QCRI
معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY

AI Chips: A Technology Race and Geopolitical Tensions

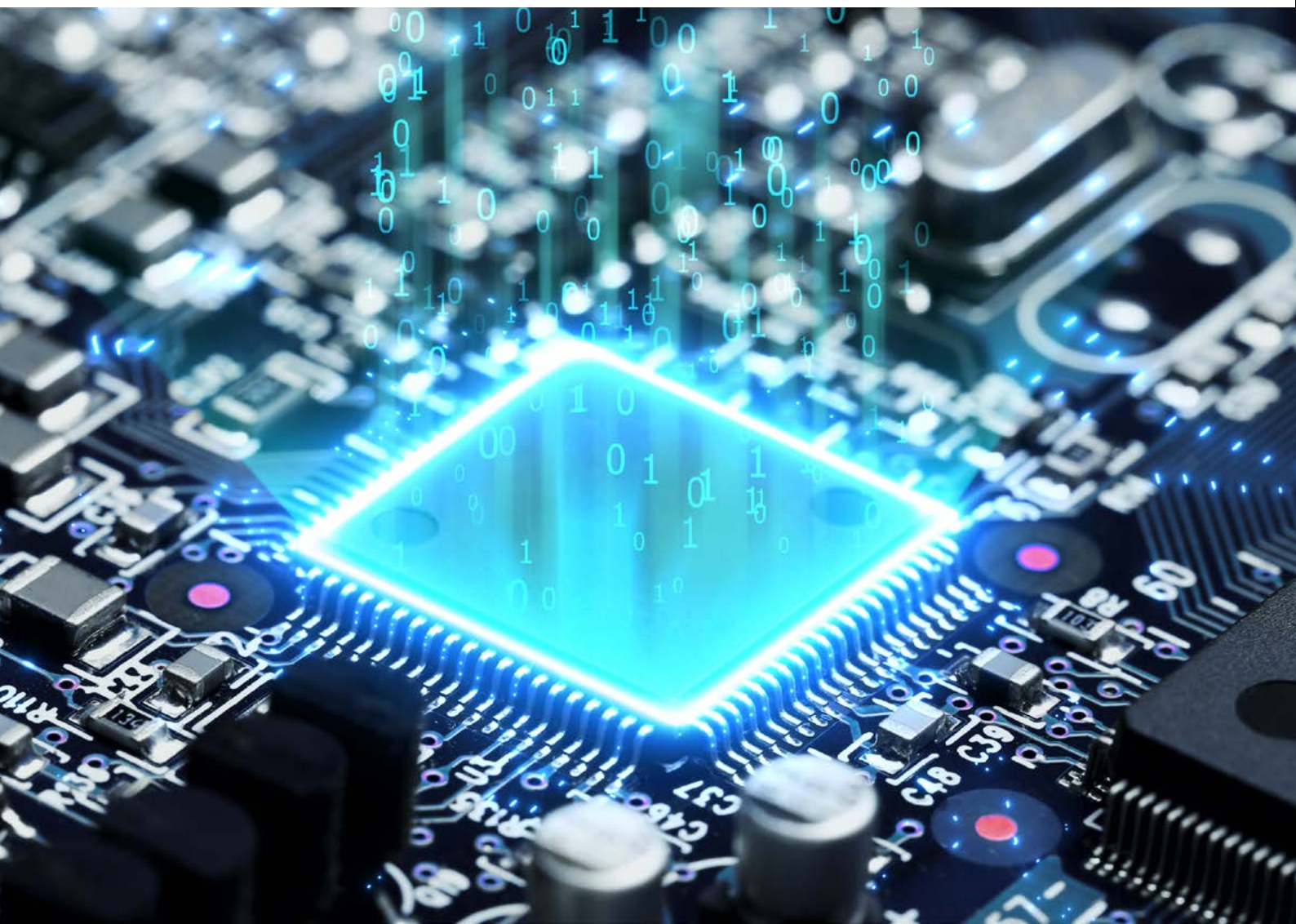


Table of Contents

Executive Summary

This white paper is one in a series of publications produced by the Qatar Computing Research Institute (QCRI) as part of its participation in the National AI Strategy of Qatar and Qatar Digital Government Strategy 2023-2025. The paper discusses AI chips, their role in speeding up AI tasks, and the geopolitical dynamics around them. We provide a comprehensive study of AI chips, which are defined as hardware processors specifically designed to accelerate AI tasks. AI chips play a crucial role in various AI applications, including image recognition, natural language processing, and machine learning. AI chips can be classified as General-Purpose Graphics Processing Units (GPGPUs), Application-Specific Integrated Circuits (ASICs), and Field Programmable Gate Arrays (FPGAs).

The paper discusses in detail the evolution of GPUs, highlighting their transition from initial usage as graphics processing units to becoming powerful computing engines. GPUs have proven to be highly effective in AI applications, particularly in deep learning algorithms, due to their highly parallel architecture and ability to process large amounts of data simultaneously.

The race to develop faster and more affordable AI chips is also discussed as a key area of focus. The growing need for processing power in AI applications has led to the emergence of specialized hardware chips such as Google's Tensor Processing Unit (TPU). TPUs are designed to handle the unique properties and requirements of AI algorithms.

The paper highlights China's rapid rise as an AI superpower and the Chinese government's strong support for AI development. Their significant investment and favorable regulatory policies have contributed to the growth of the AI chip industry in China. Moreover, the active involvement of Chinese tech giants and startups, such as Huawei and Cambricon, has boosted China's goal of becoming a global leader in AI by 2030. Geopolitical tensions surrounding AI chips, particularly between the United States and China, are discussed with a focus on recent developments in the field. The United States, concerned about China's technological advancements, has taken steps to regain its semiconductor manufacturing capabilities.

We discuss the US Chips and Science Act and the EU Chips Act, aimed at promoting semiconductor manufacturing and innovation. The implications and challenges of these acts are discussed, including the need for substantial investments and collaboration between governments, academia, and industry. Regional initiatives in countries like Saudi Arabia, the United Arab Emirates, and India are also mentioned, showcasing their efforts to establish themselves as leaders in AI technology. These countries are investing in research and development, talent acquisition, and infrastructure to support their AI initiatives.

Finally, the paper provides strategic recommendations for Qatar to respond to the AI chip race effectively. These recommendations include hosting a reference AI data center to attract chip manufacturers, starting at the assembly, testing, and packaging (ATP) stage of chip production, leveraging low energy costs as an incentive, and investing in semiconductor education and talent acquisition.

Abstract

Artificial Intelligence (AI) chips are specialized hardware designed to speed up the workload required by AI tasks. These chips play a crucial role in reducing the time of AI model training, enhancing inference efficiency, and enabling real-time data processing in various applications.

This white paper explores the significance of AI chips and the increasing race among technology companies to design AI chips. The paper also examines the geopolitical dynamics around the race to develop or acquire AI chips. It highlights China's rise as an AI superpower, which is mainly driven by governmental support, talent acquisition strategies, and contributions from Chinese tech giants and startups.

Given the significance of AI chips and the resulting geopolitical dynamics, we make some recommendations for Qatar to ensure its technical competitiveness by enhancing computer power capabilities, investing in education and research, and focusing on a single stage in the supply chain.

Introduction

The term "AI Chip" is relatively new, and we use it in this paper to refer to any hardware processor used to accelerate AI tasks including generative AI, image recognition, natural language processing and machine learning in general.

AI chips can be categorized into three types: General-Purpose Graphics Processing Units (GPGPUs), Field Programmable Gate Arrays (FPGAs), and Application-Specific Integrated Circuits (ASICs) designed specifically to optimize the computationally intensive AI workloads. The AI chips enable parallel processing, handle massive datasets, and execute complex mathematical computations with improved energy efficiency [1].

While GPUs have long been the dominant processing unit for accelerating machine learning workloads, recent breakthroughs in AI, particularly the increasing emphasis on the deep learning approach, have sparked new competition among tech giants, with several companies starting to produce AI specific processors. This AI chip competition and the quest for supremacy in AI have created new geopolitical dynamics that may have an impact on the future of AI and its applications [2].

Given this backdrop, the Qatar Computing Research Institute (QCRI) at Hamad Bin Khalifa University (HBKU) has produced this white paper as part of its participation in the National AI Strategy of Qatar and Qatar Digital Government Strategy 2023 – 2025. The paper aims to emphasize the significance of AI chips and assist policymaking in the context of Qatar's AI initiatives.

The rest of the paper is structured as follows. Section 3 investigates the evolution of GPUs and their role in AI; Sections 4 and 5 focus on the AI Chips race and China's rise as an AI superpower; Section 6 discusses some geopolitical dynamics and regional initiatives; Section 7 provides recommendations on how Qatar can strategically respond to the AI Chips race to ensure national technological competitiveness; and Section 8 concludes with a summary.

The Role of GPUs in AI

GPUs Evolution

GPUs evolved from general graphics processors in the 1980s to versatile and powerful computing engines. These processors were initially designed to handle complex graphics rendering, and now play vital role in transforming industries such as gaming, cryptocurrency mining, scientific research, and AI applications [3]. The introduction of 3D graphics in the 1990s marked a big shift in the gaming industry. This development led to the introduction of programmable shaders, which allowed developers to create more realistic and immersive visual experiences [4]. An example of this development was the NVIDIA GeForce 256, a groundbreaking GPU that integrated hardware transform and lighting capabilities, marking a significant milestone in GPU development [5].

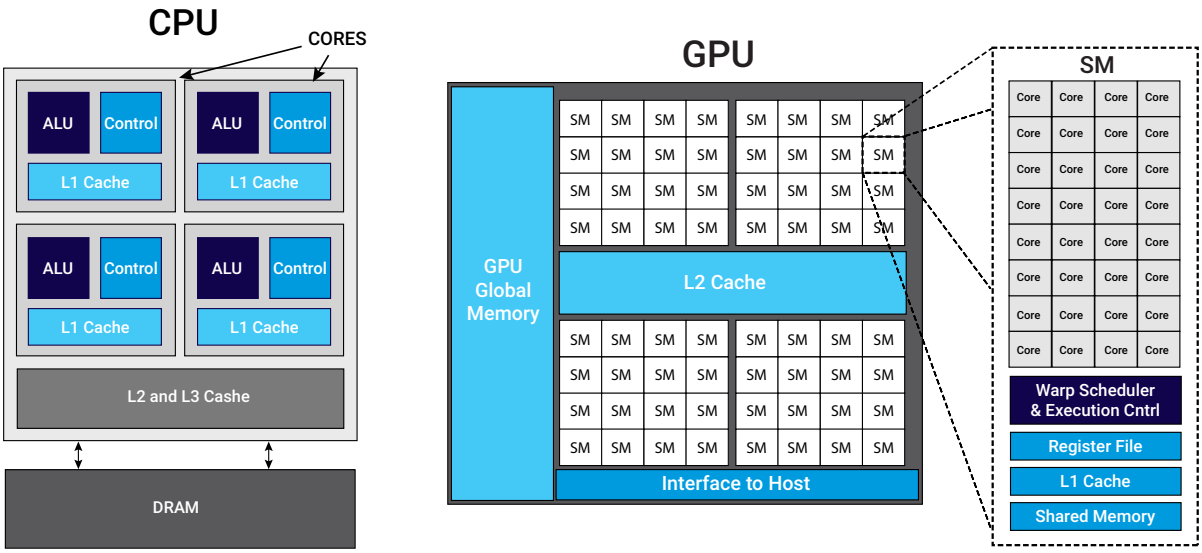
In the early 2000s, researchers and application developers noticed the promise of GPUs for general-purpose computing. This understanding gave rise to General-Purpose GPU (GPGPU) computing, in which GPUs were used to do non-graphics-related tasks [6]. The launch of NVIDIA's CUDA (Compute Unified Device Architecture), for example, revolutionized GPU computing by providing a programming model and tools for using GPUs' massive parallel processing capabilities. Meanwhile, this framework is specifically designed to run computations on NVIDIA's GPUs. As an alternative to CUDA developers use OpenCL (Open Computing Language). OpenCL is an open-source, cross-platform programming language developed by the Khronos Group, and designed to enable code to run efficiently on heterogeneous systems that include both CPUs and GPUs. However, CUDA has formed a strong software ecosystem, making it currently the dominant GPGPU programming model and NVIDIA cards the dominant GPU accelerators used in servers/clusters for AI, especially in the cloud.

[1]https://services.hukoomi.gov.qa/en/about-hukoomi/qatar-digital-government-2020-strategy

GPU Architecture

For decades, successive generations of computing systems demonstrated exponential growth in performance. A combination of factors contributed to this performance growth such as reduction in transistor sizes, improvements in hardware architecture, and enhancements in algorithms and compiler technology [7].

Figure 1 displays a simplified architecture of a CPU and a GPU in block diagram format. Both processors are crucial microprocessors in modern computer systems, capable of performing high-performance processing tasks for a variety of applications.



Block diagram of quad-core CPU

Simplified GPU Architecture with 64SMx32Core units (2,048 cores)

Figure 1: CPU/GPU Architecture Comparison

CPUs were designed to perform sequential operations and have some common features such as: **Cores** - modern computers commonly have from 2 to 64 cores; **Simultaneous multithreading** – refers to the process of delegating operations to several software threads rather than a single core; Cache - CPUs have built-in ultra-fast memory arranged in levels from L1-L3, with L1 being the fastest; **Memory Management Unit (MMU)** - All memory and caching operations are handled by the MMU; **Control Unit** - The control unit directs the CPU's functions. It instructs the RAM, logic unit, and I/O devices on how to respond to the instructions received.

On the other hand, a GPU design targets massive parallelization and allows for large computations to be completed in a fraction of the time required by a CPU. To this end, a GPU consists of hundreds or thousands of compute cores (NVIDIA A100, for example, has over 6000 cores), designed to run single-instruction multiple-thread (SIMT) programs efficiently. These cores are arranged in sectors and referred to as Streaming Multiprocessors (SMs), or Compute Units [7]. A GPU unit could contain over 100 Streaming Multiprocessors (SMs), each consisting of multiple tensor cores and a layer-0 instruction cache. In addition, tensor cores in each SM share an L-1 cache and shared memory (separate or integrated), both significantly faster but smaller than the GPU global memory, which is shared across SMs. Such a "wider" and "flatter" architecture enables high-throughput computation when data moved to the cache/shared memory can be efficiently reused, or when the global memory load/store latency can be hidden by scheduling among groups of threads ("warps" in NVIDIA term) [8]. However, when computation is less intensive or streamlined, the execution easily becomes memory-bound and delivers a much lower FLOPS than the hardware peak capacity.

[2]The current state of the art is 3nm and Intel is building a plant in Ohio that will develop chips at 18 Angstrom. The physical limit is 2 Angstrom

AI Computation with GPUs

The integration of GPUs into AI computation can be attributed to the pioneering work by Geoffrey Hinton's group at the University of Toronto, who used GPUs to train a deep convolutional neural network (CNN) on the ImageNet dataset, which contains millions of images and thousands of classes [9]. Hinton's group realized that the highly parallel architecture of GPUs could be harnessed to accelerate the training process, which until then had been a bottleneck in the field of AI. By adapting their neural network algorithms to utilize the massive number of GPU cores, Hinton's group achieved remarkable speedups in training times. Hinton's breakthrough enabled deep neural networks to be trained in a fraction of the time it previously took. The parallel processing capabilities of GPUs allowed for simultaneous execution of multiple computations, resulting in faster training iterations [10]. This reduction in training time from weeks to mere days was a game-changer for the AI community, as it accelerated the pace of research and development significantly. Researchers could now experiment with more extensive datasets, larger models, and complex architectures, pushing the boundaries of AI advancement.

The success of Hinton's group in utilizing GPUs for training deep neural networks paved the way for the widespread adoption of GPUs in the AI community. Researchers and developers quickly recognized the immense benefits of GPUs in reducing training times and improving overall performance. GPU manufacturers, such as NVIDIA, responded to this demand by developing specialized GPUs tailored for AI workloads, equipped with even more cores and improved memory bandwidth.

The adoption of GPUs in AI computations extended beyond research labs, finding applications in various industries. Companies started leveraging GPU-accelerated AI models for tasks like computer vision, natural language processing, recommendation systems, and more. GPUs enable real-time processing of vast amounts of data, empowering applications like autonomous driving, medical image analysis, and personalized advertising. The parallel processing capabilities of GPUs proved instrumental in handling the complex computations required by deep learning algorithms, making AI-powered solutions more accessible and impactful.

The demands from large AI model training (and more recently, inference as well) have in turn pushed for GPGPU hardware/software design eying scalability of distributed GPU computation. Sample products include the fast NVLink network and the NCCL library, both NVIDIA proprietary systems for inter-GPU communication.

For large language models (LLMs) a simple heuristic has emerged to estimate the time needed to train the model. For example, if P is the number of parameters and D is the number of data tokens and T is the number of TFLOP/S that a GPU system can support then the time taken to finish the computation is 6PD/T. A detailed survey of LLMs and the computing power used in training and inference can be found in [24].

AI Chips Race

The rising demand for processing power for various AI applications has accelerated the race to develop and acquire faster and more affordable AI chips. Tech giants from around the world have successfully entered the race. In this section, we will explore AI chips and their manufacturing process, as well as provide some insight into the AI chip rivalry.

AI Chips

The term "AI chips" first emerged in the mid-2010s, when tech companies started designing specialized hardware chips targeted for AI applications. These chips were designed to handle the unique properties of AI algorithms, such as their high parallelism and matrix operations, to enable faster and more efficient computations.

The AI accelerating chips can be broadly classified into three categories: General Purpose Graphics Processing Units (GPGPUs), Field Programmable Gate Arrays (FPGAs), and AI-specific Application-Specific Integrated Circuits (ASICs) [3]. The GPUs, as explained in the next section, have found extensive use in AI due to their highly parallel processing capabilities. GPUs excel at performing multiple calculations simultaneously, making them ideal for training and inference tasks in deep learning.

Field-Programmable Gate Array (FPGA), which offers flexibility in customizing hardware architectures, can be reprogrammed to adapt to different AI models and algorithms, making them versatile for a wide range of applications. Companies like AMD-Xilinx have released FPGA-based AI accelerators, such as the Alveo series, which provide high-performance computing for AI workloads. However, FPGAs are mainly used for experimentation and testing purposes.

The AI-specific Application-Specific Integrated Circuits (ASICs), on the other hand, are custom-designed chips built specifically for AI tasks, offering exceptional efficiency and performance. These chips are tailored to accelerate specific AI algorithms and provide significant speedups compared to general-purpose processors. Examples of ASIC chips include Google's Tensor Processing Units (TPUs), which are designed to maximize deep learning activities, and Intel's Neural Network Processor series (NNP-I for inference and NNP-T for training). Most AI chip designers and tech giants such as NVIDIA, IBM, AMD, Amazon, Apple, Google, and many others are focusing on ASICs now to accelerate AI training and inference tasks.

The importance of AI chips cannot be overstated. Their specialized design and optimization enable faster training and inference times, driving the progress of AI research and applications. AI chips significantly reduce the time required for training complex neural networks, which otherwise could take weeks or even months. Moreover, AI chips enhance the efficiency of AI inference, enabling real-time processing of data. This is crucial for applications like autonomous vehicles, medical diagnostics, and natural language processing. AI chips empower these applications by providing the computational power needed to process vast amounts of data and make real-time predictions. It has been reported in the state-of-ai report that the compute requirements for large-scale AI experiments have increased >300,000x in the last decade [4].

[3]For example the convolution operation, an important computation for computer vision AI is known to be "embarrassingly parallel" and can be sped up with AI Chips

Chips Production

Electronic chips are the most complicated devices ever created by humanity - IBM has recently announced the design of cutting-edge semiconductor chips with the tiniest transistors (2nm) ever created, allowing for a staggering 50 billion transistors to be packed into a chip the size of a fingernail.

Chip manufacturing is a complex process involving numerous companies from different parts of the world. This process is divided into three major stages: Design, Fabrication, and Assembly, Test, and Packaging (ATP). The manufacturing process is supported by a broad network of suppliers for materials, design tools, and manufacturing equipment. Figure 2 depicts this process along with the research and development (R&D) component which is vital for all stages.

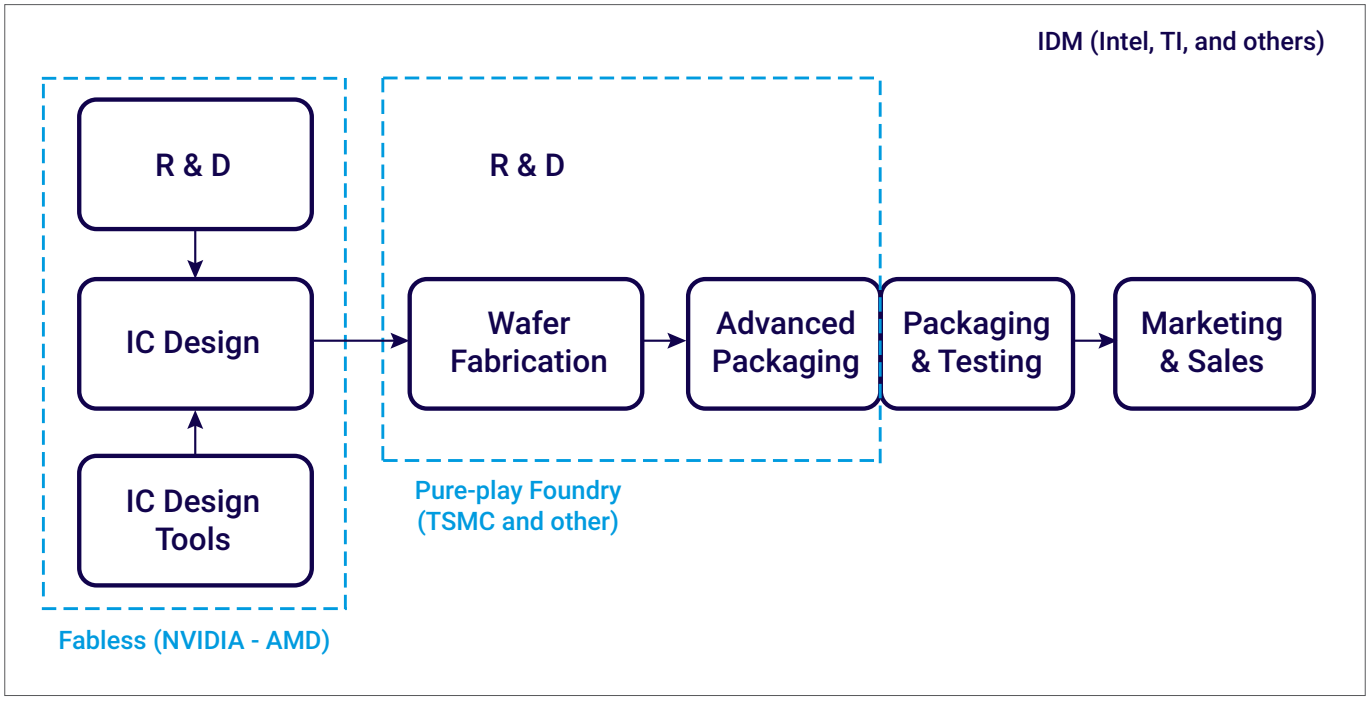


Figure 2: The Foundry Model (pioneered by Morris Chang)

In this model, Fabless companies refer to semiconductor companies that innovate, design, and market microchips while outsourcing the fabrication, packaging, and testing to third-party partners. Fabless companies partner with foundries, such as TSMC and GlobalFoundries, to print chip designs on wafers, and outsource testing and packaging to Outsourced Semiconductor Assembly and Testing (OSAT) service providers. Clients of fabless companies include original equipment manufacturers (OEMs) and end-user device makers who incorporate microchips into their devices. Integrated Device Manufacturers (IDMs) such as Intel and Texas Instruments are firms that both design and manufacture their chips.

Each stage in the production process gets input from various suppliers and produces output that can be used for the next stage or as a final product. ASML in the Netherlands, for example, is a major provider of semiconductor production equipment with its biggest customers being TSMC, Intel, and Samsung.

[4]https://newsroom.ibm.com/2021-05-06-IBM-Unveils-Worlds-First-2-Nanometer-Chip-Technology,-Opening-a-New-Frontier-for-Semiconductors

[5]https://www.youtube.com/watch?v=r_8XClnnvIk

Table 1 summarizes the main inputs and outputs for each phase:

Phase/IO	Input	Output
Design	▸ Design requirements and specifications ▸ Design software tools ▸ Expertise and knowledge of chip design principles	▸ Detailed design layouts, and schematics ▸ Design verification and validation details ▸ Intellectual property ▸ Design files for fabrication
Fabrication	▸ Detailed design layouts, and schematics ▸ Manufacturing equipment ▸ Manufacturing materials, such as silicon wafers, chemicals, and gasses ▸ Cleanroom facilities and infrastructure ▸ Quality control and testing protocols	▸ Manufactured chips or ICs ▸ Packaged chips ready for assembly and testing ▸ Testing data ▸ Process documentation and records
ATP	▸ Packaged chips from the fabrication phase ▸ Assembly and packaging materials, such as lead frames, substrates, and bonding wires ▸ Testing equipment and software ▸ Test specifications and procedures	▸ Ready-to-ship devices for distribution and use ▸ Tested and validated chips with known good die (KGD) ▸ Quality control and reliability reports

Table 1: Chip Production I/O

Table 2 shows some top companies in the semiconductor sector by country:

Country/Type	Fabless	Foundry/IDM	OSAT
United States	Nvidia, AMD/ Xilinx, Qualcomm	Intel, TI	Amkor
Taiwan	Realtek, MediaTek, Novatek	TSMC, UMC	ASE, SPIL, PTI, ChipMOS, KYEC, Chipbond
China	HiSilicon, Bitmain, Cambricon	SMIC, Sapphire	JCET, TFME, HUATIAN
Japan	–	Sony, Renesas	Shinko
South Korea	–	Samsung Electronics	STATS, ChipPAC
United Kingdom	ARM	–	–
Malta	–	GlobalFoundries	–
Singapore	–	–	UTAC
Malaysia	–	–	Globetronics
Philippines	–	–	IME

Table 2: Key Players in Chips Production

[7]https://www.fitchratings.com/research/corporate-finance/fitch-affirms-asml-at-a-outlook-stable-05-04-2023

From GPUs to AI Chips

GPUs have been dominant hardware tools for accelerating AI systems, especially for training deep neural network (DNN) models, which are computationally intensive by design. Nvidia has unveiled its H100 GPUs, claiming a 7X performance boost over its preceding A100 GPU [11]. The H100 GPUs triple the floating-point operations per second (FLOPS) of double-precision Tensor Cores, delivering 60 teraflops of FP64 computing for HPC. AI-fused HPC applications can also leverage the H100's TF32 precision to achieve one petaflop of throughput for single-precision matrix-multiply operations while requiring no code changes.

Other tech giants have been significantly investing in AI-ASICs (AI chips) since they are specifically designed for AI computations. Google, for example, has developed the Tensor Processing Units (TPU) which can handle massive matrix operations used in neural networks at fast speeds. Google began discussing the development of their own AI processor in 2008, but the serious work began in 2013. They recently released the TPU v5e chip, which beats the TPU v4 by up to 2x in terms of generative AI model training and inference performance per dollar [4].

Amazon unveiled Trainium, the second generation of its own AI processor, claiming significantly lower costs than Nvidia's A100 chips [12]. These specialized chips are used in their own AI research and applications, as well as in their computing cloud, which provides third-party customers with cloud services.

AMD-Xilinx and Intel dominate the FPGA boards, which are more customizable than ASICs [12]. However, employing FPGAs requires a lengthy and complicated development procedure. AMD offers a comprehensive multi-node portfolio to address requirements across a wide set of applications. Intel has recently released the Agilex M-Series FPGAs, claiming that they are the fastest FPGAs with in-package high bandwidth memory. ASICs and FPGAs often outperform GPUs in terms of inference efficiency [14].

In China and other countries such as South Korea, tech companies have also joined the race. Baidu has recently released the second generation of Kunlun AI chips [15]. Huawei also revealed two AI chips, the Ascend 310 and Ascend 910. LG has developed its own AI Chip with a proprietary LG Neural Engine to improve the processing of deep learning algorithms. Samsung also plans to develop a 4 nm AI accelerator, according to a press article.

The global AI chip race is not limited to tech giants alone. Startups worldwide are actively contributing to the development of advanced AI chips, driving innovation, and pushing the boundaries of artificial intelligence. Chinese startups like Cambricon and Biren, US startups like SambaNova and Cerebras, and growing players in other countries like the United Kingdom, Japan, Europe, and South Korea are participating in this race.

[7]https://www.nvidia.com/en-us/data-center/h100/#:~:text=H100%20triples%20the%20floating%2Dpoint,of%20FP64%20computing%20for%20HPC.

[8]https://www.kedglobal.com/korean-chipmakers/newsView/ked202311210015

Table 3 shows some of the top startup companies and their AI focus with the funding they accumulated.

Company	Country	Funding	AI focus
SambaNova	USA	$1.1B	Full-stack AI platforms
Cerebras	USA	$720M	The pioneering Wafer-Scale Engine (WSE)
Graphcore	UK	$692M	AI accelerators and machine learning
Tenstorrent	Canada	$334.5M	AI processors for faster training algorithms
Cambricon	China	$200M	AI chips
Prophesee	France	$111.4M	Computer Vision processing
Rebellions.ai	South Korea	$86.5M	AI accelerators
Axelera	Netherlands	$68.7M	AI acceleration cards

Table 3: AI World Leading Startups - Source: ai-startups.org

The massive investments gathered by these startups demonstrate the realization of the critical role AI chips will play in shaping the future of artificial intelligence. As these startups continue to innovate and produce specialized AI chips, the global AI chip ecosystem will become more diverse and competitive.
These achievements will pave the path for enhanced AI capabilities, increasing AI adoption across industries and opening new avenues for AI-driven applications.

From a business perspective, the commercialization of AI chips has been subject to their degrees of general-purpose functionality. GPUs and FPGAs have long been widely commercialized, while ASICs are mainly used in-house due to their specialized design and low production volume [3].

China's Rise as AI Superpower

China's rapid elevation as a global leader in AI technologies is reshaping the landscape of the industry. This section will explore the key drivers behind China's rise as an AI superpower, including government support, talent acquisition, and the significant role of tech giants and startups.

National Strategies and Government Support

National strategies and support from the government are boosting China's AI progress. On the strategy side, the ambitious "Next Generation Artificial Intelligence Development Plan" of the Chinese government seeks to make China a global leader in AI by 2030 [16]. The plan emphasizes the importance of AI technologies and the need to develop indigenous capabilities. AI businesses and research institutes have benefited from government-backed financing programs and efforts such as the National Laboratory for Parallel Computing and the National Supercomputing Center. This supports the growth and global competitiveness in AI hardware and technologies. China comes at the top worldwide in supercomputers as shown in Figure 3.

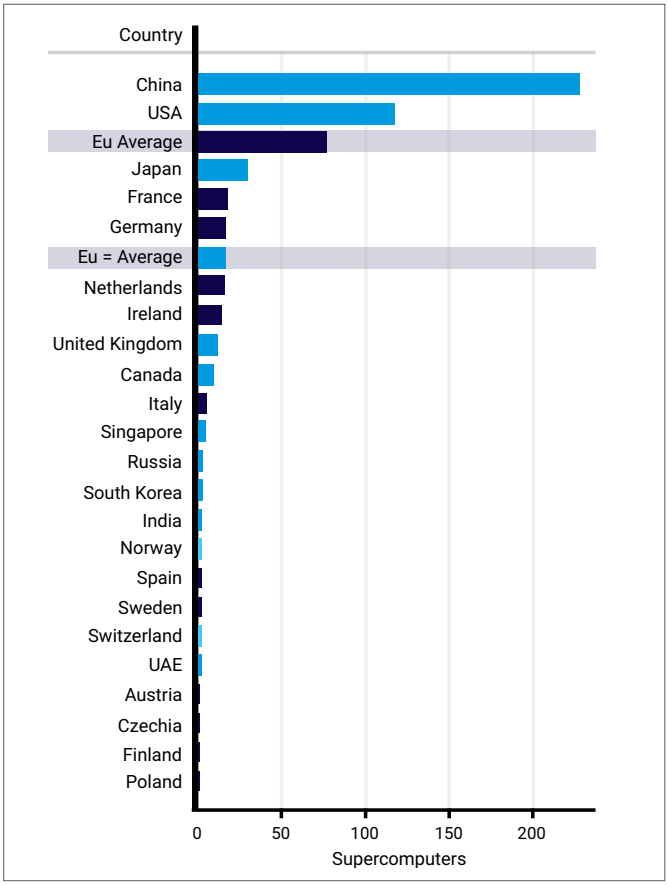


Figure 3 Supercomputers by country
Source: stateof.ai 2022

Tech giants and startups

Chinese tech giants such as Huawei and SMIC have made substantial advances in the semiconductor industry, further pushing China's rise as an AI superpower. Huawei has extended its interests to encompass AI chip design. Huawei's Ascend series of AI chips has received praise for its remarkable performance and energy efficiency. The Ascend 310 is designed to give an unrivaled performance that outputs 16 TOPS@INT8 and 8 TOPS@FP16 while consuming only 8 W of power in a typical setup; thereby making it a serious competitor to established players such as Nvidia. Huawei has made substantial investments in AI research and development, focusing on areas like AI infrastructure, cloud computing, and AI applications.

Semiconductor Manufacturing International Corporation (SMIC), on the other hand, has established itself as an important participant in the semiconductor sector. The superior chip manufacturing capabilities of the corporation have enabled the production of cutting-edge AI processors.

China's vibrant startup ecosystem is another key driver of its progress in the AI industry. Startups such as Biren, a rising star in the AI chip market, have grown in popularity due to their unique solutions and technology developments. Biren focuses on edge computing and AI inference. The company's AI processors provide excellent performance and energy efficiency, meeting the growing need for AI applications across multiple industries. Cambricon and Horizon Robotics, two other AI chip firms, have also made substantial contributions. Cambricon's MLU series AI processors have acquired commercial momentum because of their high-performance processing capabilities. Horizon Robotics works on AI chips for autonomous driving, robotics, and other applications. However, the US government's sanctions on Chinese companies, including restrictions on access to American technologies, have posed challenges to their AI chip development plans.

Talent Acquisition

China's rise as an AI superpower is facilitated by its ability to attract and retain top talent in the field. The country's investment in education and research institutions has cultivated a pool of skilled professionals, strengthening its technological capabilities.

Chinese tech giants and startups have also strategically collaborated with international partners to enhance their expertise and gain access to cutting-edge technologies. These collaborations have allowed Chinese companies to tap into global knowledge networks and accelerate their AI and AI chip development.

[9]https://e.huawei.com/en/products/computing/ascend

Geopolitical Issues

Background

Semiconductors have become a cornerstone in the global economy, powering everything from appliances and automobiles to smartphones, computers, and advanced AI systems. The importance of semiconductors in the global economy became even more apparent during the COVID-19 pandemic when supply chain irregularities caused a global chip shortage.

Given the global nature of supply chains and the important role that semiconductors play in critical technologies, it is not surprising that they have become the focus of the most recent geopolitical tension between the United States and China. However, this kind of geopolitical tension is not new. In the 1970s, the Soviet Union responded to the U.S. technological advancement by establishing a domestic chips industry and building Zelenograd, a duplicate of Silicon Valley [2].

After several years of heavy investments, they failed to catch up with the U.S. for various reasons:

1. **Their inability to develop indigenous new technologies**
2. **Huge challenges in terms of mass production of chips**
3. **The rapid technology advancements of the U.S. based firms [3]**

In the 1980s, Japan entered the race for semiconductor supremacy. Their project was built on a solid strategy targeting mass-consumer products. Hitachi, Toshiba, NEC, and Fujitsu emerged as Japanese tech giants with significant governmental support and large R&D investments. The Japanese rapid advancement in the semiconductor industry raised concerns in the United States. The US government perceived Japan's rise and market dominance as a threat to its semiconductor industry and took several actions to address the Japanese dominance and trade imbalance, including imposing trade restrictions, investigating Japan's trade practices, and negotiating agreements with Japan to open its markets. For example, in 1986 the U.S. imposed a 100% tariff on certain Japanese semiconductor products.

Taiwan is currently at the heart of the chip conflict. The tiny island produces about 90% of the world's most advanced semiconductors. Taiwan Semiconductor Manufacturing Company (TSMC) is the world's top chip foundry, focusing on efficient wafer fabrication. The company produces chips for major chip designers such as Apple, NVIDIA, AMD, and many others. The TSMC fab model was very successful due to the focus on wafer fabrication supported by heavy investment in research and development.

These changing global dynamics have resulted in unprecedented regional concentration in chip production, particularly in the semiconductor industry's high-end spectrum. American and European policymakers are now attempting to reverse this trend and reintroduce chip manufacturing to the United States and Europe. Next, we will shed some light on the US and European Chip Acts.

US Chips and Science Act

The US Chips and Science Act, which was signed into law in August 2022, is part of President Biden's "Investing in America" agenda. The Chips Act aims to promote manufacturing and innovation in the United States, strengthen economic and national security, and enhance the U.S. competitiveness in the semiconductor industry. It allocates $280 billion in spending over the next ten years, with the majority ($200 billion) going towards scientific research and commercialization. Additionally, $52.7 billion is designated for semiconductor manufacturing, research and development (R&D), and workforce development, while $24 billion is allocated for tax incentives for chip production [26]. The act also includes $3 billion for programs focused on cutting-edge technologies and wireless supply chains. The Department of Commerce receives most of the $52.7 billion fund, with $39 billion allocated for the CHIPS Program Office and $11 billion for the CHIPS Research and Development Office. The Defense Department, Department of State, and National Science Foundation also receive a portion of the funding. Furthermore, the act provides a 25% investment tax credit for qualified investments in advanced manufacturing facilities that primarily manufacture semiconductors or semiconductor manufacturing equipment. The Department of Commerce has set strategic goals to be achieved by 2030, including the establishment of large-scale logic chip fabs, advanced packaging facilities, and increased production capacity for memory chips and current-generation chips.The Department of Commerce agenda also highlights some guardrails to strengthen national security, such as restrictions on using funds or engaging in joint research with other countries including China, North Korea, and Iran [27].

The Chips Act Implications:

The U.S. CHIPS and Science Act has several implications for the American semiconductor industry that can be briefly summarized as follows:

1. **Boosting U.S. Semiconductor Production:**
The Chips Act aims to increase the U.S. semiconductor manufacturing capacity and reducing dependence on foreign sources, particularly East Asia. Priorities in manufacturing are given to critical semiconductor devices such as AI chips.

2. **Encouraging Private Sector Investment:**
The act provides significant incentives, like tax credits, to attract private sector investment in semiconductor research, development, and manufacturing. Semiconductor giants such as Intel and TSMC have already started building new Fab facilities in the U.S.

3. **Advancing Research and Development:**
One of the main goals of the Chips Act is to ensure the United States remains at the forefront of technological advancements in the industry. A substantial portion of the allocated budget is geared to R&D. This will support the development of new technologies like AI, edge computing, and wireless communication, enhancing the U.S. competitiveness in the global market.

4. **Addressing Labor Challenges:**
The act focuses on STEM education and workforce development to address the shortage of skilled workers in the semiconductor industry. By investing in education and training programs, the act aims to create a pipeline of talent to support the growing semiconductor manufacturing sector.

The Chips Act Challenges:

The CHIPS and Science Act has several positive implications, as stated in the previous section, and it is also expected to face many challenges. While it is outside the scope of this paper to address all possible challenges, we highlight some of them as follows:

1 **Implementation Time:**
Building semiconductor manufacturing capacity is a complex and time-consuming process, therefore the impact of the Act will not be immediate. Establishing new manufacturing facilities, training staff, and increasing production will take several years.

2 **Global Trade Relations:**
One of the proclaimed goals of the CHIPS Act is to bring semiconductor manufacturing back to the United States. At the same time, policymakers in the United States are taking steps to limit other countries' access to cutting-edge technologies, including China. This would prompt retaliatory moves from China and other countries that currently dominate the semiconductor industry, potentially disrupting the global supply chain and causing market volatility in the coming years. [28]

3 **Effectiveness of Incentives:**
While the act offers incentives like tax credits to stimulate private-sector investment, its efficacy is not guaranteed. It is unclear if these incentives are adequate to compensate for the advantages provided by offshore manufacturing. Furthermore, it remains to be seen whether the incentives will attract foreign investment and boost U.S. semiconductor manufacturing.

4 **Impact on Global Competitiveness:**
Some critics are concerned that the emphasis of the Act on domestic production may limit the capacity of the United States to compete globally. They suggest that collaborating with overseas partners would be more effective in terms of technological advancement and supply chain stability.

The EU Chips Act

The EU Chips Act, which came into effect in September 2023, has three main pillars:

1. **The Chips for Europe Initiative**
2. **Security of Supply**
3. **Monitoring and Crisis Response** [25]

- The Chips for Europe Initiative aims to enhance Europe's technological leadership by bridging the gap between research and industrial activities. This initiative will be carried out through the Chips Joint Undertaking and will receive direct EU funding, which is expected to be supplemented by monies from Member States. The funding will be used to construct innovative manufacturing lines, develop a cloud-based design platform, establish competence centers, and set up funds to ensure start-ups and SMEs have access to financing.

- The second pillar focuses on the security of the supply of chips. It establishes a framework for Integrated Production Facilities and Open EU Foundries, which contribute to the security and resilience of the EU's semiconductor manufacturing ecosystem.

- The third pillar includes a coordination structure between Member States and the Commission to improve collaboration, monitor semiconductor supply and demand, forecast shortages, and implement crisis measures as needed. A semiconductor alert system has already been established to detect problems in the semiconductor supply chain. More details can be found on the EU Digital Strategy website.

India Initiatives

Recognizing the significance of semiconductors and electronics in the national and global economies, the Indian government released several initiatives in the past decade such as: Make in India announced in 2014; The Production Linked Incentive (PLI) scheme, which aims to promote domestic manufacturing of electronic products; the Electronics Manufacturing Clusters (EMCs), which offer infrastructure and facilities for electronics manufacturing; and the National Policy on Electronics (NPE), which sets a target of creating a $400 billion electronics manufacturing industry by 2025.

The National Policy on Electronics was announced in 2019 (NPE 2019) and is one of the very ambitious initiatives in the field. The main objective of this policy is to position India as a global hub for Electronics System Design and Manufacturing (ESDM) and to foster an environment in which this industry can compete globally. One of the key strategies of NPE 2019 is oriented towards establishing semiconductor wafer fabrication facilities and the ecosystem for chip design and manufacture. The NPE 2019 is accompanied by several initiatives and a detailed policy and procedural system.

On September 21, 2022, the Indian Cabinet approved significant revisions to India's Semiconductor and Display Manufacturing Ecosystem Development Program. The updated program offers consistent fiscal support of 50% of project costs across all technology nodes for the building of semiconductor factories in India. More details on this can be found on India's Ministry of Electronics and IT website.

[10] https://digital-strategy.ec.europa.eu/en/factpages

[11] https://www.meity.gov.in/esdm/policies

Regional Initiatives

For the regional initiatives, we can refer to the recent developments in Saudi Arabia and the UAE.

Saudi Arabia

Saudi Arabia has launched an ambitious strategy with a big objective to becoming a global leader in AI by 2030.

Their ambitions include training 20,000 specialists, launching 300 AI firms, and securing $20 billion in domestic and international investment. The Saudi Data and Artificial Intelligence Authority (SDAIA) was founded in 2019 to oversee the country's data and AI initiatives. SDAIA is working with Huawei on its National AI Capability Development Program, with the goal of tapping into the country's estimated $4 to $5 billion data and AI economy. According to information from their website, SDAIA is investing $18 billion in data center building, with the goal of reaching 1300 MW of capacity by 2030.

The Saudi energy infrastructure is highly stable and cost-effective, making it an ideal option for hosting data centers to meet the expanding global demand for data center facilities. On the other hand, the Green Saudi project, which aims to meet 50% of domestic energy demand with renewable energy, is seen as a big move for reducing carbon emissions, and may attract more investments in data centers. More on this can be found in the special report "The Future of Data Centers in the Middle East and Africa".

Saudi Arabia is also investing in future skills development through initiatives like the Future Skills Initiative, which has already trained over 40,000 individuals in disciplines like IT, AI, and cybersecurity.

United Arab Emirates

The United Arab Emirates (UAE) has been in the forefront of adopting and integrating AI into numerous sectors, with plans to become a worldwide AI leader by 2031.

The "UAE Strategy for Artificial Intelligence" was launched in 2017 with the goal of leveraging AI to improve people's quality of life, increase government efficiency, and accelerate economic growth.

G42, Abu Dhabi based firm, and Cerebras Systems announced a strategic partnership in July 2023, with a focus on AI and Large Language Models (LLMs). The two companies are building a cluster of nine supercomputers, Condor Galaxy, with a total capacity of 36 exaFLOPs. The first of the nine AI supercomputers, Condor Galaxy 1 (CG-1), has already been deployed and performs at 4 exaFLOPs with 54 million cores, making it one of the world's largest AI supercomputers.

CG-1 was used to train Jais, a 13 billion parameters large language model for Arabic. Cerebras uses two technologies to accelerate the data-intensive training of large AI models. The first is a Wafer-Scale cluster, a novel design that connects up to 192 CS-2 units and allows them to operate as a single logical accelerator. The Wafer-Scale cluster separates memory from computing to increase the memory available to AI models from gigabytes (with graphics processing units) to terabytes. The second technology is Weight Streaming, which uses the hardware's computation and memory features to distribute work by streaming a model one layer at a time in a purely data-parallel fashion.

[12]https://www.meity.gov.in/esdm/Semiconductors-and-Display-Fab-Ecosystem

[13]https://ai.sa/index-ar.html

[14]https://sdaia.gov.sa/en/SDAIA/about/Pages/About.aspx

[15]https://argaamplus.s3.amazonaws.com/d96e6f66-ad76-45b8-8e4f-20fe746124b9.pdf

[16]https://www.cerebras.net/press-release/cerebras-and-g42-complete-4-exaflop-ai-supercomputer-and-start-the-march-towards-8-exaflops#:~:text=%E2%80%9CG42's%20strategic%20partnership%20with%20Cerebras,require%20extensive%20AI%20training%20capacity.

Recommendations for Qatar

Given that microchips technology in general, and AI chips in particular, have become an essential component of modern life, Qatar should carefully evaluate where it may fit into the microchips supply chain other than as a technological consumer. Qatar can respond strategically to the AI/Chips competition by pursuing technological competitiveness in the field of AI technology.

Recommendation

We highlight four critical areas and strongly recommend their consideration

I. Host a reference AI data center:

Qatar should commission the establishment of a world-class AI-centric data center that will serve as a reference for the region and the rest of the world. While the majority of the data center's cutting-edge physical components will be imported, Qatari local talent is critical to the data center's overall design and operation. One might ask how Qatar's expertise can be incorporated into the design of the data center.
This is a fair question that can be further researched; nonetheless, we can provide two simple answers:

- **Qatar's experience with "district cooling" can be applied to the physical design of the data center.**
- **The data center should be designed to support applications of significant relevance to Qatar, such as energy, media, logistics, and sports.**

II. Start at the ATP stage:

Start at the ATP stage: As already mentioned, the semiconductor ecosystem is divided into three stages: **design, fabrication, and ATP (Assembly, Testing and Packaging).** ATP is as equally significant as the other two stages in chip manufacturing and does not necessitate cutting-edge technologies. Many countries, like India, Singapore, and Indonesia, have started down this path, and prominent semiconductor companies are setting up ATP facilities in these locations. Micron, for example, is setting up a packaging facility in India. Qatar may do the same by collaborating with a prominent firm to build an ATP facility. On the other hand, they can look at chips used in severe environments (such as solar system controllers in the desert), which may necessitate specialized packaging, which Qatar can provide. Also, we must emphasize that ATP does not have to begin with cutting-edge chips. In fact, chips designed for 20nm to 40nm technologies are in high demand.

III. Attracting Semiconductor Players with Low Energy Costs:

Semiconductor manufacturing is an energy-intensive process, with electricity expenses accounting for a significant portion of a manufacturer's overhead costs. High energy costs can severely impact the profitability and competitiveness of semiconductor companies, making countries with lower energy costs particularly attractive for establishing manufacturing facilities. Qatar is strategically positioned to leverage its abundant natural gas and low energy costs to attract major players in the semiconductor industry. With electricity costs approximately one-third of Taiwan and one-ninth of the United States, Qatar possesses a significant competitive advantage that can make it an appealing destination for semiconductor manufacturers seeking to optimize their energy-intensive operations.

IV. Forging Qatar's Path to Semiconductor Excellence in Education:

Qatar University and Hamad Bin Khalifa University (HBKU) are both esteemed institutions in Qatar, and at least one of them must have an internationally recognized program in computer hardware and semiconductor studies. There are several programs at leading universities including at Purdue University that can serve as an excellent example to inspire and guide the development of such programs in Qatar. The semiconductor program at Purdue University is renowned for its comprehensive and cutting-edge approach to semiconductor research, education, and collaboration with industry. This program has gained recognition for its strong interdisciplinary focus and its ability to address the entire spectrum of semiconductor research and development, from materials and design tools to circuit design, architecture, fabrication, and advanced packaging integration. One of the key strengths of the Purdue program is its emphasis on research. The university possesses a critical mass of researchers dedicated to semiconductor studies, enabling them to undertake impactful and large-scale interdisciplinary research projects. Purdue leads multiple Semiconductor Research Corporation-funded interdisciplinary research programs, which demonstrates its strong connections to industry and its commitment to pushing the boundaries of semiconductor innovation.

[17]https://engineering.purdue.edu/semiconductors/semiconductors-at-purdue

Qatar Competitive Advantages

Qatar, a rapidly growing economy in the Middle East, is strategically positioned to leverage its modern infrastructure and natural resources to enter the microchips industry. We can identify several competitive advantages that Qatar possesses, that set it apart from other regional countries and making it an ideal location for this business:

I. Supportive Government Initiatives:

The Qatari government has launched various initiatives to diversify the economy and promote high-tech industries. Programs such as Qatar National Vision 2030, Qatar Industrial Strategy 2018-2022, and Qatar AI Strategy provide a roadmap for economic development and prioritize sectors like AI, Technology, and Innovation.

II. Advanced Infrastructure:

Qatar has invested significantly in developing world-class infrastructure, including state-of-the-art transportation networks, modern logistics facilities, advanced telecommunications systems, and reliable utilities. This infrastructure supports the efficient movement of goods and enables seamless business operations.

III. Industry Friendly Policies:

Qatar has implemented business-friendly policies and regulations, fostering a favorable investment climate. The government actively supports and encourages foreign direct investment, offering incentives, tax exemptions, and streamlined procedures for establishing and operating businesses in the country.

IV. Strategic Geographic Location:

The location of Qatar in the heart of the Gulf region makes it a gateway between the East and the West, facilitating easy access to global markets and enhancing trade opportunities for semiconductor and microchip products.

V. Geopolitical Stability:

Qatar's geopolitical stability and investor-friendly policies provide a secure and conducive environment for long-term investments. This stability, coupled with the low energy costs, is an attractive proposition for semiconductor manufacturers looking for a reliable base of operations.

VI. Abundant Energy Resources:

Qatar possesses significant energy reserves, particularly natural gas. This abundant and reliable energy supply can be leveraged to power semiconductor manufacturing facilities, ensuring a cost-effective and sustainable source of energy.

Potential Challenges and Mitigation

While Qatar's competitive advantages present a compelling opportunity, some challenges need to be addressed to attract semiconductor players successfully:

Skilled Workforce

The semiconductor industry requires a highly skilled workforce. Qatar needs to invest in education and training programs to develop a local talent pool capable of meeting the industry's demands or to recruit a talent pool. Collaborations with established semiconductor hubs and universities can facilitate knowledge transfer and skill development.

Supply Chain Integration

The semiconductor industry relies on a complex global supply chain. Qatar needs to develop partnerships and establish strong logistics networks to ensure a seamless flow of raw materials, equipment, and finished products.

Innovation and Intellectual Property Protection

To attract semiconductor partners, Qatar must demonstrate robust intellectual property protection and promote a culture of innovation. Strong legal frameworks and a reliable enforcement system will instill confidence in companies considering investments in the country.

Conclusions

In conclusion, this report highlights the significance of AI chips in the field of artificial intelligence and the intensifying competition among tech giants to develop their own AI accelerators. The evolution of GPUs from graphics rendering to powerful tools for AI computation has revolutionized the AI industry, enabling faster training times and more advanced AI applications.

The geopolitical tensions surrounding AI chips are influenced by the global nature of the supply chain and the scarcity of compute resources. China's rise as an AI superpower, driven by government support and the contributions of tech giants and startups, poses challenges and opportunities for other countries. The US-China tensions and the global chip shortage further complicate the geopolitical landscape.

To ensure technological competitiveness, Qatar should focus on AI technologies and data centers, invest in education and research, and focus on a niche part of the supply chain. By prioritizing these areas, Qatar can contribute to the global AI technologies ecosystem and enhance its technological capabilities.

Overall, by strategically responding to the race for AI technologies, Qatar can position itself as a key player in the AI industry, foster innovation, and contribute to the advancements in AI technologies. These measures will strengthen the technological competitiveness of Qatar and its ability to harness the potential of artificial intelligence.

References

1. D. Owens, M. Houston, D. Luebke, S. Green, J. Stone, and J. Phillips, "GPU Computing," Proceedings of the IEEE, Vol. 96, pp. 879-899, May 2008.

2. C. Miller, Chip War: The Fight for the World's Most Critical Technology, Simon & Schuster, Oct 2022.

3. S. Khan and A. Mann, "AI Chips: What They Are and Why They Matter," Center for Security and Emerging Technology (CSET), Georgetown University, Issue Brief, April 2020.
Available: https://doi.org/10.51593/20190014

4. N. Benaich and I. Hogarth, "State of AI Report,"[Online], Oct 2022.
Available: https://www.stateof.ai/2022

5. B. Samel, S. Mahajan, A. Ingole, "GPU Computing and Its Applications," Int. Research Journal of Engineering and Technology (IRJET), Vol. 03, Issue 4, April 2016.

6. D. Blythe, "The Rise of the Graphics Processors," Proc. of the IEEE, Vol. 96, Issue 5, May 2008.

7. M. Levinas. (2021, March 23). GPU Architecture Explained: Everything You Need to Know and How It Has Evolved [Online].
Available: https://www.cherryservers.com/blog/everything-you-need-to-know-about-gpu-architecture

8. NVIDIA Publications, "NVIDIA TESLA V100 GPU ARCHITECTURE", White Paper WP-08608-001_v1.1 [Online], Aug. 2017.
Available: https://images.nvidia.com/content/volta-architecture/pdf/volta-architecture-whitepaper.pdf

9. A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," Com. of the ACM, Vol. 60, Issue 6, pp: 84–90, May 2017.

10. S. R. Punyala, "Throughput Optimization and Resource Allocation on GPUs under Multi-Application Execution" [Online], 2017. Theses 2255.
Available: http://opensiuc.lib.siu.edu/theses/2255

11. NVIDIA Publications, "NVIDIA H100 PCIe GPU," Product Brief PB-11133-001_v02 [Online], Nov 2022.
Available: https://www.nvidia.com/content/dam/en-zz/Solutions/gtcs22/data-center/h100/PB-11133-001_v01.pdf

12. AWS News. (2021, Nov. 30). AWS Announces Three New Amazon EC2 Instances Powered by AWS-Designed Chip [Online].
Available: https://shorturl.at/jpV09

13. E. Nurvitadhi et. al., "Can FPGAs Beat GPUs in Accelerating Next-Generation Deep Neural Networks?", in FPGA'17: Proc. of the 2017 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays, Feb. 22-24, 2017, Monterey, CA, USA.
DOI: http://dx.doi.org/10.1145/3020078.3021740

14. T. Hamada, K. Benkrid, K. Nitadori, and M. Taiji, "A comparative study on ASIC, FPGAs, GPUs, and general-purpose processors in the O(N2) gravitational N-body simulation," in NASA/ESA Conference on Adaptive Hardware and Systems, 2009.

15. J. Ouyang, et. al., "Baidu Kunlun an AI processor for diversified workloads," In 2020 IEEE Hot Chips 32 Symposium (HCS), pp. 1–18, 2020.

16. Deloitte, "Anchor of global semiconductor-Asia Pacific Takes Off," Deloitte China, 2021. Available: **https://www2.deloitte.com/content/dam/Deloitte/cn/Documents/technology-media-telecommunications/deloitte-cn-tmt-semiconductor-report-en-211001.pdf**

17. C. Qi and X. Yuanrui, "Comparison of International AI Strategies," in the International Security and Strategy Studies Report, Vol. 7, No. 7, Center for International Security and Strategy (CISS), Tsinghua University, 2019.

18. S. Khan, A. Mann, and D. Peterson, "The Semiconductor Supply Chain: Assessing National Competitiveness," Center for Security and Emerging Technology, Georgetown University, Jan. 2021. Available: **https://doi.org/10.51593/20190016**

19. B. Buchanan, "The AI Triad and What It Means for National Security Strategy," Center for Security and Emerging Technology (CSET), Georgetown University, Aug. 2020. Available: **https://cset.georgetown.edu/publication/the-ai-triad-and-what-it-means-for-national-security-strategy**

20. J. Vipra and S. M. West, "Computational power of AI," AI Now Institute, Sep. 27, 2023. Available: **https://ainowinstitute.org/publication/policy/compute-and-ai**

21. J. Nickolls, W. J. Dally, "The GPU Computing Era," IEEE Micro, May 2010.

22. US Department of Commerce, (2023, March 21). National Security Guardrails for CHIPS for America Incentives Program [Online]. Available: **https://www.commerce.gov/news/press-releases/2023/03/commerce-department-outlines-proposed-national-security-guardrails**

23. S. Keckler, W. Dally, B. Khailany, M. Garland, D. Glasco, "GPUs and the Future of Parallel Computing," IEEE Micro , Vol. 31, pp. 7-17, Nov. 2011.

24. W. X. Zhao, et al. "A survey of large language models," arXiv preprint, arXiv:2303.18223. March 2023. Available: **https://arxiv.org/pdf/2303.18223.pdf**

25. The European Commission, (2023, Sept. 21). The Chips Act for Europe [Online]. Available: **https://www.european-chips-act.com/**

26. J. VerWey, "Betting the House: Leveraging the CHIPS and Science Act to Increase U.S. Microelectronics Supply Chain Resilience," Policy Brief, Center for Security and Emerging Technology (CSET), Georgetown University, Jan. 2023.

27. US Department of Commerce, (2023, Dec. 29). Key 2023 Accomplishments [Online], Available: **https://www.commerce.gov/news/press-releases/2023/12/secretary-commerce-gina-raimondo-highlights-key-2023-department**

28. G. Allen, "China's New Strategy for Waging the Microchip Tech War," Center for Security and Emerging Technology (CSET), Georgetown University, March 2023.

Contact Informatiom

qcri@hbku.edu.qa

General inquiries:

P.O. Box: 34110
Doha – Qatar

QatarComputing
qcri_hbku
Qatar Computing Research Institute