-a-intel/intel-buys-israeli-ai-startup-habana-labs-for-2-billion-idUSKBN1YK1BU/#:~:text=JERUSALEM%20(Reuters)%20%2D%20Intel%20Corp,bolster%20its%20data%2Dcenter%20business.

Schreiner, Maximilian. 2023. *Nvidia's H100 GPU sells like hot cakes with high profit margins.* Augst 18. https://the-decoder.com/nvidias-h100-gpu-sells-like-hot-cakes-with-high-profit-margins/.

Scientific. 2023. *The true cost of AI innovation.* https://www.scientific-computing.com/analysis-opinion/true-cost-ai-innovation#:~:text=The%20growth%20and%20proliferation%20of,%C2%A375.4%20billion)%20in%202023.

Semiconductor Industry Association. 2021. "2021 State of The U.S. Semiconductor Industry." https://www.semiconductors.org/wp-content/uploads/2021/09/2021-SIA-State-of-the-Industry-Report.pdf.

Semiconductor Industry Association. 2022. "2022 State of the U.S. Semicondcutor Industry." https://www.semiconductors.org/wp-content/uploads/2022/11/SIA_State-of-Industry-Report_Nov-2022.pdf.

Semiconductor Industry Association. 2023. "2023 State of The US Semiconductor Industry." https://www.semiconductors.org/wp-content/uploads/2023/07/SIA_State-of-Industry-Report_2023_Final_072723.pdf.

—. 2021. "Notice of Request for Public Comments on Risks in the Semiconductor Supply Chain." November 8.

Shah, Agam. 2022. *Why Nvidia sees a future in software and services: Recurring revenue.* March 10. https://www.theregister.com/2022/03/10/nvidia_software_services/#:~:text=Nvidia%20is%20repositioning%20itself%20as%20a%20software%20and,extract%20revenue%20from%20those%20using%20its%20graphics%20processors.

Shane, Neagle. 2023. *AMD sinks despite solid results, AI commentary: What lies ahead?* August 3. https://www.investing.com/analysis/amd-fails-to-rally-despite-solid-results-ai-commentary-what-lies-ahead-200640611.

Sheikh, Aamir. 2023. *Can Huawei's AI gpus truly rival Nvidia's A100 in performance?* August 27. https://www.cryptopolitan.com/huaweis-ai-gpus-rival-nvidias-a100/.

Shu, Catherine, and Rita Liao. 2023. *All the Nvidia news annouced by Jensen Huang at Computex.* May 29. https://techcrunch.com/2023/05/28/NVIDIA-computex-jensen-huang/?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAAG7WGvAOKDKn8eAOXGfp2SCF3KobuABsIls4oWkQBB3_4Pl2DuedlPQmJbvI136UH3UXW88gvYCM9bmsRF_twZqy9Cb6JF5r14g4oH60xfRM0e.

Singh, Nidhi. 2022. *The Timeline of Artificial Intelligence – From the 1940s.* November 7. https://verloop.io/blog/the-timeline-of-artificial-intelligence-from-the-1940s/.

Smolaks, Max. 2022. *AMD completes record ~$50 billion acquisition of Xilinx.* February 14. https://www.datacenterknowledge.com/business/amd-completes-record-50-billion-acquisition-xilinx.

SoftmaxAI. 2022. *Edge AI vs Cloud AI.* https://softmaxai.com/edge-ai-vs-cloud-ai/.

Spherical Insights. 2022. *Global Artificial Intelligence in Healthcare Market Insights Forecasts to 2032.* https://www.sphericalinsights.com/reports/artificial-intelligence-in-healthcare-market.

Spy Newsletter. 2023. *Home Competitive Analysis for Top Companies NVIDIA Competitive Analysis 2023 – Business Analysis.* October 22. https://spynewsletter.com/company/nvidia/#Which_companies_are_major_competitors_for_Nvidia_in_the_professional_visualization_market.

Statista. 2022. "Artificial Intelligence (AI) in healthcare."

Statista. 2023. "Artificial Intelligence: in-depth market analysis."

—. 2023. *Major countries in silicon production worldwide in 2022(in 1,000 metric tons).* January. https://www-statista-com.eu1.proxy.openathens.net/statistics/268108/world-silicon-production-by-country/.

—. 2023. *Nvidia revenue worldwide from 2017 to 2023, by specialized market.* February. https://www-statista-com.eu1.proxy.openathens.net/statistics/988040/nvidia-revenue-by-specialized-market/.

—. 2020. *Semiconductor Equipment Revenue Worldwide from 2017 to 2019, by Supplier (in Billion U.S. Dollars).* https://www-statista-

com.eu1.proxy.openathens.net/statistics/532224/worldwide-semiconductor-wafer-level-manufacturing-equipment-vendor-revenue/.

—. 2023. *Semiconductor foundries revenue share worldwide from 2019 to 2023, by quarter.* September 5. https://www.statista.com/statistics/867223/worldwide-semiconductor-foundries-by-market-share/.

StocksBNB. 2023. *Nvidia Corporation - AI is the future.* March 29. https://www.stocksbnb.com/reports/nvidia-corporation-ai-is-the-future/#:~:text=The%20market%20leadership%20is%20attributed,list%2C%20demonstrating%20their%20energy%20efficiency.

Stone, Louis. 2023. *Cerebras unveils successor to the world's largest processor: Ai Business.* July 31. https://aibusiness.com/verticals/cerebras-unveils-successor-to-the-world-s-largest-processor.

Straits Research. 2023. *Artificial Intelligence (AI) in Healthcare Market.* https://straitsresearch.com/report/artificial-intelligence-in-healthcare-market.

Strubell, Emma, Ananya Ganesh, and Andrew McCallum. 2019. "Energy and Policy Considerations for Deep Learning in NLP." *College of Information and Computer Sciences* (University of Massachusetts Amherst).

Sverdlik, Yevgeniy. 2018. *Why nvidia gifted Elon Musk's AI non-profit its latest supercomputer.* November 14. https://www.datacenterknowledge.com/archives/2016/08/18/why-nvidia-gave-musks-ai-non-profit-openai-a-supercomputer.

Tarasov, Katie. 2022. *ASML is the only company making the $200 million machines needed to print every advanced microchip. Here's an inside look.* March 23.

https://www.cnbc.com/2022/03/23/inside-asml-the-company-advanced-chipmakers-use-for-euv-lithography.html.

—. 2023. *How Arm is gaining chip dominance with its architecture in Apple, Nvidia, AMD, Amazon, Qualcomm and more.* November 9. https://www.cnbc.com/2023/11/09/how-arm-gained-chip-dominance-with-apple-nvidia-amazon-and-qualcomm.html.

—. 2023. *Nvidia CEO Jensen Huang's big bet on A.I. is paying off as his core technology powers ChatGPT.* https://www.cnbc.com/2023/03/07/nvidia-grew-from-gaming-to-ai-giant-and-now-powering-chatgpt.html.

Taylor, Petroc. 2022. *Size of the big data analytics market worldwide from 2021 to 2029.* October 7. https://www.statista.com/statistics/1336002/big-data-analytics-market-size/.

technavio. 2023. *Gaming Graphic Processing Unit (GPU) Market by End-user, Type and Geography - Forecast and Analysis 2023-2027.* June 01. https://www.technavio.com/report/gaming-gpu-market-analysis.

TechPowerUp. 2023. *Intel discontinues brand new MAX 1350 Data Center GPU, successor targets alternative markets.* April 11. https://www.techpowerup.com/307085/intel-discontinues-brand-new-max-1350-data-center-gpu-successor-targets-alternative-markets.

The Motley Fool. 2018. *A Look at NVIDIA's Quadro Business.* October 8. https://www.nasdaq.com/articles/look-nvidias-quadro-business-2018-10-08.

The Physics arXiv Blog. 2022. *AI Machines Have Beaten Moore's Law Over The Last Decade, Say Computer Scientists.* February 21. https://www.discovermagazine.com/technology/ai-machines-have-beaten-moores-law-over-the-last-decade-say-computer.

Thormundsson, Bergur. 2022. *Do you agree or disagree with the statement: "AI will displace more jobs than it creates in the long run"?* March 17. https://www.statista.com/statistics/1024145/influence-of-ai-on-jobs-according-to-ceos-in-luxembourg/.

Tilley, Aaron. 2016. *Went From Powering Video Games To Revolutionizing Artificial Intelligence.* November 30. https://www.forbes.com/sites/aarontilley/2016/11/30/nvidia-deep-learning-ai-intel/?sh=4225be857ff1.

Toews, Rob. 2023. *The Geopolitics Of AI Chips Will Define The Future Of AI.* May 7. https://www.forbes.com/sites/robtoews/2023/05/07/the-geopolitics-of-ai-chips-will-define-the-future-of-ai/?sh=499c853b5c5c.

Toh, Michelle. 2023. *Huawei wants to go all in on AI for the next decade | CNN business.* September 21. https://edition.cnn.com/2023/09/21/tech/huawei-ai-strategy-us-china-intl-hnk/index.html.

Tracxn. 2023. *Sambanova Systems - company profile.* December 19. https://tracxn.com/d/companies/sambanova-systems/__TvzSIBPeKIVjGmWovhh4uYJ9eAl6FPYF0Iwfg8H-Yjg.

Tyson, Jeff, V. Tracy Wilson, and Talon Homer. 2017. *How Graphics Cards Work.* July 27. https://computer.howstuffworks.com/graphics-card.htm.

Value, Montreal. 2023. *Nvidia's Dominance In Data Centers And Potential For AI Operating Systems.* June 20. https://seekingalpha.com/article/4612514-nvidias-dominance-in-data-centers-and-potential-for-ai-operating-systems.

Vayuvegula, Ravi. 2018. *Understanding the AI Ecosystem.* July 31. https://medium.com/@ravivayuvegula/understanding-the-ai-ecosystem-6de271b1467.

VDA. 2023. *Artificial Intelligence Act.* VDA German Association of the Automotive Industry, Berlin: German Association of the Automotive Industry. https://www.safetywissen.com/object/A11/A11.8os7388238rhp3hfn9955915bjpl6i63834276715/safetywissen?prev=%2Fnews%2FSAFETYNEWS%2F.

Verge, Jan Jason. 2015. *Amazon buys stealthy Israeli chip startup Annapurna Labs.* January 23. https://www.datacenterknowledge.com/archives/2015/01/23/amazon-buys-stealthy-israeli-chip-startup-annapurna-labs#close-modal.

VideoCardz. 2014. *NVIDIA GeForce 256 DDR Graphics Card.* July 12. https://videocardz.net/nvidia-geforce-256-ddr.

Vlastelica, Ryan. 2023. *Nvidia stock surge powered by AI results in $1 trillion market valuation (NVDA).* May 30. https://www.bloomberg.com/news/articles/2023-05-30/nvidia-surge-results-in-historic-1-trillion-market-valuation.

Wang, Brian. 2023. *IBM Research's new prototype AI chip with 14 times energy efficiency.* September 26. https://www.nextbigfuture.com/2023/09/ibm-researchs-new-prototype-ai-chip-with-14-times-energy-efficiency.html#:~:text=September%2025%2C%202023%20by%20Brian,efficient%20speech%20recognition%20and%20transcription.

Wassen Mohammad, Adel Elomri and Laoucine Kerbache. 2022. "The Global Semiconductor Chip Shortage: Causes, Implications, and Potential Remedies." *IFAC- Papers Online* 476-483.

Weiss, Todd R. 2022. *IBM Watson Health Finallt Sold by IBM After 11 Months of Rumors.* January 21. https://www.hpcwire.com/2022/01/21/ibm-watson-health-finally-sold-by-ibm-after-11-months-of-rumors/#:~:text=IBM's%20interest%20in%20selling%20Watson,cloud%20computing%20and%20other%20markets.

White, Monica. 2023. *Nvidia's outrageous pricing strategy is exactly why we need AMD and Intel.* March 2022. https://www.digitaltrends.com/computing/nvidias-pricing-strategy-is-why-we-need-amd-and-intel/.

Wikiwand. 2023. *Quadro.* https://www.wikiwand.com/en/Quadro.

Young, Katie. 2020. *How organizations can transform their workforce into an AI powerhouse.* October 1. https://blogs.nvidia.com/blog/deep-learning-dli-training-organization/#:~:text=The%20DLI%20has%20trained%20more,and%20interaction%20with%20the%20instructors.

ZDNET. 2023. *Ai Chip Startup cerebras nabs $250 Million series F round at over $4 billion valuation.* December 19. https://www.zdnet.com/article/ai-chip-startups-cerebras-nabs-250-million-series-f-round-at-over-4-billion-valuation/.

Zhu, Feng, and Marco Iansiti. 2019. *Why Some Platforms Thrive and Others Don't.* January. https://hbr.org/2019/01/why-some-platforms-thrive-and-others-dont.

APPENDICES

Appendix 1

The History of AI

Source (adapted):

Jotrin. 2022. A Brief History of the Development of AI Chips. January 04. https://www.jotrin.com/technology/details/a-brief-history-of-the-development-of-ai-chips.

Anyoha, Rockwell. 2017. The History of Artificial Intelligence. August 28. https://sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence/.

Singh, Nidhi. 2022. The Timeline of Artificial Intelligence – From the 1940s. November 7. https://verloop.io/blog/the-timeline-of-artificial-intelligence-from-the-1940s/.

"Ai Chip Startup Cerebras Nabs $250 Million Series F Round at over $4 Billion Valuation." ZDNET. Accessed December 19, 2023. https://www.zdnet.com/article/ai-chip-startups-cerebras-nabs-250-million-series-f-round-at-over-4-billion-valuation/.

Ai Startup Sambanova launches new chip designed for higher quality AI, September 19, 2023. https://www.reuters.com/technology/ai-startup-sambanova-launches-new-chip-designed-higher-quality-ai-2023-09-19/.

Byford, Lyndal. "Quantum Computer Does in 200 Seconds What a Supercomputer Takes 10,000 Years to Do." Scimex, October 24, 2019. https://www.scimex.org/newsfeed/quantum-computer-does-in-200-seconds-what-a-supercomputer-takes-10,000-years-to-do#:~:text=Researchers%20at%20Google%20have%20created,outperforms%20the%20fastest%20classical%20supercomputer.

Carr, Austin, and Ian King. "Nvidia's AI Chips Power CHATGPT-and Multibillion-Dollar Surge." Bloomberg.com, June 15, 2023. https://www.bloomberg.com/news/features/2023-06-15/nvidia-s-ai-chips-power-chatgpt-and-multibillion-dollar-surge.

"EU AI Act: First Regulation on Artificial Intelligence: News: European Parliament." EU AI Act: first regulation on artificial intelligence | News | European Parliament, December 19, 2023. https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence.%C2%A0.

Exclusive: CHATGPT-owner OpenAI is exploring making its own AI chips, October 6, 2023. https://www.reuters.com/technology/chatgpt-owner-openai-is-exploring-making-its-own-ai-chips-sources-2023-10-06/.

Freund, Karl. "Intel Acquires Habana Labs for $2 Billion." Forbes, October 5, 2023. https://www.forbes.com/sites/moorinsights/2019/12/16/intel-acquires-habana-labs-for-2b/.

Freund, Karl. "Microsoft Announces Maia AI, ARM CPU, AMD MI300, & New Nvidia for Azure." Forbes, November 23, 2023. https://www.forbes.com/sites/karlfreund/2023/11/16/microsoft-announces-maia-ai-arm-cpu-amd-mi300--new-nvidia-for-azure/.

Fung, Joey. "Amazon's AI Leap with Inferentia and Trainium Chips." TheMilSource (TMS), August 16, 2023. https://themilsource.com/2023/08/16/amazons-ai-leap-with-inferentia-and-trainium-chips/.

Hernandez, Daniela. "IBM Unveils a 'brain-like' Chip with 4,000 Processor Cores." Wired, August 7, 2014. https://www.wired.com/2014/08/ibm-unveils-a-brain-like-chip-with-4000-processor-cores/.

"Intel Discontinues Brand New MAX 1350 Data Center GPU, Successor Targets Alternative Markets." TechPowerUp, April 11, 2023. https://www.techpowerup.com/307085/intel-discontinues-brand-new-max-1350-data-center-gpu-successor-targets-alternative-markets.

Intel gives details on future AI chips as it shifts strategy, May 22, 2023. https://www.reuters.com/technology/intel-gives-details-future-ai-chips-it-shifts-strategy-2023-05-22/.

Kharpal, Arjun. "Alibaba Unveils Its First A.I. Chip as China Pushes for Its Own Semiconductor Technology." CNBC, September 25, 2019. https://www.cnbc.com/2019/09/25/alibaba-unveils-its-first-ai-chip-called-the-hanguang-800.html.

Kharpal, Arjun. "Huawei Launches A.I. Chip as It Looks to Defy Us Pressure, Pitting It against Giants like Qualcomm and Nvidia." CNBC, August 23, 2019. https://www.cnbc.com/2019/08/23/huawei-launches-ai-chip-ascend-910-pitting-it-against-nvidia-qualcomm.html.

Laird, Jeremy. "Nvidia Sells Half a Million AI Chips and Bags $14.5 Billion in Just Three Months." pcgamer, November 28, 2023. https://www.pcgamer.com/nvidia-sells-half-a-million-ai-chips-in-just-three-months-netting-dollar10-billion-plus/#:~:text=With%20that%20in%20mind%2C%20the,according%20to%20research%20outfit%20Omdia.

Lardinois, Frederic. "AWS Launches Trainium, Its New Custom ML Training Chip." TechCrunch, December 2, 2020. https://techcrunch.com/2020/12/01/aws-launches-trainium-its-new-custom-ml-training-chip/.

"Nvidia Announces Financial Results for Third Quarter Fiscal 2024." NVIDIA Corporation - NVIDIA Announces Financial Results for Third Quarter Fiscal 2024. Accessed December 19, 2023. https://investor.nvidia.com/news/press-release-

details/2023/NVIDIA-Announces-Financial-Results-for-Third-Quarter-Fiscal-2024/default.aspx#:~:text=%2A%20Announced%20record,faster%20than%20the%20previous%20record.

"Nvidia Corporation - AI Is the Future." StocksBNB, March 29, 2023. https://www.stocksbnb.com/reports/nvidia-corporation-ai-is-the-future/#:~:text=The%20market%20leadership%20is%20attributed,list%2C%20demonstrating%20their%20energy%20efficiency.

Nvidia results show its growing lead in AI chip race | Reuters. Accessed December 19, 2023. https://www.reuters.com/technology/nvidia-results-show-its-growing-lead-ai-chip-race-2023-02-23/.

Salian, Isha. "Nvidia Wins Neurips Awards for Research on Generative AI, Generalist AI Agents." NVIDIA Blog, April 20, 2023. https://blogs.nvidia.com/blog/nvidia-neurips-research/#:~:text=Two%20NVIDIA%20Research%20papers%20%E2%80%94,of%20AI%20and%20machine%20learning.

"Sambanova Systems - Company Profile." Tracxn. Accessed December 19, 2023. https://tracxn.com/d/companies/sambanova-systems/__TvzSIBPeKIVjGmWovhh4uYJ9eAl6FPYF0Iwfg8H-YJg.

Samsung to manufacture chips from AI chip startup Tenstorrent, October 2, 2023. https://www.reuters.com/technology/samsung-manufacture-chips-ai-chip-startup-tenstorrent-2023-10-02/.

Sheikh, Aamir. "Can Huawei's AI Gpus Truly Rival Nvidia's A100 in Performance?" Cryptopolitan, August 27, 2023. https://www.cryptopolitan.com/huaweis-ai-gpus-rival-nvidias-a100/.

Smolaks, Max. "AMD Completes Record ~$50 Billion Acquisition of Xilinx." Data Center Knowledge | News and analysis for the data center industry, February 14, 2022. https://www.datacenterknowledge.com/business/amd-completes-record-50-billion-acquisition-xilinx.

Smolaks, Max. "AMD Completes Record ~$50 Billion Acquisition of Xilinx." Data Center Knowledge | News and analysis for the data center industry, February 14, 2022. https://www.datacenterknowledge.com/business/amd-completes-record-50-billion-acquisition-xilinx.

Stone, Louis. "Cerebras Unveils Successor to the World's Largest Processor: Ai Business." Cerebras unveils successor to the world's largest processor | AI Business, July 31, 2023. https://aibusiness.com/verticals/cerebras-unveils-successor-to-the-world-s-largest-processor.

Sverdlik, Yevgeniy. "Why Nvidia Gifted Elon Musk's AI Non-Profit Its Latest Supercomputer." Data Center Knowledge | News and analysis for the data center industry, November 14, 2018. https://www.datacenterknowledge.com/archives/2016/08/18/why-nvidia-gave-musks-ai-non-profit-openai-a-supercomputer.

Toh, Michelle. "Huawei Wants to Go All in on AI for the next Decade | CNN Business." CNN, September 21, 2023. https://edition.cnn.com/2023/09/21/tech/huawei-ai-strategy-us-china-intl-hnk/index.html.

Verge, Jan Jason. "Amazon Buys Stealthy Israeli Chip Startup Annapurna Labs." Data Center Knowledge | News and analysis for the data center industry, January 23, 2015. https://www.datacenterknowledge.com/archives/2015/01/23/amazon-buys-stealthy-israeli-chip-startup-annapurna-labs#close-modal.

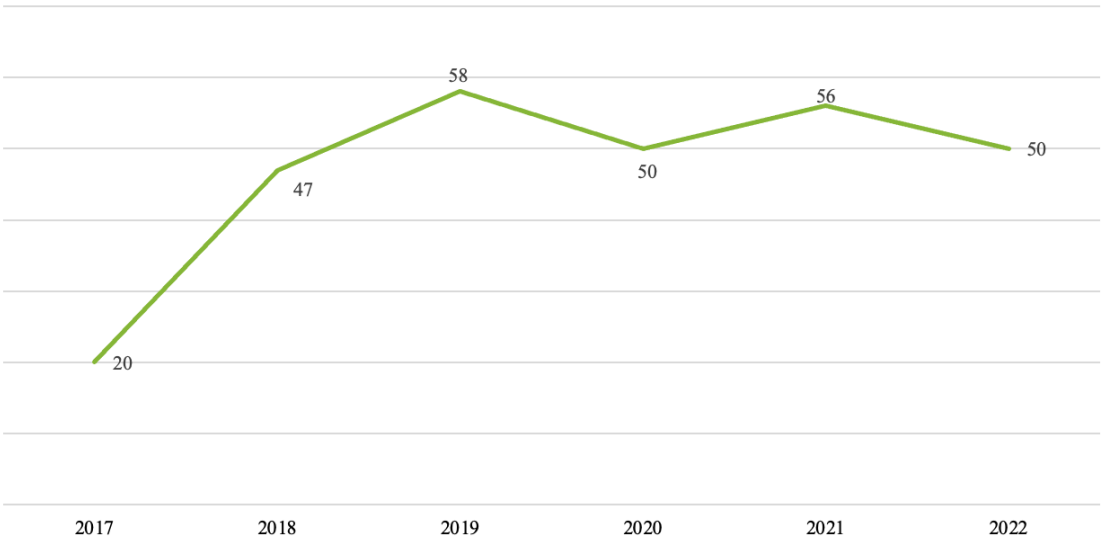
Vlastelica, Ryan. "Nvidia Stock Surge Powered by AI Results in $1 Trillion Market Valuation (NVDA)." Bloomberg.com, May 30, 2023. https://www.bloomberg.com/news/articles/2023-05-30/nvidia-surge-results-in-historic-1-trillion-market-valuation.

Wang, Brian. "IBM Research's New Prototype AI Chip with 14 Times Energy Efficiency." NextBigFuture.com, September 26, 2023. https://www.nextbigfuture.com/2023/09/ibm-researchs-new-prototype-ai-chip-with-14-times-energy-efficiency.html#:~:text=September%2025%2C%202023%20by%20Brian,efficient%20speech%20recognition%20and%20transcription.

Young, Katie. "How Organizations Can Transform Their Workforce into an AI Powerhouse." NVIDIA Blog, October 1, 2020. https://blogs.nvidia.com/blog/deep-learning-dli-training-organization/#:~:text=The%20DLI%20has%20trained%20more,and%20interaction%20with%20the%20instructors.

Appendix 2

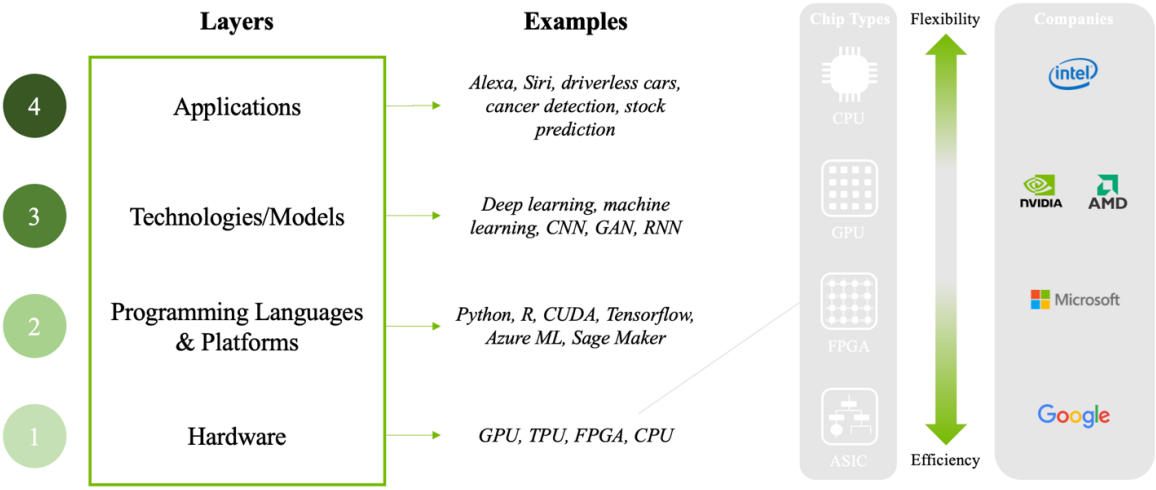
Global Adoption Rate of AI in Businesses (in %)



Source (adapted): Chui, Michael, Bryce Hall, Alex Singla, Helen Mayhew, and Alex Sukharevsky. 2022. "McKinsey & Company." The State of AI in 2022—and a Half Decade in Review. December 6. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2022-and-a-half-decade-in-review.

Appendix 3

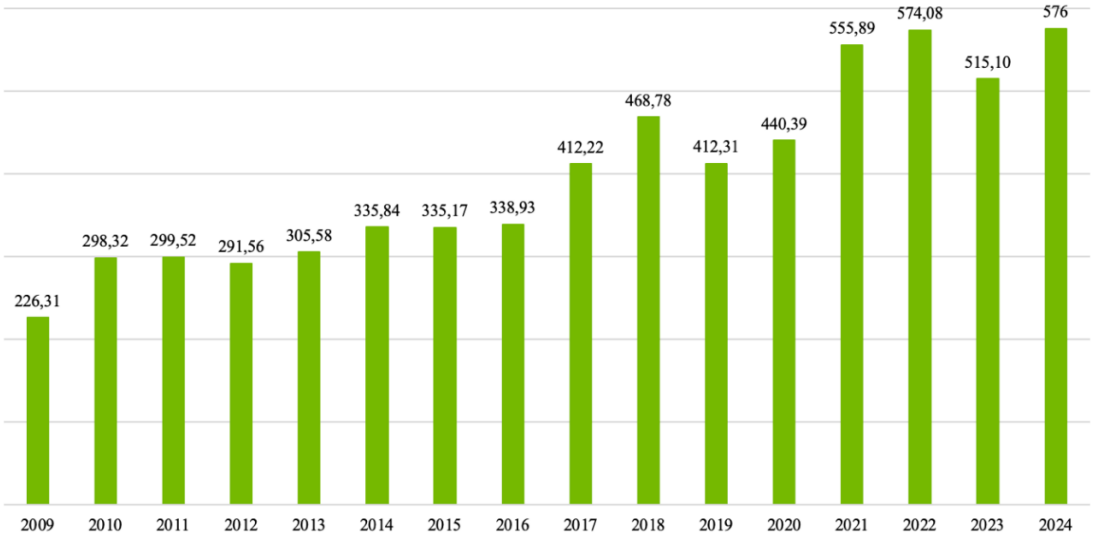
AI Ecosystem



Source (adapted): Vayuvegula, Ravi. 2018. Understanding the AI Ecosystem. July 31. https://medium.com/@ravivayuvegula/understanding-the-ai-ecosystem-6de271b1467.

Appendix 4

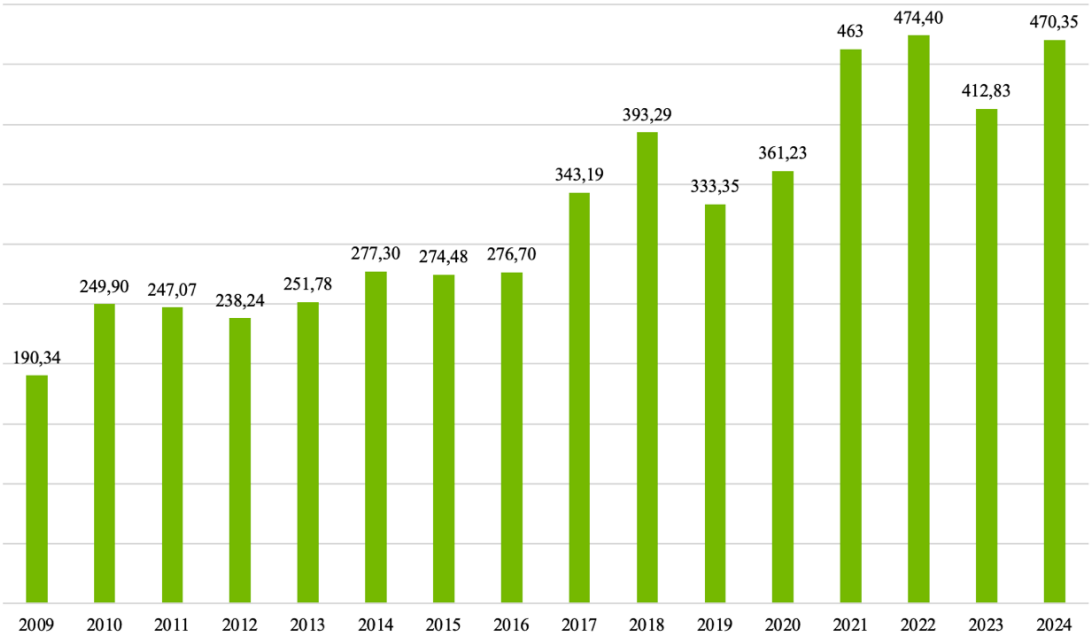
Global Semiconductor Market Revenue from 2009 to 2024 (in billion U.S. dollars)



Source (adapted): Alsop, Thomas. 2023. Semiconductor market revenue worldwide from 1987 to 2024. October 17. https://www.statista.com/statistics/266973/global-semiconductor-sales-since-1988/.

Appendix 5

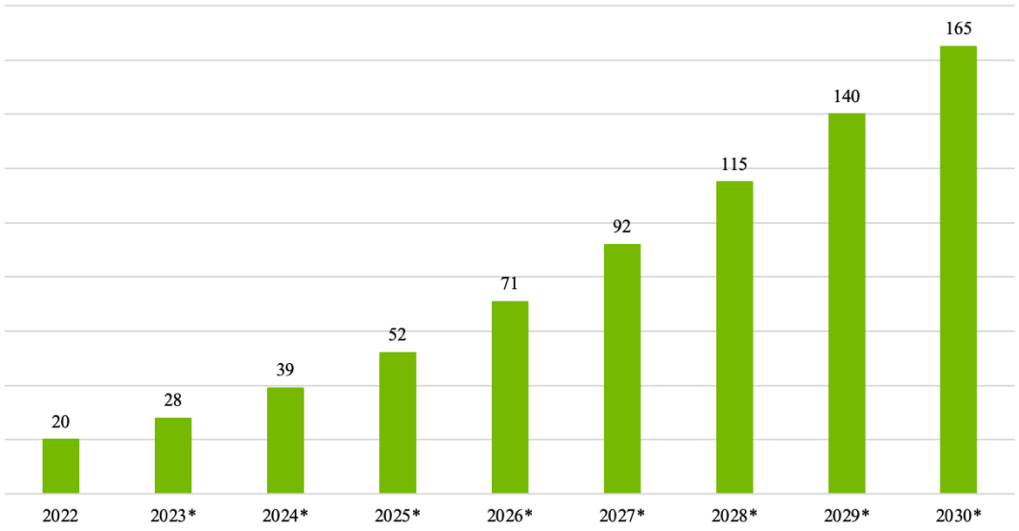
Global Integrated Circuit (IC) Market Revenue from 2009 to 2024 (in billion U.S. dollars)



Source (adapted): Alsop, Thomas. 2023. Integrated circuit (IC) market revenue worldwide worldwide from 2009 to 2024. October 17. https://www.statista.com/statistics/519456/forecast-of-worldwide-semiconductor-sales-of-integrated-circuits/.

Appendix 6

AI Chip Market Revenue from 2022 to 2030 (in Billion U.S. Dollars)



Source (adapted): Alsop, Thomas. 2023. Artificial intelligence (AI) chip market revenue from 2022 to 2027. November 22. https://www.statista.com/statistics/1283358/artificial-intelligence-chip-market-size/.

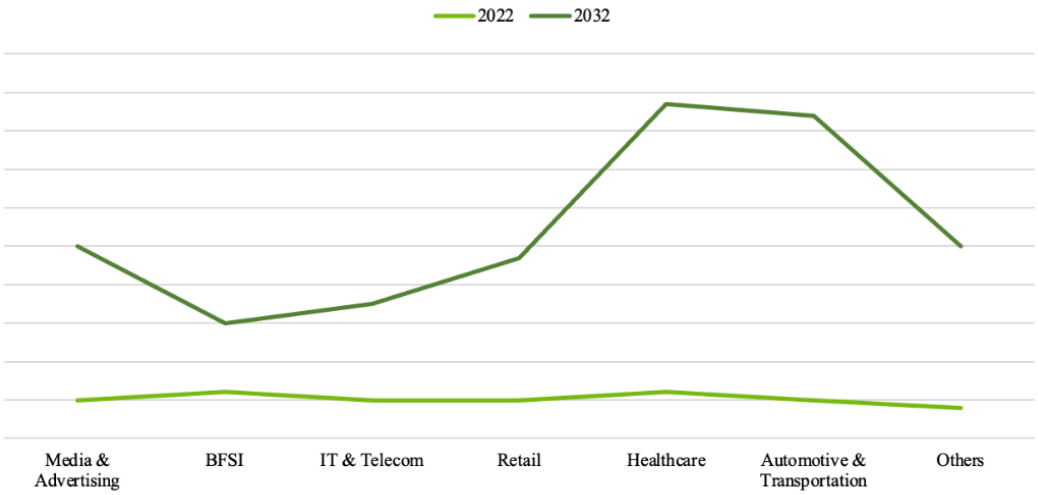
Appendix 7

AI Chip Market, by Region, Application and Chip Type 2022 (in %)

Source (adapted): Maximize Market Research. 2023. Artificial Intelligence Chip Market: Global Industry Analysis and Forecast (2023-2029) Trends, Statistics, Dynamics, Segment Analysis. https://www.maximizemarketresearch.com/market-report/artificial-intelligence-chip-market/185676/.

Appendix 8

AI Chip Market, by Industry Vertical (in %)



Source (adapted): Himanshu, J, and K Vineet. 2023. Artificial Intelligence Chip Market Research, 2032. Allied Market Research, Allied Market Research.

Appendix 9

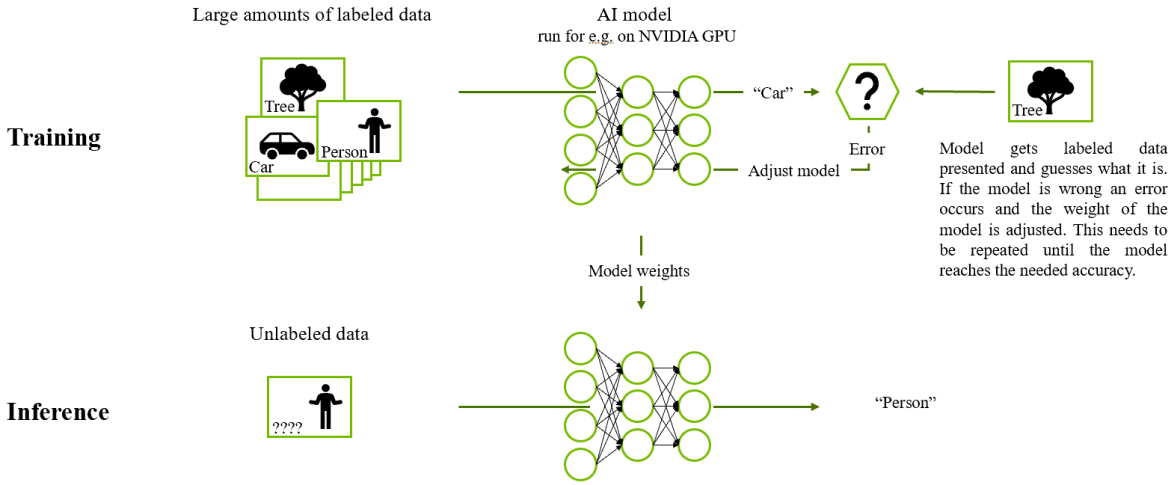
AI Chip Market Segments Overview

Aspects	Details
By Chip Type	•CPU, GPU, ASIC, FPGA
By Industry Vertical	•Media and Advertising, Financial Services, IT and Telecom, Retail, Healthcare, Automotive and Transportation, Others
By Application	•Nature Language Processing, Robotics, Computer Vision, Network Security, Others
By Processing Type	•Edge, Cloud
By Region	•**North America** (U.S., Canada, Mexico) •**Europe** (UK, Germany, France, Russia, Rest of Europe) •**Asia-Pacific** (China, Japan, India, Australia, Rest of Asia-Pacific) •**LAMEA** (Latin America, Middle East, Africa)

Source (adapted): Himanshu, J, and K Vineet. 2023. Artificial Intelligence Chip Market Research, 2032. Allied Market Research, Allied Market Research.

Appendix 10

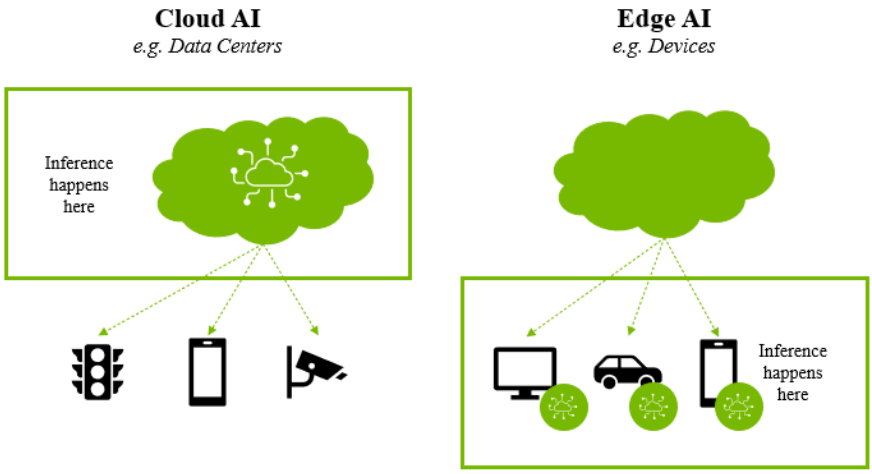
Inference vs Training



Source (adapted): Robins, Mark. 2023. Learning Training and Inference. November 5. https://www.linkedin.com/pulse/difference-between-deep-learning-training-inference-mark-robins-mdq8c/.

Appendix 11

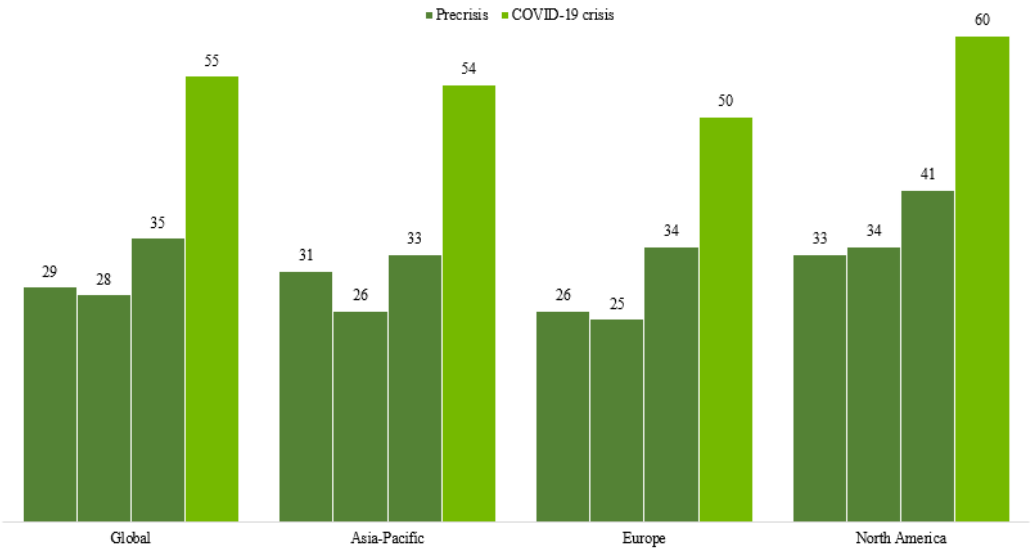
Cloud vs Edge Computing



Source (adapted): SoftmaxAI. 2022. Edge AI vs Cloud AI. https://softmaxai.com/edge-ai-vs-cloud-ai/.

Appendix 12

Average Share of Products and/or Services that are Partially or Fully Digitized (in %)



Source (adapted): LaBerge, Laura, Clayton O'Toole, Jeremy Schneider, and Kate Smaje. 2020. "McKinsey & Company." How COVID-19 has pushed companies over the technology tipping point—and transformed business forever. October 5. https://www.mckinsey.com/capabilities/strategy-and-corporate-finance/our-insights/how-covid-19-has-pushed-companies-over-the-technology-tipping-point-and-transformed-business-forever.

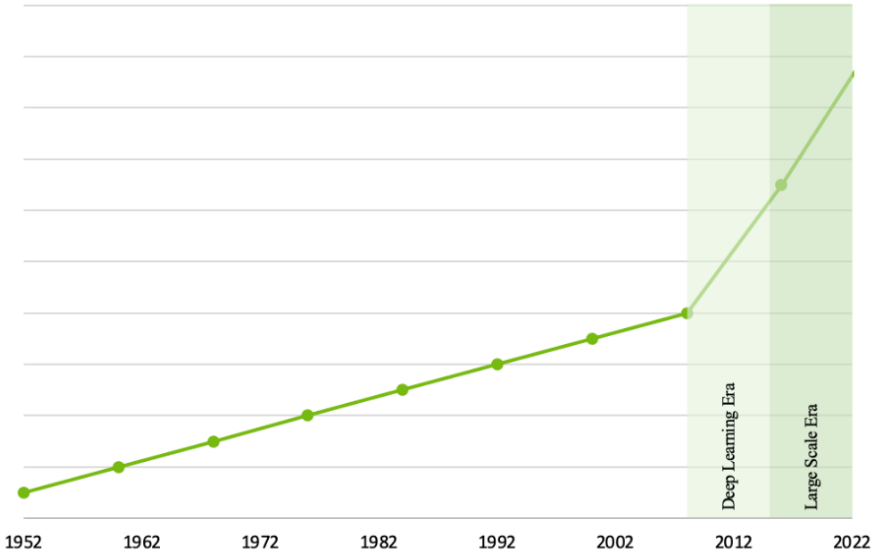
Appendix 13

Global Size of the Big Data Analytics Market from 2021 to 2029 (in billion U.S. dollars)

Source (adapted): Taylor, Petroc. 2022. Size of the big data analytics market worldwide from 2021 to 2029. October 7. https://www.statista.com/statistics/1336002/big-data-analytics-market-size/.

Appendix 14

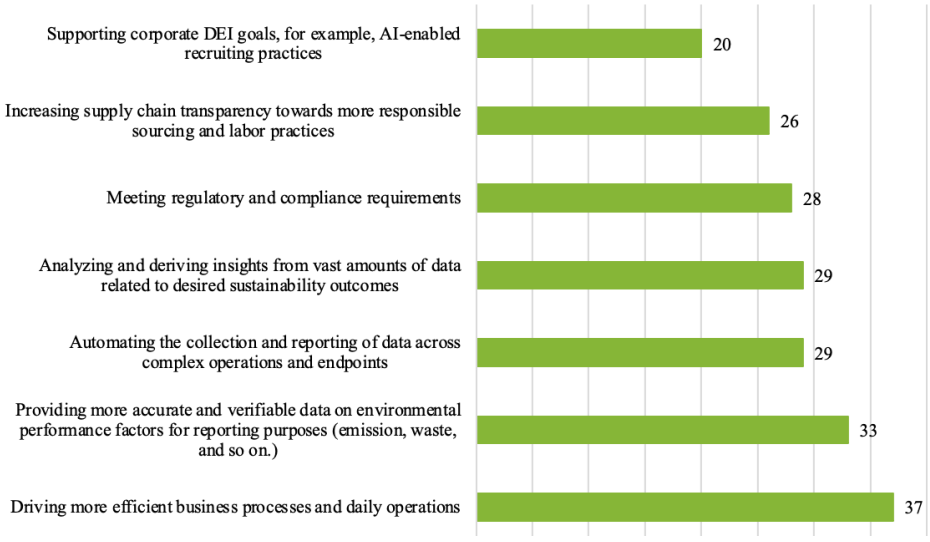
Training Compute of Machine Learning Systems Outperforming Moore's Law



Source (adapted): The Physics arXiv Blog. 2022. AI Machines Have Beaten Moore's Law Over The Last Decade, Say Computer Scientists. February 21. https://www.discovermagazine.com/technology/ai-machines-have-beaten-moores-law-over-the-last-decade-say-computer.

Appendix 15

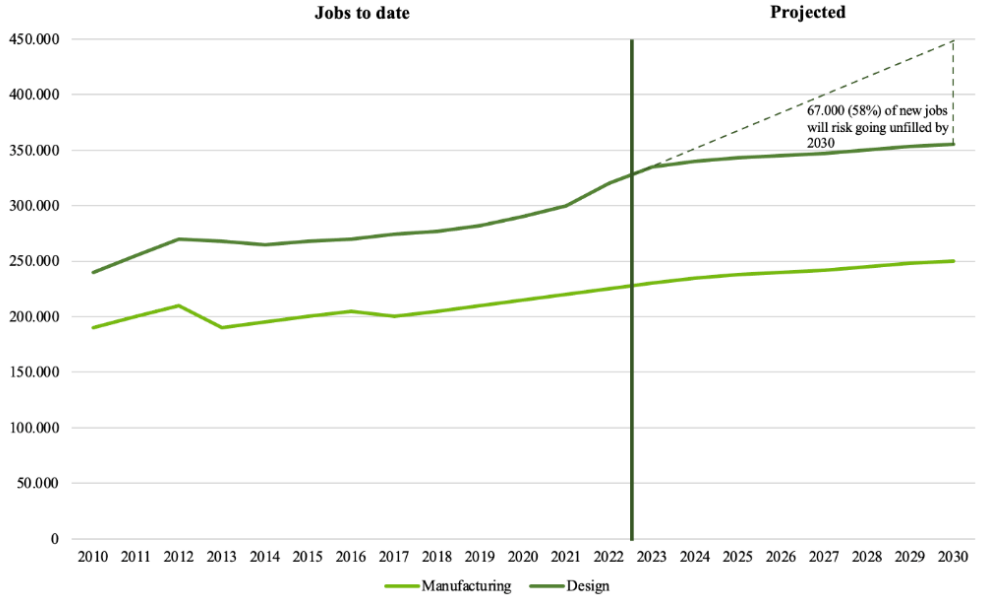
Solvability of ESG Challenges Through the Use of AI in Organizations Worldwide 2022 (in %)



Source (adapted): Thormundsson, Bergur. 2023. Which ESG or sustainability challenges do enterprises think AI has the greatest potential to help solve in 2022? May 3. https://www.statista.com/statistics/1378751/esg-solvability-through-ai-worldwide/.

Appendix 16

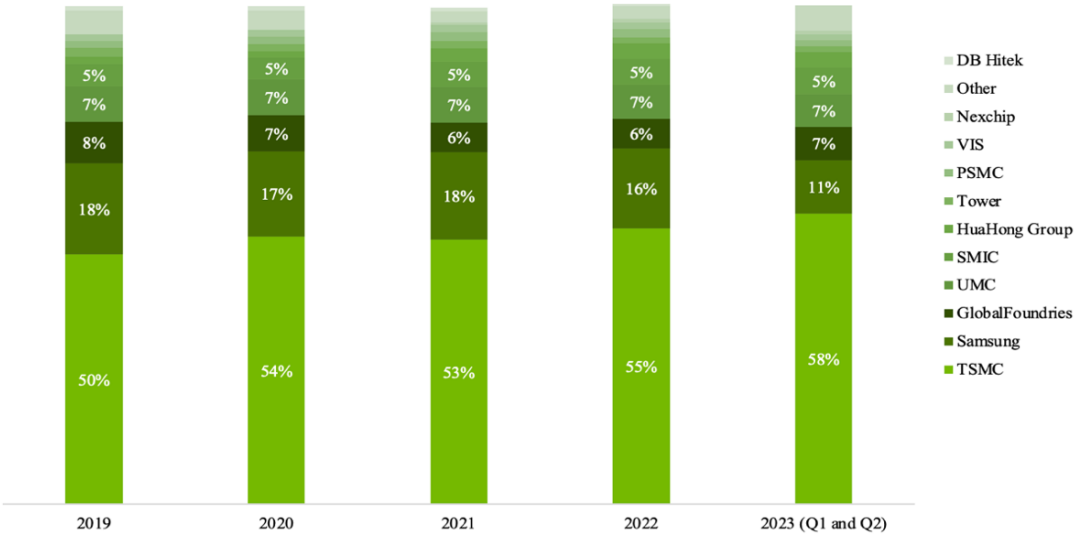
Historical Semiconductor Workforce and Projected 2023-2030 Gap



Source (adapted): Oxford Economics. 2023. Chipping Away: Assessing and Addressing the Labor Market Gap Facing the U.S. Semiconductor Industry. Exford Economics, Semiconductor Industry Association. https://www.semiconductors.org/chipping-away-assessing-and-addressing-the-labor-market-gap-facing-the-u-s-semiconductor-industry/.

Appendix 17

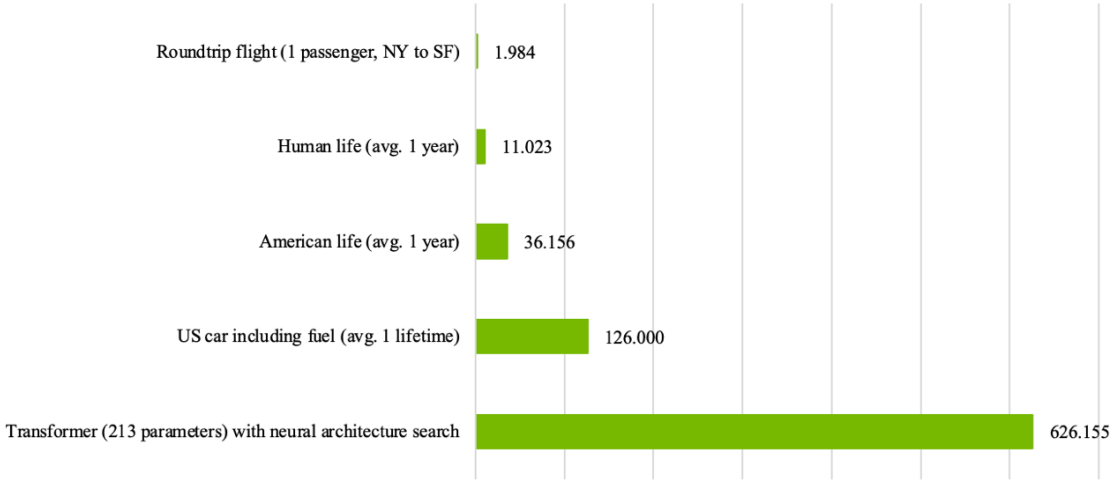
Global Semiconductor Foundries Revenue Share from 2019 to 2023



Source (adapted): Statista. 2023f. Semiconductor foundries revenue share worldwide from 2019 to 2023, by quarter. September 5. https://www.statista.com/statistics/867223/worldwide-semiconductor-foundries-by-market-share/.

Appendix 18

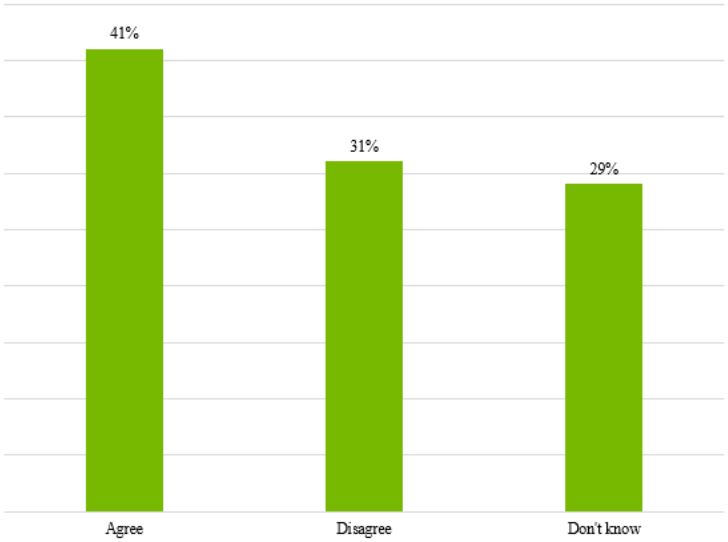
Estimated CO2 Emissions from Training Common NLP Models, Compared to Familiar Consumption (in lbs of C02 Equivalent)



Source (adapted): Strubell, Emma, Ananya Ganesh, and Andrew McCallum. 2019. "Energy and Policy Considerations for Deep Learning in NLP." College of Information and Computer Sciences (University of Massachusetts Amherst).

Appendix 19

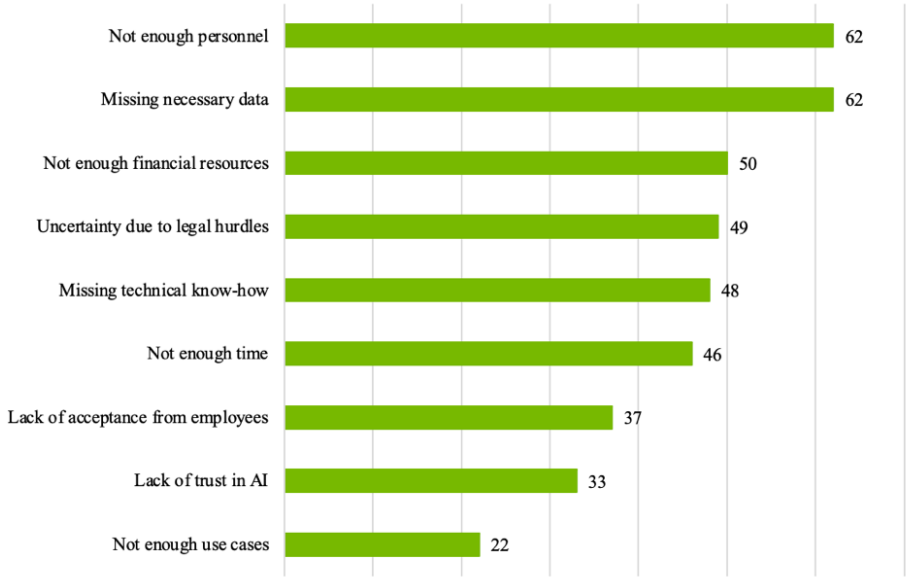
*Survey CEOs in Luxembourg on the **Statement**: "AI will Displace more Jobs than it Creates in the Long Run"? (2018)*



Source (adapted): Thormundsson, Bergur. 2022. Do you agree or disagree with the statement: "AI will displace more jobs than it creates in the long run"? March 17. https://www.statista.com/statistics/1024145/influence-of-ai-on-jobs-according-to-ceos-in-luxembourg/.

Appendix 20

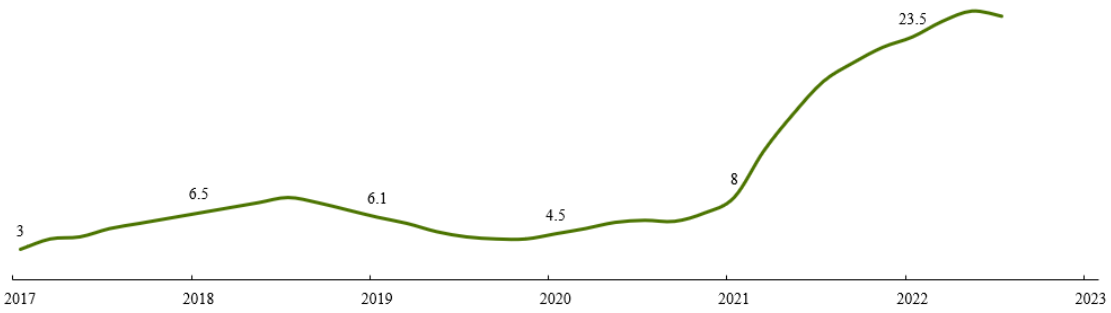
Survey on "What are the Biggest Obstacles when it Comes to Implementing AI?" (in %)



Source (adapted): Davies, Kasia. 2023. What are the biggest obstacles when it comes to implementing AI? June 19. https://www.statista.com/statistics/1393469/obstacles-implementing-ai-germany/.

Appendix 21

Waiting Time for Semiconductors (in Weeks)



Source (adapted): King, Ian. 2022. Wait Times for Chips Grow Again in March as Shortages Drag On. April 5. https://www.bloomberg.com/news/articles/2022-04-05/wait-times-for-chips-grow-again-in-march-as-shortages-drag-on.

Appendix 22

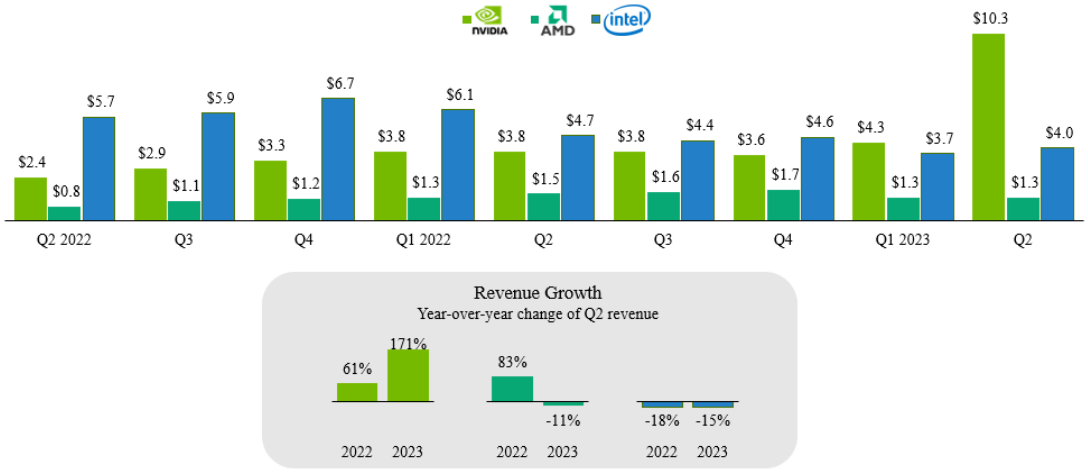
Simplified Ecosystem of AI Applications Using Cloud Services

Based on AI Chips from Fabless Chip Companies

Source (adapted): Own interpretation

Appendix 23

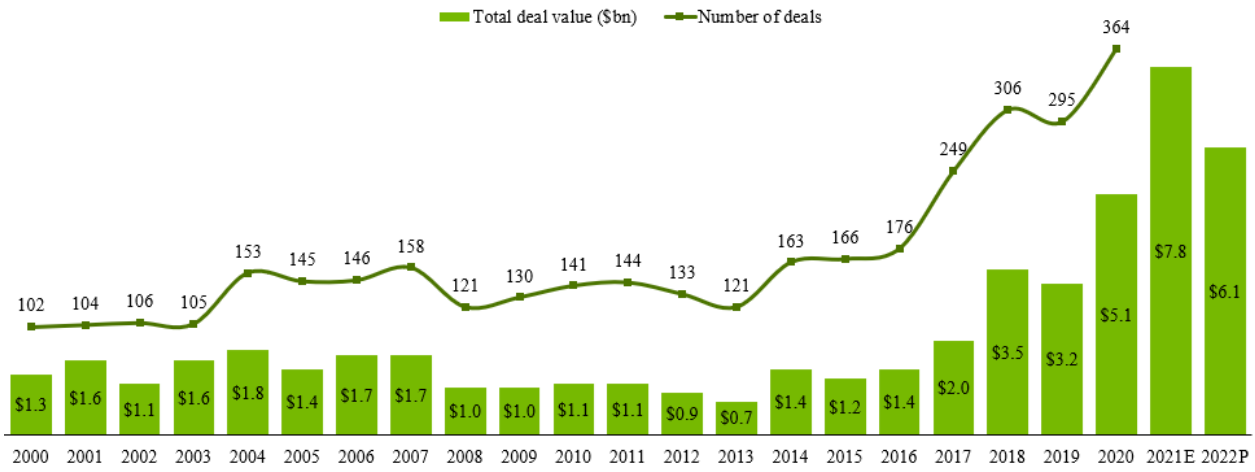
Data Center Revenue (in U.S. dollars)



Source (adapted): "Comparing AI Chip Sales: NVIDIA vs. AMD vs. Intel." Visual Capitalist. Accessed December 2, 2023. https://www.visualcapitalist.com/NVIDIA-vs-amd-vs-intel-comparing-ai-chip-sales/

Appendix 24

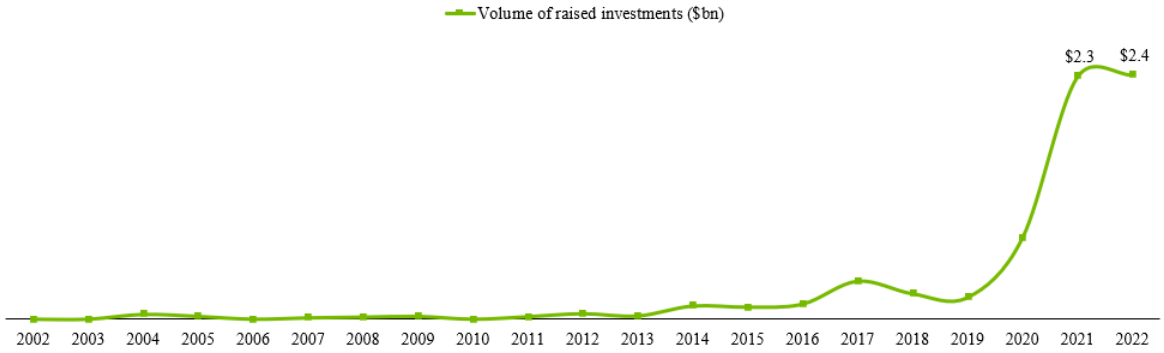
Global VC Investment in Semiconductors



Source (adapted): "Semiconductor Investors and Venture Capital Predictions 2022." Deloitte Insights. Accessed December 2, 2023. https://www2.deloitte.com/us/en/insights/industry/technology/technology-media-and-telecom-predictions/2022/semiconductor-investors-venture-capital.html.

Appendix 25

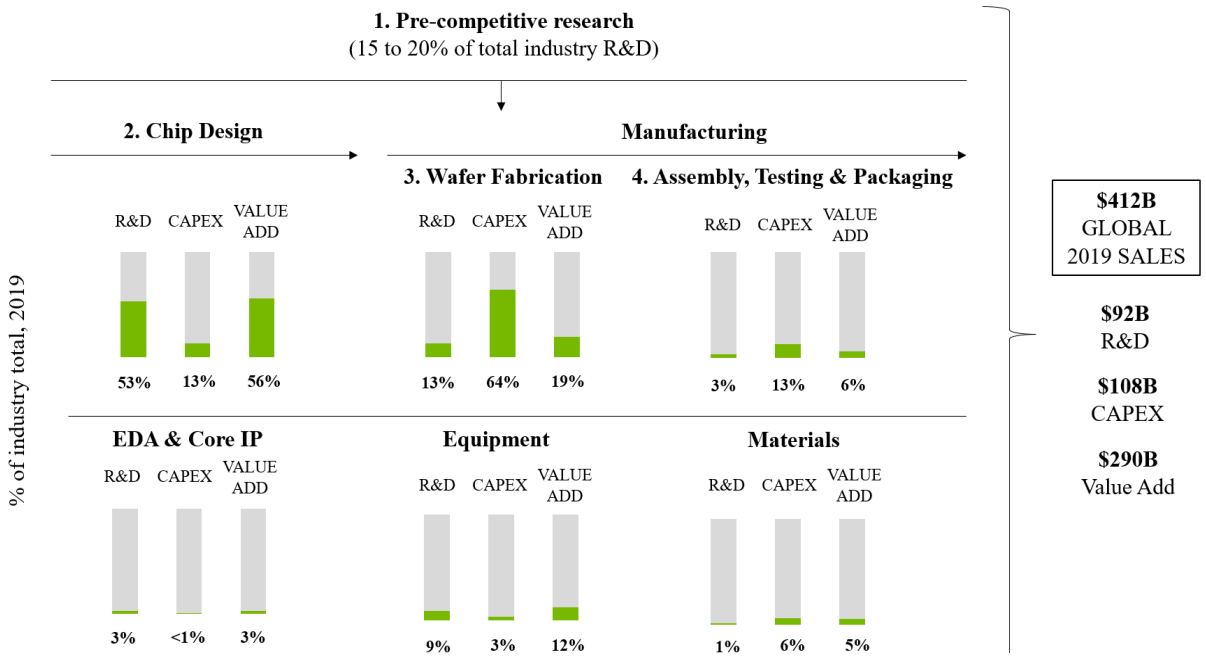
Investments in Quantum Computing



Source (adapted): "Quantum Technology Sees Record Investments, Progress on Talent Gap." McKinsey & Company. Accessed December 2, 2023. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/quantum-technology-sees-record-investments-progress-on-talent-gap.

Appendix 26

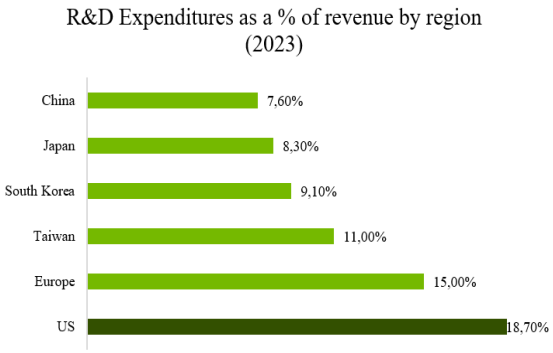
Semiconductor Industry Value Added, R&D And CAPEX by Activity 2019 (in %)

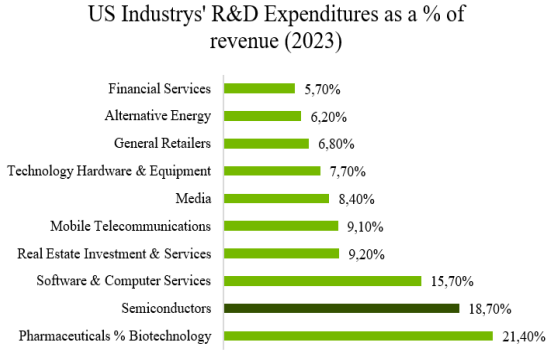


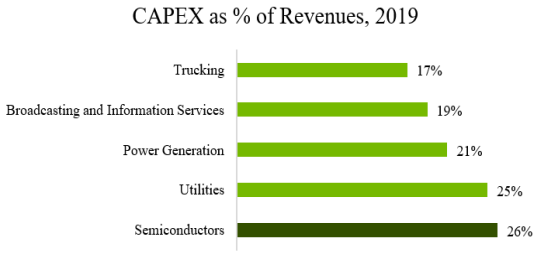
Source (adapted): Antonio Varas, R. V. (2021). *Strengthening The Global Semiconductor Supply Chain in an Uncertain Era.* BCG X SIA. Retrieved from https://www.semiconductors.org/wp-content/uploads/2021/05/BCG-x-SIA-Strengthening-the-Global-Semiconductor-Value-Chain-April-2021_1.pdf. Semiconductor Industry Association. (2021). *2021 State of The U.S. Semiconductor Industry.* Retrieved from https://www.semiconductors.org/wp-content/uploads/2021/09/2021-SIA-State-of-the-Industry-Report.pdf.

Appendix 27

R&D And CAPEX Investment (in %)



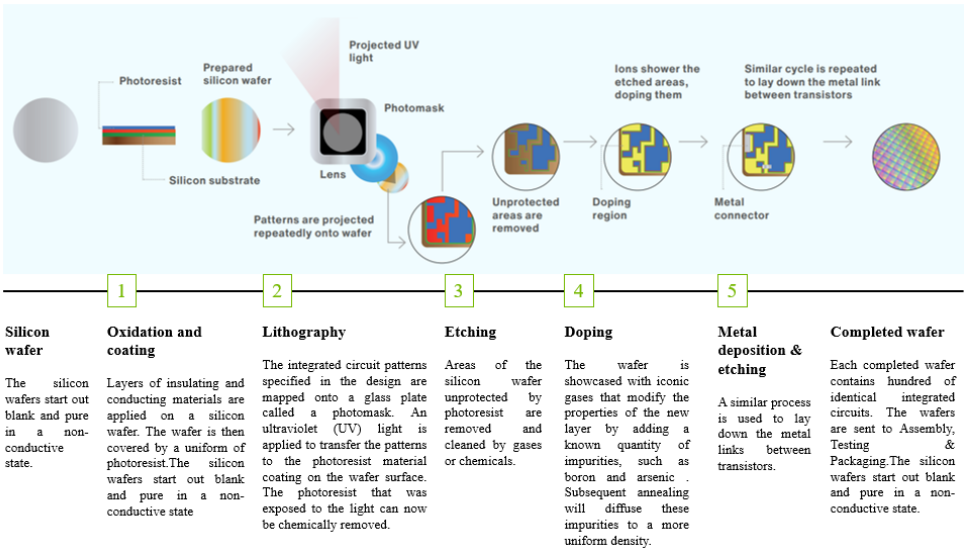




Source (adapted): Antonio Varas, R. V. (2021). *Strengthening The Global Semiconductor Supply Chain in an Uncertain Era.* BCG X SIA. Retrieved from https://www.semiconductors.org/wp-content/uploads/2021/05/BCG-x-SIA-Strengthening-the-Global-Semiconductor-Value-Chain-April-2021_1.pdf. Semiconductor Industry Association. (2021). *2021 State of The U.S. Semiconductor Industry.* Retrieved from https://www.semiconductors.org/wp-content/uploads/2021/09/2021-SIA-State-of-the-Industry-Report.pdf. Semiconductor Industry Association. (2023). *2023 State of The US Semiconductor Industry.* Retrieved from https://www.semiconductors.org/wp-content/uploads/2023/07/SIA_State-of-Industry-Report_2023_Final_072723.pdf

Appendix 28

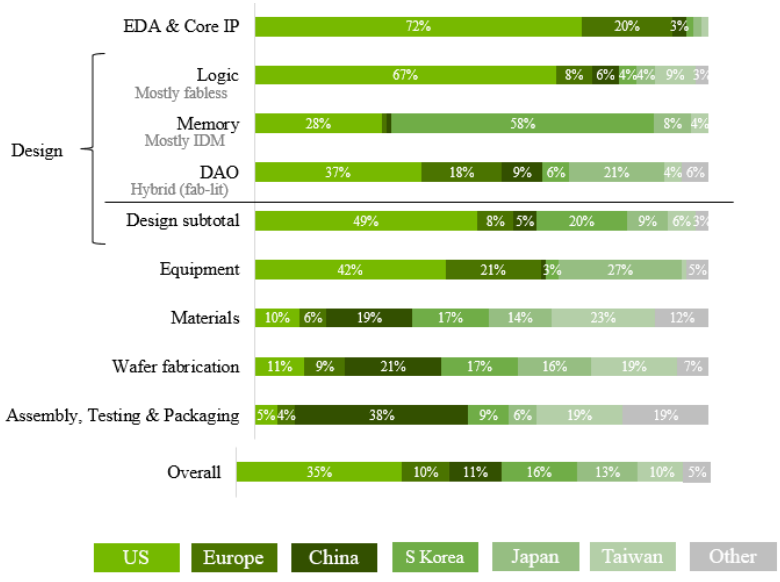
Wafer Manufacturing Phases



Antonio Varas, R. V. (2021). *Strengthening The Global Semiconductor Supply Chain in an Uncertain Era*. BCG X SIA. Retrieved from https://www.semiconductors.org/wp-content/uploads/2021/05/BCG-x-SIA-Strengthening-the-Global-Semiconductor-Value-Chain-April-2021_1.pdf.

Appendix 29

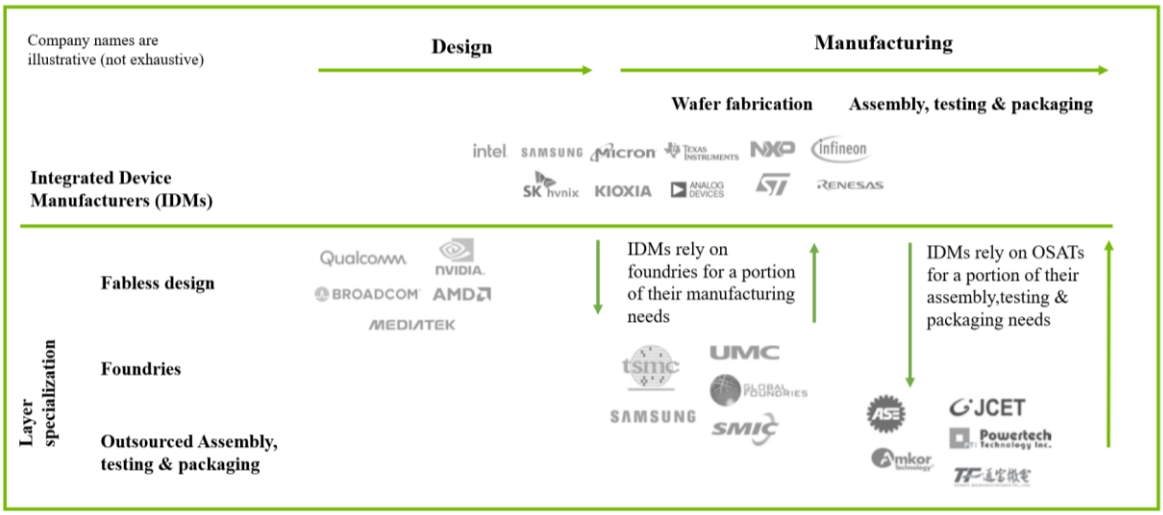
Semiconductor Industry's Distribution by Region (in %)



Source (adapted): Semiconductor Industry Association. (2022). *2022 State of the U.S. Semicondcutor Industry.* Retrieved from https://www.semiconductors.org/wp-content/uploads/2022/11/SIA_State-of-Industry-Report_Nov-2022.pdf.

Appendix 30

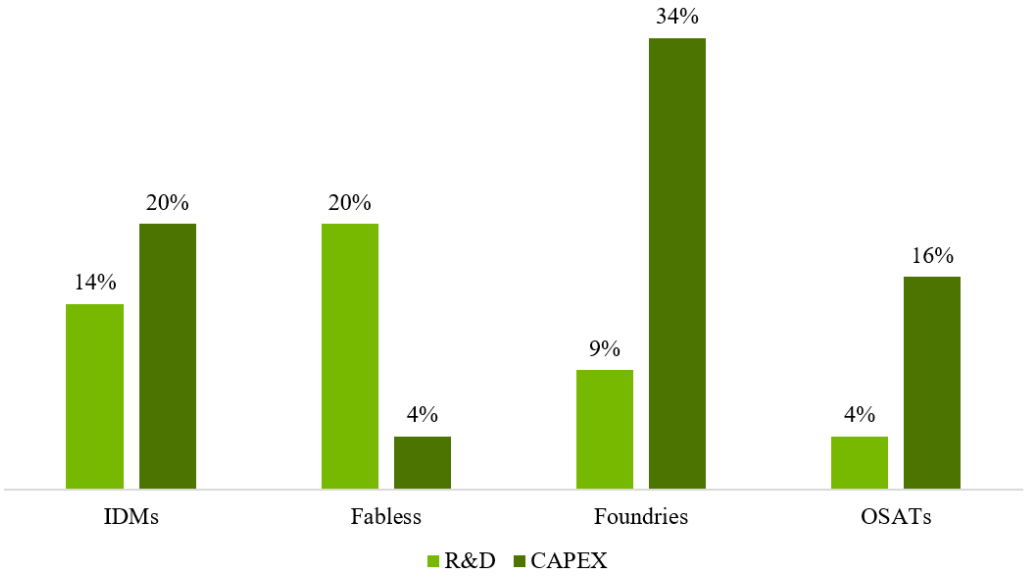
Semiconductor Industry Value Chain Structure



Antonio Varas, R. V. (2021). *Strengthening The Global Semiconductor Supply Chain in an Uncertain Era.* BCG X SIA. Retrieved from https://www.semiconductors.org/wp-content/uploads/2021/05/BCG-x-SIA-Strengthening-the-Global-Semiconductor-Value-Chain-April-2021_1.pdf.

Appendix 31

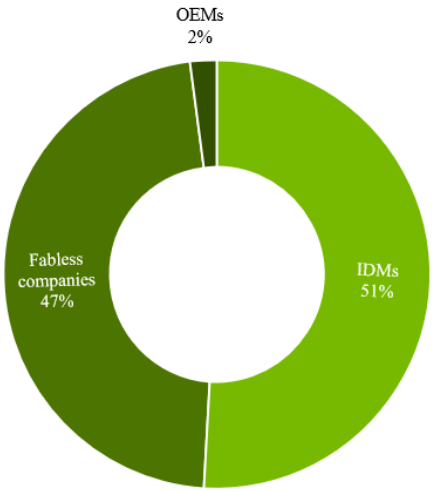
Semiconductor Companies' R&D and CAPEX Spending as a % of Revenue



Source (adapted): Antonio Varas, R. V. (2021). *Strengthening The Global Semiconductor Supply Chain in an Uncertain Era.* BCG X SIA. Retrieved from https://www.semiconductors.org/wp-content/uploads/2021/05/BCG-x-SIA-Strengthening-the-Global-Semiconductor-Value-Chain-April-2021_1.pdf.

Appendix 32

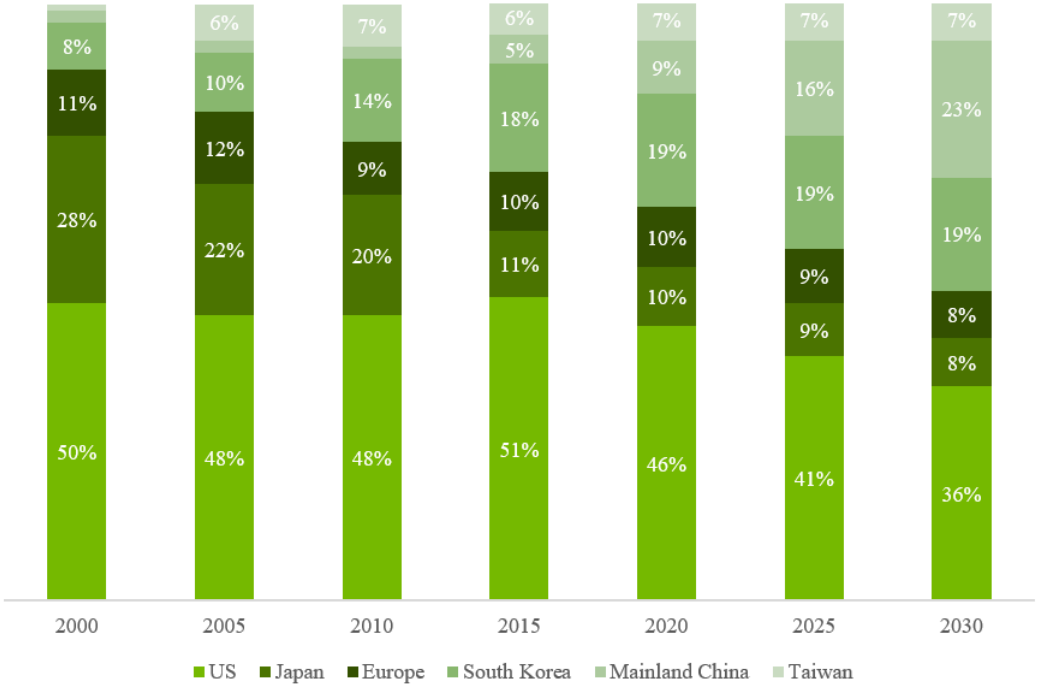
Semiconductor Design Business Model Distribution (2021)



Source (adapted): Semiconductor Industry Association. (2022). *2022 State of the U.S. Semicondcutor Industry*. Retrieved from https://www.semiconductors.org/wp-content/uploads/2022/11/SIA_State-of-Industry-Report_Nov-2022.pdf.

Appendix 33

Design Market Share by Region of Company HQ (in %)



Source (adapted): Semiconductor Industry Association. (2022). *2022 State of the U.S. Semicondcutor Industry*. Retrieved from https://www.semiconductors.org/wp-content/uploads/2022/11/SIA_State-of-Industry-Report_Nov-2022.pdf.

Appendix 34

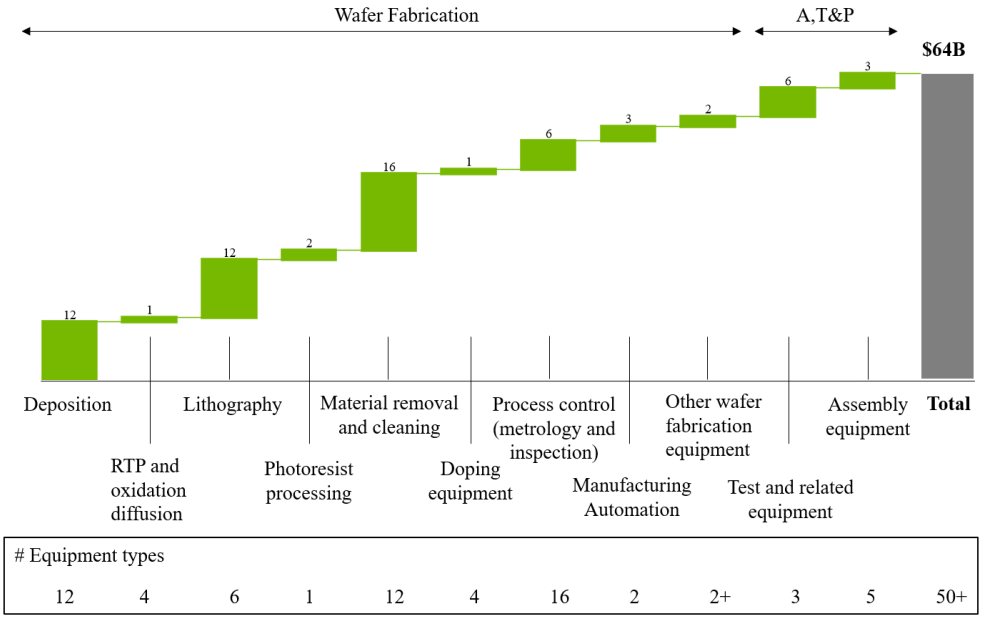
Revenue Ranking of Top 10 Design Companies (2020)

Ranking	Company
1.	**Intel**
2.	Samsung
3.	SK Hynix
4.	**Micron**
5.	**Qualcomm**
6.	**Broadcom**
7.	**NVIDIA**
8.	**Texas Instruments**
9.	**Apple**
10.	Infineon

Source (adapted): Ramiro Palma, R. V. (2022). *The Growing Challenge of Semiconductor Design Leadership.* BCG X SIA. Retrieved from https://www.semiconductors.org/wp-content/uploads/2022/11/2022_The-Growing-Challenge-of-Semiconductor-Design-Leadership_FINAL.pdf

Appendix 35

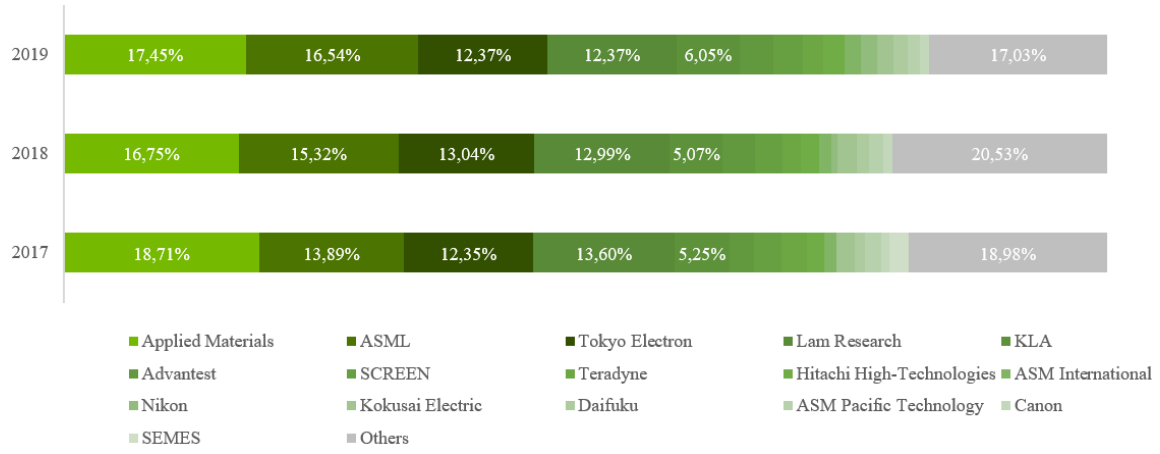
Breakdown of Market Size of Semiconductor Manufacturing Equipment by Major Families, 2019 ($ Billion)



Source (adapted): Ramiro Palma, R. V. (2022). *The Growing Challenge of Semiconductor Design Leadership*. BCG X SIA. Retrieved from https://www.semiconductors.org/wp-content/uploads/2022/11/2022_The-Growing-Challenge-of-Semiconductor-Design-Leadership_FINAL.pdf

Appendix 36

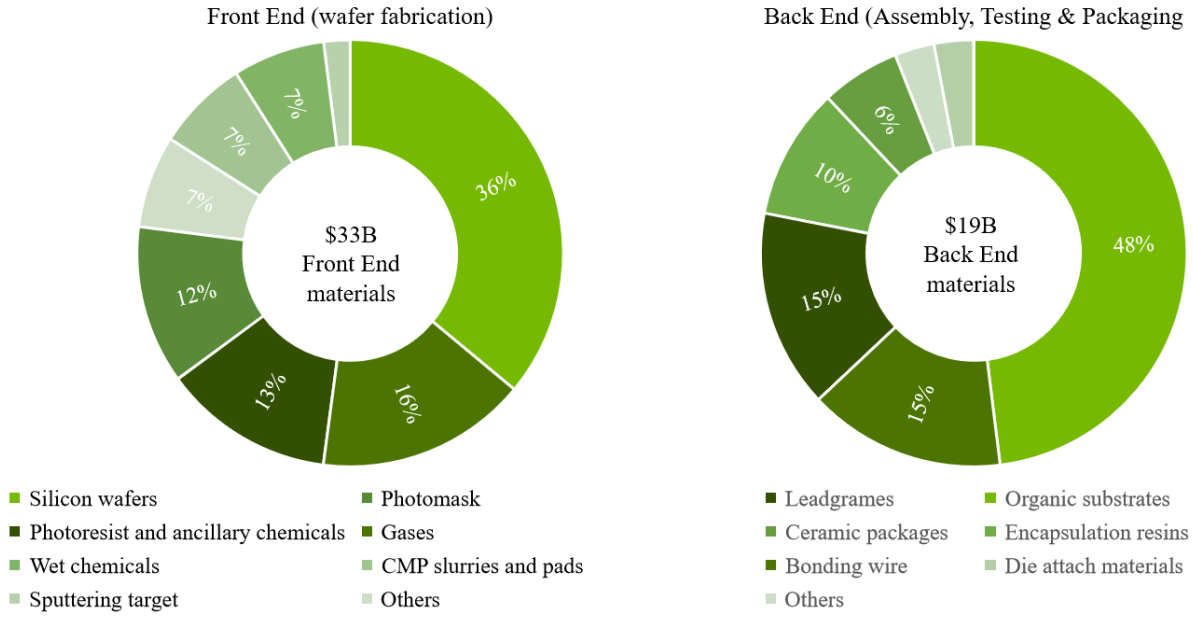
Semiconductor Equipment Revenue Worldwide, by Supplier (in %)



Source (adapted):. Statista. (2020). *Semiconductor Equipment Revenue Worldwide from 2017 to 2019, by Supplier (in Billion U.S. Dollars)*. Retrieved from https://www-statista-com.eu1.proxy.openathens.net/statistics/532224/worldwide-semiconductor-wafer-level-manufacturing-equipment-vendor-revenue/

Appendix 37

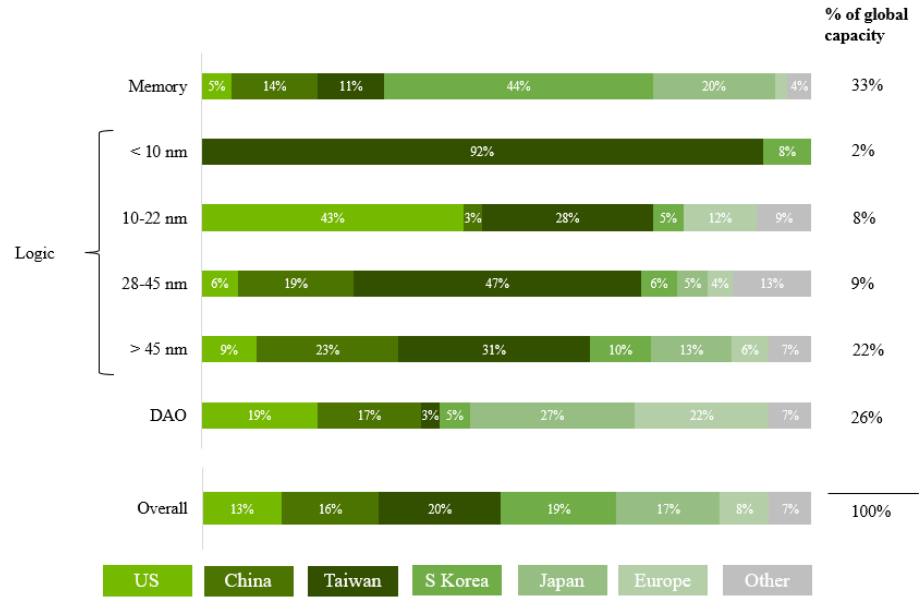
Breakdown of Market Size of Semiconductor Manufacturing Materials, 2019 (in % of $ Billion)



Source (adapted): Ramiro Palma, R. V. (2022). *The Growing Challenge of Semiconductor Design Leadership.* BCG X SIA. Retrieved from https://www.semiconductors.org/wp-content/uploads/2022/11/2022_The-Growing-Challenge-of-Semiconductor-Design-Leadership_FINAL.pdf

Appendix 38

Breakdown of the Global Wafer Fabrication Capacity per Region, 2019 (in %)



Source (adapted): Ramiro Palma, R. V. (2022). *The Growing Challenge of Semiconductor Design Leadership.* BCG X SIA. Retrieved from https://www.semiconductors.org/wp-content/uploads/2022/11/2022_The-Growing-Challenge-of-Semiconductor-Design-Leadership_FINAL.pdf

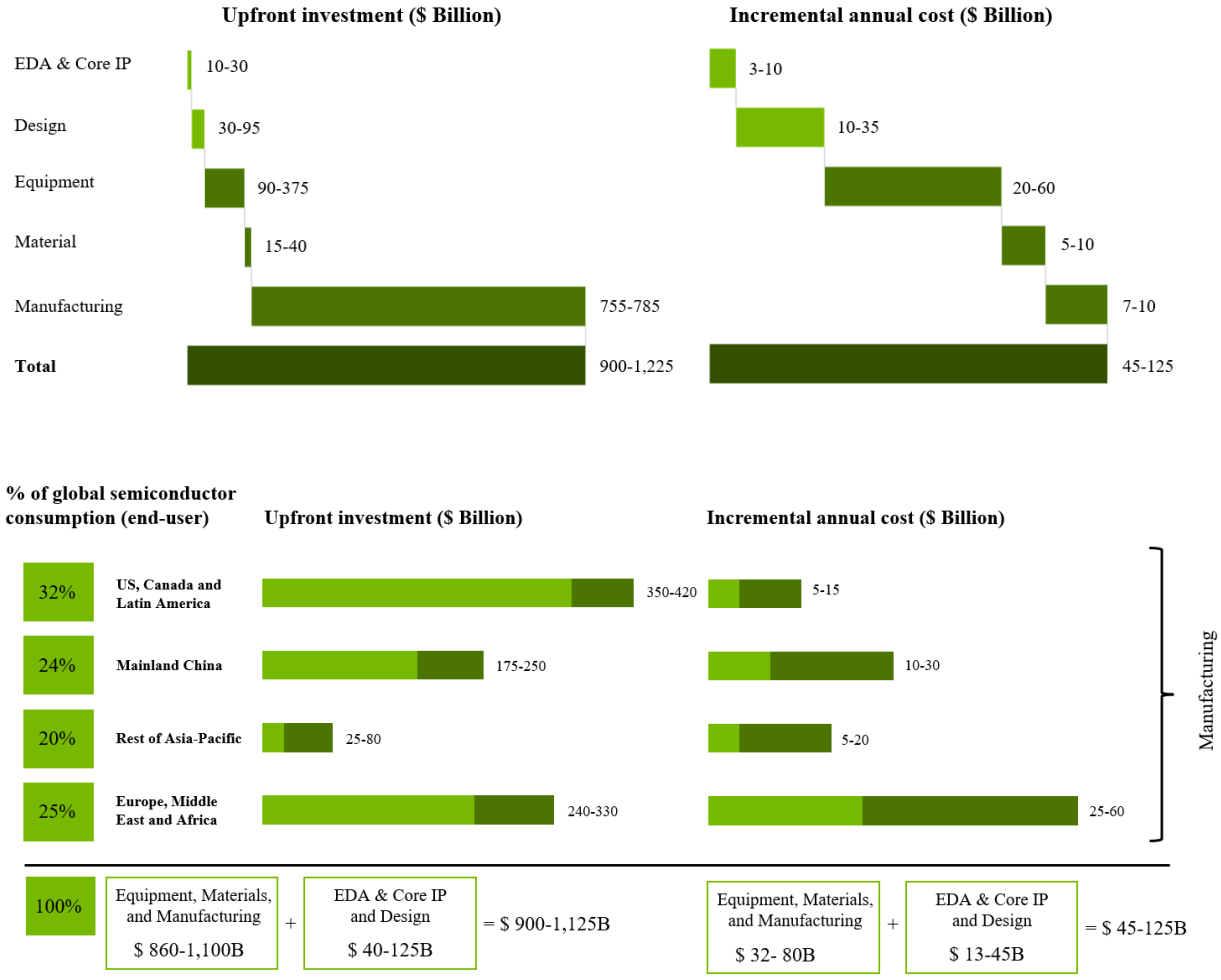
Appendix 39

Foundries' Utilization Rate During Chip Shortage (in %)

Semiconductor Industry Association. (2022). *2022 State of the U.S. Semicondcutor Industry*. Retrieved from https://www.semiconductors.org/wp-content/uploads/2022/11/SIA_State-of-Industry-Report_Nov-2022.pdf.

Appendix 40

Incremental Cost to Cover 2019 Demand with Fully "Self-sufficient" Localized Semiconductor Supply Chains

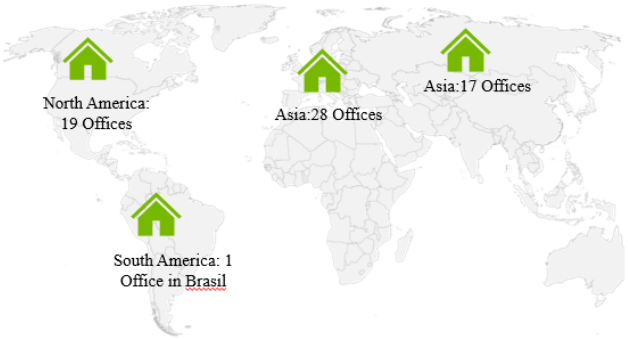


Source (adapted): Ramiro Palma, R. V. (2022). *The Growing Challenge of Semiconductor Design Leadership.* BCG X SIA. Retrieved from https://www.semiconductors.org/wp-content/uploads/2022/11/2022_The-Growing-Challenge-of-Semiconductor-Design-Leadership_FINAL.pdf

Appendix 41

NVIDIA: Global Presence and Operational Highlights in 2022

50 Offices worldwide



Nvidia Today: Key Facts

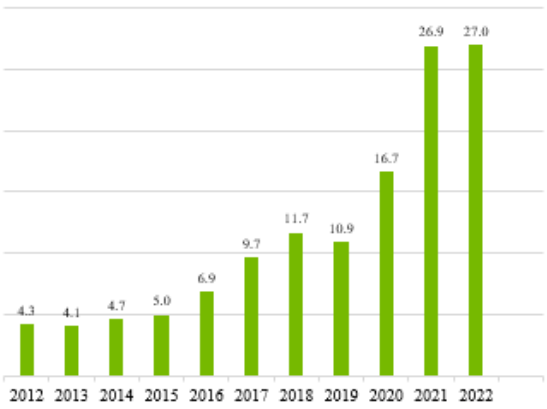
- Global workforce: 26,196 across 35 countries
- Low turnover rate: 5,3%
- R&D employees: 19,532
- SG&A: 6,664
- 2022 product launches: Over 160
- Recruitmnet:37% of hires via employee referrals

Source (adapted): Nvidia Corporation Annual Report. 2023. https://s201.q4cdn.com/141608511/files/doc_financials/2023/ar/2023-Annual-Report-1.pdf (accessed 2023).

Appendix 42

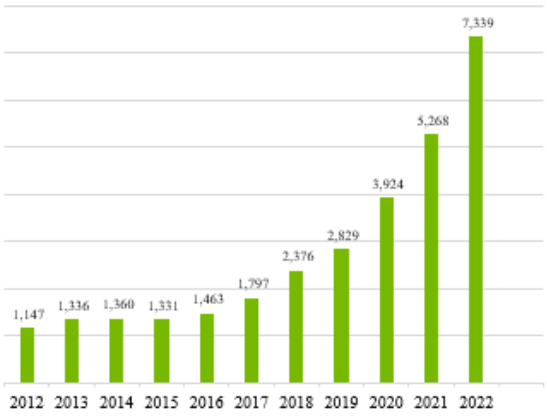
NVIDIA's Revenues and R&D Expenses from 2020 to 2022 (in billion U.S. dollars)

Revenue from 2012 to 2022 (in million U.S. dollars)



2022 Gross margin: 57% and Net Income: $4.4 billion

R&D expenses from 2012 to 2022 (in million U.S. dollars)

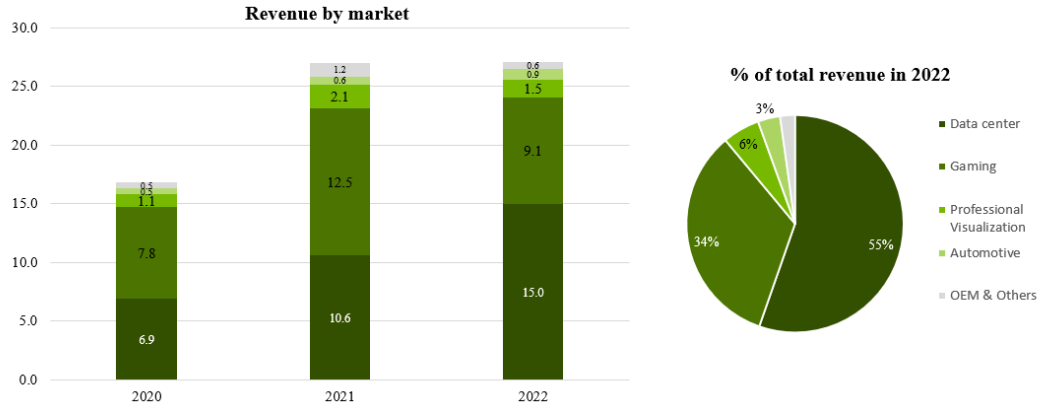


Source (adapted): NVIDIA. NVIDIA Annual Reports and Proxies. n.d. https://investor.nvidia.com/financial-info/annual-reports-and-proxies/default.aspx.

NVIDIA. NVIDIA Annual Reports and Proxies. n.d. https://investor.nvidia.com/financial-info/annual-reports-and-proxies/default.aspx.

Appendix 43

NVIDIA's Revenues by Market from 2020 to 2022 (in billion U.S. dollars)



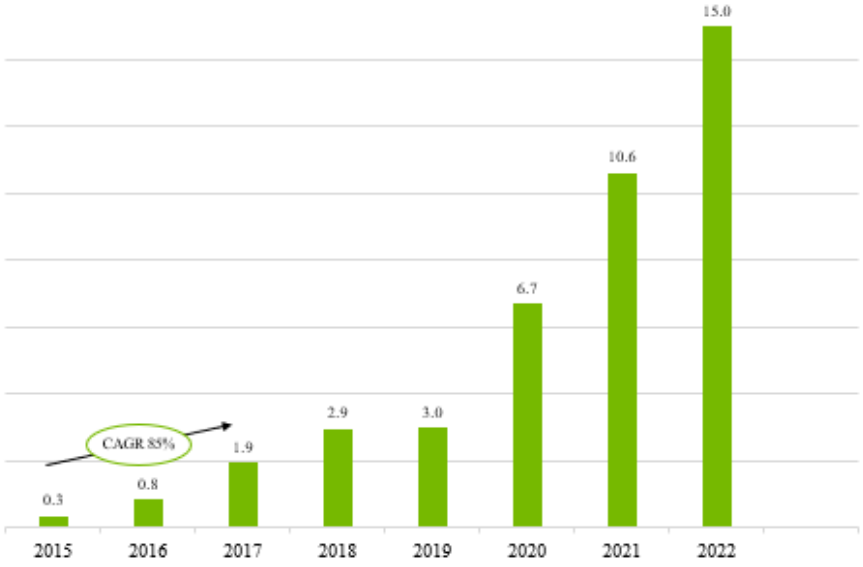
Source (adapted): Nvidia Corporation Annual Report. 2023. https://s201.q4cdn.com/141608511/files/doc_financials/2023/ar/2023-Annual-Report-1.pdf (accessed 2023).

NVIDIA Corporation Annual Report. 2022. https://s201.q4cdn.com/141608511/files/doc_financials/2022/ar/2022 Annual-Review.pdf.

NVIDIA Corporation Annual Report. 2021. https://s201.q4cdn.com/141608511/files/doc_downloads/2021/04/2021-Annual-Review.pdf.

Appendix 44

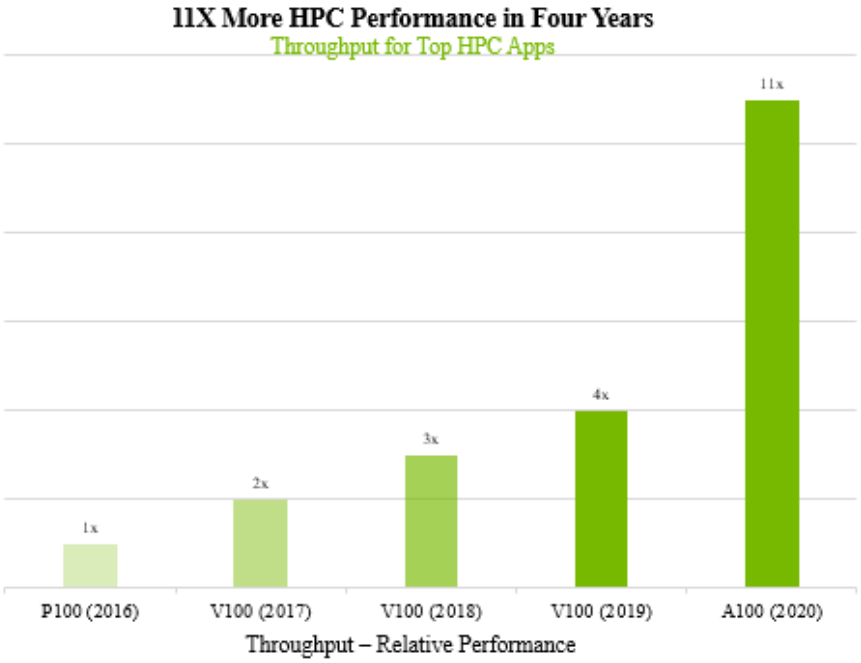
NVIDIA's Data Center Revenues 2012 to 2022 (in billion U.S. dollars)



Source (adapted): NVIDIA. NVIDIA Annual Reports and Proxies. n.d. https://investor.nvidia.com/financial-info/annual-reports-and-proxies/default.aspx.

Appendix 45

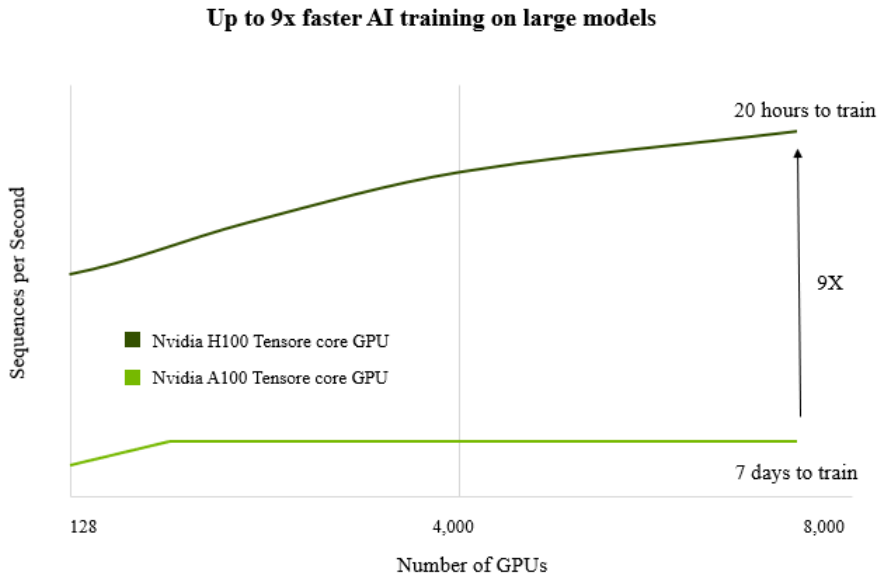
Performance Comparison of NVIDIA's P100 and A100



Source (adapted): NVIDIA. 2021. NVIDIA A100 Tensor Core GPU . n.d. https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/a100/pdf/nvidia-a100-datasheet-nvidia-us-2188504-web.pdf.

Appendix 46

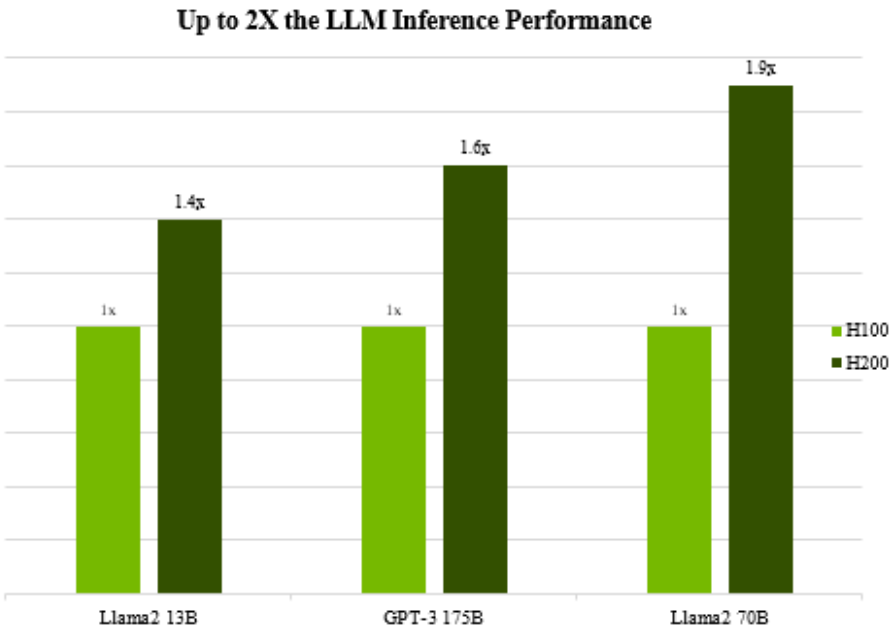
Performance Comparison of NVIDIA's A100 and H100



Source (adapted): NVIDIA H100 Tensor Core GPU. n.d. https://www.nvidia.com/de-de/data-center/h100/

Appendix 47

Performance Comparison of NVIDIA's H100 and H200



Source (adapted): NVIDIA H200 Tensor Core GPU. n.d. https://www.nvidia.com/en-us/data-center/h200/.

Appendix 48

NVIDIA's Quarterly Revenues from 2019 to 2023(in billion U.S. dollars)

Source (adapted): NVIDIA Quarterly Results. n.d. https://investor.nvidia.com/financial-info/quarterly-results/default.aspx.

Appendix 49

Largest Companies by Market Cap (November 2023)

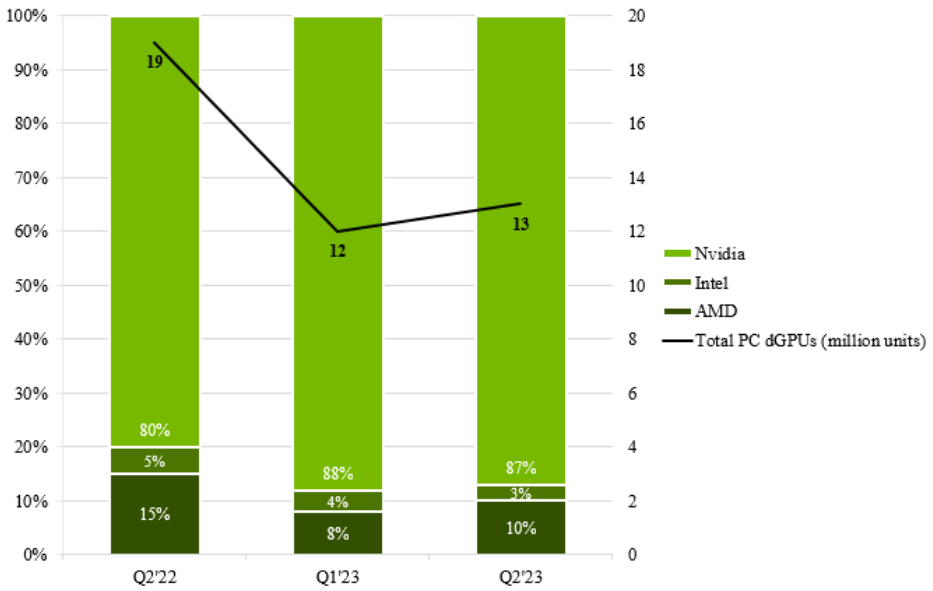
In 2021 NVIDIA ranked at position 24

Rank	Name	Marketcap (in trillion USD)	Price (USD)
1	Apple	2.98	191.3
2	Microsoft	2.77	372.9
3	Saudi Aramco	2.14	8.9
4	Alphabet (Google)	1.66	133.0
5	Amazon	1.51	146.5
6	NVIDIA	1.16	469.8
7	Meta Platforms (Facebook)	0.83	323.7
8	Berkshire Hathaway	0.77	355.6
9	Tesla	0.76	239.0
10	Eli Lilly	0.56	591.8
11	Visa	0.53	256.5
12	TSMC	0.51	98.5
⋮	⋮	⋮	
53	AMD	0.20	120.7
54	Reliance Industries	0.19	28.8
55	Pinduoduo	0.19	145.6
56	Thermo Fisher Scientific	0.19	498.9
57	Alibaba	0.19	73.4
58	SAP	0.19	159.6
59	Intel	0.18	43.6

Source (adapted): CompaniesMarketCap. Largest Companies by Market Cap. n.d. https://companiesmarketcap.com/.
PriceWaterhouseCooper (PWC). Global Top 100 Companies by Market Capitalization. n.d. https://www.pwc.com/gx/en/audit-services/publications/assets/pwc-global-top-100-companies-2021.pdf.

Appendix 50

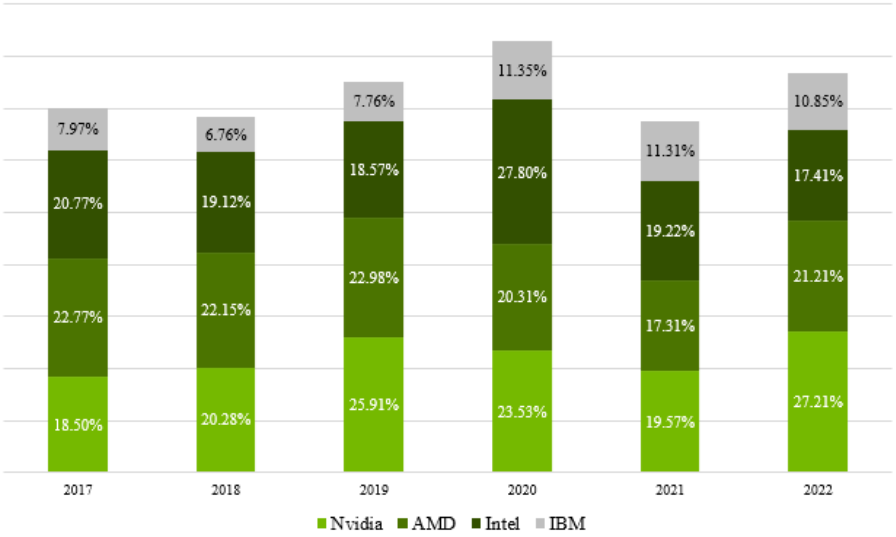
GPU Market Share from Q2 2022 to Q2 2023



Source (adapted): Reddit. GPU Marketshare in Q2 2023: NVIDIA 87%, AMD 10% and Intel 3%. n.d. https://www.reddit.com/r/NVIDIA/comments/165tw5b/gpu_marketshare_in_q2_2023_NVIDIA_87_amd_10_and/

Appendix 51

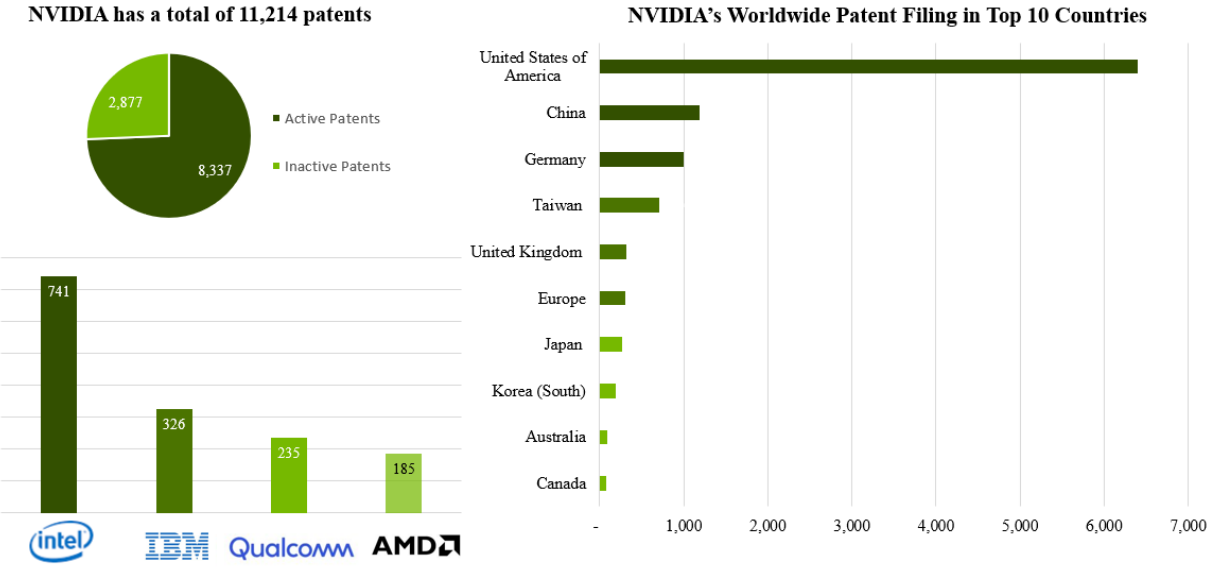
Expenditure of Selected AI Chip Companies (% of revenue), 2017-2022



Source (adapted): NVIDIA. NVIDIA Annual Reports and Proxies. n.d. https://investor.nvidia.com/financial-info/annual-reports-and-proxies/default.aspx.

Appendix 52

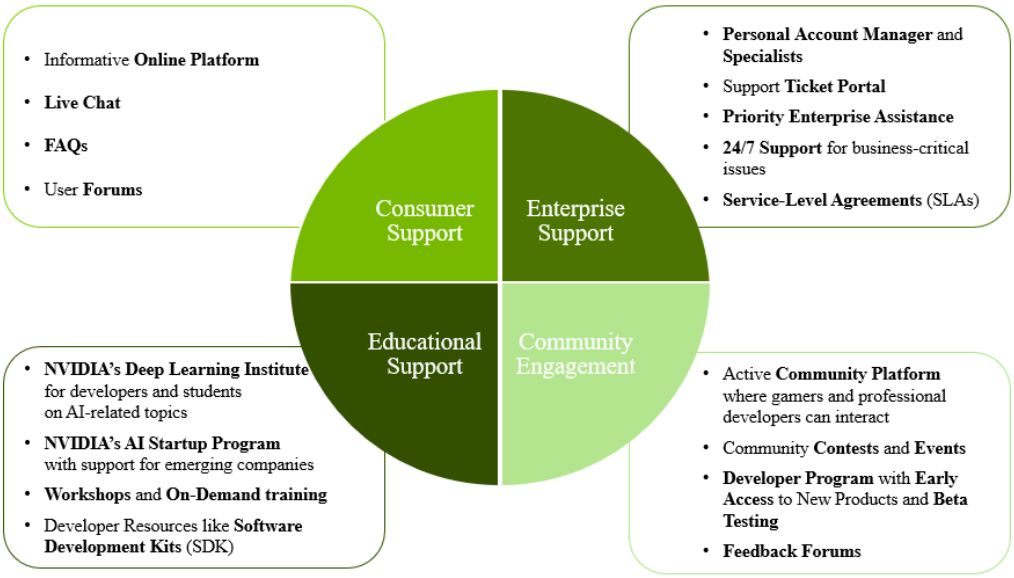
NVIDIA's and its Competitors Patent Portfolio in 2022



Source (adapted): Insights by GreyB. NVIDIA Corporation Patents - Key Insights and StatsI. n.d. https://insights.greyb.com/nvidia-corporation-patents/.

Appendix 53

NVIDIA's Support Services and Community Engagement



Source (adapted): NVIDIA. 2023. Customer Support. https://www.nvidia.com/en-us/support/consumer/.

Appendix 54

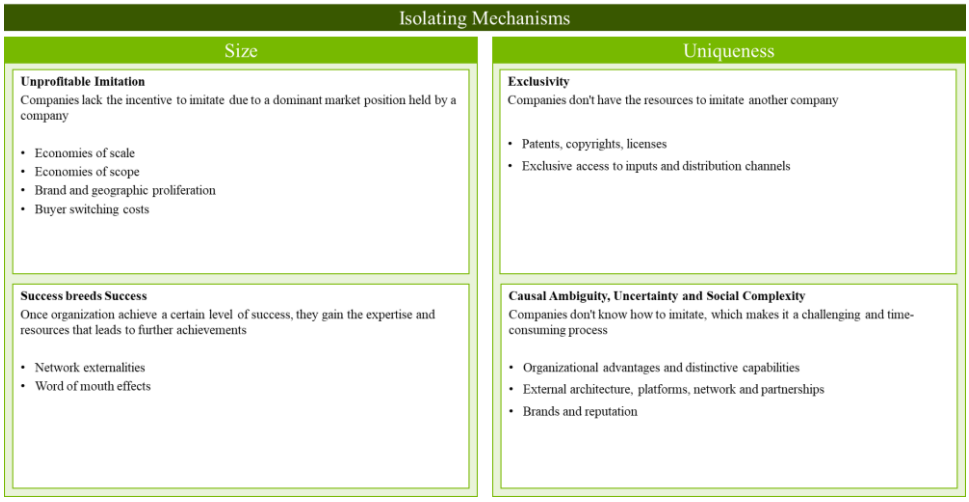
NVIDIA's Partner Types

Cloud Service Provider Offers cloud-based platforms using NVIDIA products	**Data Center Provider** Maintains facilities for data storage and IT infrastructure	**Professional Services** Delivers expert services and support for NVIDIA solutions	**Distributor** Distributes NVIDIA products to various resellers
OEM Produces original equipment incorporating Nvidia equipment	**Academic & Research Institutions** Collaborates on educational programs and research initiatives	**Solution Integrated Partner** Integrate Nvidia technologies to create specialized solutions	**Solution Provider (VAR)** Enhances NVIDIA products with additional services for resale
Governments Deploy NVIDIA solutions into public sector operations	**Global System Integrator** Implements NVIDIA-based solutions on a multinational scale	**Solution Advisor – Consultant** Offers strategic guidance on implementing NVIDIA technologies	**Independent Software Vendor** Sells enterprise-level applications enhanced by NVIDIA GPUs

Source (adapted): NVIDIA Partner Network. n.d. https://www.nvidia.com/en-us/about-nvidia/partners/. NVIDIA Partner Program Expands to 1,500 Members, Adds New Benefits. 2020. https://blogs.nvidia.com/blog/npn-expands-1500-members-new-benefits/.

Appendix 55

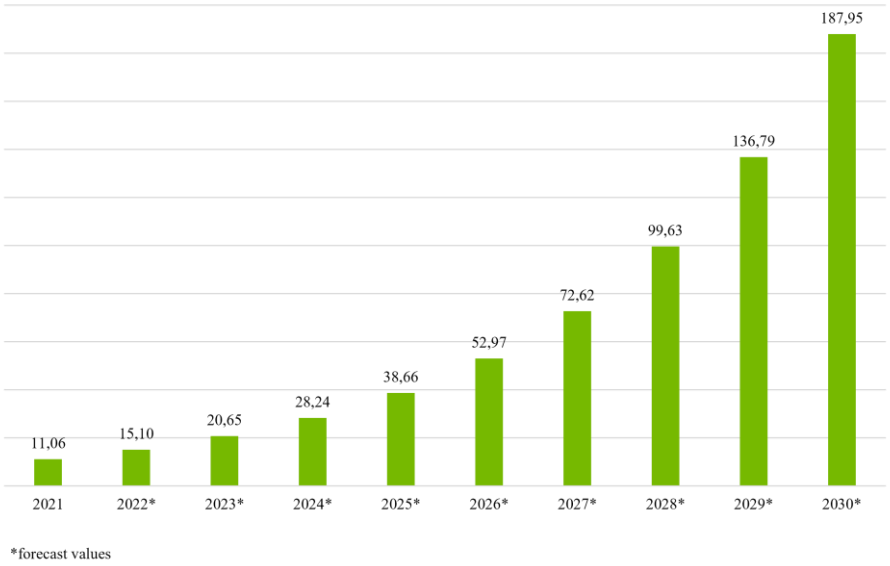
Size – Uniqueness Framework



Source (adapted): Almeida Costa, Luís. "Size – Uniqueness Framework." Framework introduced in Strategy Courses.

Appendix 56

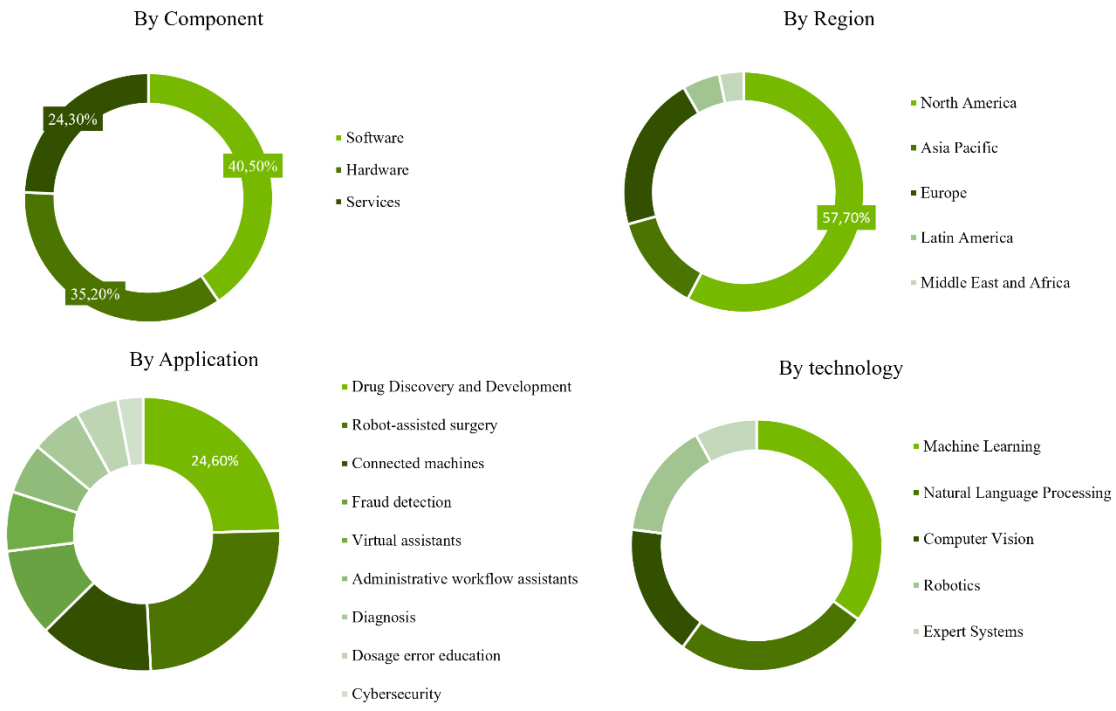
AI in Healthcare Market Size Worldwide from 2021 to 2030



Source (adapted): Statista. 2022. "Artificial Intelligence (AI) in healthcare."

Appendix 57

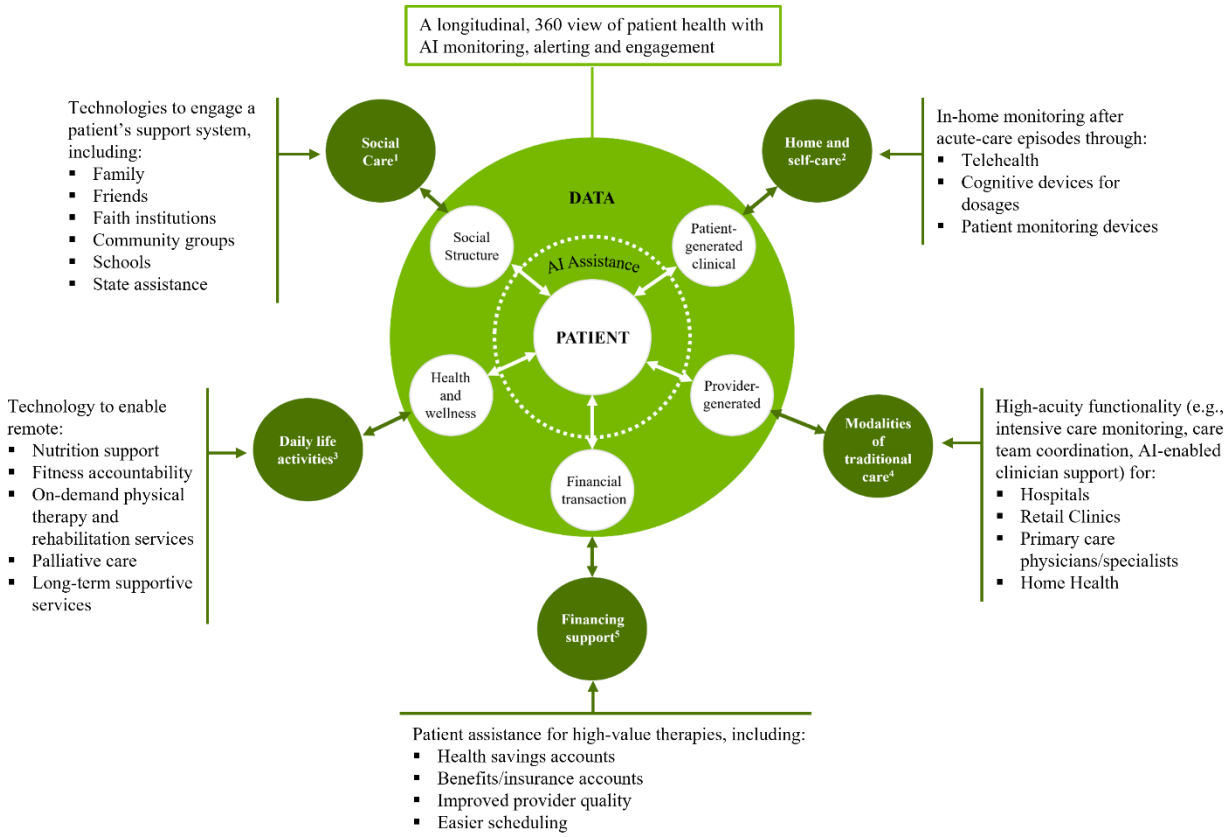
AI Healthcare Market Share (2022)



Source (adapted): Grand View Research. 2023. *AI In Healthcare Market Size, Share & Trends Analysis Report By Component (Software Solutions, Hardware, Services), By Application (Virtual Assistants, Connected Machines), By Region, And Segment Forecasts, 2024 - 2030*. https://www.grandviewresearch.com/industry-analysis/artificial-intelligence-ai-healthcare-market. Maximize Market Research. 2023. *Artificial Intelligence in Healthcare Market Size, Growth, Opportunities & Trends: Global Industry Analysis and Forecast (2023-2029).* November. https://www.maximizemarketresearch.com/market-report/global-artificial-intelligence-ai-healthcare-market/21261/. Precedence Research. 2023b. Artificial Intelligence in Healthcare Market Size, Report 2022-2030. February. https://www.precedenceresearch.com/artificial-intelligence-in-healthcare-market. Straits Research. 2023. Artificial Intelligence (AI) in Healthcare Market. https://straitsresearch.com/report/artificial-intelligence-in-healthcare-market.

Appendix 58

The Future of the Healthcare Ecosystem

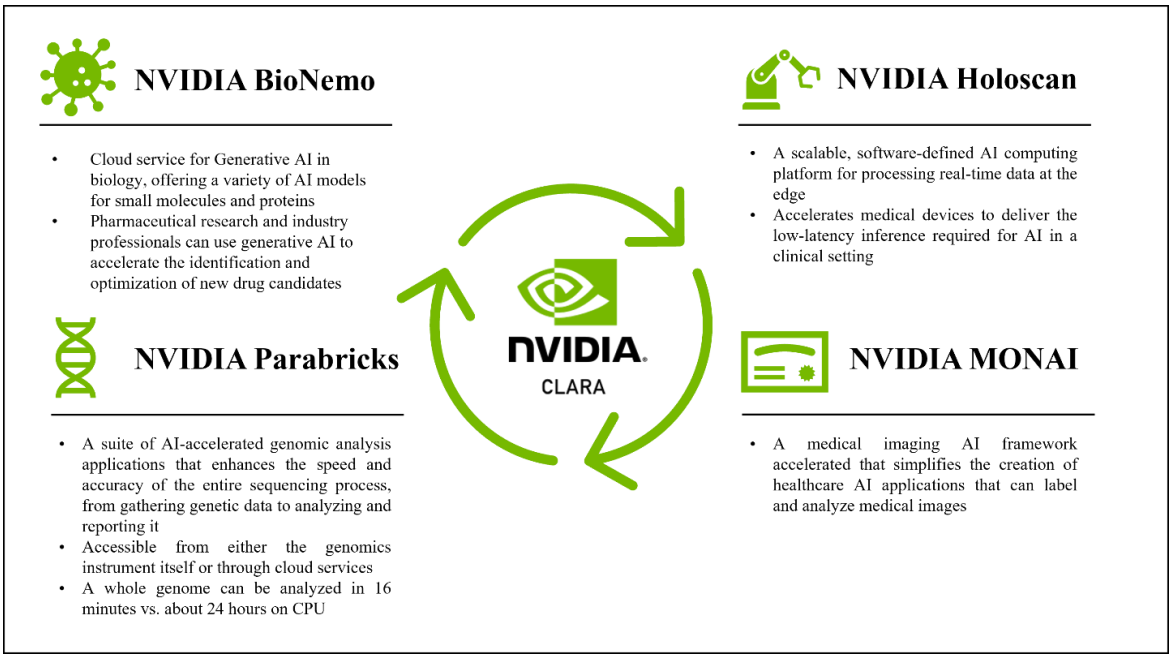


[1] Social care: Social and community networks related to a patient's holistic health.
[2] Home and self-care: Patient engagement, health-focused activities.
[3] Daily life activities: Patient actions enabling wellness, tangential to direct care delivery.
[4] Modalities of traditional care: Direct care administered by clinicians across evolving sites of care.
[5] Financing support: Operational and financial infrastructure of healthcare ecosystem.

Source (adapted): Carlton, Shubham Singhal and Stephanie. 2019. *The era of exponential improvement in healthcare?* 14 de May. https://www.mckinsey.com/industries/healthcare/our-insights/the-era-of-exponential-improvement-in-healthcare#/.

Appendix 59

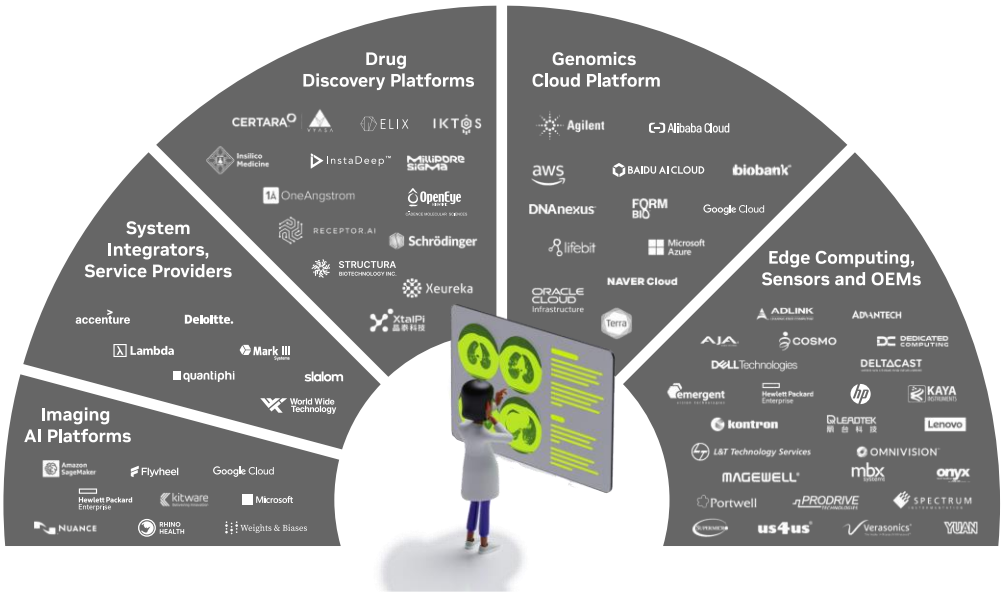
NVIDIA Clara



Source (adapted): Benemann, Kathy. 2023. *100+ Partners Bring NVIDIA Clara AI Healthcare Platform to Enterprises Worldwide.* 21 de March. https://blogs.nvidia.com/blog/nvidia-clara-ai-healthcare-enterprises/.

Appendix 60

NVIDIA Clara's Partner Network



Source (adapted): NVIDIA. n.d. NVIDIA Clara. https://www.nvidia.com/en-us/clara/.