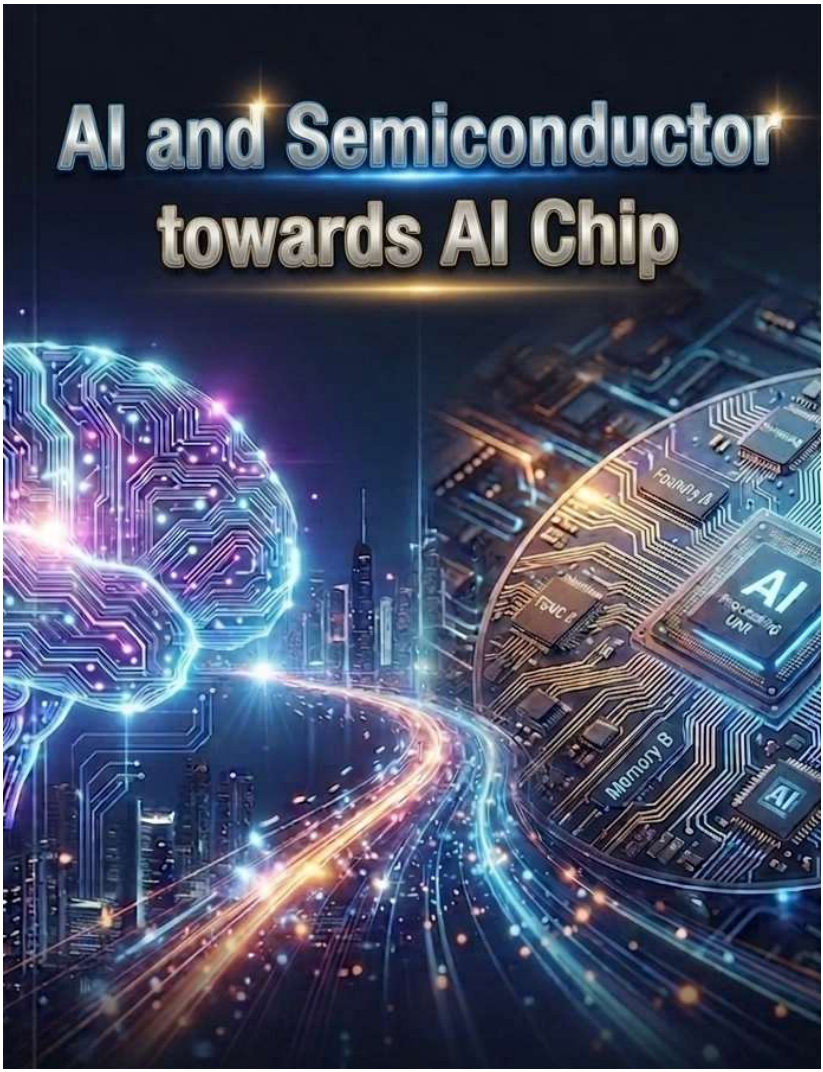

AI and Semiconductor
towards AI Chip
AI
Memory B
AI

AI and Semiconductor towards AI Chip

Preface

The modern world is currently being driven by two massive waves that have perfectly converged. The first is Artificial Intelligence (AI), the intelligent software that is revolutionizing human lifestyles and workflows. The second is semiconductor technology (Semiconductors), the hardware that serves as the foundation and crucial gear of the global computing system. As AI grows exponentially, the demand for processing power has become far too immense for traditional chip architectures to handle. This very intersection is what ushers us into a new era: the era of the AI Chip.

The inspiration and starting point for writing this book arose when the author had the opportunity to attend the international conference, 'SNU AI Semiconductor Forum (SAISF 2026): International Collaboration of Research, Education, and Strategy in Southeast and East Asia.' Held from June 1–2, 2026, at Seoul National University in the Republic of Korea, the forum was organized by the Graduate School of AI Semiconductor, one of the world's leading institutions dedicated specifically to research and human resource development in the field of AI chips. The atmosphere at SAISF 2026 was filled with the exchange of visions, innovations, latest research, and national strategies from participating countries, namely South Korea, Thailand, Vietnam, and Malaysia. Engaging in discussions with these leading researchers and experts allowed the author to clearly realize that the technology industry is transitioning toward a deep strategic integration between the domains of Computer Science and Semiconductor Engineering. Developing these two fields can no longer be separated if we wish for AI technology to achieve its ultimate capability. For this reason, the author intended to convey this essential knowledge, perspective, and

foundation through this book, 'AI and Semiconductor towards AI Chip.' The content is divided into three main chapters: **Chapter 1 : AI (Artificial Intelligence)**, which builds a fundamental understanding of the software and algorithm side, from the early days to the era of Deep Learning and Generative AI; **Chapter 2 : Semiconductor Foundation**, which discusses semiconductor technology, the semiconductor industry supply chain, and the nanometer-scale chip manufacturing process that serves as the backbone of global technology; and **Chapter 3: Towards AI Chip**, which serves as the bridge fusing AI software with the hardware that humans can create, analyzing modern chip architectures such as GPUs, TPUs, and NPUs, and peering into the future of technology that will continue to drive these intelligent brains."

The author sincerely hopes that this book will serve as a bridge of knowledge and a source of inspiration for undergraduate and graduate students, researchers, and anyone interested in the technology industry, enabling them to envision opportunities and prepare themselves for the massive wave of transformation currently taking place in the world of AI and semiconductors, both now and in the future.

Suwit Kiravittaya
6th June 2026

Note Added: This book is translated from Thai to English version of the 10th June 2026. The Thai version is onlined on 6th June 2026.

Table of Contents

1. AI (Artificial Intelligence)

1.1 History and Evolution of AI

The journey of Artificial Intelligence did not begin in a modern computer laboratory; rather, its traces have been woven into human thought for thousands of years. From ancient myths and the desire to create simulated living beings, to eventually fusing with the principles of mathematics, physics, and computer engineering today, the evolution of AI can be divided into the following significant eras:

1. The Pre-Computing Era: Logical Foundations and Mechanical Automata

Humanity has long dreamed of 'sentient inventions,' as seen in ancient Greek myths like Talos—the bronze giant—or the golden automata created by the god Hephaestus. Later, between the 17th and 19th centuries, philosophers and mathematicians such as Gottfried Leibniz, Thomas Hobbes, and René Descartes began to hypothesize that human thought processes could be systematized and mechanically calculated. Hobbes masterfully noted in his book, Leviathan, that 'For reason, ... is nothing but reckoning, that is adding and subtracting.' This foundation became more concrete when George Boole introduced Boolean Logic in 1854, followed by Alan Turing's theoretical computing machine (Turing Machine) in 1936. Turing proved that a mechanical device processing simple symbols like '0' and '1' could simulate any mathematical computation in the world.

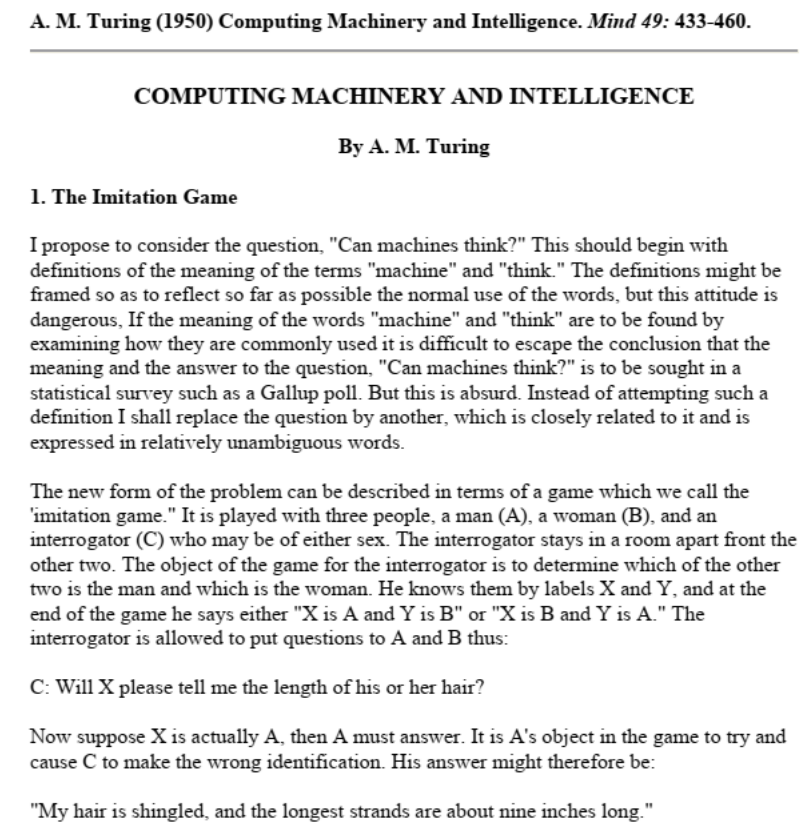
Figure 1: A portrait of Gottfried Leibniz, the philosopher who invented the binary number system and envisioned that human thought could be reduced to a calculating machine.
(Source: https://en.wikipedia.org/wiki/Gottfried_Wilhelm_Leibniz)

2. The Birth of Artificial Intelligence and the Turing Test (1941–1956)

During World War II, the world's first digital computers were built, prompting scientists to seriously debate the possibility of creating an 'electronic brain.' In 1950, Alan Turing published a historic academic paper titled 'Computing Machinery and Intelligence.' Recognizing that the term 'thinking' was too difficult to define, Turing proposed the 'Turing Test.' If a machine could engage in a text-based conversation with a human and make it indistinguishable to the human interlocutor whether they were talking to a person or a computer, then that machine could be considered to possess intelligence.

The term 'Artificial Intelligence' (AI) was first formally defined and used as a new academic discipline during a summer workshop at Dartmouth College, USA, in 1956. Led by prominent figures such as

John McCarthy, Marvin Minsky, Allen Newell, and Herbert Simon, they were highly optimistic that machines with human-level intelligence would be created within a generation.



A. M. Turing (1950) Computing Machinery and Intelligence. *Mind 49:* 433-460.

COMPUTING MACHINERY AND INTELLIGENCE

By A. M. Turing

1. The Imitation Game

I propose to consider the question, "Can machines think?" This should begin with definitions of the meaning of the terms "machine" and "think." The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous, If the meaning of the words "machine" and "think" are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, "Can machines think?" is to be sought in a statistical survey such as a Gallup poll. But this is absurd. Instead of attempting such a definition I shall replace the question by another, which is closely related to it and is expressed in relatively unambiguous words.

The new form of the problem can be described in terms of a game which we call the 'imitation game." It is played with three people, a man (A), a woman (B), and an interrogator (C) who may be of either sex. The interrogator stays in a room apart front the other two. The object of the game for the interrogator is to determine which of the other two is the man and which is the woman. He knows them by labels X and Y, and at the end of the game he says either "X is A and Y is B" or "X is B and Y is A." The interrogator is allowed to put questions to A and B thus:

C: Will X please tell me the length of his or her hair?

Now suppose X is actually A, then A must answer. It is A's object in the game to try and cause C to make the wrong identification. His answer might therefore be:

"My hair is shingled, and the longest strands are about nine inches long."

Figure 2: The opening section of the manuscript copy of the article by Alan Turing, published in 1950.

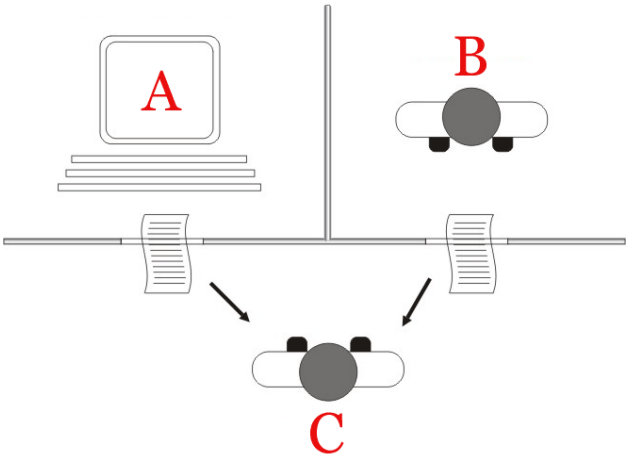


Figure 3: An illustration of the Turing Test setup.
(Source: https://en.wikipedia.org/wiki/Turing_test)

3. The First Golden Age to the AI Winter (1956–1980)

Following the Dartmouth conference, the United States government poured massive funding into AI research. Computers of that era were able to solve algebra problems, prove geometric theorems, and practice speaking English. This period also witnessed the birth of the earliest artificial neural network, known as the Perceptron.

However, researchers underestimated the complexity of AI development. When computers faced real-world problems involving immense complexity and limited data, the machinery of that era—hampered by very low memory and processing power—failed to deliver. Consequently, in 1974, the US and British governments suspended all research funding, leading to a period of disillusionment and financial starvation known as the 'First AI Winter' (1974–1980).

4. The Expert Systems Era and the Second Winter (1980–1990s)

In the 1980s, AI experienced a resurgence driven by the concept of 'Expert Systems.' This approach focused on hardcoding knowledge bases and conditional logic (If-Then Rules) from specialized experts into computers. Large corporations adopted these systems, giving birth to a billion-dollar industry. Ultimately, however, this technology hit a dead end due to high maintenance costs, difficulties in updating data, and a stark lack of flexibility. The stock market and investors were disillusioned once again, triggering the 'Second AI Winter' from the late 1980s through the 1990s.

5. The Behind-the-Scenes Era and Moore's Law (1990s–2005)

Despite being abandoned by the media and investors, AI researchers in the 1990s quietly shifted their approach. Instead of trying to build an artificial brain capable of general thinking, they turned to 'mathematics, probability logic, and statistics' to solve specific problems, laying the foundation for Machine Learning.

A historic event that the world would long remember occurred in 1997, when IBM's Deep Blue supercomputer successfully defeated Garry Kasparov, the world chess champion. This success did not stem from algorithms that thought like humans, but rather from the 'brute processing power of hardware.' Accelerated by the continuous advancement of semiconductors in accordance with Moore's Law, it enabled the computer to calculate millions of potential moves ahead within a fraction of a second.

Figure 4: (Left) The Deep Blue computer, and (Right) The historic match between Deep Blue and Garry Kasparov in 1997.
(Source: https://en.wikipedia.org/wiki/Deep_Blue_(chess_computer) and https://www.computerhistory.org/chess)

6. The Big Data Era, Deep Learning, and the AI Explosion (2005–Present)

Since 2012, the world has truly entered the third rebirth of AI, driven by three core factors:

1. **Big Data**: The rise of the internet and social media has generated massive amounts of data for training AI.
2. **Advanced Algorithms**: The development of multi-layered artificial neural networks, known as Deep Learning.
3. **Powerful Hardware**: The utilization of Graphics Processing Units (GPUs) to compute AI tasks, replacing traditional Central Processing Units (CPUs).

A major turning point occurred in 2017 with the introduction of an architecture called the Transformer model. It became the core engine driving Generative AI and Large Language Models (LLMs) such as

ChatGPT, Google Gemini, and Claude today. These models exhibit human-like capabilities in knowledge, comprehension, and creativity, leading to the most intense era of technology investment and competition in history.

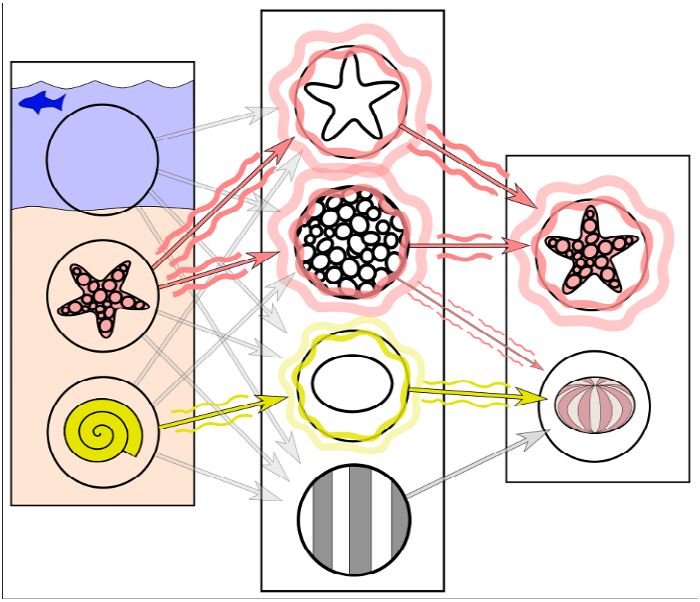


Figure 5: An example illustration of how Artificial Neural Networks, a type of Deep Learning, are utilized in object image recognition. (Source: https://en.wikipedia.org/wiki/Deep_learning)

1.2 The Core Pillars of Modern AI

Modern artificial intelligence no longer operates on human-coded, rule-based systems written line by line. Instead, the intelligence we witness today is driven by statistics, advanced mathematics, and massive data processing. This setup is built upon three core technological layers that overlap and build upon one another: Machine Learning, Deep Learning, and Generative AI—which is underpinned by the Transformer model architecture.

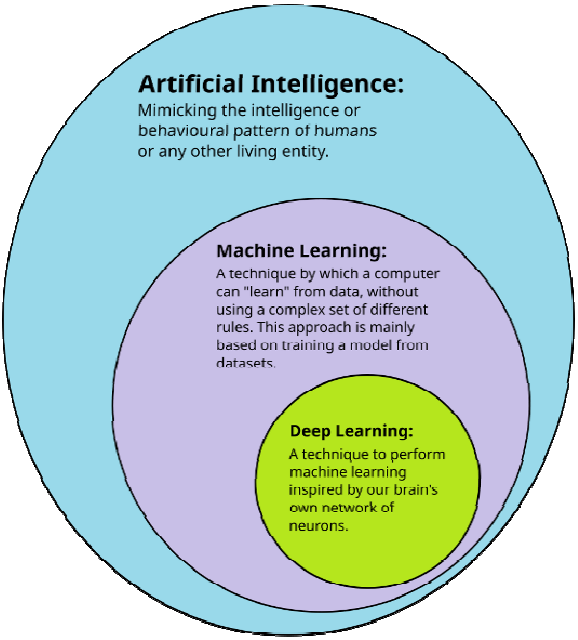


Figure 6: A diagram illustrating the relationship between AI, Machine Learning, and Deep Learning.
(Source: https://en.wikipedia.org/wiki/Deep_learning)

1. Machine Learning

In 1959, Arthur Samuel defined this term as the field of study that gives computers the ability to learn without being explicitly programmed. Instead of humans providing fixed solutions to a computer, Machine Learning algorithms build a 'model' from a sample dataset (Training Set) to flexibly discover patterns and predict outcomes on new data.

According to Wikipedia (https://en.wikipedia.org/wiki/Machine_learning), this learning process is categorized into three main paradigms: *Supervised Learning*: Training a model using predefined, labeled data—such as feeding clearly labeled images of dogs and cats so the system can differentiate their key features. *Unsupervised Learning*: Allowing the model to independently discover hidden structures and patterns within raw, unlabeled data, such as clustering customer behavior. *Reinforcement Learning*: Teaching an agent to make decisions through trial and error in a simulated environment, guided by a system of rewards or penalties to achieve the optimal goal.

2. Deep Learning

Deep Learning is currently the most powerful subfield of Machine Learning. This type of model simulates the structure and neural signaling mechanisms found in biological brains, known as Artificial Neural Networks. The word 'Deep' reflects the way 'Artificial Neurons' are stacked sequentially into multiple layers, ranging from three layers to hundreds or even thousands. Raw data is fed through these layers, with each layer extracting increasingly complex patterns or features (for example, the first layer detects simple lines, the next layer matches shapes, and the final layer can identify a human face). Key architectures of Deep Learning include: *Convolutional Neural Networks (CNN)*:

Networks utilizing convolutional filters, specifically designed to detect patterns in spatial data, such as photos and videos. *Recurrent Neural Networks (RNN)*: Networks incorporating a feedback mechanism to process sequential or time-series data, such as long text passages or audio clips.

3. Generative AI and the Transformer Architecture

The pinnacle of modern AI technology began in 2017 when Google researchers published a groundbreaking paper titled 'Attention Is All You Need,' introducing a novel artificial neural network architecture known as the Transformer.

The limitation of previous artificial neural networks was their need to process text data sequentially, word by word, which resulted in lengthy training times and a tendency to forget earlier content. The Transformer solved this problem by introducing the 'Self-Attention Mechanism.' This allows the model to see and process all words in a sentence simultaneously through parallel processing, while precisely calculating which words share the strongest relationships.

This Transformer architecture serves as the core engine driving Generative AI and Large Language Models (LLMs) across today's information technology industry—such as the GPT series (the foundation of ChatGPT), as well as Gemini and Claude. These models are pre-trained on massive amounts of internet data until they develop a deep understanding of language structure and logic. They can then generate entirely new data, whether it be text, computer code, artwork, or even audio and video.

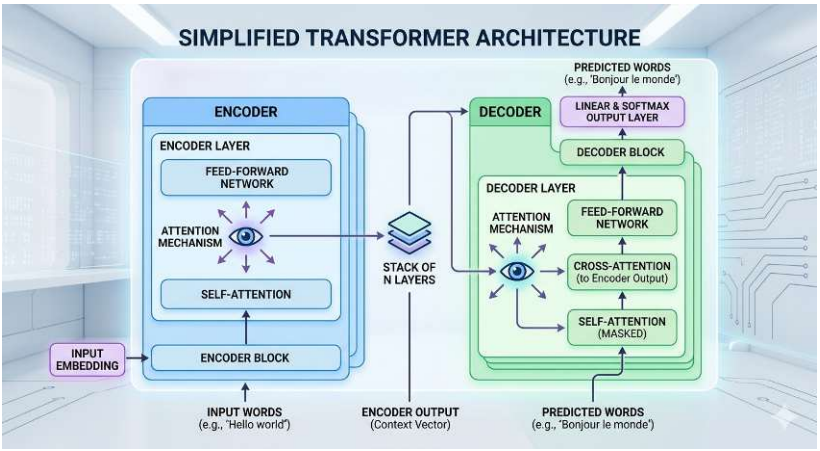


Figure 7: A diagram illustrating the Transformer model architecture, consisting of an Encoder and a Decoder, and highlighting the Attention Mechanism which serves as the core engine of this AI.

1.3 Conventional Hardware Limitations

As modern artificial intelligence models transitioned into the era of Deep Learning and the Transformer architecture, data processing demands no longer grew linearly, but exponentially. However, the computer hardware architecture we have relied on for nearly a century was not designed to accommodate this specific type of AI mathematics, ultimately leading to a major roadblock known as “The Compute Bottleneck.”

The critical limitations of conventional hardware can be classified into three main areas, based on computer architecture principles as follows:

1. The Von Neumann Bottleneck

Almost all computers in the world—from smartphones to supercomputers—rely on the Von Neumann Architecture, which was conceptualized by John von Neumann in 1945.

The fundamental principle of this system is the absolute separation of the central processing unit and the random-access memory (RAM). Every time the computer performs a calculation, the central processing unit must send a signal to 'fetch' data from the memory via a data highway called the 'bus.' Once the computation is complete, the result must be sent back to be stored in the memory. The issue for AI is that deep learning models contain parameters and weights comprising billions to hundreds of billions of variables. This massive amount of data must be constantly shuttled back and forth between the processor and the memory. As a result, the bus speed becomes a limiting factor that drags down overall performance. No matter how fast

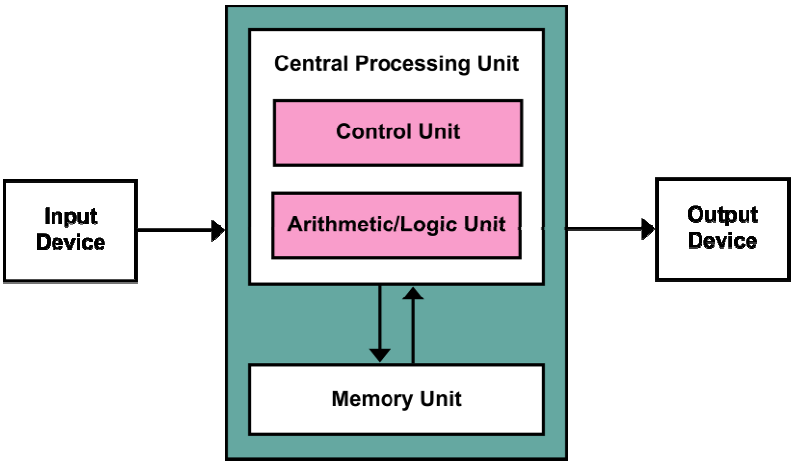


Figure 8: A structural diagram of the Von Neumann architecture. (Source: https://en.wikipedia.org/wiki/Von_Neumann_architecture)

the processing unit can compute, if it constantly has to sit idle waiting for data to arrive from the memory, the system encounters a bottleneck and suffers from high energy consumption.

2. The Disconnect Between Latency and Throughput

The Central Processing Unit (CPU) is designed to be a 'jack-of-all-trades' wizard, highly efficient at executing complex, conditional tasks rapidly one by one (focusing on speed per instruction), also known as having low latency. For this reason, most CPUs feature only a few powerful processing cores (such as 4 or 8 cores).

The issue for AI is that its tasks are not conditionally complex. Instead, they consist of massive volumes of basic mathematical operations—specifically matrix multiplication and addition—performed repeatedly and simultaneously. Using a CPU with a handful of cores for AI workloads is therefore like asking '4 or 8 university professors' to

solve one million elementary school math problems; they will inevitably finish slower than '1,000 elementary school students' working together at the same time. AI demands high throughput, meaning the capacity to process a massive volume of data simultaneously.

3. Energy and Heat Limits (The Thermal Wall)

In the past, when we wanted to make a chip more powerful, we simply ramped up its clock speed (for example, from 1 GHz to 3 GHz). However, since the early 2 0 0 0 s, the chip industry has hit a wall; increasing clock speeds any further causes transistors to release massive amounts of heat, to the point where the chip could literally melt. Consequently, computers can no longer easily boost performance by simply accelerating the clock speed of a single processor.

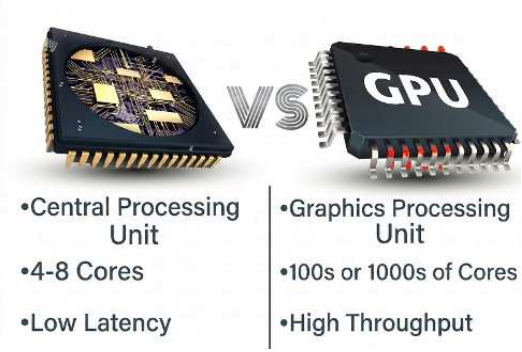


Figure 9: A comparison of features and architectures between Central Processing Units (CPUs) and Graphics Processing Units (GPUs).

As AI software advances rapidly while the Von Neumann computer architecture and CPUs struggle to push forward due to the thermal wall and memory bottlenecks, humanity's only way forward is to rethink hardware design from the ground up—starting at the level of materials science and semiconductors. In **Chapter 2 : Semiconductor Foundation**, we will explore how silicon wafers and tiny transistors are fabricated, and examine how Moore's Law might help us break through these modern computing limitations.

2. Semiconductor Foundation

2.1 The Beginning: Physics of Semiconductor Materials

Before the computer industry could manufacture processing chips packed with hundreds of billions of transistors, humanity spent more than a century studying the peculiar properties of a specific class of materials. These substances are neither non-metallic insulators nor highly conductive metals; rather, their electrical properties vary based on temperature, light, and mechanical stimulation—a behavior that completely defied the common sense of classical physicists prior to 1900.

1. Early History and the Discovery of Semiconductor Properties

In the early annals of semiconductor history, the first documentation of their unique properties was recorded by Michael Faraday, an English chemist and physicist, in 1833. Faraday discovered that heating silver sulfide (Ag_2S) caused its electrical resistance to decrease, which directly contradicted the behavior of typical metals like copper or silver, whose electrical resistance increases as they heat up. This phenomenon was the first clue indicating that this substance possessed a fundamentally different internal mechanism. Later, in 1874, German physicist Karl Ferdinand Braun discovered another crucial property known as rectification, or the point-contact effect. Braun observed that when a sharp metal wire tip came into contact with a crystal of galena (lead sulfide, PbS), electrical current flowed easily in only one direction but was blocked when reversed. This discovery was

further developed in 1894 by Sir Jagadish Chandra Bose, an Indian scientist, to create the world's first radio wave detector. Patented in 1901 as the 'cat's-whisker detector,' it became the first practical semiconductor diode ever built, finding widespread use in early crystal radio receivers. However, at the time, no one could explain why electrons flowed through galena in only a single direction. This atomic secret remained unlocked until the inception of quantum mechanics in 1925 and the subsequent development of solid-state physics in the 1930s.

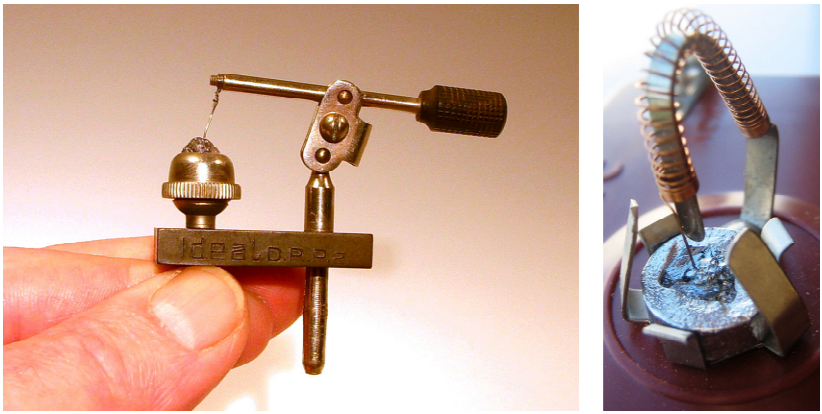


Figure 10: A pioneering radio wave detector: the cat's-whisker detector. (Source: https://en.wikipedia.org/wiki/Crystal_detector)

2. The Physics of Energy Band Structure

Semiconductor physics textbooks explain that the electrical behavior of materials arises from electron interactions governed by the Pauli Exclusion Principle. When atoms combine to form a solid crystal

structure, their previously discrete atomic energy levels merge into continuous Energy Bands. Two primary bands are essential for electrical conduction:

- *Valence Band*: The lower energy band that is densely packed with electrons tightly bound to their parent atoms. If this band is completely filled and no electrons can move, the material cannot conduct electricity.

- *Conduction Band*: The higher energy band where free electrons reside. In this band, electrons can move freely throughout the crystal lattice to conduct electrical current.

What determines whether a material acts as a conductor, an insulator, or a semiconductor is the energy gap separating these two bands, known as the Bandgap (E_g): **Conductors**: The valence band and conduction band overlap, allowing electrons to freely flow and conduct electricity instantly without requiring any external excitation energy. **Insulators**: The bandgap is exceptionally wide (typically $E_g > 5$ eV), meaning electrons cannot acquire enough energy to leap across the gap. **Semiconductors**: The bandgap is moderately sized (typically $E_g = 0.5 - 3$ eV). For instance, Silicon (Si) has a bandgap of approximately 1.12 eV at room temperature. This gap is narrow enough that thermal energy or an electrical field can excite some electrons, causing them to jump from the valence band into the conduction band. Once an electron makes this leap, it leaves behind a vacant, positively-charged vacancy in the valence band known as a 'Hole'. Both free electrons and holes serve as the vital Charge Carriers that enable electrical conduction.

3. Carrier Transport and Doping

Pure silicon contains an exactly equal number of free electrons and holes, resulting in very low electrical conductivity. In semiconductor engineering, precise control over the quantity and type of charge carriers is required. This is achieved through a process called Doping—the intentional introduction of impurity atoms into the crystal lattice structure to shift the positions of the energy bands. Engineers create two distinct types of doped semiconductors:

- *N-type Semiconductor (Negative-type)*: Created by adding Group V elements, such as Phosphorus (P) or Arsenic (As), into the silicon crystal. Because these elements possess 5 valence electrons, bonding with 4 -valence silicon leaves one extra electron that immediately becomes a free electron without requiring external energy. This process shifts the energy levels downward. Consequently, this material features electrons as the majority carriers and holes as the minority carriers.

- *P-type Semiconductor (Positive-type)*: Created by adding Group III elements, such as Boron (B) or Gallium (Ga), which possess only 3 valence electrons. When bonded into the silicon crystal lattice, a vacancy is created due to the missing electron, instantly forming a hole. This process shifts the energy levels upward. As a result, this material features holes as the majority carriers and electrons as the minority carriers.

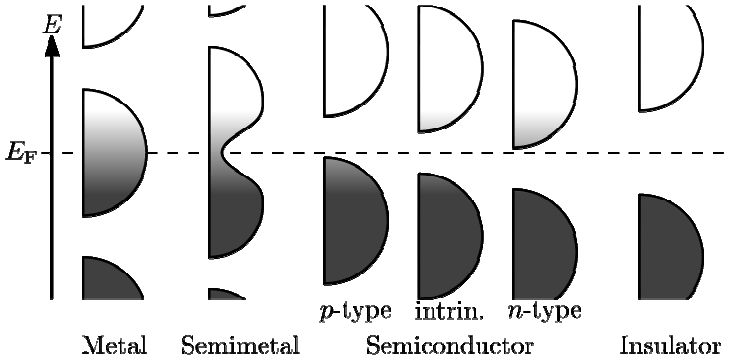


Figure 11: (From left to right) Energy band diagrams of a metal, a semimetal, a P-type semiconductor, an intrinsic (undoped) semiconductor, an N-type semiconductor, and an insulator. (Source: https://en.wikipedia.org/wiki/Semiconductor)

When these charge carriers (electrons and holes) move under the physical principles within the chip, two distinct transport phenomena occur: Drift Current, driven by the force of an electric field, and Diffusion Current, driven by gradients in charge carrier density.

These mathematical and physical behaviors form the basis for the rectification properties of diodes and the switching operations of both Bipolar Junction Transistors (BJTs) and Metal-Oxide-Semiconductor Field-Effect Transistors (MOSFETs). Ultimately, these components serve as the foundational building blocks that aggregate into complex logic gate circuits.

4. Application to Transistors and Logic Gates

Once we gained the ability to precisely control the density of both electron and hole charge carriers through doping, the next monumental milestone that reshaped human history was combining these semiconductors to construct the transistor. As a solid-state device acting as an electronic switch with no moving parts, the transistor became the smallest fundamental building block of every computer chip on Earth.

In the beginning, the Bipolar Junction Transistor (BJT) was invented in 1947 at Bell Labs by John Bardeen, Walter Brattain, and William Shockley. This device utilized electrical current to control current. However, the critical turning point leading to modern chip architecture was the invention of the Metal-Oxide-Semiconductor Field-Effect Transistor (MOSFET) in 1959 by Mohamed Atalla and Dawon Kahng. This breakthrough shifted the control mechanism to using voltage to generate an electric field for switching. This drastically reduced power consumption and made microscopic scaling significantly easier.The basic structure of an n-channel MOSFET (NMOS) consists of a P-type semiconductor substrate. Embedded on opposite sides within this substrate are two highly doped N-type semiconductor regions that serve as the charge-handling terminals, known as the Source (S) and Drain (D). Positioned between them is the control terminal, known as the Gate (G). Crucially, the gate electrode is completely isolated from the underlying substrate by an ultra-thin insulating layer made of Silicon Dioxide (SiO_2).

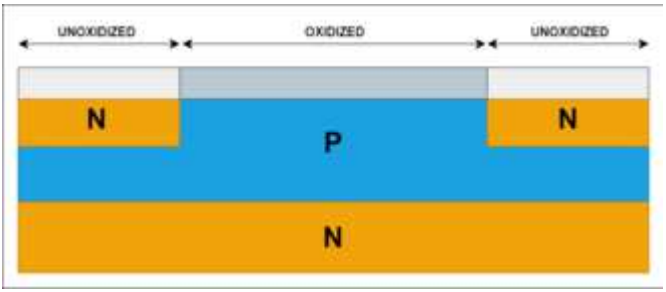


Figure 12: A diagram illustrating one of the silicon dioxide (SiO_2) transistor devices fabricated by Frosch and Derick in 1957. (Source: https://en.wikipedia.org/wiki/MOSFET)

In terms of physical mechanisms, when no voltage is applied to the gate terminal (representing an 'OFF' switch state or logic '0 '), electrons cannot flow from the source to the drain due to the potential barrier created by the P-type substrate in the middle. However, when a positive voltage is applied to the gate, the resulting electric field penetrates the insulating layer, repelling holes and attracting a high density of electrons directly beneath the insulation. This induces a phenomenon known as Inversion, which forms a conducting path connecting both sides called an n-channel. This channel allows electrons to flow freely through the device (representing an 'ON' switch state or logic '1').

In the chip design process, engineers combine these positive-voltage-activated MOSFETs (NMOS) with negative-voltage-activated MOSFETs (PMOS) to work in complementary pairs. This technology is known as Complementary MOS (CMOS). These pairs are interconnected to construct foundational mathematical Logic Gates, such as NOT, AND, OR, NAND, and NOR. These individual logic gates are then systematically linked together to form highly complex computational circuits, such as Adders and Multipliers.

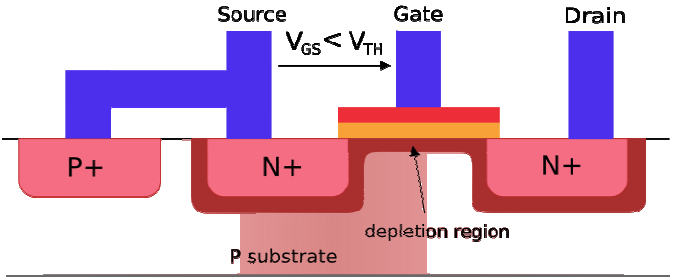


Figure 13: A cross-sectional view of a MOSFET. When the gate voltage is below the threshold level required to turn on the channel, there is no conduction between the drain and source terminals, placing the switch in the 'OFF' state. Conversely, when the gate becomes more positive, it attracts electrons toward the gate, creating an n-channel directly beneath the oxide layer. This allows electrons to flow between the N-type doped drain and source terminals, turning the switch to the 'ON' state. (Source: https://en.wikipedia.org/wiki/MOSFET)

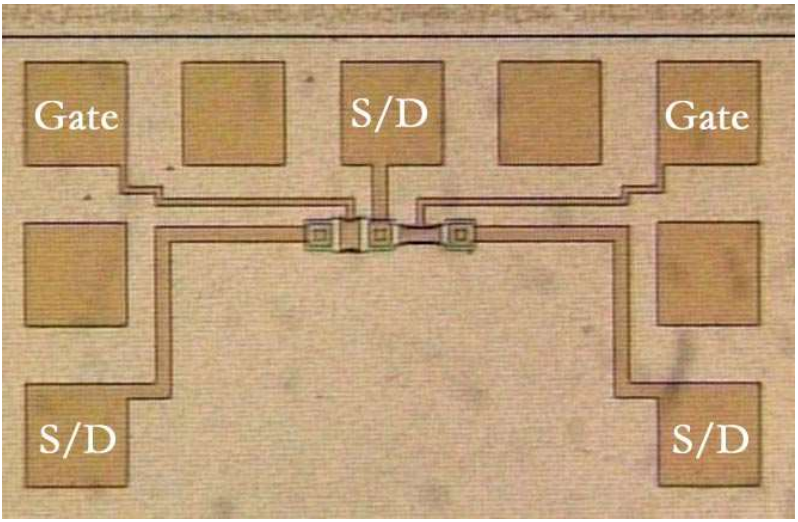


Figure 14: A microscopic photograph of two metal-gate MOSFETs, labeled with probe pads for two gates and three source/drain connections. (Source: https://en.wikipedia.org/wiki/MOSFET)

Modern AI chip architectures are firmly rooted in this very foundation. Their internal structures are engineered specifically to arrange these logic gates into massive arrays of specialized mathematical units. These units are dedicated to executing parallel Multiply-Accumulate operations on AI variable matrices simultaneously.

Consequently, the tens of billions of transistors packed onto an AI chip act as an immense network of microscopic switches, toggling simultaneously to power the artificial intelligence brain.

2.2 Moore's Law in Modern Perspective

In the world of semiconductor technology, no concept has been as influential or has shaped the industry's direction for as long as Moore's Law. This concept is not a rigid law of physics like Newton's laws or Maxwell's equations; rather, it began as a predictive economic and engineering observation. Gordon Moore, co-founder of Intel, published an article in the journal Electronics in 1965. He observed that the number of transistors and other components on an integrated circuit (IC) of a given size would double approximately every year. Later, in 1975, he revised his prediction, stating that the transistor count would double roughly every two years (24 months).

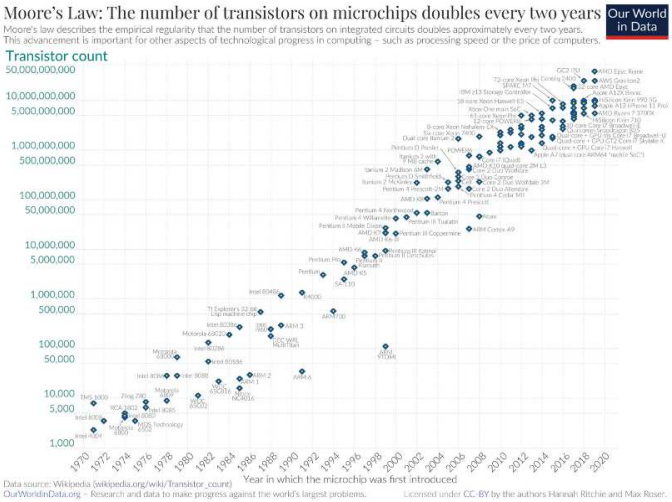


Figure 15: A graph showing microprocessor transistor counts plotted against their introduction dates, demonstrating a near-doubling of count every two years. (Source: https://en.wikipedia.org/wiki/Moore%27s_law)

For over half a century, the global computer and semiconductor industries have treated Moore's Law as a 'compass and roadmap' for shaping business strategies and engineering milestones. The ability to shrink transistor dimensions by half every two years has yielded an exponential surge in processing power while drastically driving down the manufacturing cost per transistor. This relentless scaling is what shrank computers from the size of entire rooms down to smartphones that fit in the palm of our hands. Furthermore, it served as the critical catalyst that revived artificial intelligence from its historical 'AI Winters' by finally providing hardware with sufficient computational horsepower.

However, in today's landscape and the near future, Moore's Law is rapidly approaching a dead end as it collides with severe physical limits. Industry analysts and semiconductor engineers widely agree that we can no longer downscale transistors at this traditional pace, primarily due to three core challenges:

1. The Breakdown of Dennard Scaling

In 1974, Robert Dennard proposed a physical scaling law that complemented Moore's Law, stating that as transistors shrink, their power density per unit area of the chip remains constant. This meant that smaller transistors would require less voltage and consume less power, allowing engineers to continuously increase clock speeds without overheating the chip. However, this law completely broke down around 2005–2006. As transistors grew too microscopic, the semiconductor materials could no longer adequately confine the electric charge, leading to continuous current leakage. This leakage caused power dissipation to skyrocket, pushing chips against severe thermal thresholds. Consequently, engineers could no longer increase the clock frequency of single processors beyond the 4–5 GHz barrier.

2. Quantum Tunnelling

As the industry advances into cutting-edge manufacturing nodes (such as 3 nm or 2 nm), the physical distance between the source and drain terminals, as well as the thickness of the insulating layer beneath the gate, shrinks to a mere few dozen silicon atoms. At this sub-microscopic scale, the classical laws of physics we rely on break down. Instead, electrons begin to exhibit quantum mechanical behaviors, possessing a natural probability to 'warp' or tunnel directly through the potential barriers of the insulation. This phenomenon, known as Quantum Tunneling, prevents the transistor from functioning as a perfect on/off switch. Even when the switch is theoretically turned off (commanded to logic '0 '), electrons still leak across the barrier. This leads to massive power dissipation and introduces computational instability within the chip architecture.

3. Economic Bottleneck

Beyond the brutal limits of physics, the semiconductor industry is currently bottlenecked by escalating costs. Historically, shrinking transistors inherently drove down the cost per core. Today, however, fabricating chips at single-digit nanometer nodes demands extraordinarily complex and prohibitively expensive technologies. A prime example is the Extreme Ultraviolet (EUV) lithography system, which costs hundreds of millions of dollars per unit. This astronomical price tag has caused research, development, and foundry construction costs to skyrocket. This economic reality led Jensen Huang, CEO of NVIDIA, to famously declare that 'Moore's Law is dead'—signaling that the era of obtaining double the computing performance at the exact same price point has officially come to an end.

The 'More than Moore' Era and the Path Forward for AI Chips

To bypass this imminent bottleneck, the semiconductor industry has shifted its paradigm into the 'More than Moore' era. Instead of relentlessly focusing on shrinking transistors along a flat, two-dimensional plane, engineers are now exploring alternative frontiers of innovation. The current technological landscape has pivoted toward 3D structural revolutions and domain-specific architectures.

The 3D structural revolution represents a major departure from traditional planar transistors, transitioning first to 3D fin-like structures known as FinFETs (Fin Field-Effect Transistors), and more recently to GAAFETs (Gate-All-Around FETs). In a GAAFET architecture, the gate terminal completely encloses the conducting channel from all sides, providing ultimate electrostatic control to significantly reduce current leakage.

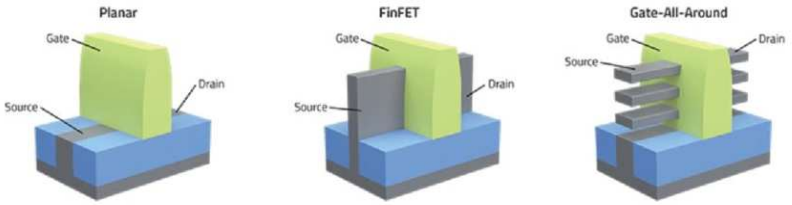


Figure 16: Structural characteristics of Planar MOSFET, FinFET, and GAAFET. (Source: Duan, H. (2024). From MOSFET to FinFET to GAAFET: The evolution, challenges, and future prospects. Proceedings of the 4th International Conference on Signal Processing and Machine Learning, 50, 113–120.)

Domain-Specific Architecture t is now widely accepted that general-purpose Central Processing Units (CPUs) can no longer act as the sole leaders of the semiconductor industry. Instead, the focus has shifted toward building specialized chips born to execute specific workloads. A prime example is the development of AI Accelerators. The internal circuitry of these chips is custom-tailored exclusively to excel at the massive matrix multiplication and addition operations required by Large Language Models (LLMs) and deep learning networks. By optimizing the hardware for these exact tasks, AI accelerators deliver a Performance-per-Watt ratio that is exponentially higher than that of general-purpose processors.

2.3 Semiconductor Supply Chain & Fabrication

The semiconductor manufacturing process—especially for artificial intelligence (AI) chips that pack hundreds of billions of transistors into an area no larger than a small coin—stands as one of the most complex, precise, and capital-intensive engineering endeavors in human history. Because of this extreme complexity, no single corporation or nation can manage the entire ecosystem independently. Instead, the industry relies on a highly interconnected Global Supply Chain comprised of specialized entities whose deep expertise makes them entirely indispensable to one another.

1. Structure and Business Models in the Global Semiconductor Supply Chain

Historically, semiconductor companies predominantly operated as Integrated Device Manufacturers (IDMs)—handling every stage from upstream to downstream in-house, including design, fabrication, testing, and packaging. Prominent examples of this model include Intel and

Samsung. However, as chip architectures shrank down to single-digit nanometer scales, the capital required to build and maintain a fabrication plant (Fab or Foundry) skyrocketed to tens of billions of dollars per facility. This economic shift catalyzed a fragmentation of the industry, dividing the supply chain into three specialized business models: (1) **Fabless** Companies (Chip Designers): These companies do not own manufacturing facilities. Instead, they focus entirely on research, development, and designing circuit architectures tailored to market demands. Once completed, their digital blueprints are sent to external foundries for manufacturing. Industry giants in this category include NVIDIA, AMD, Qualcomm, and Apple. (2) **Foundries** (Pure-Play Manufacturers): These companies do not design chips to compete with their clients. Instead, they invest colossal capital into building state-of-the-art fabs and acquiring cutting-edge machinery solely to manufacture chips on contract. The undisputed global leader in this sector is TSMC (Taiwan Semiconductor Manufacturing Company) from Taiwan, which controls over 9 0 % of the world's advanced chip manufacturing. (3) **OSAT** (Outsourced Semiconductor Assembly and Test): These downstream companies receive completed silicon wafers from the foundries. Their role is to dice the wafers into individual dies, test their functionality, and package them into finished chips ready for circuit board integration. A significant portion of Thailand's semiconductor industry operates within this OSAT sector. (Notably, Thailand houses one wafer fabrication facility, the Thai Microelectronics Center (TMEC), and one prominent fabless chip design firm, Silicon Craft Technology PLC).

2. The Chip Manufacturing Process: From Silicon Ingot to Wafer

The raw material for all AI chips originates from one of the most abundant resources on Earth: silica sand, or Silicon Dioxide (SiO_2). The journey involves the following meticulous stages: **Extraction and Purification**: The silica sand is subjected to high-temperature chemical reactions inside an electric arc furnace to strip away the oxygen, yielding metallurgical-grade silicon. This silicon then undergoes repeated, rigorous chemical refinements to transform into Electronic-grade Silicon. This resulting material achieves an ultra-high purity level of 99.999999999% (commonly referred to as 'Eleven Nines', meaning there is no more than one impurity atom for every 100 billion silicon atoms). **Ingot Growth via the Czochralski Process**: The ultra-pure silicon is melted down once again. A high-quality 'seed crystal' of silicon is dipped into the molten pool and then slowly rotated and withdrawn under incredibly precise temperature and speed controls. As the

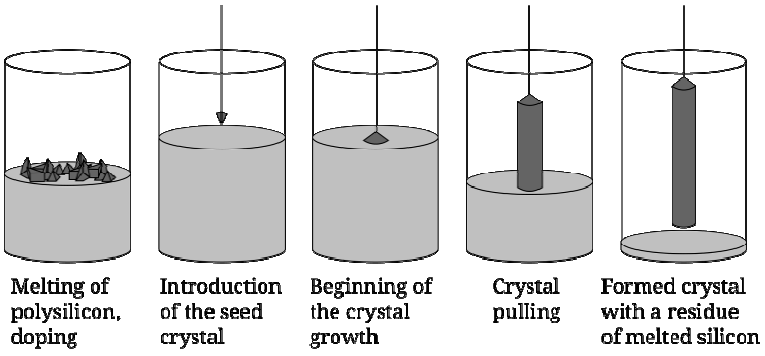


Figure 17: A diagram illustrating the single-crystal silicon ingot growth via the Czochralski process.
(Source: https://en.wikipedia.org/wiki/Wafer_(electronics))

molten silicon cools and solidifies around the seed, it crystallizes into a massive, single-crystal cylindrical block known as a Silicon Ingot. **Wafer Slicing**: The silicon ingot is ground and polished to a standardized diameter (most commonly 300 millimeters or 12 inches). Next, a high-speed diamond wire saw slices the cylinder into ultra-thin discs with a thickness of less than 1 millimeter, called Wafers. These raw wafers are then chemically etched and mechanically polished until they achieve a flawless, mirror-like finish, ready to be sent to the advanced fabrication facility.

3. Photolithography & Extreme Ultraviolet (EUV) Lithography Process

Inside the semiconductor fabrication plant—which operates under the most stringent dust-controlled cleanroom standards (Cleanroom Class 100, meaning there are no more than 100 particles larger than 0.5 microns per cubic foot of air)—silicon wafers undergo hundreds of repeated cycles to print circuit patterns. The definitive "heart and bottleneck" of this entire industry is the Photolithography process.

To print transistor switch patterns at single-digit nanometer scales onto the wafer, fabs must deploy the pinnacle of modern technology: Extreme Ultraviolet (EUV) Lithography. These machines are manufactured exclusively by a single company in the world: ASML, based in the Netherlands. An EUV machine operates by firing a high-power CO_2 laser at microscopic droplets of molten tin falling at extreme speeds. This impact generates a high-temperature plasma that emits light at an ultra-short wavelength of just 13.5 nanometers. Because light at this wavelength is completely absorbed by conventional glass lenses, the EUV light must be guided and focused

using a system of specialized mirrors polished to near-atomic smoothness.

The EUV light is projected through a photomask (or Mask) containing the circuit master blueprint, shrinking the pattern down exponentially before hitting the silicon wafer. The wafer itself is coated with a light-sensitive polymer called a Photoresist. The areas exposed to the EUV light undergo a chemical transformation, allowing them to be selectively washed away during the Developing stage, leaving behind the precise geometric layout of the circuits. Subsequently, the wafer is routed through Etching processes using plasma gases, alongside Ion Implantation stages to dope the semiconductor. This sequence is repeated dozens of times to build up the intricate, multi-layered grid of transistors and metal interconnects that constitute the final AI chip.

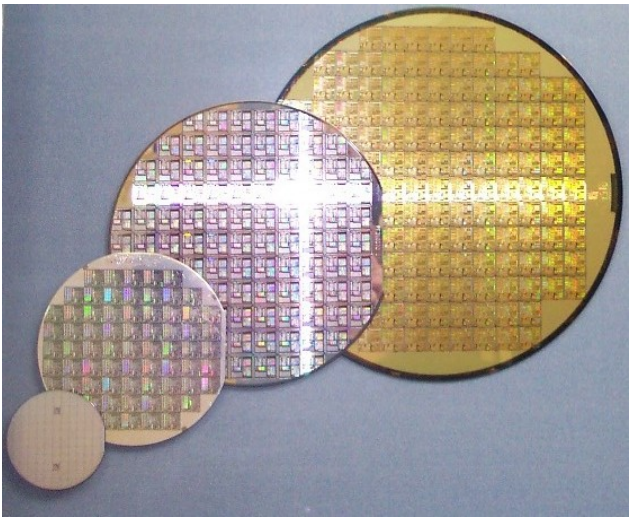


Figure 18: Silicon wafers in sizes of 2, 4, 6, and 8 inches.
(Source: https://en.wikipedia.org/wiki/Wafer_(electronics))

4. Advanced Packaging: Shattering the AI Chip Performance Bottleneck

Once the fabrication processes on the wafer are completed, hundreds or thousands of tiny individual square chips—referred to as Dies—are formed on the circular wafer and then diced apart. Historically, conventional packaging simply involved mounting a single die, wire-bonding it to metal pins, and sealing it inside a protective black plastic casing. However, for modern AI chips, this traditional approach creates an immediate data bottleneck; the processing core can compute data at breathtaking speeds, but conventional wiring cannot transfer data from the memory fast enough to keep up. Consequently, the industry has turned to Advanced Packaging, which has emerged as the deciding factor for AI hardware supremacy.

The current industry standard for this technology is TSMC's CoWoS (Chip-on-Wafer-on-Substrate). The CoWoS technique integrates the core logic die (such as a GPU or TPU) side-by-side with High Bandwidth Memory (HBM) modules on a shared silicon routing layer known as a Silicon Interposer. This interposer embeds tens of thousands of microscopic interconnecting wires, which drastically shortens physical travel distance to near-zero and exponentially amplifies the data transfer rates between the AI compute engine and memory. This advanced packaging process is one of the most critical missing jigsaw pieces, enabling modern superchips like the NVIDIA H100, B200, or Google TPU to achieve the massive throughput required to train and deploy Large Language Models (LLMs) efficiently.

When we look back from the atomic foundations of semiconductor physics, the electrical on/off switching mechanisms of MOSFETs, and 3D GAAFET structures, to the global supply chain's engineering triumphs in printing circuits with ultra-short EUV light onto

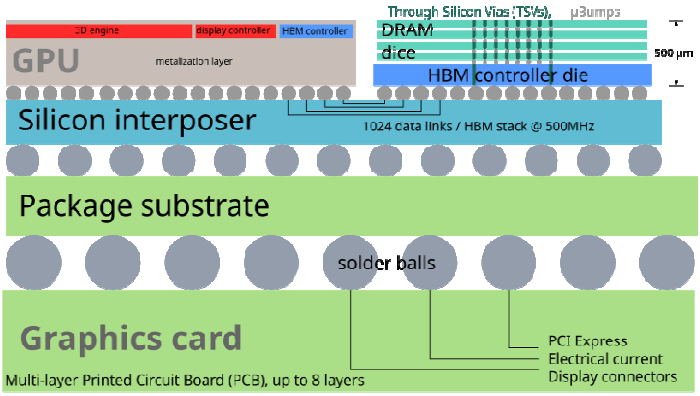


Figure 19: A cross-sectional view of a graphics card utilizing High Bandwidth Memory (HBM) with 3D chip stacking technology via Through-Silicon Vias (TSVs). (Source: https://en.wikipedia.org/wiki/Three-dimensional_integrated_circuit)

hyper-pure silicon—it all proves one thing: 'Computing power' is not an abstract concept. Rather, it is a tangible, physical manifestation derived from mastering the laws of physics and chemistry. Yet, as traditional scaling under Moore's Law collides with insurmountable physical walls of heat dissipation and quantum tunneling, the industry's survival no longer hinges on making smaller switches in general-purpose processors. Instead, it demands a total revolution in how we design the layout and architecture of these billions of transistors. Advanced packaging innovations like CoWoS and domain-specific architectures optimized for parallel acceleration have become the definitive bridge. They are shifting our trajectory from the physical constraints of materials science into a new epoch of cognitive hardware architecture—one that will write the next chapter of technological history.

3. Towards AI Chip

3.1 Why AI Chips?

In traditional computing, the Central Processing Unit (CPU) was designed as a 'generalist.' The internal architecture of a CPU is meticulously optimized for high-speed, sequential instruction processing. It features massive cache memories and complex control units to seamlessly handle operating system tasks, multi-tasking, and highly conditional logic (such as If-Else statements and Branch Prediction). However, as the world transitions into the era of artificial intelligence—specifically dominated by deep learning algorithms and Large Language Models (LLMs)—the nature of software computation has shifted radically. AI algorithms do not require processors that excel at complex logical branchings. Instead, they demand massive volumes of basic linear algebra calculations, specifically Matrix Multiplication and accumulation operations executed simultaneously. This computational behavior is known as Massively Parallel Processing. A standard CPU, which typically possesses only a few high-performance cores (generally 4 to 24 cores), cannot process these concurrent workloads efficiently, leading to severe bottlenecks in both execution time and thermal accumulation. To overcome this limitation, the technology industry had to develop a specialized hardware architecture dedicated entirely to this task: the AI Chip.

For a clearer understanding, we can categorize the necessity and architectural differences that compel the shift to AI chips into three primary dimensions:

1. Arithmetic Logic Unit (ALU) Density

Inside a processor, the unit responsible for executing mathematical calculations is the Arithmetic Logic Unit (ALU). When comparing internal structures, a CPU dedicates the vast majority of its silicon wafer area to complex control units and massive cache memories, leaving only a small fraction of space for a few ALUs per core. In contrast, an AI chip (including GPUs repurposed for AI) strips away these non-essential control layers and large caches. Instead, it floods the entire chip area with thousands or tens of thousands of smaller, streamlined computing units. This transistor layout allows an AI chip to achieve a level of simultaneous mathematical calculation (Parallelism) hundreds of times greater than a traditional CPU.

2. Parallel Instruction Paradigms (SIMD and Matrix Cores)

Traditional CPUs process data sequentially using a Single Instruction, Single Data (SISD) paradigm—where one instruction manipulates one data point at a time. AI chips, however, utilize a Single Instruction, Multiple Data (SIMD) architecture, or have evolved further to integrate specialized Matrix/Tensor Cores. These cores are dedicated hardware blocks engineered specifically to execute the Multiply-Accumulate (MAC) operations of entire matrix rows and columns within a single clock cycle. This paradigm shift drastically reduces the overhead of repeatedly fetching individual instructions, unlocking a massive leap in neural network processing efficiency.

3. Energy Efficiency (Performance-per-Watt)

As AI models scale up to encompass hundreds of billions of parameters, attempting to process them using general-purpose CPUs would require tens of thousands of chips packed into data centers. This

would demand astronomical amounts of electricity that power grids could scarcely sustain, while the system itself would risk thermal collapse from accumulated heat. AI chips solve this critical economic and engineering challenge. Because their internal circuitry is hardwired exclusively for AI-specific mathematics, they deliver exponentially faster computing throughput while consuming a fraction of the power required by CPUs.

For these reasons, the AI chip has evolved from being an optional hardware upgrade into a mandatory infrastructure—an absolute prerequisite for running, scaling, and training artificial intelligence in the modern era.

3.2 Technologies Behind World-Class AI Chips

As the computational demands of artificial intelligence algorithms skyrocketed far beyond the capabilities of general-purpose processors, a fierce global tech race emerged to develop domain-specific architectures. Today, the core technologies driving world-class AI Chips can be classified into three primary families based on their engineering structures and operational use cases: GPUs, TPUs, and NPUs. Each family possesses distinct strengths and entirely different internal processing mechanisms.

1. Graphics Processing Unit (GPU)

From Graphics Pipeline to the Bedrock of AI: Originally built to render 3 D graphics and power video games, the Graphics Processing Unit (GPU) was discovered by researchers to be perfect for neural network matrix multiplications due to its internal architecture, which consists of thousands of small, highly parallel computing cores. The undisputed global market leader in this segment is NVIDIA, which

cemented its AI dominance through advanced hardware architectures, such as the cutting-edge Blackwell platform. The technological secret behind modern AI GPUs lies in integrating dedicated hardware blocks called Tensor Cores directly onto the silicon die. These blocks specialize in Reduced Precision arithmetic (e.g., swapping data formats to lower bit-widths), exponentially accelerating computational throughput without compromising model accuracy. Furthermore, by pairing these cores with High Bandwidth Memory (HBM) and high-speed interconnects like NVLink, tens of thousands of GPUs can be clustered together inside data centers, functioning seamlessly as a single, colossal superchip.

2. Tensor Processing Unit (TPU)

Unlike GPUs, which evolved from a graphics background, Google's Tensor Processing Unit (TPU) is a pure Application-Specific Integrated Circuit (ASIC)—custom-designed from scratch exclusively for machine learning workloads. First introduced in 2016 and continuously engineered across multiple generations, the beating heart of latest TPU is the Systolic Array architecture. The systolic array mimics the rhythmic pumping mechanism of a human heart. In traditional computing, data constantly travels back and forth between the processor and memory after every single calculation, creating severe bottlenecks and wasting power. The systolic architecture overrides this by allowing data to flow continuously through a hardwired grid of interconnected ALUs. As data enters, multiplication and addition results are immediately "pumped" or passed directly to adjacent cores in waves. This technology enables the Matrix Multiply Units (MXUs) to process hundreds of thousands of data points concurrently within a single clock cycle. Today, Google links these TPU chips into massive supercomputing clusters via optical circuit switches to train and deploy its flagship Gemini model family.

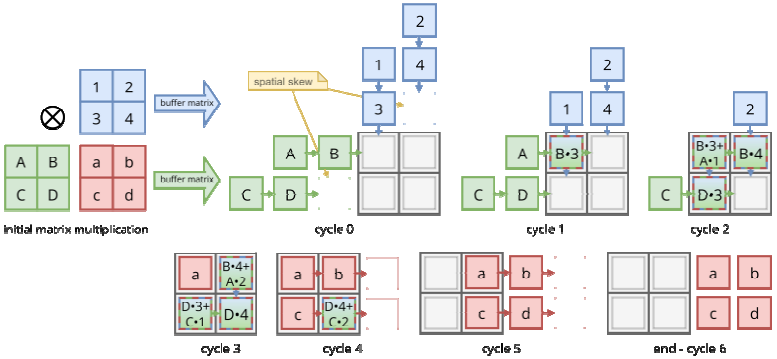


Figure 20: A systolic array algorithm performing result accumulation within processing elements.

(Source: https://en.wikipedia.org/wiki/Systolic_array)

3. Neural Processing Unit (NPU)

On-Device AI Powerhouses for Mobile Ecosystems: While GPUs and TPUs wage the AI war across massive cloud networks and hyperscale data centers, the NPU represents the miniaturization of AI chips, embedded directly into everyday client devices like smartphones, tablets, and next-generation PCs. Prime examples include Apple Silicon, Qualcomm Snapdragon, and the latest mobile processors from Intel. The ultimate goal of an NPU is not the raw, brute-force performance of a supercomputer, but rather extreme energy efficiency. NPUs are custom-tailored to run localized On-Device AI models. They handle daily tasks such as facial recognition (e.g., Face ID), real-time computational photography, and instantaneous voice-to-text transcription. Architecturally, NPUs are designed to stream low-precision integer data types rather than heavy floating-point math. This strategy slashes cache area requirements and minimizes power-hungry data

movement. Consequently, it allows mobile hardware to execute tens of trillions of operations per second (TOPS) of AI inference smoothly—all without draining a smartphone's battery in a matter of minutes.

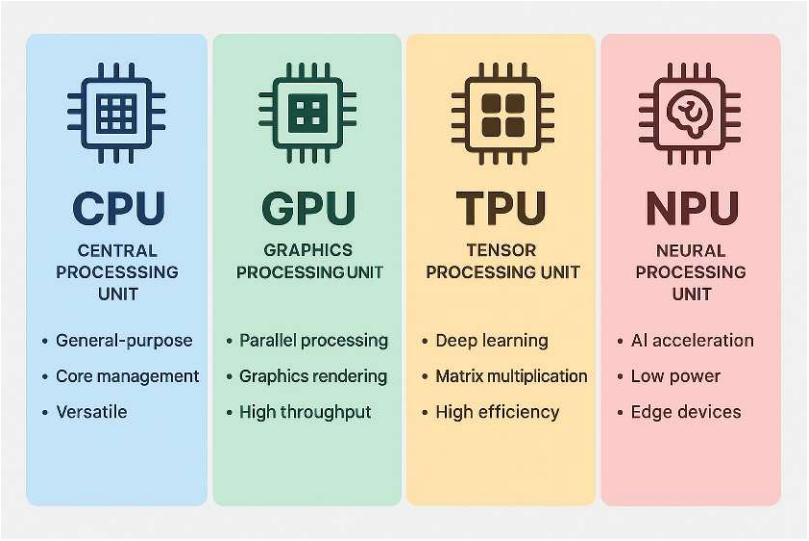


Figure 21: A comparison profile of CPU, GPU, TPU, and NPU specifications and architectures.
(Source: https://resources.l-p.com/knowledge-center/cpu-vs-gpu-vs-tpu-vs-npu-architecture-comparison-explained)

3.3 Future of AI Chips

Although current cutting-edge AI architectures—such as ultra-high-performance GPUs featuring Tensor Cores and High Bandwidth Memory (HBM)—can driven trillion-parameter models with breathtaking capability, computer architects and hardware engineers view these technologies as rapidly approaching a looming "dead end."

The fundamental crisis stems from a paradigm that has anchored the computing industry for over half a century: the Von Neumann Architecture. This design mandates a strict, physical separation between the processing unit and the memory unit, linked together by a data bus.

In AI workloads, hundreds of billions of weight parameters must be constantly shuttled back and forth across these communication channels. This triggers the infamous Von Neumann Bottleneck; no matter how fast the processing core computes, it spends critical clock cycles idling, waiting for data to arrive from memory.

Worse yet, over 80% of the total energy consumed by modern AI chips is entirely wasted as electrical resistance and heat dissipation generated during this data movement. This physical reality has catalyzed an energy crisis and pushed thermal limits to the brink in hyperscale data centers worldwide.

To avert this crisis, the engineering community—spanning electrical, computer, and semiconductor disciplines—is actively revolutionizing AI chip design. They are shifting away from legacy computing paradigms and toward three groundbreaking architectural frontiers that will entirely reshape the future of technology.

1. Neuromorphic Computing (Brain-Inspired Computing)

The human brain contains approximately 86 billion neurons and hundreds of trillions of synapses. It can analyze complex situations, recognize faces, and learn deep concepts while consuming only about 20 Watts of power—roughly equivalent to a small light bulb. In stark contrast, a modern supercomputer running an AI model of equivalent complexity demands multiple megawatts of electricity, enough to power an entire town. To bridge this massive efficiency gap, Neuromorphic Engineering was conceived to mimic the biology of the human brain. Neuromorphic chips completely discard traditional, legacy clock-driven digital systems, which continuously cycle electricity through every transistor simultaneously even when no data is being processed. Instead, they pivot to an architecture known as Spiking Neural Networks (SNNs). The internal circuitry uses transistors and capacitors to simulate biological neural behaviors. Under this architecture, the hardware switches remain in a dormant, zero-power state until an incoming data signal—or an electrical "Spike"—arrives and excites the node past a specific Threshold. This mechanism is known as Event-driven Computing. Beyond slashing power consumption by thousands of times, it enables the hardware to adapt and learn continuously in real time (On-chip Learning) by mimicking synaptic plasticity. This extreme efficiency makes neuromorphic chips ideal for advanced embedded systems, autonomous robotics, and exploration drones that must process massive sensor streams in edge environments without being tethered to a power grid. Leading milestones in this research frontier include Intel’s Loihi neuromorphic research testbed and IBM’s TrueNorth processor.

2. Optical / Photonic AI Chips

Breaking the Shackle of Electrons: When electrons flow through microscopic copper or aluminum interconnects at nanometer scales inside a traditional chip, they continuously collide with the metallic atomic structures. This friction generates electrical resistance, converting critical energy into accumulated heat. The faster data needs to be transferred, the more severe this thermal dissipation becomes. The future of AI hardware aims to liberate itself from these physical constraints of electronics by pivoting to Photons (light particles) traveling through semiconductor mediums—a pioneering field known as Silicon Photonics.

An Optical AI Chip completely replaces conventional transistor layouts with an intricate network of microscopic Optical Waveguides, splitters, and phase shifters. To execute the linear algebra calculations that form the core of AI, these photonic processors leverage the fundamental physical properties of light waves. For instance, by causing two light beams of varying intensities to undergo optical Interference, the hardware can instantly compute matrix additions and multiplications.

Because photons travel at the speed of light and possess no mass, they generate zero electrical resistance and produce virtually no heat. This allows the chip to compute hundreds of billions of matrix operations almost instantaneously—the very moment the light passes through the optical circuit. Operating at femtosecond timescales with near-zero latency, photonic chips represent the ultimate frontier in shattering the computational speed barriers for training next-generation Large Language Models (LLMs) and deep neural networks.

3. In-Memory Computing & The Memristor Revolution

If the industry cannot immediately transition to full photonics or neuromorphic biological systems, another vital survival strategy is to completely tear down the physical wall dividing processors and memory, fusing them into a single, unified entity. This paradigm is known as In-Memory Computing. The core engine driving this innovation is a breakthrough in materials science: the fourth fundamental passive circuit element (following the resistor, capacitor, and inductor)—the Memristor (Memory Resistor). A memristor possesses the unique capability to alter its internal electrical resistance based on the history of the electrical current that has previously flowed through it. Crucially, it will "remember" and retain that precise resistance state indefinitely, even after the power supply is completely turned off. This provides it with powerful non-volatile memory characteristics. By arranging millions of these memristors into a grid known as a Crossbar Array, engineers can store AI model weights directly within the hardware's physical resistance values (R). When an electrical voltage vector (V) representing data inputs passes through this array, the circuit executes Ohm's Law ($I = V/R$) and Kirchhoff's Current Law simultaneously across the entire grid. This allows matrix multiplications to be computed directly inside the memory cells themselves. By eliminating the need to move data to an external processor, this approach completely bypasses the Von Neumann bottleneck and ushers in a new era of ultra-efficient computing.

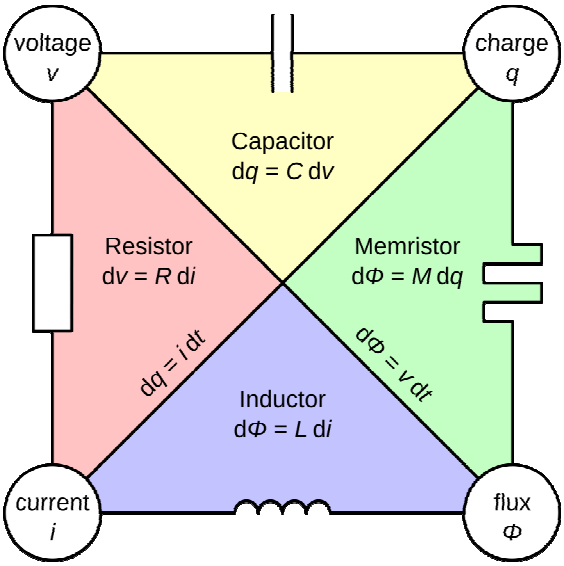


Figure 22: A diagram illustrating the conceptual symmetry among the four fundamental circuit elements: resistor, capacitor, inductor, and memristor. (Source: https://en.wikipedia.org/wiki/Memristor)

Engineers arrange millions of these memristors into a large grid known as a Crossbar Array to store the weights of an AI model in the form of electrical conductance (G = 1/R). When input data needs to be computed, it is applied to this grid structure in the form of electrical voltage (V). Based on fundamental laws of physics, specifically Ohm's Law (I = V G), the voltage passing through the conductance automatically generates an electrical current (I), which instantly executes the mathematical multiplication. Simultaneously, the currents from all converging paths flow together at the end of the wire according to Kirchhoff's Current Law, immediately achieving the

cumulative sum. This process allows the AI chip to complete millions of multiply-accumulate operations directly within its own memory cells. It entirely eliminates the waste of time and energy spent shuttling data back and forth across a computer bus, ultimately skyrocketing the system's performance-per-watt efficiency by dozens of times over legacy architectures.

Our journey through AI chip architecture—from foundational parallel processing and the clash of CPUs, GPUs, TPUs, and NPUs, to futuristic frontiers like neuromorphic, photonic, and in-memory computing—reflects that humanity has entered a "gold rush" for computational power. This battlefield is no longer confined to corporate tech labs; it has become a critical national agenda where countries compete for technological sovereignty.

The South Korean Phenomenon: A prime example is South Korea, where the AI chip industry has grown exponentially to become the backbone of the national economy. Through national initiatives like the K-AI Semiconductor Growth Forum, the country aggressively propels fabless startups to the global stage, turning AI chip engineering into one of the nation's most prestigious career paths. AI chips are far more than just electronic circuit boards. They are the ultimate, highly influential infrastructure that will dictate the direction, delegate computational power, and drive global intelligence in the next era. For a deeper dive into this macro-scale Asian rivalry, explore the analytical documentary "Can Samsung Overtake TSMC? Taiwan vs. South Korea AI Chip Rivalry Explained" on YouTube.

References and Further Reading

Deep Learning Algorithm Architecture & Computation

- Wikipedia contributors. "Deep learning." Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Deep_learning
- Wikipedia contributors. "Large language model." Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Large_language_model

Hardware Accelerators & Computer Architecture

- Wikipedia contributors. "Hardware acceleration." Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Hardware_acceleration
- Wikipedia contributors. "Graphics processing unit" (Section on Tensor Cores & AI). Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Graphics_processing_unit
- Wikipedia contributors. "Tensor Processing Unit" (Systolic Array Architecture). Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Tensor_Processing_Unit
- Wikipedia contributors. "Neural processing unit." Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Neural_processing_unit

Limitations & Future AI Chip Technologies

- Wikipedia contributors. "Von Neumann architecture" (Von Neumann Bottleneck section). Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Von_Neumann_architecture
- Wikipedia contributors. "Neuromorphic engineering" & "Spiking neural network." Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Neuromorphic_engineering
- Wikipedia contributors. "Silicon photonics." Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Silicon_photonics
- Wikipedia contributors. "Memristor" & "In-memory processing." Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Memristor

South Korea's AI Chip Industry & Current Landscapes (K-AI Semiconductor)

- Government Policies & NPU Initiatives: Driven by the Ministry of Science and ICT (MSIT), Republic of Korea, through platforms like the "K-AI Semiconductor Growth Forum" to evaluate performance indices (K-Perf).
- Market Milestones (June 2026): South Korean AI chip exports hit $30 million. Local NPU startups—including Rebellions, FuriosaAI, and Mobilint—are actively challenging NVIDIA's dominance through strategic partnerships with Samsung SDS and SK Telecom.