Introduction to AI chip making in China

Relevant background, considerations, and forecasting questions

Institute for AI Policy and Strategy (IAPS)

December 14, 2023

AUTHORS

Erich Grunewald — Associate Researcher — Institute for AI Policy and Strategy

Christopher Phenicie — Independent Researcher

Summary

The Chinese[1] government is committed to indigenizing AI chip design and manufacture, having set it as a national priority and invested many billions of dollars in related ventures. It is only with access to large numbers of AI chips – tens of thousands and growing – that the development of frontier AI systems is possible.[2] While the Chinese semiconductor industry has had some notable successes in chip design and fabrication, it faces the formidable challenge of trying to replicate a highly advanced and globalized supply chain – currently spread out over Germany, Japan, the Netherlands, South Korea, Taiwan, the United States, and elsewhere – in one country alone.

The task of indigenization is made more difficult by extensive export controls on advanced semiconductors and related tooling and materials that target or affect China:

- US export controls, introduced in October 2022 and revised in October 2023, ban exports of AI chips above a certain threshold to China, restrict Chinese AI chip makers' access to foreign chip fabrication plants ("fabs"), and restrict exports of semiconductor manufacturing equipment (SME), components, and design software to China.[3]
- Japanese export controls, announced in May 2023, restrict exports of key SME and materials.
- Dutch export controls, announced in June 2023, restrict exports of key SME.

Better understanding China's ability to indigenously produce AI chips over the next years can help solve strategically important open problems in AI governance, such as how capable AI systems developed in China will be in the future, and whether we will see major efforts to smuggle large quantities of AI chips. To that end, we have created a set of forecasting questions, published on the forecasting platform Metaculus, about Chinese progress in AI chip making. We encourage interested readers to make their own forecasts on some of these questions, with this report as a guide. (more)

[1] When we use the term "China" in this report, we refer to the People's Republic of China (including Macao and Hong Kong), not the Republic of China (Taiwan).

[2] For reference, the state of the art in large language models, OpenAI's GPT-4, was trained with about 25K AI chips (Patel & Wong, 2023). The number of AI chips used to train a typical frontier model is increasing, as money invested in AI model training outpaces improvements to AI chip price-performance.

It is possible to train AI systems without AI chips, e.g., using consumer-grade graphics cards. However, in practice all frontier AI is developed using AI chips today, and even if that were not the case, forecasts of the quality of indigenous Chinese AI chip production would still be valuable. That is because both AI chips and consumer-grade graphics cards are designed and manufactured in a similar way, with the same equipment and material, meaning their development is subject to similar constraints.

[3] Note that the controls on AI chip exports and foreign fabs are relevant to Chinese semiconductor progress in that they incentivize indigenization by cutting Chinese firms off from foreign suppliers.

This report aims to provide a useful introduction for anyone interested in indigenous Chinese AI chip making and related export controls, or interested in making forecasts about Chinese AI chip making progress. To do so, we:

- Explain four key concepts in semiconductor manufacturing: "AI chip"/"accelerator"/"server", "process node", "yield", and "non-planar transistor". (more)
- Provide an overview of the chip-making process, including design, fabrication, and assembly, test, and packaging (ATP). (more)
- Provide an overview of relevant export controls. (more)
- Describe key inputs (listed below), context, and organizations relevant to understanding Chinese chip-making progress. This includes giving overviews of chip fabrication and foundries, AI and memory chip makers, and the software, equipment, and materials necessary for AI chip making. (more)

Key inputs into the Chinese semiconductor industry include:

1. **Design software and IP cores.** Electronic design automation (EDA) software is necessary for designing AI chips and validating those designs. Semiconductor intellectual property cores ("IP cores") are reusable building blocks that can be licensed from IP core vendors.
2. **SME and materials.** SME and some materials are difficult to produce and crucial to the semiconductor manufacturing process. One particularly important type of SME is photolithography machines.
3. **Intellectual property and other explicit knowledge.** Various kinds of explicit knowledge are necessary for designing and fabricating chips, as well as developing tooling and materials to do so. Such knowledge can be acquired through research, licensing agreements, inspections, reverse engineering, and intellectual property theft.
4. **Talent and know-how.** Operating a fab and doing the research and development needed to create new chip designs, SME, materials, and software all require many researchers, technicians, and engineers, and tacit knowledge that is gained with years of experience.
5. **Investment and subsidies.** The semiconductor business is highly capital intensive. The Chinese government has provided tens of billions of dollars in subsidies to support its semiconductor industry via the China Integrated Circuit Industry Investment Fund ("Big Fund"). However, the Big Fund has to some extent suffered from corruption and waste, and it is unclear how effectively leading Chinese semiconductor companies can absorb additional funding.

Table of contents

Summary 1
Introduction 4
Key concepts 6
AI chips, accelerators, and servers 6
Process node 7
Yield 9
Non-planar transistor 9
Overview of the chip-making process 11
Design 11
Fabrication 12
Material preparation 12
Wafer fabrication 12
Assembly, test, and packaging 13
Key inputs, context, and organizations 15
Chip makers 16
Memory chip makers 17
Chip fabrication and foundries 18
Software, equipment, and materials 19
EDA software and IP cores 20
Photolithography 20
Relevant export controls 23
SME and materials controls 23
Photolithography controls 24
Forecasting questions 26
Links and resources 28
Acknowledgements 30
Bibliography 31

Introduction

On October 7th, 2022, the United States instituted wide-ranging export controls aimed at preventing China's "military modernization, including the development of weapons of mass destruction (WMD), and human rights abuses" by limiting Chinese development of advanced AI. The October 7th controls banned exports of AI chips above a certain threshold to China, restricted Chinese AI chip makers' access to foreign chip fabrication plants ("fabs"), and restricted exports of US-made semiconductor manufacturing equipment (SME), components, and electronic design automation (EDA) software to China (Allen, 2022).

A lot has happened since then. After months of negotiations with the US, Japan announced its own restrictions on key SME and materials in May 2023 (Kelly & Uranaka, 2023), and the Netherlands followed suit a month later (Government of the Netherlands, 2023). Japan and the Netherlands both dominate crucial parts of the semiconductor supply chain. And in October 2023, the Bureau of Industry and Security (BIS) – the agency tasked with administering and enforcing US export controls on dual-use goods like semiconductors – substantially updated the previous year's rules (see Relevant export controls).

The extent to which these efforts will or will not succeed has major implications for the Chinese AI industry, given the compute-intensive nature of modern AI development (Sevilla et al., 2022). The Chinese government and industry, meanwhile, are committed to indigenizing AI chip design and manufacture. The Chinese Communist Party's 14th Five-Year Plan, covering the period from 2021 to 2025, emphasizes innovation and technological self-reliance (Cheung et al., 2022) and lists semiconductors as one of seven key technological priorities (Xinhua News Agency, 2021). As of September 2023, China reportedly intends to launch a major new state fund aiming to invest $41B in domestic semiconductor companies (Zhu et al., 2023).

The Chinese semiconductor industry has seen some successes. The Huawei Ascend 910 and Biren BR100 AI chips, announced in 2019 and 2022 respectively, were in some respects competitive[4] with the then-leading Nvidia A100, but relied on foreign foundries for their fabrication, an avenue that has been shut off since October 2022. Chinese equipment and materials suppliers have also made progress – for example, SME firms like Naura and Advanced Micro-Fabrication Equipment Inc. China (AMEC) have made substantial advances in deposition and etching and cleaning systems. Another sign of Chinese progress appeared in September 2023, when Huawei launched a new line of smartphones containing a chip that had a sophisticated design and was fabricated indigenously using an advanced 7 nm process (see Process node), albeit with foreign-made SME (Savov & Wu, 2023).

[4] For example, on the MLPerf benchmarks for data center inference (v2.1), a Biren server handles 14K language model queries per second, while Nvidia A100 80GB servers handle from 12K to 19K queries per second. Note, however, that this benchmark uses BERT-large, a comparatively small language model, with three orders of magnitude fewer parameters than GPT-3; the relative performance of the two chips may be different for larger models like GPT-3 or GPT-4. On the Huawei Ascend 910, see, e.g., Kennedy (2020).

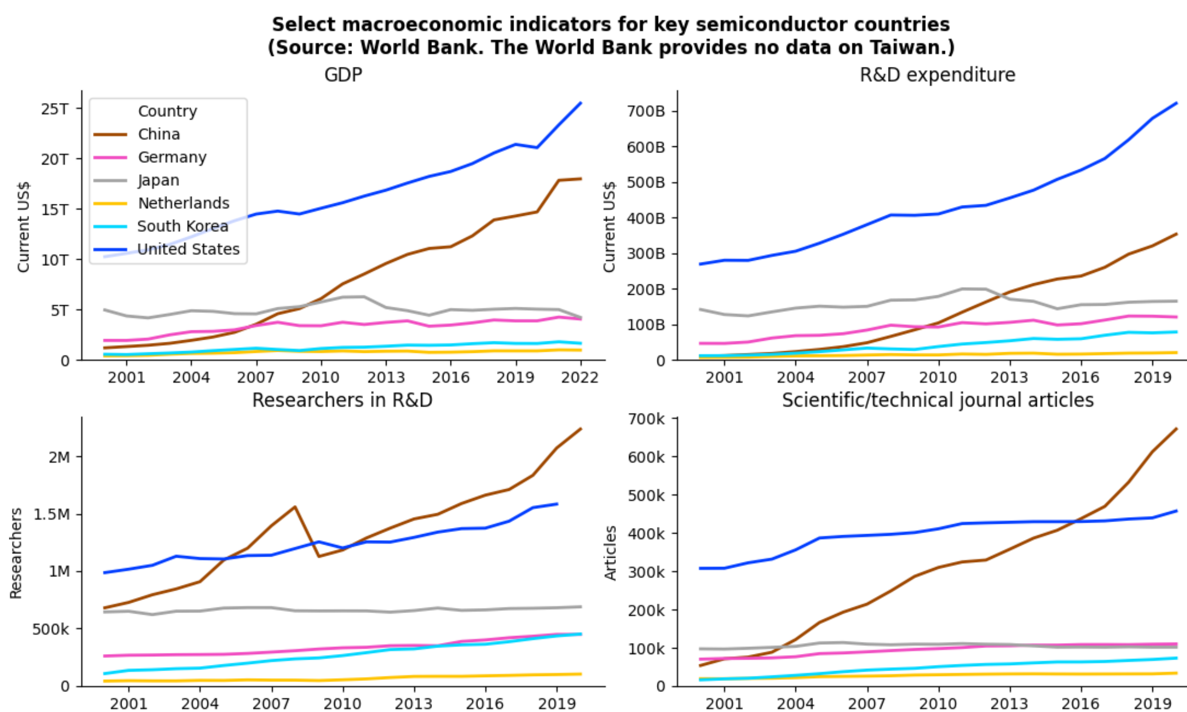

Figure 1. Select macroeconomic indicators for key semiconductor countries. Taiwan was not included in the original data. Source: World Bank.

However, many challenges remain for the Chinese semiconductor industry. For one, it is trying to replicate a highly advanced and globalized supply chain, in a sense trying to catch up to not only the American semiconductor industry, but also simultaneously those of Germany, Japan, the Netherlands, South Korea, and Taiwan (see Figure 1). That matters because semiconductor manufacturing is extremely specialized and capital-intensive, involving some of the most advanced machinery humans have invented. But there are also a number of specific technologies where Chinese companies are substantially behind the state of the art and struggling to catch up, most notably photolithography, a crucial part of semiconductor manufacturing and a highly costly and difficult technology to develop (see Photolithography).

This report has two goals:

1. It aims to be a useful primer for anyone interested in Chinese AI chip making progress and related export controls. As such, it gives an overview of the chip-making process, as well as of relevant export controls.
2. It aims to be a useful introduction and reference for anyone interested in forecasting China's ability to indigenously produce AI chips. To this end, it describes key inputs, context, and organizations relevant to understanding and forecasting Chinese chip-making progress.

Key concepts

AI chips, accelerators, and servers

An integrated circuit ("chip") is a group of electronic circuits laid out on a piece of semiconductor material. AI chips are integrated circuits specialized for AI training and/or inference workloads. There are different types of AI chips. Some, like Google's Tensor Processing Units (TPUs), are application-specific integrated circuits (ASICs), chips customized for highly specific workloads such as tensor[5] operations used in deep learning training/inference. Other AI chips, like those made by Nvidia, are graphical processing units (GPUs), chips designed to more generally execute mathematical operations in a highly parallel manner.[6]

Increasingly, AI chips are in fact composed of several smaller, modular chips ("chiplets") mounted on a larger "interposer" chip. For example, the H100 consists of the GH100 GPU chiplet as well as six high-bandwidth memory (HBM) chiplets[7] (Figure 2, left).[8] The move toward chiplets is motivated by an increasing need for high-speed communication between compute and memory, limitations on how large you can make increasingly advanced chips, flexibility in using the most advanced processes only for the parts of a chip that need it, and considerations around yield (Moore, 2019; McWilliams, 2022). (For more, see Assembly, test, and packaging.)

Although this report takes "AI chips" as the main unit of analysis, individual chips are typically not what is sold to end users. Rather, once an AI chip has been fabricated (see Overview of the chip-making process), the chip maker assembles it into a device ("AI accelerator"), as seen in Figure 2 (middle).[9] AI accelerators are either sold directly to end users or distributors, or to original equipment manufacturers (OEMs), which integrate those accelerators into servers (Figure 2, right). A server typically contains about four to eight AI accelerators, a central processing unit (CPU), additional memory, high-speed networking components, persistent storage, a cooling system, and more. For example,

[5] You can think of a "tensor" as an array of a particular dimension, e.g., a scalar, a vector, or a matrix.
[6] GPUs are so named because they were originally designed to perform computations needed to render graphics. Many GPUs, such as those in consumer-grade graphics cards, are still optimized for graphics rendering, but some GPUs are now primarily optimized to perform AI workloads (which have similarities with those in graphics rendering). Like TPUs, Nvidia's data center GPUs have features aimed at speeding up tensor operations used in deep learning, but in addition to those, and unlike TPUs, GPUs also have features unrelated to deep learning, such as for genomics and graph analysis.
[7] Note that, although HBM stacks are not always referred to as chiplets, they sometimes are, and they do fall under that category for most definitions of "chiplet".
[8] The H100 is sold in different versions. Some versions of the H100 architecture feature five, not six, HBM chiplets.
[9] Not all AI chips are produced for use in AI accelerators and therefore data centers. There are also, for example, AI chips developed purely for inference outside data centers ("edge" computing). However, this report focuses only on AI chips intended for data centers, where frontier AI models are always trained.

the Nvidia H100 is an AI accelerator that is used in Nvidia DGX H100 servers. While a chip might be the size of a postage stamp, an accelerator is the size of a rather elongated hardcover book, and a server is the size of two carry-on luggages. Customers typically buy servers, and these servers are then installed in cabinets in data centers alongside other necessary infrastructure, either by a cloud service provider or a company that rents out space in a data center ("colocation provider").[10]

Figure 2. Left, an Nvidia H100 printed circuit board, without casing, with the GH100 GPU (outlined in red) and six surrounding high-bandwidth memory chips (outlined in blue) in the center (source: CNET). Middle, an Nvidia A100 accelerator, with casing (source: Linus Tech Tips). Right, Jensen Huang, president and CEO of Nvidia, signing the first delivered Nvidia DGX server (source: Nvidia).

Process node

A term that often comes up when discussing semiconductor manufacturing is "process node" (or "technology node"). A key driver of chip improvements is to make the physical features on integrated circuits ("chips" or "ICs") ever smaller in order to fit more of them on a given area.[11] "Process node" is a measure of that progress – the smaller the process node, the more transistors per area. However, it is an ambiguous and for that reason often misinterpreted term.

The ambiguity is partly due to the measure no longer corresponding to any physical dimension on the chips. Originally, "process node" followed the size of the devices (e.g., transistors) on a chip, measured in micrometers (µm) and later nanometers (nm). However, over time the process node measure stopped tracking physical dimensions, and it should now be better understood as "generation", in the same way that 4G and 5G describe new generations of cellular network technology. (Transistors are still shrinking,

[10] There is sometimes confusion about the difference between a data center and a supercomputer. A supercomputer is a set of computers, e.g., AI accelerators, connected together in a cluster. A data center is a physical facility housing computers. A supercomputer is always located in a data center.
[11] This is desirable because, with some caveats, the more devices (transistors, etc.) you can fit on a given area, the more performant or capable the chip will be. For example, more transistors can allow logic chips to perform more computations, and more capacitors can allow memory chips to store more information. So density matters, because chips cannot be made arbitrarily large due to various physical constraints, and because there are various other costs in compensating from worse chip performance by using more chips.

with modern ones now having some features as small as 5 nm; for reference, the distance between neighboring atoms in silicon is about one twentieth of that.[12]) In recognition of this, companies are adopting new naming conventions, for example the N3 process by Taiwan Semiconductor Manufacturing Company Limited (TSMC), Samsung's 3GAE process, and the Intel 3 process (all roughly belonging to the 3 nm generation).

The most advanced process today is TSMC's N3E, belonging to the 3 nm class, which can fit an estimated 216 million transistors per square millimeter (MTr/mm^2) (Schor, 2023). (However, note that measures of transistor density vary for different chip architectures.) The most advanced process node achieved by a Chinese company is the Semiconductor Manufacturing International Corporation (SMIC) N+2 process, which is in the 7 nm class and reached volume production in 2023. Equivalent processes were achieved by TSMC in 2018, Samsung in 2019, and Intel in 2021 (WikiChip, 2022). (These dates refer to volume production. Usually, a company will begin by producing small volumes of chips on a new process, called "risk production", and only later scale up once the process has been refined.) Figure 3 shows transistor density improvements by different manufacturers over the last decade.

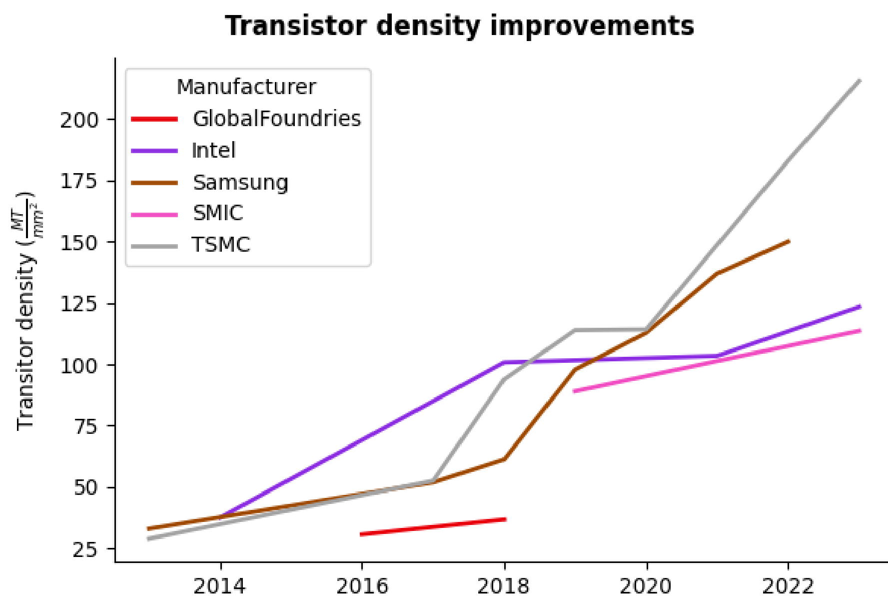

Figure 3. Transistor density improvements over time. This data is incomplete and based largely on relatively uncertain estimates. Sources: Wikipedia, WikiChip.

[12] The 2021 International Roadmap for Devices and Systems describes the 3 nm process node as having transistors with a fin width of 5 nm. Silicon crystals have a diamond cubic structure, with a lattice constant of about 0.54 nm, meaning the nearest-neighbor distance between silicon atoms is about 0.24 nm.

Yield

Not every chip makes it through the fabrication process without defects. The proportion of working chips to the number of chips that could potentially have been fabricated given the same inputs is called "yield". It is given as a percentage, where 80% means that, on average, one fifth of chips were defective. Defects can be caused by contaminants, environmental conditions, equipment issues, and more. Yield is a problem in semiconductor manufacturing because it involves hundreds of sequential steps and extreme precision and environmental control. (See Overview of the chip-making process for more on the semiconductor fabrication process.)

Increasing yield is a priority for fabs, because increases in yield directly translate to increases in profitability. New process nodes will typically start out with low yield, and yield is then gradually improved as the fab refines the process. Yield can vary significantly for different chip sizes, process nodes, and companies. Take, for example, the introduction of 3 nm processes: While Samsung's 3GAE process had a reported initial yield of 10-20% after it began mass production in 2022, TSMC's N3 process, which started mass production a few months later, had a reported yield of 60-80% (Shilov, 2022).[13]

Non-planar transistor

The transistor, invented in 1947, is an electrical switch that controls the flow of electrical current in a circuit, either letting current pass through, or blocking it, depending on an input voltage (the "gate"). You can think of it like a faucet or tap, where a handle (gate) lets you block or free the flow of water (current). The transistor is the basis for all modern logic chips, and most modern electronics.

Each new process node generation incorporates new breakthroughs in order to shrink transistors to ever smaller sizes. One important area of innovation is in the design of transistors (in particular "field effect transistors", or FETs), and the movement from planar, "two-dimensional" transistors to non-planar, "three-dimensional" geometries. As transistors have gotten smaller, planar designs have started running into problems, necessitating new, non-planar designs (Chatterjee, 2020). There are two types of non-planar transistors used in commercial AI chips:

- FinFETs (named for the fin-shaped structure protruding from the semiconductor base), first introduced by Intel at its 22 nm process (WikiChip, 2020) and later by others at the 16/14 nm process nodes (WikiChip, 2023).
- Gates-all-around FET (GAAFET), the successor of FinFET, which was introduced at the 3 nm process node by Samsung (WikiChip, 2022), and which will be used at the 2 nm nodes by Intel and TSMC (Wikipedia, 2023).[14]

[13] Yield figures depend on which design is being fabricated. Typically yield numbers are given for a standardized test chip.

[14] Specifically, Samsung is using multi-bridge channel FETs (MBCFETs), a variant of GAAFETs, for its 3 nm process. Intel is developing RibbonFETs for its 2 nm process, another variant of GAAFETs.

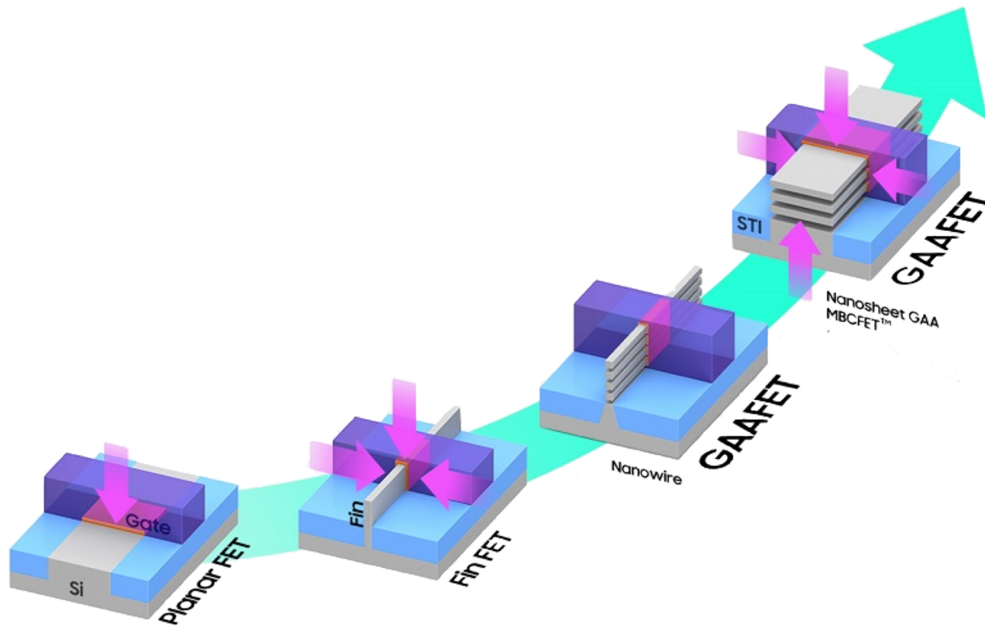

Figure 4. Planar and non-planar transistor designs. Source: Samsung.

Manufacture of non-planar transistors is substantially harder than planar transistors but necessary for leading-edge processes.[15] Switching to a new transistor design requires substantial changes across the design and manufacturing process. As a result, some export controls, e.g., those on design software, explicitly refer to non-planar transistors.

[15] Similarly to advanced logic chips, memory chips are also fabricated with non-planar processes (Patel et al., 2023), though these are different from the techniques described in this section.

Overview of the chip-making process

At a high level, AI chips are made in three steps: design, fabrication, and packaging. The chips are then assembled into devices ("AI accelerators"), which in turn are typically integrated into servers that can be mounted in data center racks. The semiconductor design and manufacturing process is highly demanding and sophisticated. This section provides an overview of that process.

Design

Chip designers specify high-level requirements, turn those into a detailed abstract model of the circuit logic, and use electronic design automation (EDA) software to translate that model into a blueprint of the physical circuit. The high-level requirements are typically based on earlier research carried out to find novel ways of improving performance. The design process involves defining the structure ("microarchitecture") of the circuit's logic components, designing each component or licensing a ready-made component ("IP core") from a vendor, and specifying the connections between components. It is an iterative process. The chip's specifications can be revised in order to trade off cost and performance. The design is also tested and validated to help ensure that it will behave as expected once fabricated, for example by simulating the design, running checks on code to scan it for errors, and making formal proofs about parts of the circuit logic.

An AI chip is designed for a particular process node. The fabrication plant ("fab") will provide the chip maker with a process design kit (PDK), which provides information needed to design and validate chips for one of the fab's processes. The end result of the design process is design files containing a detailed description of the chip's physical layout, which is shared with the fab. Chip design is a difficult and laborious process. Chip designers like Nvidia and AMD have tens of thousands of employees (Wikipedia, 2023; Wikipedia, 2023), and the design process for cutting-edge AI chips lasts for a year or longer.[16]

[16] We base the claim about how long AI chip design takes on conversations with people familiar with the chip design process. In addition, according to Jouppi et al. (2021), it took 27 months between the deployments of TPUv1 and TPUv2, 15 months between TPUv2 and TPUv3, and 15 months between TPUv3 and TPUv4, though it is possible that the design processes overlap, or that there are pauses between them. Making a new chip design from scratch, rather than iterating on previous designs, is more involved and takes longer.

Fabrication

Material preparation

Chips are fabricated on thin, circular slices of semiconductor[17] called wafers, typically 300 mm (12 inches) in diameter.[18] The most common semiconductor used is silicon. Modern silicon wafers, which are grown in furnaces and then sliced, polished, cleaned, doped, and inspected, need to be extremely pure and smooth in order to avoid defects in chip fabrication.

Prior to fabrication, photomasks are produced from the description of the physical layout provided by the chip designer. A photomask holds a pattern that will be imprinted onto the wafers, like a custom stencil.[19] Photomasks are created using direct-write ("maskless") lithography equipment, involving multiple stages of optical and electron beam lithography, etching, and inspection to ensure there are no defects.

In addition to wafers and photomasks, semiconductor manufacturing also requires various specialized chemicals and gasses.

Wafer fabrication

Supplied with silicon wafers and photomasks, the fab is ready to manufacture the chips. The wafer fabrication process involves imprinting patterns onto the silicon wafer in layers of different materials, including insulators (dielectrics), conductors (metals), and semiconductors. It is a highly difficult and complex process, as it demands extreme precision, involves hundreds of individual steps, and must at the same time achieve high economic efficiency. What follows is a simplified description of that process.

First, the transistors and other devices (resistors, capacitors, etc.) are formed by repeatedly depositing, patterning, and etching layers of materials. The wafer is always processed as a whole, allowing trillions of transistors to be created in a short amount of time. The reason why the wafer can be processed in one go is patterning techniques. To form patterns, a light-sensitive chemical ("photoresist") is deposited on top of a material and, using a photolithography machine, exposed to ultraviolet light passing through the photomask, and is then developed, leaving a pattern of gaps in the photoresist. (This is akin to how light reacts on photographic film in a camera.) The developed photoresist

---

[17] A "semiconductor" is a solid material that conducts a moderate amount of electricity – more than insulators (like rubber), and less than most metals (like copper). Sometimes "semiconductor" is also loosely used to refer to a chip made with a semiconductor material.
[18] Other wafer sizes, such as 200 mm (8 inches) are also used, but advanced process nodes typically employ the larger 300 mm wafers, partly because the fixed costs of processing each wafer are higher for more advanced processes. Indeed, the trend over the last six decades has been one of increasingly large wafers being introduced for increasingly more advanced processes.
[19] Unlike stencils, photomasks do not look exactly like the physical patterns they help print. That is because they must compensate for distortions caused by things like light diffraction, using techniques like optical proximity correction.

now acts as a mask, allowing the material underneath to be partially etched away, after which the photoresist is washed away, leaving behind only the patterned material. These steps are repeated until the devices are fully formed. In this process, it is also necessary to modify the electrical properties of parts of the semiconductor; to do this, dopants are introduced using a technique called ion implantation. When excess material needs to be removed from the wafer's surface, it is polished in a chemical slurry using a technique called chemical mechanical planarization (CMP). To remove contaminants and residue, the wafer is periodically cleaned using chemicals, plasmas, and a variety of physical techniques like sound waves or jets.

After the devices (transistors, etc.) have been formed, layers of metals and insulators are deposited and etched with chemicals to form connections between the devices. The steps used to create the metal layers resemble those used to form the devices underneath, as they also involve photolithography, deposition, etching, and planarization, and are also repeated with different patterns over multiple layers. Finally, a protective layer is deposited on top of the metal layers, with gaps etched in it for connections to external circuitry.

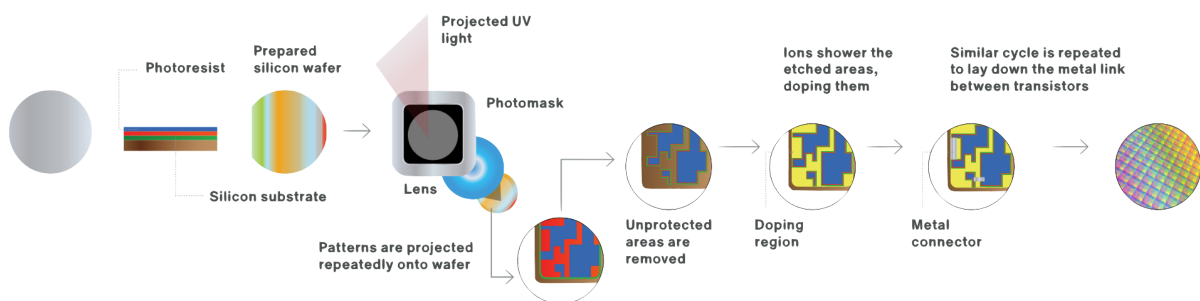

Figure 4. A simplified illustration of the wafer fabrication process. Source: Boston Consulting Group & Semiconductor Industry Association [(2021)](#).

The entire process is almost fully automated. Wafer handler tools transport and align wafers between and during process steps, and process control tools regularly monitor the wafers for defects, as detecting defects early on is critical to keeping costs down. See Software, equipment, and materials for a description of some of the highly specialized tools and materials the fabrication process relies on.

Assembly, test, and packaging

After fabrication, the patterned wafers are typically sent to a different facility for assembly, test, and packaging (ATP).[20] First, the silicon wafer is cut into individual, rectangular chips ("dies"). Those dies are then tested, and good dies are encased in a protective covering. Once packaged, the chips undergo final functional testing and "burn-in", where they are tested at extreme temperatures and voltages. ATP is

[20] These vendors are grouped under the term outsourced semiconductor assembly and test (OSAT). Some chip manufacturers do carry out ATP in-house, at least for some of their chips.

sometimes collectively referred to as "backend" manufacturing, with wafer fabrication being "frontend" manufacturing. Today, a single package sometimes combines multiple modular chips ("chiplets"), which communicate via die-to-die interconnects (see AI chips, accelerators, and servers for more on chiplets).

Although the packaging process is less complex than the fabrication process, it is not trivial. ATP is a relative strength of China, which has a 14% share of the market (Khan et al., 2021). According to the Semiconductor Industry Association (SIA), 22% of the world's ATP facilities are located in China, though advanced packaging is still dominated by Taiwanese companies (Semiconductor Industry Association, 2021).

While the main purpose of packaging has traditionally been to protect the chip from corrosion and other damage and to add connectors for other chips, advanced packaging techniques are now an increasingly important driver of semiconductor progress (Shivakumar & Borges, 2023). Advanced packaging is a term for a group of techniques for arranging multiple fabricated chips and chiplets together in a single package. Two relevant variants of advanced packaging are 2.5D, where multiple chips are arranged on an "interposer" chip that facilitates communication, and 3D, where chips are stacked on top of each other. Such techniques include:

- Through-silicon vias (TSVs), used to connect vertically stacked dies. TSV is used to create high-bandwidth memory, and therefore necessary for modern AI accelerators.
- Chip-on-Wafer-on-Substrate (CoWoS), used to stack chiplets on wafers. TSMC uses CoWoS to make Nvidia H100 chips, by packaging the GPU alongside six high-bandwidth memory chiplets.
- Embedded multi-die interconnect bridge (EMIB), used for die-to-die interconnect. This accomplishes similar goals to 2.5D techniques like CoWoS, but without requiring, or benefitting from, the interposer, which is costly and consumes space (LaPedus, 2018).
- Hybrid bonding, used to stack chips or wafers on wafers using direct copper-to-copper interconnects.

For more on advanced packaging, see Patel (2021, 2022, 2022).

Key inputs, context, and organizations

In order to understand how capable Chinese indigenous AI chip production is, and to better predict how capable it will be, it is useful to think in terms of inputs. For some inputs, the Chinese semiconductor industry has limited supply such that its ability to produce cutting-edge AI chips may be substantially restricted. Key inputs include:

1. **Design software and IP cores.** Electronic design automation (EDA) software is necessary for designing AI chips and validating those designs. Semiconductor intellectual property cores (IP cores) are reusable building blocks that can be licensed from IP core vendors.
    - See Design above for an overview of the chip design process, and Chip makers below for an overview of key context and organizations.
2. **Semiconductor manufacturing equipment (SME) and materials.** SME and some materials are crucial to the semiconductor manufacturing process. Building advanced SME itself depends on crucial components that SME firms often procure from third parties.
    - See Fabrication above for an overview of the chip fabrication process, and Chip fabrication and foundries below for an overview of key context and organizations.
3. **Intellectual property and other explicit knowledge.** Various kinds of explicit knowledge are necessary or useful in designing and fabricating chips, as well as developing tooling and materials to do so. Such knowledge can be acquired through research, licensing agreements, inspections, reverse engineering, and intellectual property theft.
4. **Talent and know-how.** Operating a fab, and doing the research and development needed to create new chip designs, SME, materials, and software all require many researchers, technicians, and engineers, and tacit knowledge that is gained with years of experience. Such talent can be grown indigenously, and also acquired by headhunting foreign workers or incentivizing Chinese workers to return from abroad.[21] SME firms can also gain know-how by collaborating with fabs, though leading fabs like Intel, Samsung, and TSMC largely do not collaborate with Chinese SME firms (Hunt et al., 2021).
5. **Investment and subsidies.** The semiconductor business is highly capital intensive. Of the top 25 companies in worldwide R&D spending, five are semiconductor firms (Irwin-Hunt, 2023).[22] A modern fab takes about three years, $10B, and 6K construction workers to build (Intel, 2023), and some SME has cost billions of dollars to develop (ASML, 2022). China has provided tens of billions of dollars in subsidies to support its semiconductor industry via the China Integrated Circuit Industry Investment Fund ("Big Fund"), and is rumored to

[21] Hunt et al. (2021) estimates about 1.1K Chinese nationals working for leading SME firms outside China, of whom very few return.
[22] They are Huawei (6th), Samsung (8th), Intel (9th), Qualcomm (19th), and Nvidia (25th). TSMC is ranked 37th, AMD is 41st, and Broadcom is 43rd. Other companies ranking highly represent the software, automotive, and biotechnology industries.

launch a new initiative aiming to invest another $41B (Zhu et al., 2023). However, Big Fund investments have to some extent suffered from corruption and waste (White & Liu, 2022), and it is unclear how effectively leading Chinese semiconductor companies can absorb additional funding.

The remainder of this section gives an overview of key organizations and context for some of these inputs.

Chip makers

The majority of AI chip performance improvements come from improvements in chip architectures rather than physical shrinkage (i.e., fitting more transistors per area).[23] Such improvements are gained through innovations in chip design carried out by design firms (often just called "chip makers"). Most design firms these days are "fabless", meaning they outsource the actual manufacturing of chips to foundries, though some – notably Intel and Samsung – still also manufacture the chips that they design.

The dominant AI chip maker is Nvidia, a fabless design firm based in the United States. There has been a striking increase in demand for Nvidia products since ChatGPT was released at the end of 2022, with revenue for AI chips in Q2 2023 nearly tripling year-on-year (Nvidia, 2023), and then increasing by another 40% from Q2 to Q3 (Nvidia, 2023); Nvidia has an estimated 80-95% share of the global AI chip market (Nellis & Mehta, 2023). Part of Nvidia's strength lies in the software ecosystem it provides for AI researchers and engineers, but it also has a lead in hardware performance. Nvidia's flagship AI accelerators are the A100 (2020) and H100 (2022). When the October 2022 export controls de facto banned these accelerators from being exported to China, Nvidia developed alternatives – the A800 and H800 – specifically for the Chinese market, with slightly degraded performance so as to pass the export control thresholds. In October 2023, BIS revised the restrictions to also cover these alternatives, although Nvidia is now planning a new line of AI chips targeting the Chinese market (Nellis et al., 2023).

Meanwhile, Nvidia's competitors endeavor to catch up. One competitor is Google, which develops its own AI chips, Tensor Processing Units (TPUs), together with Broadcom (Patel et al., 2023). TPUs are not sold, but rather used by Google internally and rented out to customers via Google's cloud services. Like Nvidia, AMD and Intel develop retail AI chips, as do a number of start-ups, e.g., Cerebras and SambaNova. But

[23] Nvidia has estimated a 2.5x performance improvement for AI inference from physical shrinkage over the past decade, versus a 400x improvement from changes in chip architecture (e.g., moving to lower-precision numbers) over the same period (Moore, 2023). Ruschhaupt (unpublished) roughly estimates, for recent Nvidia AI chips, and with high uncertainty, a 4.9x improvement in half-precision floating-point (FP16) performance via physical shrinkage, versus a 7.2x improvement in FP16 performance via chip architecture. (Note that the latter estimate compares two chips on FP16 performance, meaning it does not capture performance gained from the move to lower-precision numbers, which explains most of the discrepancy between 7.2x and 400x.) A semiconductor expert we spoke with also expressed the view that most performance improvements come from architectural changes.

with the exception of Google, these competitors have so far largely been unable to compete with Nvidia, at least when it comes to real-world AI training workloads.

The companies mentioned above are all American. Chinese AI chip makers include Huawei's subsidiary HiSilicon, Biren Technology (壁仞科技), and Moore Threads (摩尔线程).[24] Many Chinese chip designers are formidable, but they are limited by US export controls preventing them from outsourcing fabrication to foreign fabs. (There are no leading-edge fabs in China; see Chip fabrication and foundries.) Specifically, a so-called foreign-direct product rule (FDPR, defined in § 734.9(h)), restricts cutting-edge chips fabbed outside China from being exported into China if they have been made using American tooling. In addition, Huawei, Biren, and Moore Threads are all placed on BIS's Entity List, adding a license requirement with a presumption of denial for additional exports to these companies.

Memory chip makers

AI chips are logic chips: they perform computations on data. However, data must be stored somewhere when it is not being processed. Hence, AI accelerators also contain dynamic random access memory (DRAM) dies – circuits storing data temporarily, optimized for fast retrieval and storage. In the last decade, chip makers have started stacking DRAM dies on top of each other, a technique known as high-bandwidth memory (HBM) and common in AI accelerators. As the name suggests, HBM allows the AI chip to read/write data from/to memory at a faster rate. However, progress in memory bandwidth has not kept pace with progress in processor performance, putting a limit on the performance of the overall system, a phenomenon referred to as the "memory wall" (Woo, 2019).

The leading DRAM makers are SK Hynix (South Korea), Samsung (South Korea), and Micron (US). SK Hynix produces the HBM used in Nvidia's flagship accelerators (Shilov, 2022). The leading Chinese competitor is ChangXin Memory Technologies (CXMT). As of 2022, CXMT was able to manufacture DDR4 memory (Tyson, 2022), a type of DRAM that was first mass-produced in 2014 (Wikipedia, 2023). When Huawei launched its new flagship smartphone in September 2023, it was not with CXMT's products but memory chips by SK Hynix, likely stockpiled since SK Hynix has not sold chips to Huawei since the latter was added to the Entity List in 2019 (Liu, 2023). As of 2023, CXMT plans to expand using American equipment (Liu et al, 2023), and in October received a $2B investment from the Big Fund (Mo et al., 2023). Because memory chips are easier to design than logic chips, memory chip makers design the chips that they fabricate themselves.

[24] Other notable Chinese AI chip makers include Cambricon Technologies (寒武纪) and Horizon Robotics (地平线), though these have largely focused on AI chips used outside data centers, e.g., for inference on smartphones or in cars. Cambricon does not quite seem to be living up to initial expectations, as its share price is down 40% since the first-day open price in July 2020 (adjusted for inflation), and a founding investor recently sold its entire stake, worth $200M (Cao, 2023). Horizon Robotics is not publicly listed.

Chip fabrication and foundries

Two kinds of companies fabricate chips: integrated device manufacturers (IDMs), and "pure-play" foundries. IDMs design the chips that they manufacture, whereas pure-play foundries solely fabricate chips for others. See Table 1 for some notable semiconductor manufacturers.

Company	Business model	Product	Country	Foundry market share[25]
ChangXin Memory Technologies (CXMT)	IDM	Memory chips	China	Not a foundry
Hua Hong Semiconductor	Pure-play foundry	Logic chips	China	3.0%
Intel	IDM	Logic chips	United States	Not a foundry
Micron	IDM	Memory chips	South Korea	Not a foundry
Samsung	IDM	Memory and logic chips	South Korea	12.4%
Semiconductor Manufacturing International Corporation (SMIC)	Pure-play foundry	Logic chips	China	5.3%
SK Hynix	IDM	Memory chips	South Korea	Not a foundry
Taiwan Semiconductor Manufacturing Company Limited (TSMC)	Pure-play foundry	Logic chips	Taiwan	60.1%

Table 1. Notable semiconductor manufacturers.

TSMC (Taiwan), Samsung (South Korea), and Intel (US) are market leaders. SMIC is the leading Chinese foundry, having achieved volume production with a 7 nm process, previously reached by TSMC in 2018, Samsung in 2019, and Intel in 2021 (WikiChip,

[25] The market share shown is of chip foundries, from Chiao & Chung (2023). IDMs generally are not listed since they generally do not sell fabrication services to other companies. Samsung is an exception since it both fabricates its own chip designs and also sells chip fabrication as a service. The remaining 19% belong to other foundries, which do not fabricate advanced chips.

2022). (See Process node.) Like SMIC, Hua Hong too is Chinese, but in Q1 2023 only reported revenue at 55 nm process nodes and above (Hua Hong, 2023). As of early 2023, SMIC and Hua Hong held 5.3% and 3.0% of the global foundry market respectively (Chiao & Chung, 2023), though the numbers would be substantially lower if looking only at advanced process nodes.

Software, equipment, and materials

The AI chip manufacturing process involves a large number of essential, specialized, and highly sophisticated tools and materials, including:

- Electronic design automation (EDA) software and IP cores, key dependencies for chip design. (See the section on EDA software and IP cores below.)
- Wafer manufacturing tools and wafers.
- Wafer handling tools, used to transport and align wafers between and during process steps.
- Ion implanters, used to introduce dopant impurities into the wafer, to control its electrical properties.
- Deposition tools, used to deposit materials onto the wafer.
- Etching and cleaning tools, used to remove deposited materials and unwanted particles or residue. These machines rely on some important gasses and chemicals, which need to be highly pure.
- Photolithography tools, used to form patterns on a wafer by exposing a deposited photoresist material to ultraviolet light through a photomask; the patterned photoresist is then developed and used to create patterns in the underlying material. (See the section on Photolithography below.)
- Process control tools, used to monitor the fabrication process to maximize yield, improve throughput, ensure quality, and reduce costs. Examples include metrology tools (these make precise measurements to ensure everything has the right dimensions, is perfectly aligned, and so on) and inspection tools (these scan wafers, photomasks, or packaging to detect defects).

Though there are many SME manufacturers, there are still several parts of the supply chain that are dominated by just one or a few suppliers. The most important SME suppliers are American, Dutch, and Japanese. While Chinese SME firms have made substantial advances for some tools, they lag far behind market leaders in other areas. Most major SME firms are located outside China:

- Applied Materials (US) makes deposition, ion implanting, etch/clean, and process control tools among other things.
- ASML (Netherlands) makes photolithography tools.
- KLA (US) makes process controls and deposition tools.
- Lam Research (US) makes deposition and etch/clean tools.
- Nikon (Japan) makes photolithography tools.
- Tokyo Electron (Japan) makes deposition, etch/clean, and photoresist processing tools.

There are four notable Chinese SME firms:

- Advanced Micro-Fabrication Equipment Inc. China (AMEC) makes etch/clean tools.
- Jiangsu Nata Opto-electronic Material Company makes gasses and photoresists.
- Naura Technology (China) makes deposition and etch/clean tools.
- Shanghai Micro Electronics Equipment (SMEE) makes photolithography tools.

See the Emerging Technology Observatory's Supply Chain Explorer for more companies relevant to the semiconductor supply chain (Emerging Technology Observatory, 2023).

EDA software and IP cores

Chip design has a number of critical dependencies. Most notably, it relies on electronic design automation (EDA) software for translating high-level designs into circuit layouts, and for simulating and validating chip architectures before they are fabricated. The EDA industry is dominated by three companies: Cadence (US), Synopsys (US), and Siemens EDA (US-based but German-owned). Chip designers also make use of IP cores, pre-designed and verified components that can be licensed from vendors and integrated into designs. IP cores are offered as part of EDA suites, and also by dedicated IP core firms, the most prominent of which is Arm (UK). (See Design for a high-level description of the chip design process.) Chinese AI chip designers likely rely on foreign EDA software and IP cores, both of which can, in some circumstances at least, be pirated.

The Chinese EDA industry is small but growing (Khan et al., 2021). Notable companies offering EDA suites include Empyrean Technology and Huawei. Chinese EDA firms have received major subsidies, and some have seen sales grow by 30% in the first three quarters of 2023 (Shilov, 2023). Empyrean claims that its EDA suite supports designs for chips at 7 nm (Lu, 2023), and Huawei claims its software supports designs for chips at 14 nm and above (Kirton et al., 2023). However, Cadence, Synopsys, and Siemens EDA still combine for more than 80% of the Chinese market (Shilov, 2023). To comply with export controls, American EDA firms must customize their tools to exclude support for gate-all-around field-effect transistor (GAAFET) designs (see Non-planar transistors) when sold to Chinese customers.

Photolithography

The most complex, expensive, and therefore important type of SME is photolithography. Photolithography machines expose a photomask (which holds the pattern that needs to be etched onto the wafer) to ultraviolet light; the patterned light then reaches a photoresist material on the wafer. The exposed parts of the photoresist react to the light, after which the wafer is developed, leaving patterned gaps for the next deposition or etching steps that form the chip's circuits. There are five main types of photolithography machine:

- i-line, introduced in 1984 (Kato, 2007), employs 365 nm ultraviolet light from mercury lamps.
- Krypton fluoride (KrF), introduced in 1987 (Kato, 2007), employs 248 nm deep ultraviolet (DUV) light from an excimer laser.
- Argon fluoride (ArF), or ArF dry, introduced in 1998 (Kato, 2007), employs 193 nm DUV light from an excimer laser.
- Argon fluoride immersion (ArFi), introduced in 2004 (Kato, 2007), employs 193 nm DUV light from an excimer laser, with an immersion medium (typically water).
- Extreme ultraviolet (EUV), introduced in 2013 (ASML, 2023) and used in mass production from 2020.

Photolithography, and in particular EUV machines, is a critical bottleneck for Chinese semiconductor manufacturing. ASML is the leading photolithography machine maker, and the only one capable of producing EUV machines. It took ASML and others billions of dollars and about three decades to develop EUV (ASML, 2023). Chinese fabs lack access to EUV machines due to export restrictions (see Photolithography controls), and are instead relying on imported DUV machines. Two Japanese companies, Nikon and Canon, also make DUV machines, with Canon marketing i-line and KrF machines and Nikon marketing i-line, KrF, ArF, and ArFi machines. The only Chinese company in the photolithography space is SMEE, which is planning to deliver a DUV machine capable of 28 nm production (Bloomberg News, 2023) and is now planning a domestic stock market listing (Xiang, 2023). There are alternatives to EUV that could plausibly enable continued scaling below 7 nm nodes, e.g., nanoimprint lithography, but these are not used in volume production today.

Figure 5 shows the wavelength of light used at different process nodes, as well as additional techniques that are used to continue to shrink to smaller nodes while using 193 nm light (DUV). EUV machines are crucial for fabrication at nodes below 5 nm, and for economically competitive fabrication at 5 nm.[26] EUV photolithography is now also used to make advanced DRAM chips (Micron, 2023). (The SME export controls focus specifically on EUV and immersion DUV machines. The US has also placed SMEE on the Entity List, preventing it from importing some supplies.)

[26] It is possible that fabs can realize high-volume 5 nm processes without EUV, but this would involve substantially lower yield. See Patel (2023) for an argument that Chinese fabs can reach 5 nm without EUV.

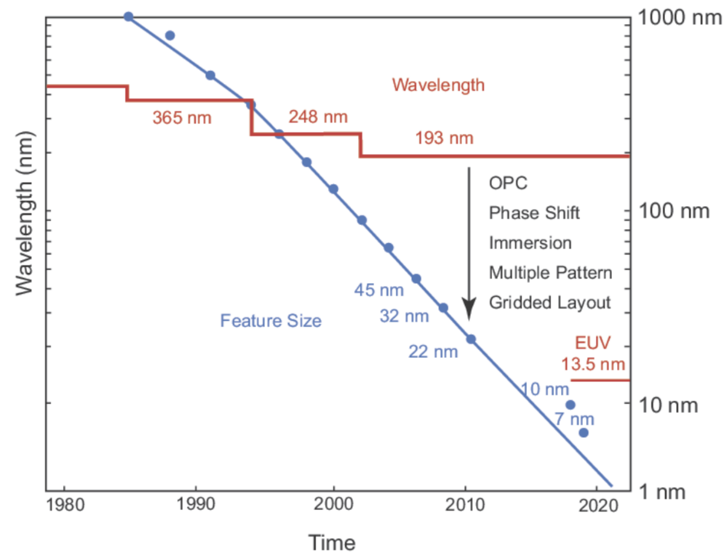

Figure 4. Historical progression of IC feature size and photolithography technologies.

Figure 5. Wavelength of light used at different process nodes. Source: Newport [(2023)](#).

Photolithography itself has several key dependencies. One is different components used in the photolithography machines, like custom mirrors (supplied for ASML by Zeiss) and light sources (supplied for ASML by its subsidiary Cymer). Others are materials like photomasks, the manufacturing of which in turn requires maskless lithography tools, and photoresists. Photoresists are customized to suit different lithography technologies (EUV lithography in particular requires photoresists with special properties), dimensions, materials, and processes. Photomask and photoresist manufacture is largely dominated by Japanese firms (Emerging Technology Observatory, 2023).

Relevant export controls

In October 2022, the US introduced sweeping export controls aiming to prevent China from developing advanced semiconductors, including restrictions on SME, EDA, and related intellectual property. In particular, these regulations aim to limit China's ability to produce advanced logic, DRAM, and NAND memory chips. Logic and DRAM chips are key to AI computing, while NAND chips are less important (and therefore not discussed in detail here). For the purpose of American SME controls, relevant logic chips ("advanced-node integrated circuits") are those that have non-planar transistors (see Non-planar transistors) or were produced at a ≤16 nm process node (see Process node), and relevant DRAM chips are those that were produced at a process node of ≤18 nm half-pitch.[27]

The US export controls are defined in the Export Administration Regulations (EAR). They received a significant update in October 2023, concerning both advanced chips themselves, and also SME used to make such chips. This section gives a shallow overview of the SME export controls, building on the overview of technologies and limiting factors involved, in particular discussions of Chip makers, Chip fabrication and foundries, and Software, equipment, and materials. The aim is not to give an exhaustive overview of all the relevant export controls, but rather to provide context necessary for understanding the controls and to highlight key parts of them.

SME and materials controls

The US export controls, including the revisions announced in October 2023, include restrictions on:

- **Epitaxial growth** tools (ECCN 3B001.a in the CCL), used to make layers of silicon and silicon germanium on wafers (to make the most advanced non-planar transistors, i.e., GAAFETs and its variants).
- **Ion implantation** tools (3B001.b).
- **Etch and deposition** tools (3B001.c). The updated controls on etch tools focus on two different types of tools: those used to make GAAFETs, and those used for making very high aspect ratio holes, which are most relevant for producing NAND memory.
- **Deposition** tools (3B001.d). These controls target several different deposition techniques, described in detail in Patel et al. (2023). They include electroplating (putting metals on planar surfaces and vertical channels), physical vapor deposition (PVD, used for barriers, liners, and seeds for metal growth), chemical vapor deposition (CVD, used for insulators and metals), and atomic layer deposition (ALD, a finer but slower version of CVD).

---

[27] This definition was used in the October 2022 controls and codified under the "advanced-node integrated circuits" term in the October 2023 update. Note that this definition is different from the one used to prevent exports on AI chips. The AI chip restriction is defined under ECCN 3A090 in the CCL (§ 774).

    - A full treatment of all the controlled deposition tools is beyond the scope of this report, but to get a sense of the type of tools that are targeted, see Applied Materials' Cobalt product family, and especially the Endura.
- **Wafer handling** tools (3B001.e).
- **Masks** (3B001.g-i), including for EUV photolithography.
- **Wafer annealing, cleaning and mask inspection** (3B001.o-p and 3B002.c). These tools are used for both logic and memory chips, and they can be specialized for different processes nodes. For example, Lam Research's EOS product family is specialized to clean wafers with non-planar transistors (FinFETs).
- **Lithography** tools (3B001.f), including photolithography, described in Photolithography controls below.

Some of the controlled items mentioned above are partly or primarily made outside the US. Japan in particular produces much SME, including "advanced crystal epitaxial growth equipment, photomask coating equipment, deposition equipment, etching equipment, heat treatment equipment, cleaning equipment, photomask inspection equipment, and [ArF photolithography]" (Allen et al., 2023). Japan announced its own semiconductor export controls in May 2023 after negotiations with the US (Kelly & Uranaka, 2023).

Photolithography controls

As mentioned in an earlier section, photolithography is perhaps the most important part of semiconductor fabrication (see Photolithography). The photolithography market is dominated by the Dutch company ASML, the only company capable of producing EUV machines. **EUV photolithography** (3B001.f.1.a) has been controlled since 2019 via the multilateral Wassenaar Arrangement, of which both the US and the Netherlands are members (Armasu, 2019). As a result, there are no EUV machines in China.[28] While Wassenaar restricted EUV *machines*, the October 2023 revision extended that to also include other components necessary for the process, such as photoresist, pellicles, and mask-making tools.

EUV lithography is typically used for process nodes at and below 7 nm (see Process node). Therefore, in order to restrict access to nodes as high as 16 nm, the US export controls also needed to cover some types of **DUV photolithography** machines (3B001.f.1.b), in particular immersion technology (ArFi). The Netherlands announced its own semiconductor export controls in June 2023 (Government of the Netherlands, 2023) after negotiations with the US. One Japanese company, Nikon, also makes immersion DUV machines, which are now restricted by the Japanese export controls.[29]

[28] SME, and especially photolithography machines, is hard to smuggle. The machines are large, bulky, few in number, and typically require extensive support to set up and service.
[29] Note that "restricted" here does not necessarily imply a ban. Photolithography makers could perhaps also ship the machines with the built-in software modified to limit performance in certain ways, in order to comply with export controls. It is still unclear whether this will happen, and whether such software restrictions can be circumvented.

For more details on the export controls and immersion photolithography, see Welch (2023).

Forecasting questions

We have created a set of forecasting questions intended to provide strategically valuable information on China's ability to indigenously produce AI chips at scale over the next decade. These questions, listed below, are published on the forecasting platform Metaculus. They were selected according to three criteria:

1. Each question should capture some aspect of China's ability to indigenously produce AI chips at scale, i.e., it should be correlated with that question.
2. Each question should be largely uncorrelated with all other questions, to avoid redundancy.
3. Each question should straightforwardly and unambiguously resolve to a definitive answer.

We encourage interested readers to make their own forecasts on some of these questions. We expect more forecasts from more people to improve the aggregate forecasts. Solid predictions on these questions can help answer strategically important open problems in AI governance such as:

- How effective are the US export controls at achieving their goals?
- How much computing power will Chinese AI labs have access to in the future?
- How capable will AI systems developed in China be in the future?
- Will we see major efforts to smuggle large quantities of AI chips into China?

Another reason why it might be a good idea to make forecasts on these questions is that it is an effective way of learning about the Chinese semiconductor industry. Getting expertise in AI hardware can be a great way to help with AI governance, and knowledge about semiconductor design and manufacturing and export controls is highly relevant. We think readers who are interested in getting relevant expertise in AI hardware would benefit from making forecasts on these questions.

The forecasting questions we have written and published are:

- Will Biren or Moore Threads be publicly traded before 2028?
- Will the following major cloud providers {AWS, Google, Azure, OVH} buy AI accelerators designed by a Chinese firm before 2028?
- How much of Nvidia's quarterly 2024 revenue (FY 2025) will come from the Chinese market?
- Will Huada Empyrean be publicly traded in 2026?
- How much of major electronic design automation (EDA) companies' revenue will come from the Chinese market in the last quarter of 2025?
- Will a Chinese firm market an argon fluoride immersion (ArFi) photolithography machine by 2025?
- Will a Chinese-made ArFi lithography machine be used in volume production anywhere before 2028?

- Will a Chinese firm market an extreme ultraviolet (EUV) photolithography machine before 2033?
- Will the following Chinese SME firms {AMEC, Nata, Naura, SMEE} have a higher market cap than any of ASML, Applied Materials, or Lam Research in 2033?
- Will any Chinese semiconductor foundry have at least 20% of the market cap that TSMC has in 2028?
- Will any Chinese semiconductor foundry have 20% or more global semiconductor market share before 2033?
- Will a Chinese fab achieve volume production at <N nm nodes {<5 nm, <3 nm, <2 nm} before 2030?
- Will a Chinese fab achieve a process with a density of greater than 150M transistors per square millimeter by 2027?
- Will a Chinese fab have a volume production process that uses EUV photolithography before 2033?
- Will the following major cloud providers {AWS, Google, Azure, OVH} buy AI accelerators fabricated in China before 2028?
- Will any of the following Western AI chip makers {Nvidia, Google, AMD} have any of their AI chips fabricated in China before 2033?
- How many 8"-equivalent wafers will SMIC ship in each quarter of 2024?
- How much quarterly revenue will SMIC earn from 12" wafers in 2024?

We may also add additional forecasting questions in the future. If so, these will be visible under the AI in China project.

Links and resources

This section collects some links and resources that can be useful when trying to better understand or forecast the progress of China's semiconductor industry.

The Chinese semiconductor industry and related export policy are ongoing processes, and as such, traditional news media is one important source of information. English-language sources covering the semiconductor industry include Nikkei Asia, Bloomberg, Reuters, and the Financial Times. The best way of following these sources is probably to use a news alert service like Google Alerts. The Chinese-language news site Jjiwei (爱集微) is a good source of information about the Chinese semiconductor industry.[30]

Primary sources are useful for fully understanding US, Dutch, and Japanese export controls. The US export controls are defined in the Export Administration Regulations (EAR), which are available in the Code of Federal Regulations. (See Relevant export controls for an overview of the EAR.) The Dutch export controls announced in 2023 are available in Dutch on a Dutch government website. The Japanese export controls announced in 2023 have been translated into English in Allen (2023). Important parts of the EAR include:

- The Steps for Using the EAR (§ 732) defines the process used to determine whether a given export needs a license or not. (The EAR does not outright ban exports; however, it can require a license with a "presumption of denial" which is similar to a ban in practice.)
- The Definitions of Terms (§ 772).
- The Commerce Control List (CCL, § 774), which lists all the different types of goods that are controlled, categorized according to Export Control Classification Numbers (ECCNs). The CCL is organized into different sections, with all relevant SME, materials, software, and intellectual property included under category 3 ("Electronics"). Category 3 also includes AI and memory chips, and category 4 ("Computers") includes AI accelerators and servers.

The Center for Security and Emerging Technology (CSET) has published some of the most comprehensive in-depth reports assessing how capable the Chinese semiconductor industry is and the bottlenecks it faces. See, e.g., Khan (2020) for background on US export controls, Khan (2021) on the global semiconductor supply chain, and Hunt et al. (2021) on Chinese progress in semiconductor manufacturing equipment. Note, however, that these reports are somewhat outdated, having been written prior to the October 2022 export controls and other important developments. CSET also provides a supply chain explorer showing the key parts of the chip-making supply chain and which companies and countries dominate those parts.

[30] There are surely other good Chinese-language sources too, but we are less familiar with those than with English-language sources.

The Center for Strategic and International Studies (CSIS) publishes analyses of export controls and their effects, such as Allen (2022) on the US export controls announced in October 2022, Allen (2023) on the implications of Huawei's new smartphone chip, and Benson (2023) on the US export controls announced in October 2023. Other notable sources of analysis include the ChinaTalk blog and podcast, and SemiAnalysis.

Acknowledgements

We are grateful to the following people for discussion and input: Onni Aarne, Bill Anderson-Samways, Renan Araujo, Marie Buhl, Ying Yi Dang, Péter Drótos, Amanda El-Dakhakhni, Oliver Guest, Lennart Heim, Sihao Huang, and Peter Ruschhaupt. We are also grateful to Adam Papineau for copyediting.

Bibliography

Allen, G. C. (2022, October 11). *Choking off China's Access to the Future of AI*. Center for Strategic and International Studies. Retrieved November 28, 2023, from https://www.csis.org/analysis/choking-chinas-access-future-ai

Allen, G. C. (2023, July 18). *CSIS Translation: Updated Japanese Export Controls on High-Performance Semiconductor Manufacturing Equipment*. Center for Strategic and International Studies. Retrieved November 28, 2023, from https://www.csis.org/analysis/csis-translation-updated-japanese-export-controls-high-performance-semiconductor

Allen, G. C. (2023, October 6). *In Chip Race, China Gives Huawei the Steering Wheel: Huawei's New Smartphone and the Future of Semiconductor Export Controls*. Center for Strategic and International Studies. Retrieved November 28, 2023, from https://www.csis.org/analysis/chip-race-china-gives-huawei-steering-wheel-huaweis-new-smartphone-and-future

Allen, G. C., Benson, E., & Putnam, M. (2023, April 10). *Japan and the Netherlands Announce Plans for New Export Controls on Semiconductor Equipment*. Center for Strategic and International Studies. Retrieved November 28, 2023, from https://www.csis.org/analysis/japan-and-netherlands-announce-plans-new-export-controls-semiconductor-equipment

Armasu, L. (2019, November 7). *Report: China Chip Production Hits Road Bump With EUV Shipment Delayed Over US Tensions*. Tom's Hardware. Retrieved November 28, 2023, from https://www.tomshardware.com/news/asml-euv-china-chip-production-delayed-us-tensions

ASML. (2023). *EUV lithography systems*. ASML. Retrieved November 28, 2023, from https://www.asml.com/en/products/euv-lithography-systems

Benson, E. (2023, October 18). *Updated October 7 Semiconductor Export Controls*. CSIS. Retrieved November 28, 2023, from https://www.csis.org/analysis/updated-october-7-semiconductor-export-controls

Bloomberg News. (2023, August 1). *China Chip Firms Soar on Report of Advance in Fabrication Tech*. Retrieved November 28, 2023, from https://www.bloomberg.com/news/articles/2023-08-02/china-chip-firms-surge-on-report-of-technology-breakthrough

Boston Consulting Group & Semiconductor Industry Association. (2021, April 1). *Strengthening the Global Semiconductor Supply Chain in an Uncertain Era*. Retrieved November 28, 2023, from https://www.semiconductors.org/wp-content/uploads/2021/05/BCG-x-SIA-Strengthening-the-Global-Semiconductor-Value-Chain-April-2021_1.pdf

Cao, A. (2023, September 25). *Founding investor in China's leading AI chip firm Cambricon dumps almost all of its stake as losses mount*. South China Morning Post. Retrieved November 29, 2023, from https://www.scmp.com/tech/tech-war/article/3235757/founding-investor-chinas-leading-ai-chip-firm-cambricon-dumps-almost-all-its-stake-losses-mount

Chatterjee, I. (2020, November 2). *RADNEXT*. RADNEXT. Retrieved November 28, 2023, from https://radnext.web.cern.ch/blog/from-mosfets-to-finfets/

Cheung, T. M., Naughton, B., & Hagt, E. (2022, July 12). *China's Roadmap to Becoming a Science, Technology, and Innovation Great Power - IGCC*. UC Institute on Global Conflict and Cooperation. Retrieved November 28, 2023, from https://ucigcc.org/publication/chinas-roadmap-to-becoming-a-science-technology-and-innovation-great-power/

Chiao, J., & Chung, E. (2023, June 12). *Top 10 Foundries Report Nearly 20% QoQ Revenue Decline in 1Q23, Continued Slide Expected in Q2, Says TrendForce*. TrendForce.

Retrieved November 28, 2023, from https://www.trendforce.com/presscenter/news/20230612-11719.html

Emerging Technology Observatory. (2022, October 16). *Supply Chain Explorer*. Retrieved November 28, 2023, from https://chipexplorer.eto.tech/

Government of the Netherlands. (2023, June 30). *Government publishes additional export measures for advanced semiconductor manufacturing equipment*. Government.nl. Retrieved November 28, 2023, from https://www.government.nl/latest/news/2023/06/30/government-publishes-additional-export-measures-for-advanced-semiconductor-manufacturing-equipment

Hofman, S. (2022, March 30). *Making EUV: from lab to fab*. ASML. Retrieved November 28, 2023, from https://www.asml.com/en/news/stories/2022/making-euv-lab-to-fab

Hunt, W., Khan, S. M., & Peterson, D. (2021, March). *China's Progress in Semiconductor Manufacturing Equipment*. Center for Security and Emerging Technology. Retrieved November 28, 2023, from https://cset.georgetown.edu/publication/chinas-progress-in-semiconductor-manufacturing-equipment/

Intel. (2023, September 22). *What Does it Take to Build a Fab?* Intel. Retrieved November 28, 2023, from https://www.intel.com/content/www/us/en/corporate/usa-chipmaking/news-and-resources/what-does-it-take-to-build-a-fab.html

Irwin-Hunt, A. (2023, June 19). *Top 100 global innovation leaders*. fDi Intelligence. Retrieved November 28, 2023, from https://www.fdiintelligence.com/content/feature/global-innovation-leaders-2022-edition-82527

Jouppi, N. P., Hyun Yoon, D., Ashcraft, M., Gottscho, M., Jablin, T. B., Kurian, G., Laudon, J., Li, S., Ma, P., Ma, X., Norrie, T., Patil, N., Prasad, S., Young, C., Zhou, Z., & Patterson, D. (2021). Ten Lessons From Three Generations Shaped Google's TPUv4i. *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)*, 1-14. 10.1109/ISCA52012.2021.00010

Kato, A. (2007, May). *Chronology of Lithography Milestones*. Retrieved November 28, 2023, from https://www.lithoguru.com/scientist/litho_history/Kato_Litho_History.pdf

Kelly, T., & Uranaka, M. (2023, March 31). *Japan restricts chipmaking equipment exports as it aligns with US China curbs*. Reuters. Retrieved November 28, 2023, from https://www.reuters.com/technology/japan-restrict-chipmaking-equipment-exports-aligning-it-with-us-china-curbs-2023-03-31/

Kennedy, P. (2020, July 29). *MLPerf 0.7 Training Results When Winners Only Apply*. ServeTheHome. Retrieved November 28, 2023, from https://www.servethehome.com/mlperf-0-7-results-when-winners-only-apply-nvidia-google-intel-huawei-inspur/

Khan, S. M. (2020, October). *U.S. Semiconductor Exports to China: Current Policies and Trends*. Center for Security and Emerging Technology. Retrieved November 28, 2023, from https://cset.georgetown.edu/publication/u-s-semiconductor-exports-to-china-current-policies-and-trends/

Khan, S. M., Peterson, D., & Mann, A. (2021, January). *The Semiconductor Supply Chain: Assessing National Competitiveness*. Center for Security and Emerging Technology. Retrieved November 28, 2023, from https://cset.georgetown.edu/publication/the-semiconductor-supply-chain/

Kirton, D., Horwitz, J., & Goh, B. (2023, March 23). *Huawei makes breakthroughs in design tools for 14nm chips -media*. Reuters. Retrieved November 28, 2023, from

https://www.reuters.com/technology/huawei-makes-breakthrough-design-tools-14nm-chips-media-2023-03-24/

LaPedus, M. (2018, July 12). *Bridges Vs. Interposers*. Semiconductor Engineering. Retrieved December 4, 2023, from https://semiengineering.com/using-silicon-bridges-in-packages/

Liu, Q., Leng, C., Olcott, E., & Sevastopulo, D. (2023, May 9). *Chinese chipmaker plans listing after clearing US export controls*. Financial Times. Retrieved November 28, 2023, from https://www.ft.com/content/bea78fe8-e2ae-4fe9-9a0e-4f6dafa3d91f

Liu, Z. (2023, September 7). *SK hynix Investigates Sanctions-Breaking 'Ghost RAM' In Huawei Phone*. Tom's Hardware. Retrieved November 28, 2023, from https://www.tomshardware.com/news/sk-hynix-digs-deep-into-sanctions-breaking-ghost-ram-in-huawei-phone

Lu, M. (2023, May 22). *Chinese EDA champion Empyrean can now "fully support" 7nm; sees 1Q23 sales growth*. Digitimes. Retrieved November 28, 2023, from https://www.digitimes.com/news/a20230522VL203/7nm-china-eda-ic-design-distribution.html

McWilliams, P. (2022, September 30). *Moore's Law is Dead – Long-live the Chiplet!* SemiWiki. Retrieved December 1, 2023, from https://semiwiki.com/chiplet/318859-moores-law-is-dead-long-live-the-chiplet/

Micron. (2023, May 17). *Micron to Bring EUV Technology to Japan, Advancing Next-Generation Memory Manufacturing | Micron Technology*. Micron Investor Relations. Retrieved November 28, 2023, from https://investors.micron.com/news-releases/news-release-details/micron-bring-euv-technology-japan-advancing-next-generation

Mo, Y., Liu, R., & Goh, B. (2023, October 31). *China's semiconductor state fund invests $2 billion in memory chip firm*. Reuters. Retrieved November 28, 2023, from

https://www.reuters.com/technology/chinas-semiconductor-state-fund-invests-2-billion-memory-chip-firm-2023-10-31/

Moore, S. K. (2019, April 12). *Intel's View of the Chiplet Revolution*. IEEE Spectrum. Retrieved December 1, 2023, from https://spectrum.ieee.org/intels-view-of-the-chiplet-revolution

Moore, S. K. (2023, September 7). *The Secret to Nvidia's AI Success*. IEEE Spectrum. Retrieved November 29, 2023, from https://spectrum.ieee.org/nvidia-gpu

Nellis, S., Cherney, M. A., & Potkin, F. (2023, November 9). *Nvidia plans to release three new chips for China - analysts*. Reuters. Retrieved November 28, 2023, from https://www.reuters.com/technology/nvidia-plans-release-three-new-chips-china-local-media-2023-11-09/

Nellis, S., & Mehta, C. (2023, June 13). *With no big customers named, AMD's AI chip challenge to Nvidia remains uphill fight*. Reuters. Retrieved November 28, 2023, from https://www.reuters.com/technology/amd-likely-offer-details-ai-chip-challenge-nvidia-2023-06-13/

Newport. (2023). *Deep UV Photolithography*. Newport Corporation. Retrieved November 28, 2023, from https://www.newport.com/n/deep-uv-photolithography

Nvidia. (2023, August 25). *Form 10-Q for the quarterly period ended July 30, 2023*. Retrieved November 28, 2023, from https://s201.q4cdn.com/141608511/files/doc_financials/2024/q2/19771e6b-cc29-4027-899e-51a0c386111e.pdf

Nvidia. (2023, November 21). *Form 10-Q for the quarterly period ended October 29, 2023*. Retrieved November 28, 2023, from https://s201.q4cdn.com/141608511/files/doc_financials/2024/q3/NVIDIA-10Q.pdf

Patel, D. (2021, December 15). *Advanced Packaging Part 1 – Pad Limited Designs, Breakdown Of Economic Semiconductor Scaling, Heterogeneous Compute, and Chiplets*.

SemiAnalysis. Retrieved December 1, 2023, from https://www.semianalysis.com/p/advanced-packaging-part-1-pad-limited

Patel, D. (2022, January 6). *Advanced Packaging Part 2 - Review Of Options/Use From Intel, TSMC, Samsung, AMD, ASE, Sony, Micron, SKHynix, YMTC, Tesla, and Nvidia*. SemiAnalysis. Retrieved December 1, 2023, from https://www.semianalysis.com/p/advanced-packaging-part-2-review

Patel, D. (2022, January 18). *Advanced Packaging Part 3 – Intel's Curious Bet on Thermocompression Bonding, ASM Pacific, Kulicke and Soffa, and Besi TCB Tool Landscape*. SemiAnalysis. Retrieved December 1, 2023, from https://www.semianalysis.com/p/advanced-packaging-part-3-intels

Patel, D. (2023, January 29). *The Gaps In The New China Lithography Restrictions – ASML, SMEE, Nikon, Canon, EUV, DUV, ArFi, ArF Dry, KrF, and Photoresist*. SemiAnalysis. Retrieved November 28, 2023, from https://www.semianalysis.com/p/the-gaps-in-the-new-china-lithography

Patel, D., Nishball, D., & Xie, M. (2023, August 30). *Broadcom's Google TPU Revenue Explosion, Networking Boom, VMWare Integration*. SemiAnalysis. Retrieved November 28, 2023, from https://www.semianalysis.com/p/broadcoms-google-tpu-revenue-explosion

Patel, D., & Wong, G. (2023, July 10). *GPT-4 Architecture, Infrastructure, Training Dataset, Costs, Vision, MoE*. SemiAnalysis. Retrieved November 30, 2023, from https://www.semianalysis.com/p/gpt-4-architecture-infrastructure

Patel, D., Xie, M., & Chu, W. (2023, October 15). *Going Vertical: Gate All Around, 3D DRAM, 3D NAND - Kokusai Electric IPO*. SemiAnalysis. Retrieved November 28, 2023, from https://www.semianalysis.com/p/going-vertical-gate-all-around-3d

Savov, V., Wu, D., Marks, S., Drozdiak, N., & Nardelli, A. (2023, September 4). *Look Inside Huawei Mate 60 Pro Phone Powered by Made-in-China Chip*. Bloomberg.com. Retrieved November 28, 2023, from

https://www.bloomberg.com/news/features/2023-09-04/look-inside-huawei-mate-60-pro-phone-powered-by-made-in-china-chip

Schor, D. (2023, May 27). *TSMC N3, And Challenges Ahead – WikiChip Fuse*. WikiChip Fuse. Retrieved November 28, 2023, from https://fuse.wikichip.org/news/7375/tsmc-n3-and-challenges-ahead/

Semiconductor Industry Association. (2021, December 1). *Submission of the Semiconductor Industry Association Regarding Index of Comments on the Possible Reinstatement of Previously Extended Exclusions*. Retrieved November 28, 2023, from https://www.semiconductors.org/wp-content/uploads/2021/12/SIA-Section-301-Product-Exclusions-Reinstatement-Request-FINAL-12011.pdf

Sevilla, J., Heim, L., Ho, A., Besiroglu, T., Hobbhahn, M., & Villalobos, P. (2022, February 16). *Compute Trends Across Three Eras of Machine Learning*. Epoch AI. Retrieved November 28, 2023, from https://epochai.org/blog/compute-trends

Shilov, A. (2022, June 9). *SK Hynix Kicks Off HBM3 Mass Production, Ships to Nvidia*. Tom's Hardware. Retrieved November 28, 2023, from https://www.tomshardware.com/news/sk-hynix-kicks-off-hbm-3-mass-production-ships-to-nvidia

Shilov, A. (2022, December 31). *Analysts Estimate TSMC's 3nm Yields Between 60% and 80%*. Tom's Hardware. Retrieved November 28, 2023, from https://www.tomshardware.com/news/analysts-estimate-tsmc-n3-yields-between-60-and-80-percent

Shilov, A. (2023, November 6). *Chinese Chip Design Software Devs Receive Massive Government Support*. Tom's Hardware. Retrieved November 28, 2023, from https://www.tomshardware.com/tech-industry/semiconductors/chinese-chip-design-software-devs-receive-massive-government-support

Shivakumar, S., & Borges, C. (2023, June 26). *Advanced Packaging and the Future of Moore's Law*. Center for Strategic and International Studies. Retrieved November 28,

2023, from https://www.csis.org/analysis/advanced-packaging-and-future-moores-law

Tyson, M. (2022, February 28). *China's CXMT Aims to Tackle DDR5 Market Later This Year*. Tom's Hardware. Retrieved November 28, 2023, from https://www.tomshardware.com/news/cxmt-ddr5-memory-china

Welch, N. (2023, November 1). *Litho World & Commerce: Lost in Translation?* ChinaTalk. Retrieved November 28, 2023, from https://www.chinatalk.media/p/litho-world-and-commerce-lost-in

White, E., & Liu, Q. (2022, September 28). *China's Big Fund corruption probe casts shadow over chip sector*. Financial Times. Retrieved November 28, 2023, from https://www.ft.com/content/8358e81b-f4e7-4bad-bc08-19a77035e1b4

WikiChip. (2020, November 25). *22 nm lithography process*. WikiChip. Retrieved November 29, 2023, from https://en.wikichip.org/wiki/22_nm_lithography_process

WikiChip. (2022, November 5). *7 nm lithography process*. WikiChip. Retrieved November 28, 2023, from https://en.wikichip.org/wiki/7_nm_lithography_process

WikiChip. (2023, May 25). *14 nm lithography process*. WikiChip. Retrieved November 29, 2023, from https://en.wikichip.org/wiki/14_nm_lithography_process

Wikipedia. (2023, November 16). *DDR4 SDRAM*. Wikipedia. Retrieved November 28, 2023, from https://en.wikipedia.org/wiki/DDR4_SDRAM

Wikipedia. (2023, November 24). *Nvidia*. Wikipedia. Retrieved November 29, 2023, from https://en.wikipedia.org/wiki/Nvidia

Wikipedia. (2023, November 27). *AMD*. Wikipedia. Retrieved November 29, 2023, from https://en.wikipedia.org/wiki/AMD

Wikipedia. (2023, November 27). *2 nm process*. Wikipedia. Retrieved November 28, 2023, from https://en.wikipedia.org/wiki/2_nm_process

Woo, S. (2019, September 12). *Breaking Down The AI Memory Wall*. Semiconductor Engineering. Retrieved November 28, 2023, from https://semiengineering.com/breaking-down-the-ai-memory-wall/

Xiang, R. (2023, November 12). 中国光刻机巨头拟*A*股上市？. Ijiwei. Retrieved November 28, 2023, from https://laoyaoba.com/n/884282

Xinhua News Agency. (2021, May 13). *CSET Original Translation: China's 14th Five-Year Plan*. Center for Security and Emerging Technology. Retrieved November 28, 2023, from https://cset.georgetown.edu/publication/china-14th-five-year-plan/

Zhu, J., Huang, K., Mo, Y., & Liu, R. (2023, September 5). *Exclusive: China to launch $40 billion state fund to boost chip industry*. Reuters. Retrieved November 28, 2023, from https://www.reuters.com/technology/china-launch-new-40-bln-state-fund-boost-chip-industry-sources-say-2023-09-05/