Semiconductors & Chip Production

Figures current as of October 2026

Semiconductors — what a chip is and how it is made, the nodes and the end of scaling, the companies and their business models, TSMC and Taiwan, Nvidia and the AI accelerator, the memory boom of 2026, the equipment and packaging chokepoints, the US-China chip war, China's catch-up, the subsidies, and the demand behind a $1.5 trillion industry.

MODULE 1 — Foundations: What a Chip Is

Goal: Establish the device, the categories, and the industry's scale.

Transistors and how they compute; the chip categories with their 2026 sales; the industry's scale and memory's share of it; and why a chip is the most complex object made.

MODULE 2 — How a Chip Is Made

Goal: Walk through the fab, step by step, with the economics.

The thousand-step process in a table, the fab's scale and cost, yield and the learning curve, and a wafer's economics worked at the 3nm node.

MODULE 3 — Moore's Law, the Nodes, and the End of Scaling

Goal: Explain the cadence that built the industry and what replaced it.

The nodes, past and planned, in a table, what a 'nanometer' means now, the physics that ended classical scaling, and the three substitutes: architecture, packaging, and specialization.

MODULE 4 — The Business Models and the Companies

Goal: Map who does what and who earns what.

Designers, foundries, memory makers, equipment makers, and the rest; the largest companies by market value, with their latest quarters; and where the profit sits in 2026.

MODULE 5 — TSMC and the Taiwan Concentration

Goal: Profile the company the world's computing runs on.

The numbers through the second quarter of 2026, the American commitment, why TSMC won, the Taiwan risk, and a leading-edge fab's economics worked.

MODULE 6 — Nvidia and the AI Accelerator

Goal: Cover the company that became the industry's center.

The numbers through the July 2026 quarter, the product cadence from Hopper to Rubin, the four-part moat, the rivals, and the circular deals.

MODULE 7 — Intel, Samsung, and the Foundry Race

Goal: Cover the companies trying to be the second leading-edge foundry.

Intel's fall and rescue, 18A and 14A; Samsung's 2nm and its memory comeback; Japan's Rapidus; and how many leading-edge makers the world will have.

MODULE 8 — Memory: DRAM, NAND, and the Supercycle

Goal: Cover the commodity chips that became the industry's largest business.

The makers and their shares, the 2026 price explosion and its causes, the cycle's history, the new fabs, and China's memory makers.

MODULE 9 — Equipment and Materials: The Chokepoints

Goal: Map the layer where one company can stop an industry.

ASML and its lithography machines, the equipment makers with their latest results, the materials from photoresist to quartz, and the chokepoint map.

MODULE 10 — Advanced Packaging: The New Bottleneck

Goal: Explain why the constraint moved from the wafer to the package.

TSMC's packaging capacity, chiplets and memory stacking, the packaging technologies in a table, and who is building more of it.

MODULE 11 — Analog, Power, RF, and the Mature Nodes

Goal: Cover the chips that are not at the leading edge and the market China is flooding.

Texas Instruments and the analog world, silicon carbide and Wolfspeed, the 2021 shortage, China's mature-node surge, the Nexperia crisis, and the tariffs.

MODULE 12 — Chip Design: EDA, IP, and Custom Silicon

Goal: Cover the software and architectures chips are designed with.

Design software, Arm and RISC-V, Arm's first chip of its own, and the custom-chip business that lets the cloud companies design their own.

MODULE 13 — Export Controls and the US-China Chip War

Goal: Trace the controls, the countermeasures, and the 2026 arrangement.

The timeline in a table, the logic of the controls, the sale-for-a-cut deals for Nvidia's China chips, the 2026 summits and the truce, China's countermeasures, and the US-Japan precedent of the 1980s.

MODULE 14 — China's Chip Industry

Goal: Assess how far China has come and where it is stuck.

SMIC's 5nm-class process, Huawei's newest AI chip, the memory makers and the listing wave, the equipment gap, and an assessment of how far behind China is.

MODULE 15 — Industrial Policy: The CHIPS Act and the World's Subsidies

Goal: Cover the money governments are spending and what it has built.

What the CHIPS Act built and what it became, the tariff lever, and the programs of Taiwan, Korea, Japan, Europe, China, and India.

MODULE 16 — Demand: Where the Chips Go

Goal: Cover the end markets and the AI capital-spending cycle.

The end markets in 2026, the cloud companies' capital spending, and the case for and against the demand's durability.

MODULE 17 — Where Things Stand, and What to Watch

Goal: Close with the argument, the contradictions, and a dated watch list.

The three forces that shape the industry, six contradictions, the signals from October 2026 to January 2027, and a last word.

The State of the Art, as of October 2026

The Leading Chips, Processes, and Machines, and When They Shipped

This table is the course's reference point: the most advanced product or process in each part of the industry as of October 8, 2026, who makes it, and when it shipped. Of the terms that recur, a node is a generation of chipmaking process, named in nanometers (nm), though the names no longer measure anything (Module 3). A graphics processing unit (GPU) is the parallel processor that runs most AI work, and high-bandwidth memory (HBM) is the memory stacked beside it. Extreme-ultraviolet (EUV) lithography prints a chip's finest features, while deep-ultraviolet (DUV) lithography is the older generation. Dates in the future are the companies' own plans, and figures marked 'reported' come from press or analyst reports rather than from the companies.

Category	Leading product or process	Maker	Status and date	Next
Leading-edge chipmaking	N2, a 2nm-class node and TSMC's first with gate-all-around transistors, whose gate wraps around the channel	Taiwan Semiconductor Manufacturing Company (TSMC)	In volume since late 2025; first revenue in Q2 2026	A16, with power wiring beneath the transistors, then A14, A13, and A12 (Module 3)
US leading edge	18A, a 1.8nm-class node, the first in volume with that backside power wiring	Intel	High-volume production for Panther Lake laptop chips and Xeon 6+ server chips (2026)	14A, using High-NA (high-numerical-aperture) EUV (Module 3)
Korean leading edge	SF2, a 2nm-class node	Samsung Foundry	The Exynos 2600 phone chip in the Galaxy S26 (early 2026); a second-generation 2nm ramping in the second half	Tesla's AI chips at Taylor, Texas
China's leading edge	N+3, a 5nm-class node made with DUV lithography	Semiconductor Manufacturing International Corporation (SMIC)	Confirmed by TechInsights in Huawei's Kirin 9030 phone chip (late 2025)	Limited without EUV
AI accelerator	Rubin GPU, with HBM4 (fourth-generation HBM); specifications in Module 6	Nvidia	In full production; running at the large clouds from August 2026	Rubin Ultra and Feynman (Module 6)
Rival AI accelerators	MI400 series (MI455X) and the Helios rack	Advanced Micro Devices (AMD)	Launched; Helios ramping from the second half of 2026	OpenAI's order (Module 6)
Custom AI chips	TPU v7 'Ironwood' (tensor processing unit), with 192 gigabytes (GB) of HBM3e	Google, with Broadcom	Generally available since April 2026	TPU 8t (training) and 8i (inference) on TSMC's 2nm for late 2027
China's AI chip	Ascend 950PR	Huawei, made at SMIC (reported)	Mass production since April 2026; volumes in Module 14	Versions with ordinary DDR (double-data-rate) memory and with HBM
Memory	HBM4 stacks of 12 or more DRAM (working-memory) dies	SK Hynix, Samsung, Micron	Samsung shipping since February 2026; SK Hynix in volume from Q2 2026; Micron qualifying with customers	HBM4E samples shipped in the first half of 2026

Category	Leading product or process	Maker	Status and date	Next
Phone processor	A20 Pro	Apple, made by TSMC	In the iPhone 18 Pro, announced September 9, 2026 and on sale September 18; widely reported to use TSMC's 2nm	—
Data-center processor (CPU)	Vera (88 Arm-based cores), paired with Rubin; Arm's first chip of its own, the AGI CPU	Nvidia; Arm, with Meta	Shipping in 2026	—
Lithography	EUV scanners; High-NA EUV, with a wider aperture for finer lines	ASML	Shipping; 2026 volumes in Module 9	Capacity plans in Module 9
Advanced packaging	CoWoS-L (chip-on-wafer-on-substrate), which joins compute dies and HBM on one base	TSMC	In volume (capacity and expansion plans in Module 10)	—

Company announcements and reports; TechInsights for SMIC's process.

Who's Who

The Companies, Labs, and Agencies This Course Names

Name	What it is	Based in
Absolics	Maker of glass substrates for chip packages, in Georgia; part of Korea's SKC	US
Advanced Micro Devices (AMD)	Designer of CPUs and GPUs	US
Advantest	The largest maker of chip-testing equipment	Japan
AGC; Hoya	Makers of the glass blanks for EUV photomasks	Japan
Air Liquide; Air Products; Linde	Industrial-gas suppliers to fabs	France; US; Britain and US
Ajinomoto	Food company that makes the build-up film (ABF) used in chip substrates	Japan
Alchip; Global Unichip (GUC)	Custom-chip design firms	Taiwan
Alibaba	Internet and cloud company; designs chips through its T-Head unit; maker of the Qwen AI models	China
Alphabet (Google)	Google's parent; designs the TPU AI chips and runs Google Cloud	US
Amazon (Amazon Web Services, AWS)	Retailer and the largest cloud provider; designs the Trainium and Graviton chips	US
AMEC (Advanced Micro-Fabrication Equipment)	Maker of etch and deposition tools	China
Amkor	Outsourced packaging and test company	US
Analog Devices	Analog chipmaker	US
Anthropic	AI lab, maker of the Claude models	US
Apple	Device maker that designs its own processors and modems, made by TSMC	US
Applied Materials	The broadest equipment maker: deposition, etch, implant, polishing, inspection	US
Arm	Designer of processor architectures licensed to chipmakers, and now of a chip of its own; mostly owned by SoftBank	Britain
ASE	The largest outsourced packaging and test company	Taiwan
ASM International	Maker of atomic-layer deposition tools	Netherlands
ASML	Maker of lithography machines; the only maker of EUV tools	Netherlands
Axcelis	Maker of ion implanters	US
Baidu	Internet company; designs the Kunlun AI chips	China
Bell Labs	AT&T's research laboratory, where the transistor was invented	US
Big Fund (China Integrated Circuit Industry Investment Fund)	China's state chip-investment fund	China
Biren	AI-chip designer	China
Bosch	Engineering group; maker of sensors and a partner in TSMC's Dresden fab	Germany
Broadcom	Designer of networking chips and custom AI accelerators, including Google's TPU; also a software company	US
ByteDance; JD.com; Tencent	Internet companies	China
BYD	Carmaker that makes its own power chips	China

Name	What it is	Based in
Cadence	Maker of chip-design (EDA) software	US
Cambricon	AI-chip designer	China
Canon; Nikon	Makers of DUV lithography tools	Japan
Cerebras	Maker of wafer-sized AI chips	US
Coherent	Maker of lasers and optical components	US
Commerce Department	The US agency that runs export controls and keeps the 'entity list'; it also administers the CHIPS Act, the chip-subsidy law	US
CoreWeave	Company that rents out GPU computing in the cloud	US
Counterpoint Research	Market-research firm tracking memory shares	Hong Kong
CXMT (ChangXin Memory Technologies)	DRAM maker	China
DeepSeek	AI lab	China
Disco	Maker of dicing and grinding tools	Japan
DNP (Dai Nippon Printing); Toppan	Photomask makers	Japan
DuPont	Chemicals maker, including photoresists	US
Ebara	Maker of wafer-polishing tools	Japan
Elpida; Qimonda	Former memory makers	Japan; Germany
Empyrean	China's largest maker of chip-design software	China
Entegris	Maker of materials and contamination-control products for fabs	US
European Commission	The European Union's executive, which runs the Chips Act	EU
Fujifilm	Maker of photoresists and process chemicals	Japan
GigaDevice	Designer of microcontrollers and memory	China
GlobalFoundries	Foundry for mature nodes, spun out of AMD	US
GlobalWafers; Siltronic; SK Siltron	Silicon-wafer makers	Taiwan; Germany; Korea
Groq	Inference-chip start-up	US
Hitachi	Conglomerate whose high-tech unit makes inspection tools	Japan
House Foreign Affairs Committee	The US House committee that advanced the export-control bills	US
Hua Hong	Foundry for mature nodes and power chips	China
Huawei	Telecom-equipment and phone maker under US sanctions; designs chips through HiSilicon, including the Kirin phone chips and the Ascend AI chips	China
Hygon	Designer of server processors and AI chips	China
Hyundai; LG	Korean conglomerates; Hyundai's chip arm became SK Hynix	Korea
IBM	Computing company	US
Ibiden; Unimicron	Makers of chip-package substrates	Japan; Taiwan
IEEE	Engineering society that publishes the industry's device roadmap	US
IMEC	Shared semiconductor research center	Belgium
India Semiconductor Mission	India's chip-subsidy program	India
Infineon	Maker of power, automotive, and security chips	Germany
Intel	Maker of PC and server processors; Intel Foundry makes chips for others; part-owned by the US government	US
JCET	Outsourced packaging company	China

Name	What it is	Based in
JSR; Tokyo Ohka	Photoresist makers; JSR was bought by a Japanese state fund	Japan
Kioxia	NAND flash maker, formerly Toshiba's memory business	Japan
KLA	The leader in inspection and measurement tools	US
Kokusai; Screen	Makers of furnace and wafer-cleaning tools	Japan
Lam Research	Maker of etch and deposition tools	US
Marvell	Designer of custom and networking chips	US
MediaTek	Designer of phone processors	Taiwan
Mellanox	Networking-chip company Nvidia bought	Israel
Meta	Social-media company; designs its MTIA chips	US
MetaX; Moore Threads	GPU designers	China
Micron	Memory maker (DRAM, HBM, NAND)	US
Microchip	Microcontroller maker	US
Microsoft	Software and cloud company (Azure); designs the Maia chips; OpenAI's partner	US
Mitsubishi Electric; Rohm	Power-chip makers	Japan
Mobileye	Maker of driver-assistance chips, majority-owned by Intel	Israel
Mostek; Motorola	American chipmakers	US
Naura	China's largest equipment maker	China
NEC; Fujitsu; Toshiba	Japanese electronics groups	Japan
Nexchip	Foundry for mature nodes, in Hefei	China
Nexperia	Maker of basic chips for cars, owned by China's Wingtech	Netherlands
Nvidia	Designer of GPUs and AI systems and of the CUDA software platform	US
NXP	Maker of automotive and industrial chips	Netherlands
Office of the US Trade Representative (USTR)	The US trade agency that runs Section 301 cases, the tariff remedy for unfair foreign trade practices	US
onsemi	Maker of power and sensor chips	US
OpenAI	AI lab, maker of ChatGPT	US
Oracle	Software and cloud company	US
Photronics	Photomask maker	US
Powertech	Memory packaging and test company	Taiwan
Qorvo; Skyworks	Radio-frequency chipmakers	US
Qualcomm	Designer of phone processors and modems	US
Rambus	Designer of memory-interface circuits	US
Rapidus	Government-backed foundry start-up aiming at 2nm	Japan
Renesas	Maker of microcontrollers and automotive chips	Japan
Resonac; Stella Chemifa	Makers of specialty gases and chemicals	Japan
Samsung Electronics	The largest memory maker; also makes phones and, through Samsung Foundry, chips for others	Korea
SambaNova; Tenstorrent	AI-chip start-ups	US; Canada
SanDisk	NAND flash maker, spun out of Western Digital in 2025	US
SEMI	The equipment and materials industry's association and data source	US

Name	What it is	Based in
SemiAnalysis	Research firm that models chip supply and yields	US
Sematech	Government-industry research consortium	US
Semiconductor Industry Association (SIA)	The American chip industry's association, publisher of monthly sales data	US
Shin-Etsu; SUMCO	The two largest silicon-wafer makers	Japan
Sibelco	Minerals company that mines high-purity quartz at Spruce Pine, North Carolina	Belgium
Siemens EDA	The chip-design software unit of Siemens	Germany
SK Hynix	Memory maker and the leader in HBM; owns Solidigm	Korea
SMEE (Shanghai Micro Electronics Equipment)	China's maker of lithography tools	China
SMIC (Semiconductor Manufacturing International Corporation)	China's largest foundry	China
SoftBank	Investment group; Arm's majority owner and an investor in Intel	Japan
Sony	Electronics maker and the leader in image sensors	Japan
SpaceX; Tesla	Musk-led rocket and car companies	US
StarPower	Power-chip maker	China
State Administration for Market Regulation	China's market and antitrust regulator	China
STMicroelectronics (STMicro)	Franco-Italian maker of automotive, industrial, and power chips	Switzerland
Synopsys	The largest maker of chip-design (EDA) software	US
Tata Electronics	Part of India's Tata group, building the Dholera fab	India
TDK	Electronics-components maker, including sensors	Japan
TechInsights	Firm that takes chips apart to identify how they were made	Canada
Teradyne	Maker of chip-testing equipment	US
Texas Instruments (TI)	The largest analog chipmaker	US
Tokyo Electron	Equipment maker: coater-developers, etch, deposition	Japan
TrendForce	Market-research firm tracking memory prices and shares	Taiwan
TSMC (Taiwan Semiconductor Manufacturing Company)	The largest foundry, making most leading-edge chips	Taiwan
UMC (United Microelectronics Corporation)	Foundry for mature nodes	Taiwan
Wingtech	Electronics group that owns Nexperia	China
Wolfspeed	Silicon-carbide maker	US
World Semiconductor Trade Statistics (WSTS)	The industry's statistics body, which counts and forecasts chip sales	US
YMTC (Yangtze Memory Technologies)	NAND flash maker	China
Zeiss	Optics company that makes the mirrors in ASML's EUV tools	Germany
ZTE	Telecom-equipment maker	China

Organizations grouped in one row share a role. Company figures and dates are in the modules.

Module 1 — Foundations

What a Chip Is

Recommended Reading / Reference

Book: Chris Miller, *Chip War: The Fight for the World's Most Critical Technology*

Reference: World Semiconductor Trade Statistics (WSTS), the industry's own sales count and forecast (spring and autumn editions)

1a. The Transistor

A semiconductor chip is a slice of silicon carrying billions of transistors, each a switch that a small voltage on its gate turns on or off, and wired into circuits, transistors store and process information as ones and zeros. The transistor was invented at Bell Labs in 1947, and the integrated circuit, many transistors on one piece of silicon, followed in 1958-59, invented separately by Jack Kilby and Robert Noyce. The industry's history since then is the shrinking of the transistor: Intel's first microprocessor (Module 3) held 2,300 of them, and Nvidia's newest graphics processor, Rubin, holds hundreds of billions (Module 6). Each generation's transistors were smaller, faster, and cheaper than the last, and a leading-edge transistor today is a few dozen atoms across, built in three dimensions with the gate wrapped around the channel, and it switches billions of times a second. Machines print its features with light no eye can see, at sizes smaller than a virus. A chip is the most complex object humans manufacture, and the world makes about a trillion of them a year.

1b. The Types

Category (WSTS)	What it includes	2025 sales	2026 forecast	Growth 2026	Leaders
Memory	DRAM (working memory, including the high-bandwidth memory stacked beside AI chips) and NAND flash (storage)	$230B	$804B	+250%	Samsung, SK Hynix, Micron; Kioxia, SanDisk; China's CXMT and YMTC
Logic	Graphics processors and AI accelerators, phone systems-on-chip, custom chips, and other digital logic	$300B	$411B	+37%	Nvidia, Broadcom, AMD, Qualcomm, MediaTek, Apple (for itself)
Micro	Microprocessors for PCs and servers, microcontrollers, and signal processors	$85B	$102B	+20%	Intel, AMD, Arm's licensees; NXP, Renesas, Microchip, STMicro, Infineon
Analog	Power management, signal conversion, and the interfaces to the physical world	$87B	$95B	+10%	Texas Instruments, Analog Devices, Infineon, STMicro, NXP
Optoelectronics	Image sensors, lasers, light-emitting diodes (LEDs), and optical parts for networks	$43B	$44B	+3%	Sony, Broadcom, Coherent
Discretes	Power transistors and diodes, including silicon carbide and gallium nitride	$31B	$33B	+8%	Infineon, onsemi, STMicro; China's BYD and StarPower
Sensors	Motion, pressure, and other sensors	$21B	$22B	+3%	Bosch, STMicro, TDK
Total		$796B	$1,511B	+90%	Memory more than half of the 2026 total

World Semiconductor Trade Statistics, spring forecast of June 2, 2026. Company revenues such as Nvidia's include boards, systems, and networking, so they do not add up to these chip categories.

1c. The Scale

The industry is on course for its largest year by far. World Semiconductor Trade Statistics (WSTS), the industry's statistics body, forecast the 2026 total in the table above, a rise that would be the fastest since its records began. By the end of August, sales for the year had already passed $1 trillion, and the Semiconductor Industry Association (SIA), the American industry group, expected the full year to top $1.6 trillion. Two forces did it, and the first is the AI build-out: the accelerators that Nvidia and its rivals sell to data centers, and the memory beside them. The second is the price of memory itself, because memory makers had turned their capacity toward the high-bandwidth memory AI chips need, and ordinary memory ran short, so contract prices for server memory soared (Module 8). Memory, in the table's first row, is both the fastest-growing category and the largest. Logic, which includes the AI chips, grows fast too, while the analog, power, sensor, and microcontroller chips that go into tens of billions of devices a year grow about 10%. The concentration runs through every module: one company designs most AI chips and one company makes them, one company builds the machines that print the finest layers, and three companies make the memory. One island, Taiwan, holds about 90% of the most advanced chipmaking capacity, which is why an industry that was once a component supplier became the object of export controls, tariffs, subsidies, and the capital spending of Module 16.

Module 1 Key Points

Foundations

• A chip is billions of transistors, switches a few dozen atoms across, printed on silicon; it is the most complex object humans make.

• WSTS forecasts 2026 sales of $1.51 trillion, up 90% from 2025: the fastest growth in the industry's recorded history.

• Memory is more than half of 2026 sales, at $804 billion, because AI demand drove its prices up.

• Almost all the growth comes from AI chips and memory; analog, power, sensor, and microcontroller chips grow about 10%.

• One designer, one manufacturer, one lithography-tool maker, three memory makers, and one island hold the industry's chokepoints.

• That concentration is why a component industry became the object of export controls, tariffs, and subsidies.

Module 1 Quiz

Check Your Understanding

1. What did WSTS forecast in June 2026 for the industry's 2026 sales?

A) $1.51 trillion, up 90%
B) About $900 billion
C) Roughly $1.1 trillion, up 40%
D) Close to $2 trillion

2. Why did memory sales more than triple in 2026?

A) A surge in smartphone and PC sales after two weak years
B) HBM for AI squeezed supply, and prices soared
C) Tariffs on Chinese memory raised prices around the world
D) Memory makers built many new fabs at once and sold more chips

3. How does the rest of the industry, outside memory and logic, fare in 2026?

A) It shrinks, because AI spending crowds out every other kind of chip.
B) It grows as fast as memory.
C) It grows about 10%, a normal pace for the analog, power, sensor, and microcontroller chips.
D) It doubles, carried along by the same price increases as memory.

4. Where is about 90% of the most advanced chipmaking capacity?

A) South Korea
B) The United States, with Arizona's new fabs
C) Japan
D) Taiwan

5. Why did an industry that was once a component supplier become the object of export controls and subsidies?

A) Because a few companies and one island hold each critical step.
B) Because chips became the largest export of every advanced economy.
C) Because chip prices rose too fast for the buyers of phones, PCs, and cars.
D) Because governments own most of the world's fabs and set their prices.

Module 2 — How a Chip Is Made

The Fab, Step by Step, With the Economics

Recommended Reading / Reference

Book: Peter Van Zant, *Microchip Fabrication: A Practical Guide to Semiconductor Processing*

Reference: SEMI's *World Fab Forecast* (quarterly), the industry's count of fabs, capacity, and fab spending

2a. The Process

A chip is printed, layer by layer, onto a wafer of silicon: a disc cut from a single silicon crystal, usually grown and polished in Japan. Inside the fab the wafer cycles, once per layer, through the same sequence, in which a thin film is deposited and a light-sensitive resist is coated on top. Ultraviolet light shone through a patterned mask prints the circuit into the resist, the exposed resist is washed away, the pattern is etched into the film, and the surface is polished flat for the next layer. The finest layers, where the transistors sit, are printed with extreme-ultraviolet light from an ASML machine the size of a bus, and the coarser wiring layers use older deep-ultraviolet tools. Ion implanters dope the silicon to set its electrical properties, and metrology tools inspect every step, because a defect a few nanometers across ruins a die. When the last layer is done, the wafer carries a few hundred chips, which are tested, cut apart, and packaged (Module 10). A leading-edge fab, at the cost and scale in the last row of the table below, runs in a cleanroom about a thousand times cleaner than a hospital operating room, and it must run every hour of the year to earn back its cost.

Stage	What happens	Key equipment (maker)	Duration	Note
Wafer	A 300-mm disc of ultra-pure single-crystal silicon, polished to atomic flatness	Shin-Etsu, SUMCO (Japan) grow and slice the crystals	Weeks (upstream)	Japan's share in Module 9
Deposition	Thin films of insulators, metals, and semiconductors laid down, sometimes one atomic layer at a time	Applied Materials, Lam, Tokyo Electron, ASM (atomic-layer deposition)	Hundreds of steps	The first step of a cycle repeated 80-100 times
Photoresist coat	A light-sensitive polymer spun onto the wafer	Tokyo Electron (coater-developers, about 90%); JSR, Shin-Etsu, Tokyo Ohka (resist)	Per layer	A Japanese chokepoint
Lithography	Light projected through a mask prints the circuit pattern into the resist: extreme ultraviolet (EUV, 13.5-nm light) for the finest layers, deep ultraviolet (DUV, 193-nm light, often through water, called immersion) for the rest	ASML (all EUV; about 80% of DUV); Nikon, Canon (DUV)	Per layer; about 20 EUV layers at 3nm	The chokepoint of chokepoints; Module 9
Etch	Plasma removes material where the resist was exposed, carving the pattern	Lam, Tokyo Electron, Applied	Per layer	Atomic precision
Ion implantation and doping	Atoms fired into the silicon to create the transistor's electrical regions	Applied, Axcelis	Dozens of steps	

Stage	What happens	Key equipment (maker)	Duration	Note
Chemical-mechanical polishing	The wafer flattened for the next layer	Applied, Ebara	Per layer	
Metrology and inspection	Every layer measured and checked for defects	KLA (the leader), Applied, Hitachi	Continuous	
Back-end-of-line wiring	Ten to fifteen layers of copper wiring connect the transistors	Applied, Lam	Hundreds of steps	The wiring now limits speed as much as the transistor
Wafer test	Every die (one chip's area on the wafer) probed electrically	Advantest, Teradyne	Days	Good dies mapped
Dicing, packaging, and final test	The wafer cut into dies; each die mounted, connected, and sealed; for AI chips, advanced packaging that bonds dies and memory stacks together	ASE, Amkor, TSMC (CoWoS), Intel (Foveros); Disco (dicing)	Days to weeks	Module 10; the AI chip's bottleneck
Total	About 1,000 process steps		3-4 months from bare wafer to finished chip	A leading-edge fab: $20-30B, 100,000 wafers a month, about 1,500 machines

2b. Yield and the Learning Curve

Yield is the share of dies on a wafer that work, and it decides a fab's economics. A new process starts at 20-50% yield, because a thousand steps interact in ways no simulation predicts, and over one to two years it climbs toward 80-95% as engineers find and fix each source of defects. A fab that cannot climb the curve loses money on every wafer: Samsung's 3nm process struggled for years, Intel's 10nm arrived years late (Module 7), and Huawei's Ascend chips, made at SMIC, are reported by analysts to yield 20-40%. A fab that climbs fast prints money, as TSMC's 3nm and 2nm lines have. The learning curve is why leading-edge manufacturing concentrates: the company that ran the last node has the engineers, the process recipes, the equipment relationships, and the customers' trust to run the next one. A rival starting cold must spend years and billions to reach the yield the leader started from. It is also why the industry's history is a list of companies that fell off the curve: the Japanese memory makers, Motorola, AMD's own fabs (spun off as GlobalFoundries, which stopped at 14nm), and then Intel, which is why so few companies now attempt the leading edge (Module 3).

Worked example — A 3nm Wafer's Economics

A 300-mm wafer at TSMC's 3nm node costs a large customer about $20,000 to process, and a 2nm wafer about $30,000, according to industry reports. A smartphone processor of 100 square millimeters fits about 600 dies on a wafer, and at 85% yield, 510 of them work, so each costs about $39 to make before packaging and test, against the $120-150 the phone maker pays for it. An AI accelerator die of 800 square millimeters, near the largest size a lithography tool can print, fits about 70 to a wafer, and at 75% yield, 52 work, or about $385 of silicon each. Turning that die into a product takes several thousand dollars of high-bandwidth memory stacks and advanced packaging, and the finished Blackwell-class GPU sells for $30,000-40,000. The arithmetic says three things. The silicon is most of a phone chip's cost but a small share of an AI chip's price, and memory and packaging, once rounding errors, are now among an AI chip's largest costs. And the margins sit with the makers of the scarce parts: TSMC, Nvidia, and the memory makers (Modules 4 to 6).

Module 2 Key Points

How a Chip Is Made

• A chip is printed in 80-100 cycles of deposition, lithography, etch, and polishing: about a thousand steps over three to four months.

• Yield decides a fab's economics: new processes start at 20-50% and climb toward 80-95% over one to two years.

• The learning curve concentrates the industry: whoever ran the last node runs the next, so only three companies make leading-edge chips.

• In a 3nm wafer's arithmetic, silicon is most of a phone chip's cost but a small share of an AI chip's price.

• Memory and packaging, once rounding errors, are now among an AI chip's largest costs.

Module 2 Quiz

Check Your Understanding

1. How many times does a wafer cycle through the core process sequence?

A) About a dozen times
B) Several hundred times
C) 80 to 100 times, about once per layer
D) 1,000 times

2. Why does leading-edge manufacturing concentrate in so few companies?

A) Patents on each new node lock rivals out of the process for twenty years.
B) The learning curve rewards the incumbent.
C) Each country licenses one maker.
D) Silicon wafers are scarce, so the largest buyers take all of the supply.

3. In the worked example, why is the silicon a small share of an AI chip's price?

A) Because the AI die is tiny
B) Because TSMC sells its AI wafers below cost to win Nvidia's business
C) Because the yield on AI dies is nearly 100%
D) Because about $385 of silicon becomes a $30,000-40,000 product

4. What yield does a new process typically start at?

A) About 80-95%
B) Between 20% and 50%
C) Nearly 100%, since the steps are simulated in advance
D) Under 5%

5. What happened to GlobalFoundries, AMD's former fabs, at the leading edge?

A) It became TSMC's main rival at the 2nm node.
B) It merged with Intel's foundry.
C) It moved all production to China, then closed its American fabs.
D) It stopped at 14nm and left the race.

Module 3 — Moore's Law, the Nodes, and the End of Scaling

The Cadence That Built the Industry, and What Replaced It

Recommended Reading / Reference

Book: Chris Miller, *Chip War* (the scaling chapters)

Reference: The IEEE *International Roadmap for Devices and Systems*, the industry's shared forecast of the nodes ahead

3a. Moore's Law

In 1965 Gordon Moore, later a founder of Intel, observed that the number of transistors on a chip was doubling every year, and in 1975 he revised the pace to every two years, and the observation became a law because the industry planned to it. Each new generation of manufacturing process, called a node, shrank the transistor by about 30% in each direction, which doubled the density, roughly halved the cost per transistor, and raised the speed. The two-year cadence held from 1971 to about 2015 and built the PC, the phone, the internet, and the cloud, because a product too costly this year was affordable two years later. It also concentrated the industry, because each node's fab cost roughly doubled: a fab cost $100 million in 1980, $1 billion in 2000, $5 billion in 2010, and $20-30 billion at 2nm. The number of companies able to build at the leading edge fell from dozens to three, and the leader's advantage compounded through the learning curve of Module 2.

Node (marketing name)	Year (volume)	Transistor density (million per mm²)	Structure	Companies at the node	Note
10 µm	1971	~0.001	Planar	Intel, TI, and dozens	The Intel 4004
1 µm	1985	~0.1	Planar	Twenty-plus	The Japanese memory era
180 nm	1999	~0.4	Planar	Fifteen	The internet build-out
90 nm	2004	~1.5	Planar (strained silicon)	Ten	The last easy node
45/40 nm	2008	~4	Planar (high-k metal gate)	Eight	Intel's lead at its peak
28 nm	2011	~10	Planar (the last planar node)	Six	Still a workhorse: about 20% of wafers
16/14 nm	2014-15	~30	FinFET (the channel raised into a fin)	Five (TSMC, Samsung, Intel, GlobalFoundries; UMC stopped here)	GlobalFoundries quit the race (2018)
7 nm	2018	~90	FinFET	Three (TSMC, Samsung; Intel late)	SMIC reached 7nm-class with DUV (Module 14)
5 nm	2020	~140	FinFET	Three	TSMC's largest node by revenue (Module 5)

Node (marketing name)	Year (volume)	Transistor density (million per mm²)	Structure	Companies at the node	Note
3 nm	2022-23	~200 (TSMC N3)	FinFET (TSMC); gate-all-around (Samsung)	Three (Samsung's yields troubled)	TSMC's second-largest node by revenue (Module 5)
2 nm	2025-26 (State of the Art)	~250-300	Gate-all-around nanosheets for all three; backside power (Intel 18A first)	Three	TSMC's newest node
A16	2027 (TSMC; slipped from 2026 at the April 2026 symposium)	—	Nanosheets plus backside power (TSMC's 'Super Power Rail')	One to three	TSMC's first backside-power node
14A / A14	2028 (Intel 14A, with High-NA EUV; TSMC A14, without it)	—	Nanosheets	Two or three	Customers in Module 7
A13 / A12	2029 (TSMC)	—	Second-generation nanosheets; A12 with backside power	Two or three	Announced April 2026
Beyond 2030	2030s	—	Stacked transistors (CFET)	—	IMEC's roadmap runs to 2039

Density figures approximate. Dates after 2026 are the companies' own roadmaps, and TSMC quotes future density only relative to earlier nodes (A14 about 20% denser than N2; A13 about 6% denser than A14).

3b. What Ended, and What a Node Means Now

Classical scaling ended around 2005: under 'Dennard scaling,' smaller transistors had used proportionally less power, so chips got faster without getting hotter, until current began leaking through transistors too small to switch cleanly. Voltages could not fall further, and clock speeds stalled near 4 GHz, or 4 billion cycles a second, while density kept rising but more slowly, from two years per node to about three after 2015. Around 5nm the cost per transistor stopped falling, because EUV lithography and repeated patterning raised wafer costs faster than density rose. The node names also stopped meaning a dimension, since nothing in a '3nm' chip is 3 nanometers across, and the name is now a label for a step in density and performance. A TSMC 3nm, a Samsung 3nm, and an Intel 18A ('18 angstroms,' or 1.8 nanometers) cannot be compared by name. What kept changing was the transistor's shape: the flat, planar transistor gave way to the FinFET, which raised the channel into a fin that the gate grips on three sides. At 2nm the FinFET gave way to the gate-all-around nanosheet, which the gate surrounds completely, and backside power delivery, first in Intel's 18A, moves the power wiring beneath the transistors and frees room above them. The table gives the dates of TSMC's next nodes and of the stacked transistors that IMEC, the Belgian research center the industry shares, projects beyond them.

3c. The Three Substitutes

The industry replaced Moore's law with three things. The first is architecture: instead of a faster general-purpose processor, designers build specialized ones that do one job many times faster per watt, such as graphics processors with thousands of simple cores, tensor units for the matrix math of AI, and neural engines in phones. Most of Nvidia's gains over the past decade came from design and software, not from the node. The second is packaging: instead of one large die, several smaller 'chiplets' are bonded together, as AMD does (Module 10). Memory is stacked beside or on top of the processor as high-bandwidth memory (HBM), and the whole assembly sits on a silicon base called an interposer, as in TSMC's CoWoS process, so that the package does what the transistor used to. The third is specialization: the large cloud companies design their own chips for their own workloads, such as Google's TPU, Amazon's Trainium, Microsoft's Maia, and Meta's MTIA. By Nvidia's own measure, which it calls 'Huang's law,' AI performance per dollar has improved faster in the 2020s than Moore's law ever delivered, while the transistor itself improves by only 10-15% a node. That is why the industry's constraint moved from the fab to the package and the memory, and why the equipment and packaging chokepoints of Modules 9 and 10 now matter as much as lithography.

Module 3 Key Points

Moore's Law, the Nodes, and the End of Scaling

• Moore's law was an industry plan: double transistor density every two years, which cut the cost per transistor and built the digital economy.

• Each node's fab cost roughly doubled, from $100 million in 1980 to $20-30 billion at 2nm, shrinking the field to three.

• Dennard scaling ended around 2005, and the cost per transistor stopped falling around 5nm.

• Node names are labels, not dimensions: nothing in a '3nm' chip measures 3 nanometers.

• The transistor's shape kept changing, from planar to FinFET to gate-all-around; TSMC's backside-power node, A16, arrives in 2027.

• Architecture, packaging, and specialization replaced shrinking as the source of gains, moving the constraint to the package and memory.

Module 3 Quiz

Check Your Understanding

1. What did Moore's law describe?

A) Transistor counts doubling about every two years
B) A law of physics that limits how small transistors can get
C) A legal rule set by the government for transistor density in 1975
D) A pricing formula the memory makers used to set prices for each new generation of DRAM

2. What does a node name such as '3nm' mean today?

A) It is the length of the transistor's gate, measured in nanometers.
B) It is the width of the finest wire on the chip.
C) It is a marketing label for a step in density and performance, not a physical measurement.
D) It is the thickness of the finished silicon wafer.

3. When is TSMC's A16, its first node with backside power, now due?

A) In 2025, on schedule
B) 2027, after a slip
C) Already in early 2026
D) Not until 2030 or later

4. What ended around 2005?

A) The FinFET transistor, which was replaced by planar designs.
B) The use of light to print chips, which gave way to electron beams.
C) Dennard scaling, under which smaller transistors used less power
D) The industry's two-year cadence of new fabs and nodes.

5. What are the three substitutes for Moore's law?

A) Faster clocks, higher voltages, bigger wafers
B) Subsidies, tariffs, export controls
C) Silicon carbide, gallium nitride, and graphene
D) Architecture, packaging, specialization

Module 4 — The Business Models and the Companies

Who Does What, and Who Earns What

Recommended Reading / Reference

Book: Chris Miller, *Chip War* (on Morris Chang and the fabless model)

Reference: The companies' quarterly reports and filings, the source for every figure in the table below

4a. The Models

Model	What it does	Examples	Economics	Trend
Integrated device manufacturer (IDM)	Designs and makes its own chips	Samsung, SK Hynix, Micron, Intel (in part), Texas Instruments, Infineon, STMicro	Full control, and the full capital burden of fabs	Most logic makers abandoned it; the memory makers are the durable IDMs
Fabless	Designs chips and pays a foundry to make them	Nvidia, AMD, Qualcomm, Broadcom, MediaTek, Marvell, Apple (for its own use)	Asset-light; gross margins of 50-75% at the top; dependent on TSMC	The winning model since 2000
Foundry	Makes chips for other companies' designs	TSMC (about 70% of foundry revenue; Module 5 for its share of the leading edge), Samsung Foundry, Intel Foundry, GlobalFoundries, UMC, SMIC	Capital-intensive; scale and yield decide	One dominant, two challengers, the rest at mature nodes
Equipment	Makes the machines that make chips	ASML, Applied Materials, Lam Research, Tokyo Electron, KLA	High margins; cyclical; the chokepoint layer	Module 9
Materials	Wafers, resists, gases, chemicals	Shin-Etsu, SUMCO, JSR, Air Liquide, Entegris	Steady; dominated by Japan	Module 9
Assembly and test (OSAT)	Packages and tests chips (outsourced semiconductor assembly and test)	ASE, Amkor, JCET; TSMC and Intel for advanced packaging	Thin margins, except in advanced packaging	Module 10
Design software and IP	Design software and licensed circuit blocks (intellectual property, or IP)	Synopsys, Cadence, Siemens EDA; Arm	Software margins; every chip needs them	Module 12
Custom silicon	Designs chips to a customer's specification	Broadcom (Google's TPU and others), Marvell, MediaTek, Alchip	The large cloud companies' route to their own chips	The fastest-growing service

The industry's structure rests on a separation of design from manufacturing, which Morris Chang created when he founded TSMC in 1987 to make chips for anyone who designed them. Before that, a chip company was an integrated device manufacturer that did both, and the cost of fabs meant only the largest could compete. After it, a 'fabless' company could design a chip and rent TSMC's fabs, which is how Nvidia, Qualcomm, Broadcom, Apple's chip division, and every AI start-up came to exist, and how AMD survived after selling its fabs in 2009. The model also split the profit: the designers at the top earn the table's gross margins on their intellectual property, TSMC earns the margins of Module 5 on manufacturing it alone can

do, and the equipment makers earn about 50% on the machines. Memory, long the cyclical poor relation, became in 2026 the most profitable part of the chain, because prices rose faster than costs ever could. The ranking below shows the result: a designer that makes nothing is the most valuable company in the world, its manufacturer is second in the industry, and the memory makers have joined them near the top.

4b. The Companies

Company	Country	Model	Revenue, latest quarter	Growth y/y	Market value (Oct 8, 2026)	Note
Nvidia	US	Fabless (AI accelerators, networking)	$96.2B (to July 26, 2026)	+106%	$5.7T	Module 6
TSMC	Taiwan	Foundry	$40.2B (Q2 2026)	+34%	$2.4T	Module 5
Broadcom	US	Fabless (custom AI chips, networking) and software	$29.6B (to Aug 2, 2026)	+86%	$1.8T	Module 12
Samsung Electronics	Korea	IDM (memory, foundry, phones)	Chip division ₩127.5T, about $89B (Q2 2026)	—	$1.3T (whole company)	Modules 7 and 8
Micron	US	IDM (memory)	$54.2B (to Sept 3, 2026)	+379%	$1.2T	87% gross margin
AMD	US	Fabless (CPUs, GPUs)	$11.5B (Q2 2026)	+50%	$1.1T	Data center $6.7B, up 107%; the MI400 and Helios ramp
SK Hynix	Korea	IDM (memory)	₩79.3T, about $55B (Q2 2026)	+257%	$0.9T	Leader in high-bandwidth memory; about 76% operating margin
ASML	Netherlands	Equipment	€9.3B (Q2 2026)	—	$0.7T	The only maker of EUV tools
Intel	US	IDM (CPUs) and foundry	$16.1B (Q2 2026)	+25%	$0.6T	Module 7
CXMT	China	IDM (memory)	—	—	$0.5T	China's DRAM maker (Module 8)
Applied Materials	US	Equipment	$9.1B (to July 26, 2026)	+25%	$0.4T	The broadest equipment maker
Lam Research	US	Equipment	$6.7B (to June 28, 2026)	—	$0.4T	Etch and deposition; guided $8.1B for the next quarter
Arm	Britain (SoftBank)	Design IP, and now its own chip	$1.29B (to June 30, 2026)	—	$0.3T	Module 12
Texas Instruments	US	IDM (analog)	$5.5B (Q2 2026)	+23%	$0.26T	Analog's leader, on its own 300-mm fabs
Qualcomm	US	Fabless (phones, PCs, cars)	$9.9B (to June 2026)	—	$0.19T	The largest seller of phone processors and modems, by revenue
SMIC	China	Foundry	$3.0B (Q2 2026)	+36%	—	China's leading foundry; 90% of sales in China

Company reports for the latest quarter each has published; Samsung's and SK Hynix's dollar figures converted at about 1,430 won to the dollar. Market values from companiesmarketcap.com, rounded; SMIC's is omitted because its listings are valued separately. A dash means the comparison is not given in the source used.

4c. Where the Value Sits

In 2026 the industry's profit sits in three places, the first of them the AI designers: Nvidia (Module 6), and Broadcom, whose custom-chip business is the second-largest AI chip business. The second is the manufacturer they depend on, TSMC (Module 5). The third, new in 2026, is memory, whose makers' margins in the table above soared with its prices, and Samsung's guided third-quarter profit (Module 8) comes nearly all from chips. The equipment makers earn steadily on the capital spending the others must do, while the mature-node, analog, automotive, and packaging companies earn what a more ordinary market allows, under pressure from Chinese capacity. The concentration of Module 1 is the risk the rest of the course explores: one supplier, TSMC, makes nearly every AI chip, and one equipment maker, ASML, has no substitute. And four cloud companies' capital spending (Module 16) is the demand behind all of it.

Module 4 Key Points

The Business Models and the Companies

• Morris Chang's TSMC, founded in 1987, separated chip design from manufacturing and made the fabless industry possible.

• Designers earn gross margins of 50-75%, TSMC about 68%, and the equipment makers about 50%.

• In 2026 memory became the most profitable part of the chain; Micron's gross margin reached 87%.

• Nvidia, worth about $5.7 trillion, is the world's most valuable company and makes none of its own chips.

• The profit sits in three places: the AI designers, TSMC, and the memory makers.

Module 4 Quiz

Check Your Understanding

1. What did Morris Chang's founding of TSMC in 1987 make possible?

A) Fabless companies that design chips and rent a foundry's fabs to make them
B) The first memory chip
C) EUV lithography
D) Moving most chipmaking to Korea

2. Which part of the chain became the most profitable in 2026?

A) Advanced packaging, now scarce, earned the highest margins in the chain.
B) Equipment did.
C) Memory did, because its prices rose far faster than its costs.
D) The mature-node foundries in China earned the most.

3. What was Micron's gross margin in its quarter to early September 2026?

A) About 40%
B) 87%
C) Roughly 52%
D) About 67%, like TSMC's

4. Which company is second in the industry by market value, after Nvidia?

A) TSMC, Nvidia's foundry
B) Broadcom, the custom-chip designer
C) Samsung Electronics
D) Micron, after its memory boom

5. Why does the module call the concentration the risk the rest of the course explores?

A) Because the industry is too fragmented to coordinate its spending.
B) Because China now controls most of the chain, including design software
C) Because prices are falling everywhere.
D) Because one foundry makes nearly every AI chip and one toolmaker has no substitute.

Module 5 — TSMC and the Taiwan Concentration

The Company the World's Computing Runs On

Recommended Reading / Reference

Book: Chris Miller, *Chip War* (on Morris Chang and TSMC's rise)

Reference: TSMC's quarterly management reports and earnings-call transcripts (the second quarter of 2026, July 16)

5a. The Numbers

Measure	Figure	Note
Revenue	Q2 2026 in Module 4's table; Q3 guided at $44.6-45.8B; 2026 growth guided at slightly above 40%, on 2025 revenue of about $122B	Far ahead of the next foundry (Module 4)
Margins	Gross 67.7%, operating 60.3% (Q2 2026); Q3 gross margin guided at 65-67%	Unusual for any manufacturer; the 2nm ramp and overseas fabs each cost a few points
Capital spending	$60-64B in 2026, raised on July 16 from $52-56B	Most of it for 2nm and 3nm capacity and advanced packaging
Node mix (Q2 2026 wafer revenue)	5nm 33%; 3nm 30%; 7nm 11%; 2nm 3%; 16nm and older the rest	7nm and below 77% of wafer revenue
By platform (Q2 2026)	High-performance computing (HPC: AI and servers) 66%; smartphones 22%; the rest connected devices, cars, and consumer electronics	HPC passed smartphones in 2023 and is now three times as large
Customers	Nvidia, Apple, AMD, Broadcom, Qualcomm, MediaTek, Intel, and the cloud companies' chip teams	Every leading AI accelerator
Leading-edge capacity	2nm about 90,000 wafers a month by end-2026 and about 110,000 by mid-2027; 3nm above 180,000 (Taiwanese press reports, September 2026)	Taiwan's share in Module 1
CoWoS advanced packaging	Capacity and expansion plans in Module 10	The binding constraint on AI hardware
Overseas fabs	Arizona: the first fab making 4nm chips since late 2024; the second installing tools for 3nm production in 2027; the US commitment raised to $265B on July 16, 2026, adding 2nm-and-beyond fabs and packaging. Japan: Kumamoto. Germany: Dresden (automotive nodes, 2027)	—
Roadmap	N2P (2026), then A16, A14, A13, and A12	Their dates in Module 3's table

Taiwan Semiconductor Manufacturing Company (TSMC) makes about 90% of the world's leading-edge chips, including Nvidia's AI accelerators and the processors of Apple, AMD, Broadcom, Qualcomm, and the cloud companies. Its 2026 numbers, in the table above, are the industry's center of gravity, and high-performance computing, meaning AI and servers, has overtaken the phone business TSMC once depended on. The overseas expansion is real but secondary. TSMC's enlarged American commitment, $100 billion more than before, brings 2nm-class fabs and advanced packaging to Arizona, while its Japanese and German fabs make older nodes and the leading edge stays in Taiwan.

5b. Why TSMC Won

Of the three things TSMC won on, the first is trust: a pure-play foundry that designs no chips of its own can be given every customer's secrets. Samsung, which competes with its customers in phones and memory, could not offer that, and neither could Intel, which competed with everyone. The second is scale and yield: running more leading-edge wafers than anyone let TSMC climb the learning curve first on every node and fund the next one, so its 2026 spending, in the table above, is an entry fee no rival pays. The third is Taiwan itself, with an engineering culture, a government that built the science parks and the power supply, and a workforce that accepts the discipline of a fab, and until the 2020s, its chipmaking was rarely a political question. The result is a company whose customers cannot easily leave. Nvidia's Blackwell and Rubin chips exist only on TSMC's processes and packaging, and so, by wide report, does Apple's newest phone processor, the A20 Pro (State of the Art). Its rivals have not caught it, and its position has become a strategic asset of Taiwan, often called the island's 'silicon shield.'

5c. The Taiwan Risk

The concentration is the largest single-point risk in the world economy, because a Chinese blockade or invasion of Taiwan would stop the production of nearly every advanced chip on earth. Bloomberg Economics estimated in 2024 that a war over Taiwan would cost the world economy about $10 trillion in its first year. No plan in place replaces more than a fraction of that capacity within five years, whether Arizona's fabs, Japan's, the American subsidies, or Intel's foundry. The responses run in three directions, the first of them dispersal: TSMC's overseas fabs will hold a growing share of capacity but lag the Taiwan leading edge. The US-Taiwan trade agreement of January 15, 2026 ties tariff relief to building in America: Taiwanese chipmakers building American fabs may import up to two and a half times their planned American capacity free of the new chip tariffs while they build, and one and a half times once the fabs run. The second is deterrence: on the 'silicon shield' argument, Taiwan's chips make it too valuable to attack and too valuable for the United States to abandon, and the American military posture in the region backs the argument. The third is resilience: stockpiles, the cloud companies' inventory builds, and contingency plans that no one considers adequate, so the risk cannot be hedged at the scale that matters, and every party is betting it will not be tested.

Worked example — A Leading-Edge Fab's Economics

A 2nm fab in Taiwan costs about $25 billion: about $5 billion for the building and cleanroom, $18 billion for some 1,500 machines, and $2 billion for the ramp-up. A fifth of the machine budget is ASML's EUV tools, at more than $200 million each. At full output, which takes about two years to reach, the fab makes 100,000 wafers a month, and at about $30,000 a wafer its revenue is $36 billion a year. Cash costs for materials, power, labor, and maintenance run about 35% of revenue, or $12.6 billion, and depreciation, mostly the machines written off over five years, adds about $4.5 billion. That leaves operating profit of about $19 billion, a margin above 50%, which before tax would repay the $25 billion in about a year and a half of full output. Counting the two-year ramp, when output and yields are lower, the payback is closer to three or four years. The result holds only if the fab runs full: at 70% utilization, revenue falls to about $25 billion, and because most costs are fixed, operating profit falls to about $8 billion. Yield matters as much: at 60% yield, a customer pays for wafers on which four chips in ten do not work, and a foundry that cannot raise yields loses the customer, while a node that customers skip, Samsung's long problem, strands $18 billion of machines. The same fab in Arizona costs 30-50% more to build and 10-20% more to run, which the tariff relief and American subsidies are meant to offset. The arithmetic shows why only a company with customers to fill a fab can build one, and why TSMC's overseas fabs trail by a node, since the newest tools go where the yield engineers are. And it shows that every fab is a bet on demand three years out, the bet the AI buyers of Module 16 are underwriting.

Module 5 Key Points

TSMC and the Taiwan Concentration

• TSMC makes about 90% of leading-edge chips; its second-quarter 2026 revenue was $40.2 billion at a 67.7% gross margin.

• It expects 2026 revenue to grow slightly more than 40% and will spend $60-64 billion on capacity.

• TSMC won on trust, on scale and yield, and on Taiwan's engineering base; customers such as Nvidia cannot easily leave.

• Its American commitment rose to $265 billion in July 2026, but the leading edge stays in Taiwan.

• A war over Taiwan could cost the world about $10 trillion in a year, and the risk cannot be hedged at scale.

• A 2nm fab costs about $25 billion and earns its return only when it runs full at high yields.

Module 5 Quiz

Check Your Understanding

1. What gross margin did TSMC report for the second quarter of 2026?

A) 60%
B) About 75%, like Nvidia's margin on the chips TSMC makes for it
C) About 52%
D) 67.7%

2. Where does TSMC keep its leading edge?

A) In Taiwan, where its yield engineers and newest fabs are
B) In Arizona, since the July 2026 commitment
C) Split evenly between Taiwan and Japan
D) In Germany, at the Dresden plant

3. What did TSMC announce on July 16, 2026 about its American investment?

A) A cut in its US plans to $65 billion
B) $100 billion more, to $265 billion
C) A pause on building its second Arizona fab
D) A sale of its Arizona fabs to Intel

4. Why could Samsung and Intel not offer customers what TSMC offers?

A) They lacked EUV tools.
B) Their fabs were too small to win any customer's trust or orders at all.
C) They competed with their own customers, in phones, memory, and processors.
D) They were barred from selling to American companies by the CHIPS Act and the tariffs.

5. In the worked example, what operating margin does a full 2nm fab earn?

A) About 10%
B) Just above 50%
C) Roughly 25%, about $9 billion
D) Nearly 90%

Module 6 — Nvidia and the AI Accelerator

The Company That Became the Industry's Center

Recommended Reading / Reference

Book: Tae Kim, *The Nvidia Way*

Reference: Nvidia's quarterly press releases and Form 10-Q (the second quarter of fiscal 2027, August 26, 2026)

6a. The Numbers

Quarter (fiscal year)	Revenue	Data center	Gross margin	Note
Q1 FY2024 (to Apr 2023)	$7.2B	$4.3B	65%	Before the surge
Q4 FY2024 (to Jan 2024)	$22.1B	$18.4B	76%	The Hopper shortage
Q4 FY2025 (to Jan 2025)	$39.3B	$35.6B	73%	Blackwell's first shipments; the DeepSeek scare
Q1 FY2026 (to Apr 2025)	$44.1B	$39.1B	61%	The H20 write-down for China (Module 13)
Q4 FY2026 (to Jan 25, 2026)	$68.1B	$62.3B	75%	Fiscal 2026 revenue $215.9B, up 65%
Q1 FY2027 (to Apr 26, 2026)	$81.6B	$75.2B	75%	No China data-center revenue assumed
Q2 FY2027 (to Jul 26, 2026)	$96.2B	$89.0B (+117%)	75%	Vera Rubin in full production
Q3 FY2027 (guidance)	About $108B	—	About 74%	Report in Module 17's watch list

Nvidia's press releases; its fiscal year ends in late January.

Nvidia's quarterly revenue rose thirteenfold in a little over three years, from the table's first row to its July 2026 quarter, and the data-center column is the story: in effect, the capital budget of the AI industry turned into one company's income statement. Its operating profit in the July quarter was $63.7 billion, and the table gives its guidance for the quarter to October. Jensen Huang, the founder and chief executive, frames the business as 'compute is revenue': the AI industry's output is tokens, the units of text and data a model produces, and customers who sell tokens profitably keep buying chips. That is the bull case in a sentence, and Nvidia's market value (Module 4) is the market's agreement with it. Behind both is a short list of buyers: Amazon, Microsoft, Alphabet, and Meta, the four cloud companies whose capital spending Module 16 tallies, plus Oracle and the AI labs, and a large share of what they spend becomes Nvidia's revenue.

6b. The Product Cadence and the Moat

Generation	Launch	Process (TSMC)	Transistors	Memory	Peak low-precision compute per GPU	Note
A100 (Ampere)	2020	7nm	54B	40-80 GB HBM2e	About 0.3 petaflops (16-bit)	The generation ChatGPT was trained on
H100 (Hopper)	2022	4nm-class	80B	80 GB HBM3	About 2 petaflops (8-bit)	The shortage of 2023-24; H20 the cut-down version for China
B200 / GB200 (Blackwell)	2024-25	4nm-class; two dies	208B	192 GB HBM3e	About 10 petaflops (4-bit)	Sold as a 72-GPU rack, the NVL72
B300 / GB300 (Blackwell Ultra)	2025-26	4nm-class; two dies	208B	288 GB HBM3e	About 15 petaflops (4-bit)	The volume product of 2025-26
Rubin (with the Vera CPU)	2026 (State of the Art)	3nm; two dies	336B	288 GB HBM4 at 22 terabytes a second	50 petaflops (4-bit, inference); 35 (training)	In the Vera Rubin NVL72 rack
Rubin Ultra; Feynman	2027; 2028 (Nvidia's roadmap)	—	—	Next-generation HBM	—	Rubin Ultra in a 600-kilowatt rack

Nvidia's specifications; dense figures, without sparsity. A petaflop is a thousand trillion operations a second; 4-bit and 8-bit refer to the precision of the numbers, lower precision being faster.

Nvidia's moat has four parts, and the first is the chips, which it now renews every year, from Hopper to Blackwell, Blackwell Ultra, and Rubin. Each generation sharply raises performance per rack, on TSMC's best processes and packaging. The second is software: CUDA, the programming platform Nvidia has built since 2006, is what most AI models, libraries, and engineers are built on, and moving a data center to a rival's chip means rewriting much of that software. The third is the system: since Blackwell, the product is not a chip but a rack of GPUs (the NVL72 in the table above), tied together by NVLink, Nvidia's own connection fabric, with networking from its 2020 purchase of Mellanox and liquid cooling throughout. Customers buy a working slice of a data center, not a component. The fourth is the supply chain: Nvidia holds the largest share of TSMC's advanced-packaging capacity and first call on high-bandwidth memory. In December 2025 it also paid about $20 billion to license the technology of Groq, an inference-chip start-up, and hire its leaders, and a Groq-based inference chip now sits in its 2026 lineup. The rivals in the next section each attack one of these four parts, and none yet has all four.

6c. The Rivals

Rival	Product	Status (October 2026)	Edge	Weakness
AMD	MI400 series (MI455X) and the Helios rack	OpenAI's 6-gigawatt deal (October 2025, with warrants for up to 10% of AMD); launch and ramp in the State of the Art	Price; open software (ROCm) maturing	A smaller software base than CUDA; a first rack-scale product
Google (TPU, with Broadcom)	TPU v7 'Ironwood'; TPU 8t and 8i (State of the Art)	Anthropic contracted for up to a million TPUs and more than a gigawatt in 2026	Cost per token; vertical integration	Sold mainly through Google's own cloud

Rival	Product	Status (October 2026)	Edge	Weakness
Amazon (Trainium)	Trainium 2; Trainium 3 (3nm, announced December 2025)	The chips behind Anthropic's 'Project Rainier'	Price; Amazon Web Services' reach	Most of AWS's fleet is still Nvidia
Microsoft, Meta, OpenAI	Maia 200 (January 2026); MTIA; OpenAI's chip with Broadcom	In-house inference chips, mostly for their own workloads	Their own workloads	Late; Nvidia remains the training platform
Broadcom (the custom-chip designer)	Custom accelerators for Google and other cloud and AI companies	AI chip revenue in Module 12	Design services plus networking	Not a merchant platform; a few large customers
Huawei (Ascend)	Ascend 950PR	In mass production (State of the Art; volumes in Module 14)	A Chinese market closed to Nvidia	Made on SMIC's processes at low yields; memory shortages
Start-ups	Cerebras (wafer-scale), SambaNova, Tenstorrent, and others	Cerebras listed on Nasdaq on May 14, 2026	Inference speed	Scale and software

6d. The Circular Deals

Much of the 2025-26 financing of AI infrastructure runs in circles: in September 2025 Nvidia signed a letter of intent to invest up to $100 billion in OpenAI as OpenAI deployed 10 gigawatts of Nvidia systems. The letter never became a contract, and in early 2026 Nvidia moved instead to a direct investment of about $30 billion in OpenAI's next funding round, according to the Financial Times. That round closed in March at $122 billion, valuing OpenAI at $852 billion. AMD gave OpenAI warrants on its shares in exchange for the order in the rivals table above, and Oracle signed a $300 billion computing contract with OpenAI in 2025 and borrowed heavily to build it, filling the data centers with Nvidia chips. Nvidia also committed up to $10 billion to Anthropic and holds stakes in CoreWeave, a cloud company that rents out its GPUs, and in other customers. Critics call this the cycle's tell, because the chip seller is funding the chip buyer's purchases, while supporters reply that every capital-intensive build-out has been financed by its suppliers, from the railroads to the fiber-optic networks of the late 1990s. The critics' answer is that 1999's fiber build is exactly their point. The deeper fact is that Nvidia's revenue depends on a few companies' capital budgets and on the AI labs' ability to keep raising money. Whether that spending earns its cost is the question Module 16 takes up, and it will be answered by the customers' own revenues in 2027-28, not by Nvidia's.

Module 6 Key Points

Nvidia and the AI Accelerator

• Nvidia's quarterly revenue rose thirteenfold in three years, to $96.2 billion in the quarter to July 2026.

• It expects about $108 billion in the quarter to October, at a gross margin near 74%.

• Its moat has four parts: annual chip generations, CUDA software, rack-scale systems, and a captive supply chain.

• Rubin, with 336 billion transistors and 288 GB of HBM4, entered full production in 2026.

• Rivals each attack one part of the moat: AMD on price, and Google and Amazon with their own chips.

• Its revenue depends on a few customers' capital budgets and on the AI labs' ability to keep raising money.

Module 6 Quiz

Check Your Understanding

1. What revenue does Nvidia expect for the quarter to October 2026?

A) $81.6 billion
B) About $96 billion
C) About $130 billion, up 35%
D) $108 billion

2. What are the four parts of Nvidia's moat, according to the module?

A) Its chips, the CUDA software, rack-scale systems, and first call on the supply chain.
B) Patents, tariffs, subsidies, and the lobbying budget it spends in Washington.
C) Price, marketing, retail stores, and its gaming business.
D) Memory, packaging, design software, and its own fabs.

3. What happened to the letter of intent Nvidia signed with OpenAI in September 2025?

A) It was signed as written.
B) It was replaced by a merger of the two companies' chip teams under one roof.
C) It never became a contract.
D) It was blocked by antitrust regulators.

4. Why does CUDA protect Nvidia's position?

A) It is required by law for AI data centers.
B) It makes Nvidia's chips cheaper to manufacture.
C) It is the only language that runs on TSMC's chips.
D) Most AI software is built on it, so moving to a rival's chip means rewriting much of that software.

5. Which rival has a 6-gigawatt deal with OpenAI?

A) Huawei
B) AMD, which gave OpenAI warrants
C) Google's TPU business
D) Cerebras, the wafer-scale start-up

Module 7 — Intel, Samsung, and the Foundry Race

The Companies Trying to Be the Second Leading-Edge Foundry

Recommended Reading / Reference

Book: Chris Miller, *Chip War* (on Intel's decline)

Reference: Intel's quarterly earnings releases (the second quarter of 2026, July 23)

	Intel Foundry	Samsung Foundry	Rapidus (Japan)	SMIC (China)
Leading node (October 2026)	18A (with gate-all-around transistors and backside power) in high-volume production for Intel's Panther Lake PCs and Xeon 6+ servers; 18A-P in early ('risk') production; 14A, using High-NA EUV, next (Module 3)	SF2 (2nm-class); its first 2nm phone chip, the Exynos 2600, in the Galaxy S26; a second-generation 2nm for phones (dates in the State of the Art)	2nm pilot line at Chitose, Hokkaido (first prototypes 2025); mass production planned for fiscal 2027, on a process licensed from IBM	N+3 (5nm-class, made with older DUV lithography; Module 14)
External customers	Microsoft (18A); the US government's Secure Enclave program; Tesla for 14A, and perhaps Apple (7a)	Tesla (a $16.5B contract for its AI chips at Taylor, Texas, July 2025); 2nm design wins for data-center chips; base dies for high-bandwidth memory	Customers still being signed; the government as anchor	Huawei, Alibaba, and Chinese designers barred from TSMC
Latest results	Foundry revenue $5.8B in Q2 2026, up 31%, with a $2.1B operating loss	Foundry results improved in Q2 2026; double-digit growth targeted	Pre-revenue	Gross margin 25% in Q2 2026; 94% of capacity in use (revenue in Module 4)
Capacity	Arizona (Fab 52, 18A); Oregon; Ireland; Israel; Ohio delayed to the 2030s	Taylor, Texas; Hwaseong and Pyeongtaek, Korea	One fab in Hokkaido	Beijing, Shanghai, Shenzhen
The state's role	US government stake of about 10% (August 2025, from $8.9B of converted grants); Nvidia's $5B (September 2025) and SoftBank's $2B (2025) investments	Korean tax credits; the Samsung-SK Hynix fab plan of Module 8 (memory-led)	Government R&D; support of ¥2.35T after April 2026's ¥631.5B	China's 'Big Fund' and state procurement
Prospect	Turns on 14A (7a)	A second source at 2nm (7b)	Turns on its yields (7c)	Limited to DUV (7c)

7a. Intel

Intel ruled the industry for forty years: its x86 processors, built on the instruction set it created, ran nearly every PC and server, and its manufacturing was the best in the world through the 22nm node in 2012. It lost the lead within a decade, when its 10nm node, which it tried to make without EUV, arrived three years late, in 2019, and TSMC passed it at 7nm. Apple left for its own Arm-based chips in 2020, AMD took a large share of the server market with TSMC-made chips, and Nvidia took the data center's growth. Intel's revenue fell from $79 billion in 2021 to $53 billion in 2024, with a $19 billion loss that year, and the rescue became a national project. Pat Gelsinger's 2021 plan to make chips for other companies, called 'IDM 2.0,' cost him his

job in December 2024, and Lip-Bu Tan, the former head of Cadence, took over in March 2025, cut about a quarter of the workforce, and pushed back the Ohio fabs. The Trump administration converted grants into the stake in the table above, the first American government ownership of a major chipmaker, and Nvidia and SoftBank invested alongside it. The 18A node, the first with backside power delivery, is in high-volume production for Intel's own chips, and a subset of the Panther Lake laptop chips is made with ASML's High-NA EUV tools. The business has turned with the AI boom: revenue is growing again (Module 4), data-center chip sales rose 59%, and the stock has risen about sixfold in a year. The question that decides Intel's future is still 14A, the High-NA node (Module 3), which Tan has said Intel will build at scale only with committed outside customers. Tesla said in April 2026 that it will use 14A at Terafab, the chip complex it is planning in Austin with Intel and SpaceX. President Trump said in April and again on June 18, 2026 that Apple had agreed to work with Intel, though neither company has confirmed it. If 14A fills, Intel becomes the West's second leading-edge foundry, and if it does not, it remains a captive fab for its own chips.

7b. Samsung

Samsung is the world's largest memory maker and, after Nvidia, its largest chipmaker by revenue, but its foundry, long TSMC's only rival at the leading edge, lost money for years because its yields did not climb. Its 3nm gate-all-around process, launched first in 2022, never reached the yields customers needed, and Qualcomm, Nvidia, and Google moved their leading products to TSMC. The 2025-26 turn has three parts, beginning with 2nm: the Exynos 2600, its first 2nm phone chip, shipped in Galaxy S26 phones (State of the Art), and it reports 2nm design wins for data-center chips. The second is anchor customers: Tesla signed the contract in the table above for its AI chips at the Taylor, Texas fab, and the foundry also makes the base dies that sit under high-bandwidth memory stacks. The third is memory: Samsung began shipping HBM4 (State of the Art) and returned to first place in DRAM (Module 8), which has restored its profits and its standing with Nvidia. Samsung's structural problem remains: it competes with its foundry customers in phones and memory (Module 5), and its foundry is small beside a memory business many times its size. Its prospect is to be a real second source at 2nm for customers who want one, and after the Taiwan arithmetic of Module 5, that is every customer.

7c. Rapidus, and How Many Leading-Edge Makers

Rapidus is Japan's attempt to return to a leading edge it left in the 1990s, founded in 2022 by eight Japanese companies with the government as its anchor. It licensed IBM's 2nm process for the Hokkaido pilot line in the table above, with the government's research money behind it. Its odds are the learning curve's: a company that has never run a node is starting at the hardest one. The count that matters is how many companies will make leading-edge chips by 2028: TSMC certainly will, Samsung probably will at 2nm, Intel will if 14A finds its customers, and Rapidus will if its yields arrive, while SMIC remains about two generations behind (Module 14), limited to older DUV lithography. That is three to four companies, against about a dozen in 2000 and more than twenty in 1990, and for the leading edge, the geography is Taiwan, Korea, and, if the American and Japanese projects work, Arizona and Hokkaido.

Module 7 Key Points

Intel, Samsung, and the Foundry Race

• Intel lost a forty-year lead within a decade: its 10nm node arrived three years late, and TSMC passed it.

• Its rescue is a national project: the US government took about 10% in August 2025, alongside Nvidia and SoftBank.

• Intel's 18A is in high-volume production; its future turns on 14A in 2028, which Tesla says it will use.

• Samsung's foundry has 2nm and Tesla as an anchor customer, and HBM4 helped return Samsung to first place in DRAM.

• Rapidus aims to make 2nm chips in fiscal 2027, backed by ¥2.35 trillion of government research support.

• By 2028 three or four companies will make leading-edge chips, against about a dozen in 2000.

Module 7 Quiz

Check Your Understanding

1. What decides whether Intel becomes the West's second leading-edge foundry?

A) Whether the Ohio fabs open on time in 2026.
B) Whether the government sells its stake.
C) Whether 14A wins outside customers.
D) Whether its 18A node can match TSMC's 2nm on yield and cost within the next quarter.

2. What share of Intel does the US government hold?

A) About 10%
B) A majority stake
C) A fifth
D) No stake, only grants

3. What three parts make up Samsung's 2025-26 turn?

A) A merger with Intel's foundry
B) Exiting memory to focus on logic
C) A new fab in Japan for TSMC
D) A 2nm process with its own phone chip, anchor customers such as Tesla, and a comeback in memory with HBM4

4. What is Samsung's structural problem as a foundry?

A) It competes with its customers.
B) It has no EUV tools.
C) Its fabs are all in Texas, far from its engineers in Korea, so yields lag.
D) It serves too many small customers to earn a profit.

5. How many companies does the module expect to make leading-edge chips by 2028?

A) Only TSMC
B) Three or four companies
C) Two, TSMC and Intel
D) About a dozen, as in 2000

Module 8 — Memory: DRAM, NAND, and the Supercycle

The Commodity Chips That Became the Industry's Largest Business

Recommended Reading / Reference

Book: Chris Miller, *Chip War* (on the memory wars of the 1980s and 1990s)

Reference: TrendForce's quarterly DRAM and NAND revenue rankings and contract-price reports

8a. The Market and the Makers

	DRAM	NAND flash	High-bandwidth memory (HBM, a kind of DRAM)
What it is	Working memory: capacitors that hold a computer's running data, lost when the power goes off	Storage: cells that keep data without power, in phones and solid-state drives (SSDs)	DRAM dies stacked 8 to 16 high, wired through the silicon, and placed beside an AI chip to feed it data at terabytes a second
Sales	$154.7B in Q2 2026, up 59.5% in one quarter	$68.9B for the top five makers in Q2 2026, up 77%	About $55B in 2026, up 58% (Bank of America estimate): now a small share of DRAM sales, because ordinary DRAM prices rose faster
Makers and shares (Q2 2026)	Samsung 39.4%, SK Hynix 24.9%, Micron 23.3%; China's CXMT about 10% of world output	Samsung 29.3%, SK Hynix with Solidigm about 18%, Micron about 15%, Kioxia 13.6%, SanDisk about 11%; China's YMTC	SK Hynix 50%, Samsung 33%, Micron 18% (Counterpoint)
Prices	Server DRAM contract prices up 90-95% in Q1 2026, 58-63% in Q2, and 13-18% in Q3, each quarter on the last	Sharply higher on demand for SSDs in AI servers	HBM3E contracts for 2026 up about 20%; HBM4 reportedly priced well above HBM3E
Where made	Korea, Taiwan, Japan, the United States (Micron), China (CXMT)	Korea, Japan (Kioxia's Yokkaichi), China (YMTC's Wuhan), Singapore (Micron)	Korea, and Micron's plants in Taiwan and Japan
Newest generation	DDR5 (for PCs and servers) and LPDDR5X (low-power, for phones)	Layer counts above 300	HBM4, with HBM4E sampling (dates in the State of the Art)

TrendForce quarterly rankings (DRAM, September 7, 2026; NAND, August 18, 2026) and contract-price estimates; Counterpoint for HBM shares; CXMT's share from its IPO coverage (July 2026).

Memory is the industry's commodity: DRAM and NAND chips are made by a handful of companies and sold by the gigabyte, at prices that swing with the balance of capacity and demand. In 2024-25 memory became an AI bottleneck, and in 2026 it became the industry's largest business. The cause began with high-bandwidth memory, of which every AI accelerator needs 100-300 gigabytes, stacked beside the processor, and because HBM uses about three times the silicon wafer area per bit of ordinary DRAM, ordinary DRAM ran short as the makers turned their capacity toward it. At the same time, AI servers began to need large amounts of ordinary memory and storage too, and contract prices for server DRAM jumped in each quarter of 2026 (the table above gives TrendForce's estimates). Compounded, those quarterly rises give the cumulative move in the cycle table below. The makers' results changed accordingly (Module 4 gives their margins): Micron's revenue for the fiscal year to September 3, 2026 was $133 billion, against $37 billion

a year earlier, and Samsung guided to ₩107 trillion of third-quarter operating profit, the first time it has topped ₩100 trillion. The cost of the boom falls on everyone who buys memory (Module 16).

8b. The Cycle

Period	What happened	Cause	DRAM price move	Lesson
1985-86	Most US DRAM makers exited under Japanese competition	Japanese scale; a recession	−70%	The origin of the US-Japan chip war (Module 13)
1995-98	Korean entrants (Samsung, Hyundai, LG) crushed the Japanese makers	Korean investment through the downturn	−80%	Japan's exit; Korea's rise
2007-09	The 'DRAM war': Qimonda went bankrupt, Elpida followed (2012), and Micron bought it	Overcapacity into the financial crisis	−75%	Consolidation to three makers
2016-18	The first 'supercycle': prices doubled on smartphones and cloud servers	Tight supply after consolidation	+100%	Samsung's record profits
2019 and 2023	Two crashes: the trade-war inventory glut, then the post-pandemic collapse in PCs and phones	Oversupply; falling demand	−50% each	The three makers cut spending and survived
2024-26	The AI supercycle (8a)	AI accelerators; HBM's wafer intensity; years of restrained capacity spending	Server DRAM about 3.5× in nine months of 2026	Memory the industry's largest business

8c. The Bet on Capacity, and China's Memory

The makers are now building: on June 29, 2026, Samsung and SK Hynix pledged 800 trillion won for four new fabs in Korea's southwest, two each, with Samsung's in Gwangju. SK Hynix is also accelerating its M15X fab and expects the first cleanroom of its Yongin cluster in early 2027, and Micron is building in Idaho and New York. The risk is the cycle's: every past memory boom ended in a glut when the capacity built for it arrived, and the demand behind this one rests on a few companies' AI budgets. The makers argue this time is different, because they are signing multi-year agreements with their largest customers, which Micron says give it confidence in the durability of its results. China is the other variable: CXMT, China's DRAM maker, listed in Shanghai on July 27, 2026, raising about $8.6 billion in the largest IPO in Asia so far in 2026, and its shares rose 466% on the first day. It is already a major DRAM maker (market table above), but its IPO prospectus contains no high-bandwidth memory project, and one analyst firm, SemiAnalysis, estimates its HBM3 yields at about 25%. YMTC, on the US entity list (Module 13) since 2022, makes competitive NAND at Wuhan, but neither makes HBM at scale, and that gap is the memory chokepoint for Huawei's AI chips. Huawei says it has developed high-bandwidth memory of its own for the Ascend 950, but how much it can make is not public.

Module 8 Key Points

Memory: DRAM, NAND, and the Supercycle

• Memory became the industry's largest business in 2026, forecast at $804 billion, more than half of all chip sales.

• AI's high-bandwidth memory uses about three times the wafer area per bit, so ordinary DRAM ran short.

• Server DRAM contract prices rose about three and a half times between December 2025 and September 2026.

• Samsung leads DRAM with 39% of revenue; SK Hynix leads high-bandwidth memory with about half.

• Samsung and SK Hynix pledged 800 trillion won for new fabs, and every past memory boom ended in a glut.

• China's CXMT makes about a tenth of the world's DRAM but no high-bandwidth memory at scale.

Module 8 Quiz

Check Your Understanding

1. Which company led DRAM in the second quarter of 2026?

A) SK Hynix
B) Micron Technology
C) Samsung
D) China's CXMT

2. What share of the HBM market did SK Hynix hold in the second quarter of 2026?

A) Roughly a fifth
B) About half, ahead of Samsung and Micron
C) Nearly all of it
D) About a third

3. Why did ordinary DRAM run short?

A) A fire at a Korean fab that halted most of the world's DRAM output for months
B) Demand for PCs surged
C) Makers moved wafers to HBM, which needs about three times the area per bit
D) China banned exports of memory chips to the United States and its allies in 2026

4. What risk does the module see in the memory makers' new fabs?

A) The fabs will be too small to matter.
B) The government will take most of the profits through new windfall taxes.
C) The makers lack the cash to build them.
D) Every past memory boom ended in a glut when new capacity arrived.

5. What is missing from CXMT's IPO prospectus?

A) Any HBM project
B) Any DDR5 products
C) Plans for new DRAM capacity in Shanghai
D) Any mention of its shareholders and the state funds behind it

Module 9 — Equipment and Materials: The Chokepoints

The Layer Where One Company Can Stop an Industry

Recommended Reading / Reference

Book: Chris Miller, *Chip War* (on ASML and extreme-ultraviolet lithography)

Reference: SEMI's mid-year equipment forecast (July 14, 2026)

9a. The Equipment Makers

Company	Country	Revenue, latest quarter	Segment	Chokepoint
ASML	Netherlands	Latest quarter in Module 4; 2026 guided at €43-45B; China about 20% of 2026 sales	Lithography: all EUV tools and most advanced DUV	Absolute: no other company makes EUV; about 65 standard EUV tools and 4-5 High-NA tools in 2026
Applied Materials	US	Latest quarter in Module 4; China 28%	Deposition, etch, implant, polishing, inspection: the broadest range	Broad but substitutable
Lam Research	US	Latest quarter in Module 4; fiscal 2026 $23.2B; China 26%	Etch and deposition, the memory makers' supplier	Etch for 3D NAND and gate-all-around transistors
Tokyo Electron	Japan	¥732B, about $5B (April-June 2026), up 33%	Coater-developers, etch, deposition	Coater-developers: nearly every EUV layer passes through one
KLA	US	$3.7B (to June 2026)	Metrology and inspection	Yield's guardian: the tools that find the defects
The niche leaders	Japan, US, Netherlands	Advantest's market value about $190B (October 2026)	Advantest and Teradyne (test), ASM International (atomic-layer deposition), Disco (dicing), Screen (cleaning), Kokusai (furnaces)	Test is concentrated
China's equipment makers	China	Naura's market value about $66B (October 2026)	Naura, AMEC, and others (etch, deposition); SMEE (DUV lithography at older nodes)	No EUV in production; a prototype under test (below)
The market		Wafer-fab equipment a record $143.9B in 2026, up 23%; all equipment $165.9B		China the largest buyer, then Taiwan and Korea

Company reports; SEMI's mid-year forecast of July 14, 2026; market values from companiesmarketcap.com, October 8, 2026.

The machines that make chips come mostly from five companies, and one of them has no substitute. ASML, based in Veldhoven in the Netherlands, is the only maker of extreme-ultraviolet (EUV) lithography tools, without which no chip below about 7nm can be made economically. An EUV machine has about 100,000 parts and makes its light by hitting droplets of molten tin with a laser 50,000 times a second, then focuses that light with the smoothest mirrors ever made, from Zeiss, its German partner. A standard EUV tool costs more than $200 million, and the newer High-NA version, which bends the light through a wider aperture to

print finer lines, costs about $400 million, and Intel's 14A will use it, while TSMC says its nodes through A12 will not need it. ASML expects to raise its EUV capacity by about 30% in 2027, beyond the 2026 shipments in the table. Applied Materials, Lam, Tokyo Electron, and KLA make the deposition, etch, coating, and inspection tools around it, and each leads its niche, but every one of them has a rival. The Dutch government's decision in 2019, at American urging, to withhold EUV export licenses for China is the most consequential export control in the industry. China's state program has built a prototype EUV machine: Reuters reported in December 2025 that it had been assembled in Shenzhen with the help of former ASML engineers and could generate EUV light, but had not yet made working chips, and its sponsors hope for chips by 2028. Until it works, China's most advanced production (Module 14) runs on older deep-ultraviolet tools, at several times the cost.

9b. Materials

Material	What it is	Leaders	Share	Chokepoint
Silicon wafers	Single-crystal discs (Module 2)	Shin-Etsu, SUMCO (Japan), GlobalWafers (Taiwan), Siltronic (Germany), SK Siltron (Korea)	Japan about 60%	Japanese
Photoresist	The light-sensitive polymer; EUV resists are the hardest to make	JSR (bought by a Japanese state fund in 2024), Tokyo Ohka, Shin-Etsu, Fujifilm; DuPont	Japan about 90%; nearly all EUV resist	A target of Japan's restriction on Korea (below)
Photomasks	The stencils for each layer	Photronics, Toppan, DNP; the foundries' own shops	Japan and the US	EUV mask blanks (Hoya, AGC) are Japanese
Specialty gases and chemicals	Neon for lithography lasers, fluorine compounds, hydrogen fluoride, polishing slurries	Air Liquide, Linde, Air Products; Resonac, Stella Chemifa (Japan)	Diverse	Neon's 2022 shock (Ukraine made about half); hydrogen fluoride in the Japan-Korea dispute (below)
High-purity quartz	Quartz for the crucibles in which silicon crystals are grown	Sibelco (Spruce Pine, North Carolina)	US: about 80% of the world's high-purity quartz	Hurricane Helene closed Spruce Pine for weeks in September 2024
Minor metals and rare earths	Gallium, germanium, tungsten, and rare earths used in tools and packages	China (about 98% of gallium; most germanium and tungsten)	China	China's export controls, the tightest suspended under the truce (Module 13)
Substrates and packaging materials	The film and boards that chips are mounted on, including Ajinomoto build-up film (ABF)	Ajinomoto (nearly all ABF), Ibiden, Unimicron	Japan and Taiwan	The 2021 ABF shortage; glass substrates the next generation

9c. The Chokepoint Map

The chip supply chain is a set of chokepoints held in a few countries: America holds design software (Module 12), the equipment of Applied, Lam, and KLA, and the chip designs themselves, which are the basis of the export controls of Module 13. The Netherlands holds ASML's lithography, without which American controls would leak. Japan's chokepoints are the least discussed and the most numerous: it leads in the wafers and photoresist of the table above, and it makes the coater-developers that apply the resist, the mask blanks, the ABF substrate film, and a dozen specialty chemicals. Japan showed that leverage in 2019, when it restricted three chemicals to Korea, photoresist and hydrogen fluoride among them, and Korea's memory industry scrambled. Taiwan holds the fabs and the packaging, and Korea the memory, while China, in the other direction, holds the gallium, germanium, tungsten, and rare earths used in tools and packages. Its export controls are partly suspended under the US-China truce (Module 13). Washington keeps pressing on ASML's remaining sales of older DUV tools to China. A bill called the MATCH Act, which would align allied controls on chipmaking equipment with America's, cleared the House Foreign Affairs Committee in April 2026, and its backers are seeking to attach it to the 2027 defense authorization bill, though the Dutch government has been reluctant to go further. This is the chokepoint map in motion: every control on one country's link invites a control on another's.

Module 9 Key Points

Equipment and Materials: The Chokepoints

• ASML is the only maker of EUV lithography tools, without which no chip below about 7nm is economical.

• An EUV tool costs over $200 million and a High-NA tool about $400 million; ASML ships about 65 EUV tools in 2026.

• Wafer-fab equipment spending reaches a record $143.9 billion in 2026, with China the largest buyer.

• Japan holds the least discussed chokepoints: about 60% of wafers and 90% of photoresist.

• China's EUV prototype makes EUV light but no working chips yet; its sponsors hope for chips by 2028.

• Every control on one country's link in the chain invites a control on another's.

Module 9 Quiz

Check Your Understanding

1. What makes ASML the industry's absolute chokepoint?

A) It owns a controlling stake in TSMC and Samsung
B) It makes more machines than any rival
C) It alone makes EUV tools
D) It supplies nearly all of the world's photoresist and the masks that pattern it

2. Which company makes the mirrors inside ASML's EUV tools?

A) KLA
B) Tokyo Electron
C) Nikon, ASML's Japanese rival
D) Zeiss, of Germany

3. What did Reuters report about China's EUV program in December 2025?

A) A prototype built in Shenzhen with former ASML engineers can make EUV light but has not yet made working chips
B) China bought EUV tools through intermediaries
C) SMIC began volume EUV production
D) The program was abandoned

4. Which country holds the most numerous chokepoints in chip materials?

A) Korea
B) Japan, in wafers and resists
C) The Netherlands, through ASML's suppliers
D) China, through its control of gallium and every other metal used in chips

5. How large is the 2026 market for wafer-fab equipment, according to SEMI?

A) About $80 billion, flat on 2025.
B) Roughly $300 billion, double 2025's level.
C) About $144 billion, a record.
D) About $60 billion, after a slump in China's buying.

Module 10 — Advanced Packaging: The New Bottleneck

Why the Constraint Moved From the Wafer to the Package

Recommended Reading / Reference

Book: John H. Lau, *Chiplet Design and Heterogeneous Integration Packaging*

Reference: Yole Group's annual advanced-packaging market reports

Technology	What it is	Who	Capacity or status (2026)	Used for
CoWoS (chip-on-wafer-on-substrate)	Compute dies and memory stacks bonded side by side onto a silicon base, the interposer, that carries thousands of connections between them; CoWoS-L uses smaller silicon bridges to go beyond the size of one lithography field	TSMC, nearly all of it; ASE and Amkor for some steps	About 130,000 wafers a month by end-2026, with a plan to double by 2028 (Taiwanese press reports)	Every Nvidia, AMD, Broadcom, and Google accelerator; Nvidia the largest user
SoIC (system on integrated chips)	Dies stacked directly on one another and fused copper to copper ('hybrid bonding'), with no solder bumps	TSMC	Ramping	AMD's stacked-cache processors; future accelerators
Foveros and EMIB	Intel's 3D stacking and its embedded silicon bridge	Intel	Offered to outside foundry customers	Intel's own chips; a second source for CoWoS-class packaging
HBM stacking	DRAM dies stacked with through-silicon vias, the vertical wires that pass through each die	SK Hynix, Samsung, Micron	Sold out through 2026	Every AI accelerator; Module 8
Chiplets and UCIe	Several dies, for compute, input-output, and memory, in one package, linked by a standard interface called UCIe	AMD (since 2019), Intel, Apple, Nvidia (Blackwell's and Rubin's two dies)	Standard practice	CPUs and GPUs: small dies yield better than large ones
Panel-level and glass substrates	Packaging on large rectangular panels instead of round wafers; glass in place of organic boards for flatness and density	Intel, Samsung, Absolics (Georgia), and others	Pilot lines	The next density step, later in the decade
Co-packaged optics	Optical connections built into the package, replacing pluggable optical modules	Nvidia (photonic network switches, 2025-26), Broadcom, TSMC	First products shipping	Rack-to-rack links at lower power
Outsourced assembly and test (OSATs)	Conventional packaging and test for everyone	ASE (Taiwan; the largest), Amkor (US), JCET (China), Powertech	Plentiful for conventional work; scarce for advanced	Phones, PCs, cars

10a. Why the Package Became the Constraint

For fifty years packaging was an afterthought: the plastic and pins around a die, done by contract assemblers in Malaysia and the Philippines at thin margins. In the 2020s it became the binding constraint, because an AI accelerator is not one chip but an assembly. It holds two or more compute dies at the largest size a lithography tool can print, with eight or more memory stacks beside them, all bonded onto an interposer that carries connections at a density no circuit board can match, on a base the size of a hand. TSMC's CoWoS process does this for nearly every AI accelerator, and its capacity, about 35,000 wafers a month in 2024, is reported to nearly quadruple by the end of 2026 (table above). It has been short throughout, and TSMC said in April 2026 that tight supply of its advanced capacity could last into 2027. The constraint moved for the reason Module 3 gave: when the transistor stops improving quickly, performance comes from putting more silicon and more memory closer together. The technology that does that, the interposer, the bond, and the through-silicon via, has been harder to scale than lithography, because it is newer, its equipment is specialized, and only one company runs it at volume. Many of the AI chip delays and 'sold out' announcements of 2024-26 trace to packaging lines in Taiwan.

10b. Who Is Building It

The response is capacity everywhere: TSMC is expanding CoWoS on the plan in the table above, and its enlarged American commitment (Module 5) includes advanced packaging in Arizona, whose chips have had to fly to Taiwan to be packaged. Intel's Foveros and EMIB are among its most credible foundry offerings. SK Hynix is building a dedicated advanced-packaging plant for its memory stacks, and the Korean plan that came with the fab pledge of Module 8 names the Chungcheong region as its packaging hub. Amkor is building an advanced-packaging plant in Arizona, and ASE, the largest contract packager, is adding advanced lines in Taiwan and Malaysia. Spending on packaging and test equipment is rising with it: SEMI forecasts test equipment at $15.3 billion in 2026, up 31%, because AI chips need far more testing than ordinary ones. The next steps are hybrid bonding, glass substrates, and optics inside the package, and the constraint will move to whichever proves slowest. Packaging capacity should catch up with demand in 2027-28 if AI spending holds, but a second source at scale is still two to three years away, and until then the package, more than the wafer, is where the industry's next Taiwan concentration lives.

Module 10 Key Points

Advanced Packaging: The New Bottleneck

• An AI accelerator is an assembly of compute dies and memory stacks on an interposer, so packaging became the constraint.

• TSMC's CoWoS capacity is reported to reach about 130,000 wafers a month by end-2026, and to double by 2028.

• Packaging is harder to scale than lithography: the technology is newer, the tools specialized, and one company runs it at volume.

• TSMC's $265 billion US commitment adds packaging in Arizona, so American-made chips need not fly to Taiwan.

• Capacity should catch demand in 2027-28 if AI spending holds; a second source at scale is two to three years away.

Module 10 Quiz

Check Your Understanding

1. Which company runs CoWoS packaging at volume?

A) Amkor
B) TSMC
C) ASE
D) Intel Foundry

2. What CoWoS capacity is reported for the end of 2026?

A) 35,000 a month
B) About 75,000
C) 500,000 a month
D) About 130,000 a month

3. Why has packaging been harder to scale than lithography?

A) Newer technology, specialized tools, one producer
B) Tariffs on packages shipped from Taiwan to American customers
C) No demand until 2026
D) Labor shortages

4. What does an interposer do?

A) Carries dense connections between dies
B) Cools the chip
C) Stores data for the GPU between training runs
D) Replaces the transistor's gate with a wrapped nanosheet so the chip runs faster

5. When does the module expect packaging capacity to catch up with demand?

A) Already in 2025
B) Not before the mid-2030s
C) By late 2026
D) In 2027-28

Module 11 — Analog, Power, RF, and the Mature Nodes

The Chips That Are Not at the Leading Edge, and the Market China Is Flooding

Recommended Reading / Reference

Book: Chris Miller, *Chip War* (on the 2021 shortage and the auto industry)

Reference: The Office of the US Trade Representative's Section 301 determination on Chinese semiconductors (December 23, 2025)

Segment	What it does	Leaders	China's position	2026 issue
Analog and mixed-signal (sales in Module 1)	Amplifies, converts, and manages real-world signals; manages power	Texas Instruments, Analog Devices, Infineon, STMicro, NXP, Renesas, Microchip	Rising fast at the low end, with price wars	Recovery from the 2024-25 inventory correction
Power semiconductors (Module 1's discretes)	Switches and converts power: silicon transistors, and silicon carbide (SiC) and gallium nitride (GaN) for high voltages and frequencies	Infineon, onsemi, STMicro, Mitsubishi Electric, Rohm; BYD, Hua Hong, StarPower (China)	A large and growing share of silicon carbide	Wolfspeed's bankruptcy and restructuring (below); a silicon-carbide price collapse
Radio-frequency (RF) and wireless	Radio front-ends, modems, Wi-Fi, 5G	Qualcomm, Broadcom, Skyworks, Qorvo, MediaTek; Huawei's HiSilicon	Huawei designs its own	Gallium supply, which China controls
Microcontrollers	Embedded control in cars, appliances, and machines	NXP, Renesas, Microchip, STMicro, Infineon, TI; China's GigaDevice	Rising	The automotive cycle
Automotive chips	1,000-3,000 chips per car, and more in an electric vehicle	Infineon, NXP, STMicro, Renesas, TI, onsemi; Nvidia, Qualcomm, Mobileye for computing	Chinese carmakers increasingly use Chinese chips	The Nexperia crisis (below)
Mature-node foundry (28nm and above)	Makes all of the above for fabless customers	TSMC (still the largest at mature nodes), UMC, GlobalFoundries, SMIC, Hua Hong, Nexchip	A rising share of world mature-node capacity (below)	US tariffs on Chinese chips, with a further increase to come (below)

11a. The Rest of the Industry

By value, logic and memory chips, including the AI chips that make the headlines, are about four-fifths of 2026's chip sales, because AI demand and memory prices soared. By count, most chips are something else: they are made on nodes from 28 to 180 nanometers, in fabs ten to thirty years old, for every product that is not a data center or a phone. Analog chips are the amplifiers, converters, and power managers that turn the physical world's signals into digital ones and back. Made on old processes, where design know-how matters more than the node, they form a market, sized in Module 1, led by Texas Instruments and a European trio of Infineon, STMicro, and NXP. Power semiconductors switch and convert current in every electric vehicle, solar inverter, data-center power supply, and appliance, and the silicon carbide and gallium nitride versions that handle high voltages and frequencies were the decade's growth story until Chinese capacity drove prices down. Wolfspeed, the American silicon-carbide champion, filed for bankruptcy in June 2025 with about $6.5 billion of debt and emerged in late September 2025 after a restructuring that handed

much of the company to its creditors, including Renesas. Microcontrollers run every car and machine, radio chips connect them, and a car is full of these chips, which is why the 2021 shortage idled car plants worldwide (below).

11b. The 2021 Shortage

The shortage began in 2020, when carmakers cut their chip orders as the pandemic closed showrooms, just as demand for PCs, game consoles, and home electronics surged. When car sales rebounded in 2021, the foundries' mature-node capacity, planned years ahead, was already booked, and the February freeze that shut fabs in Texas and a March fire at a Renesas fab in Japan made it worse. Car plants idled around the world, and millions of cars went unbuilt for want of chips that cost a dollar or two. The responses were direct supply deals between carmakers and foundries, larger inventories, and the subsidy laws of Module 15, America's and Europe's CHIPS Acts. Then the cycle turned, and by 2023 the shortage had become a glut, which China's new mature-node capacity, the subject of the next section, made lasting. The lesson is the one the Nexperia crisis repeated in 2025: the chips that stop an economy are the cheap ones, and the response to a shortage builds the next glut.

11c. China's Mature-Node Surge

China's answer to being cut off from the leading edge was to build the trailing one. Barred from EUV and the most advanced DUV tools, Chinese foundries built capacity at 28 nanometers and above, on tools they could still buy, with SMIC, Hua Hong, Nexchip, and dozens of provincial start-ups funded by the national 'Big Fund' and local governments. By industry estimates, China's share of the world's mature-node capacity rose from about a fifth in 2020 to about a third by 2026, in the analog, power, microcontroller, display, and connectivity parts that go into cars, appliances, solar inverters, and machines. Chinese makers often sell them well below Western prices, in the same overcapacity pattern as solar panels and steel, and Western makers have felt it as a price war since 2024. The Nexperia crisis showed how tangled the trailing edge is: Nexperia, a Dutch maker of basic chips for cars, is owned by China's Wingtech. In October 2025 the Dutch government took control of it under an emergency law, China responded by blocking exports from Nexperia's Chinese plant, and carmakers faced shortages within weeks. The Netherlands suspended its order in November, but in 2026 the company remained split: its Chinese unit switched to local suppliers for wafers, and its Dutch side sought to separate from its owner. Washington's response to the flood is tariffs: Chinese chips have faced a 50% American tariff since January 2025, and in December 2025 the US Trade Representative found China's chip policies unfair and set a further tariff, at zero for now, to rise in June 2027 at a rate to be announced. Tariffs are harder to enforce on these chips than controls on the leading edge, because a $1 chip inside a Chinese-made appliance is invisible at the border. The mature-node flood is the chip war's second front, and China is winning it by volume.

Module 11 Key Points

Analog, Power, RF, and the Mature Nodes

• By value, logic and memory are about four-fifths of 2026 sales; by count, most chips are made on older nodes.

• Analog is a $95 billion market led by Texas Instruments, where design know-how matters more than the node.

• Wolfspeed, the American silicon-carbide champion, went bankrupt in June 2025 after Chinese capacity drove prices down.

• China holds about a third of world mature-node capacity, by industry estimates, and sells well below Western prices.

• The 2021 shortage and the 2025 Nexperia crisis showed that a basic chip can halt car plants within weeks.

• US tariffs on Chinese chips are 50%, with a further increase set for June 2027.

Module 11 Quiz

Check Your Understanding

1. Which American silicon-carbide maker went bankrupt in June 2025?

A) Texas Instruments
B) Microchip Technology
C) Wolfspeed, which emerged later in 2025
D) Analog Devices, after Chinese silicon-carbide prices collapsed

2. About how large is the analog chip market in 2026?

A) $9 billion
B) About $95 billion
C) $400 billion
D) Half a trillion

3. What did the Nexperia crisis of late 2025 show?

A) EUV is a chokepoint
B) China lacks mature nodes
C) That Dutch courts can block any American chip export and seize the plants that make them
D) A basic chip can halt car plants within weeks

4. Why are tariffs on Chinese mature-node chips hard to enforce?

A) They are too low to matter.
B) A $1 chip inside an appliance is invisible at the border.
C) Europe blocks the tariffs at its ports.
D) China stopped exporting such chips.

5. What did China build when it was cut off from the leading edge?

A) Mature-node capacity
B) Its own EUV production lines
C) A copy of TSMC's Arizona fab in Shenzhen, using smuggled 2nm tools and engineers
D) Leading-edge memory fabs

Module 12 — Chip Design: EDA, IP, and Custom Silicon

The Software and Architectures Chips Are Designed With

Recommended Reading / Reference

Book: John Hennessy and David Patterson, *Computer Architecture: A Quantitative Approach*

Reference: Arm's and Broadcom's quarterly results, which track the licensing and custom-chip businesses

Layer	What it is	Leaders	Economics	2025-26
Electronic design automation (EDA)	The software that designs, simulates, and verifies chips	Synopsys (which bought Ansys in 2025), Cadence, Siemens EDA; China's Empyrean	Software margins; subscriptions	The US ban on EDA sales to China, and its lifting (below)
Processor architecture and IP	Licensed instruction sets and processor designs ('IP,' intellectual property)	Arm (in nearly every phone; rising in servers and PCs); RISC-V (open); x86 (Intel and AMD only)	Royalties of a few percent per chip, plus license fees	Arm's royalties up 22% in the quarter to June 2026; its first chip of its own (below)
Interface IP	Blocks for standard connections (USB, PCIe), memory controllers, and high-speed links	Synopsys, Cadence, Rambus	Licensing	The high-speed links for AI networking
Custom silicon (ASIC design services)	Designing an application-specific integrated circuit (ASIC) to a customer's specification	Broadcom, Marvell, MediaTek, Alchip, GUC	Design fees plus a share of the chip's margin	Broadcom's AI chip growth (below); Marvell a record $2.74B quarter
The cloud companies' own chips	Accelerators and CPUs designed in-house	Google (TPU; Axion), Amazon (Trainium; Graviton), Microsoft (Maia; Cobalt), Meta (MTIA), Apple, Tesla, OpenAI	Lower cost per token; independence from Nvidia's margins	Module 6's rivals table
Chinese design	Huawei's HiSilicon (Ascend, Kirin), Cambricon, Hygon, Alibaba's T-Head, Baidu's Kunlun, Moore Threads, MetaX, Biren	Domestic customers	On SMIC's processes; dependent on foreign EDA and IP	A listing wave (Module 14)

12a. EDA and IP

No one lays out a chip by hand: electronic design automation (EDA) software turns a description of a circuit into billions of placed transistors, simulates them, checks them, and prepares the masks. Three companies dominate it, two American, Synopsys and Cadence, and one German, Siemens EDA, and EDA is the chokepoint that comes before every other, because a Chinese designer with SMIC's fabs and Huawei's engineers still needs Synopsys's or Cadence's tools. That is why the American ban on EDA sales to China in May 2025 was the control Beijing most wanted lifted, and why Washington lifted it in July 2025 in exchange for China's easing of rare-earth exports. Processor architecture is the other software layer, and an instruction set is the vocabulary of commands a processor understands. Arm's is in nearly every phone, and through Nvidia's Grace and Vera processors, Amazon's Graviton, and Apple's Macs, it is rising in servers and PCs, while x86, used only by Intel and AMD, declines with the PC. RISC-V, an open instruction set with

no license to revoke, is China's escape route and the embedded world's growing default. The layer's structural change of 2026 is Arm's: in March 2026 it launched its first chip of its own, a data-center processor designed with Meta, and it said in July that demand had passed $2 billion, which puts Arm in competition with the licensees who pay it royalties.

12b. Custom Silicon: The Customers' Own Chips

The largest structural threat to Nvidia is its customers: Google has trained its Gemini models on its own tensor processing units (TPUs) for years, and Amazon's Trainium chips power Project Rainier, one of Anthropic's largest training clusters. Microsoft and Meta run inference on their own chips, and Apple designs nearly every chip in its devices, including, since 2025, its own modem. Tesla designs its own AI chips, made at Samsung and TSMC, while OpenAI is building an accelerator with Broadcom (Module 6 lists the products). The business behind them is design services: Broadcom designs custom accelerators and the networking around them for Google and other cloud and AI companies. Its AI chip revenue was $16.7 billion in the quarter to August 2, 2026, up 221% from a year earlier, and it guided to $21.7 billion for the next quarter. Marvell and MediaTek are the other large designers, and MediaTek is co-designing Google's next inference TPU. The economics favor the customers: Nvidia's gross margin (Module 6) is a cost they can design around, and a chip built for one workload can deliver more tokens per dollar. TSMC benefits either way, because it makes nearly all of them, and the loser at the margin is the merchant chip's share, which is why Nvidia's moat rests on its systems and software as much as on its silicon.

Module 12 Key Points

Chip Design: EDA, IP, and Custom Silicon

• No chip can be designed without EDA software, and three companies, two of them American, dominate it.

• The US banned EDA sales to China in May 2025 and lifted the ban in July in exchange for rare earths.

• Arm's instruction set runs nearly every phone, and in March 2026 Arm launched its first chip of its own.

• RISC-V, an open instruction set with no license to revoke, is China's escape route.

• Broadcom's AI chip revenue was $16.7 billion in the quarter to August 2026, up 221%.

• The cloud companies' own chips are Nvidia's largest structural threat, and TSMC makes them either way.

Module 12 Quiz

Check Your Understanding

1. Why did Beijing most want the May 2025 EDA ban lifted?

A) Because EDA is cheap
B) Because China leads EDA and wanted to sell its own tools in America
C) Because China's designers still need American design software
D) Because the ban covered phones and the chips China sells abroad

2. What did Arm launch in March 2026?

A) A RISC-V design
B) A merger with Intel
C) Its own server processor
D) A new joint venture in China

3. What is RISC-V's significance for China?

A) It is open, with no license to revoke.
B) It is a Chinese invention from Tsinghua.
C) It runs only on SMIC's chips.
D) It is faster than Arm's designs in every benchmark.

4. How much AI chip revenue did Broadcom report for its quarter to August 2, 2026?

A) About $2.7 billion
B) $16.7 billion
C) About $89 billion
D) About $30 billion a quarter

5. Who benefits whether Nvidia's chips or the cloud companies' own chips win?

A) AMD
B) Samsung, which supplies the memory for both
C) Intel, which packages most of them in Arizona
D) TSMC, which manufactures nearly all of them, whoever designs them

Module 13 — Export Controls and the US-China Chip War

The Controls, the Countermeasures, and the 2026 Arrangement

Recommended Reading / Reference

Book: Chris Miller, *Chip War*

Reference: The Congressional Research Service's reports on US export controls and China's advanced semiconductors (updated through 2026)

13a. The Logic

The American theory of the chip war, stated in October 2022, was to keep China 'as far behind as possible' at the leading edge. It meant denying three things: the most advanced AI chips, so that China's labs could not train frontier models, the equipment to make advanced chips, so that SMIC could not catch up, and the American engineers and software that Chinese fabs relied on. The controls used the chokepoints of Module 9: American design software and equipment, and, through allied agreements, Dutch lithography and Japanese tools, and the 'foreign direct product rule' extended them to any chip made anywhere with American technology. The record after four years is mixed: Chinese labs trained strong models, such as DeepSeek's and Alibaba's Qwen, on stockpiled and smuggled Nvidia chips and on Huawei's, and SMIC reached 7nm and then a 5nm-class process with older DUV tools, at low yields (Module 14). China's chip self-sufficiency, the share of its chip use made at home, has risen but remains far below Beijing's goals (Module 14). By its chief executive's account, Nvidia's share of China's AI chip market fell from about 95% to almost nothing, and Huawei took much of it. Much of the controls' cost fell on the American companies they restricted, which is the argument the 2025-26 arrangements tried to answer.

Date	Action	Target	Effect
April-May 2018	An export ban on ZTE, lifted after a $1B fine; the first tariffs on Chinese tech	A Chinese telecom maker's supply chain	Proof that an American supply cutoff can stop a Chinese company in weeks
May 2019	Huawei added to the 'entity list,' the Commerce Department's register of foreign firms that need a license for any American technology	Huawei's phones and 5G gear	Google's Android services withdrawn; Huawei's phone sales collapsed by 2021
May and August 2020	The 'foreign direct product rule': any chip made with American tools or software needs a license for Huawei	TSMC's supply to Huawei	TSMC stopped; Huawei's own Kirin chips ended until 2023
October 7, 2022	The comprehensive rules: bans on advanced AI chips, on equipment for advanced logic and memory, and on Americans supporting Chinese fabs	China's AI and leading-edge manufacturing	The chip war's declaration; Nvidia's first China-only chips
2023	The Netherlands (advanced DUV) and Japan (23 tool categories) join; October 2023 rules close the workarounds; China controls gallium and germanium (July) and graphite (December)	The allies and the countermeasures	Module 9's chokepoint map activated on both sides
December 2024	Controls on high-bandwidth memory; 140 more entity listings; China bans gallium, germanium, and antimony exports to the US	Memory and equipment	The memory chokepoint on Huawei's AI chips (Module 8)

Date	Action	Target	Effect
January-May 2025	The 'AI diffusion' rule capping most countries' access to American AI chips, rescinded in May by the new administration	The world's access	Replaced by country deals, such as the Gulf's
April-August 2025	Licenses required for Nvidia's H20, the chip it designed for China ($4.5B written down); restored in July, with 15% of its China revenue to go to the US (agreed in August); EDA sales banned and restored (Module 12)	Nvidia's China business; China's chip designers	The first sale-for-a-cut arrangement
September-October 2025	China's market regulator finds Nvidia broke antitrust law (September); China's sweeping rare-earth controls (October); the October 30 truce suspends them, and an American rule on affiliates, for a year	The bargaining	Chips and rare earths as one negotiation
December 2025-January 2026	Nvidia's H200 allowed to China with 25% of each sale to the US; January rules route each chip through the US for inspection and collect the 25% through a narrow tariff under Section 232, the national-security trade law (January 14); the US-Taiwan trade deal (Module 5); a Section 301 finding on Chinese chips (Module 11)	The 2026 settlement	A managed rivalry of licenses and tariffs
2026	About ten Chinese buyers cleared for H200s, but Beijing blocks most orders (May); Nvidia's China sales under 1% of data-center revenue in its July quarter; a two-day Trump-Xi summit (May) and Xi's state visit to Washington, after which the truce was extended (below); export-control bills clear a House committee (Module 9)	The next round	No new chip curbs; deadlines pushed back

13b. The 2025-26 Arrangement

The second Trump administration replaced denial with managed sale, starting by rescinding the 'AI diffusion' rule as unworkable. The ban on Nvidia's H20 was reversed within months, on terms that made Nvidia the first company to pay the state for an export license (table above). The arrangement was later extended to the more capable H200, on the terms and inspection rules the table sets out. Then Beijing declined to buy: Chinese companies including Alibaba, Tencent, ByteDance, and JD.com were cleared to purchase H200s, but the government told them not to, to steer demand toward Huawei, and President Trump said in May that China 'chose not to' approve the purchases. Nvidia reported only a small number of H200 sales to China in its July quarter (the table gives their share). The 2026 summits produced pageantry and little change on chips: after Xi Jinping's state visit to Washington in late September, the two sides extended their truce by two months, to January 10, 2027, and deferred the larger questions. The chip war of 2026 is a managed rivalry with dated deadlines, in which each side's controls are bargaining chips and neither's are fully enforced.

13c. China's Countermeasures

China's replies run through the chokepoints it holds, beginning with materials: its controls on gallium, germanium, antimony, tungsten, graphite, and rare earths in 2023-25 were each timed to an American action, and that leverage reversed the EDA ban (Module 12). The broadest were suspended under the truce

described above, though licensing for seven medium and heavy rare earths, imposed in April 2025, was never lifted. The second is regulation: China's market regulator found against Nvidia in an antitrust review, other agencies issued 'security' guidance against the H20 and then the H200, and cybersecurity reviews have targeted Micron and Intel. The third is procurement: state guidance tells government agencies and telecom carriers to replace American chips and to buy Huawei's. The fourth is substitution: the domestic industry of Module 14, funded by the third phase of the national 'Big Fund' (Module 15) and protected by the demand the controls created. The countermeasures have worked where China's share is decisive, in materials, and where its market is the prize, in procurement, but they have not given China the machines, and that asymmetry is what the chip war still rests on.

13d. The First Chip War

The contest with China has a precedent: in the early 1980s Japan's NEC, Toshiba, Hitachi, and Fujitsu, with scale, quality, and patient capital, took most of the world market for DRAM, the working-memory chip, during an American downturn. Amid charges of dumping, Intel, Mostek, and most other American memory makers left the business (Module 8), and only Micron and Texas Instruments stayed. The 1986 US-Japan Semiconductor Agreement set price floors for Japanese memory and a target for foreign chips' share of Japan's market, and in 1987 Washington imposed tariffs when it judged Japan out of compliance and helped found Sematech, a government-industry research consortium. Japan's share peaked and then fell, but the share it lost went mostly to Korea (Module 8), while America regained its lead through design rather than memory. The lesson for the 2020s is that trade policy shifts market share between rivals more than it restores it to the country that made the policy, and the winner is often a third party.

Module 13 Key Points

Export Controls and the US-China Chip War

• The 2022 rules aimed to keep China 'as far behind as possible' by denying it chips, tools, and expertise.

• The record is mixed: SMIC reached a 5nm-class process without EUV, and Huawei took much of Nvidia's China market.

• The 2025-26 arrangement sold licenses for a cut, 15% of H20 revenue and then 25% of each H200, but Beijing blocked most purchases.

• The 1980s US-Japan chip war is the precedent: trade policy shifted memory share to Korea, not back to America.

• The US-China truce now runs to January 10, 2027, after two 2026 summits produced no new chip curbs.

• China's countermeasures work in materials and procurement but have not given it the machines.

Module 13 Quiz

Check Your Understanding

1. What was novel about the 2025 deal that restored Nvidia's H20 sales to China?

A) It transferred Nvidia's designs to a Chinese partner.
B) Nvidia paid 15% of China sales for a license.
C) It removed all tariffs on chips in both directions.
D) It banned the chip from China permanently, with no exceptions.

2. Until when does the US-China truce now run?

A) November 10, 2026, as first agreed
B) June 2027
C) January 10, 2027, after a two-month extension in September
D) The end of 2028, after the September summit

3. Why did almost no H200s reach China in 2026 despite American licenses?

A) Beijing told its companies not to buy them.
B) The 25% fee was waived.
C) Nvidia refused to ship them.
D) The chips failed China's safety tests and were sent back to the United States for repair.

4. Which restriction does the module say China never lifted, even under the truce?

A) Its October 2025 rare-earth and battery-technology controls
B) Its 2024 gallium ban
C) Every control it imposed in 2023-25, since the truce left all of them in force
D) Rare-earth licensing from April 2025

5. What did the 2026 summits produce on chips?

A) Little change, and no new curbs
B) A ban on H200s
C) Full HBM relief
D) An EUV sale

Module 14 — China's Chip Industry

How Far China Has Come, and Where It Is Stuck

Recommended Reading / Reference

Book: Chris Miller, *Chip War* (on China's long push for self-sufficiency)

Reference: TechInsights' teardowns of Huawei's chips, the most reliable public evidence of China's manufacturing level

Segment	Chinese position (October 2026)	Leaders	Gap to the frontier	Constraint
Foundry (logic)	SMIC's N+3, a 5nm-class process (14a); SMIC's results in Modules 4 and 7	SMIC, Hua Hong, Nexchip	Behind TSMC's 2nm (assessed below); lower yields on the advanced nodes	No EUV; DUV patterning repeated several times per layer; controls on servicing and spare parts
AI accelerators	Huawei's Ascend 950PR in mass production, with about 750,000 units expected this year (Reuters); Cambricon worth about $94B	Huawei, Cambricon, Hygon, Alibaba's T-Head, Baidu's Kunlun, Moore Threads, MetaX, Biren	A generation or more behind Nvidia per chip	Yields; memory; advanced packaging; SMIC's capacity
Memory	CXMT (Module 8); YMTC a major NAND maker	CXMT, YMTC	DRAM a generation behind; NAND close; HBM not at scale	Equipment controls
Equipment	Naura and AMEC competitive in etch and deposition; SMEE's DUV tools for older nodes; a state EUV prototype under test	Naura, AMEC, SMEE	Lithography a decade or more behind; metrology behind	The chokepoint that decides everything else
Design and EDA	Design world-class (Huawei's HiSilicon; the AI start-ups); domestic EDA tools good at older nodes only	HiSilicon, the AI designers; Empyrean	Design near parity; EDA behind	The EDA ban of Module 12 showed the dependence
Self-sufficiency	Roughly a quarter to a third of chips used in China made there (industry estimates), against Beijing's goal of 70% by 2025			The leading edge the gap; the mature nodes the surplus
Money	The Big Fund's third phase (Module 15); provincial funds; state procurement; a listing wave (14a)			Capital is abundant; technology is the limit

14a. How Far

China's chip industry has come further under the controls than their authors expected, and not as far as Beijing planned. SMIC, the national foundry, made Huawei's 7nm Kirin chip in 2023 with deep-ultraviolet tools and repeated patterning, for the Mate 60 phone, whose launch in August 2023, during the American commerce secretary's visit to Beijing, announced the achievement. SMIC then reached a 5nm-class process, called N+3, which TechInsights confirmed in Huawei's Kirin 9030 phone chip (State of the Art), at several times TSMC's cost and with lower yields. Its revenue grows on the captive demand the controls created (Module 4). Huawei's Ascend 950PR accelerator is now in mass production (the table gives its expected 2026 volume). It is more compatible than earlier Ascend chips with software written for Nvidia's

CUDA, and Huawei's CloudMatrix racks compete with Nvidia's per rack by using more chips and more power. Chinese labs have trained strong models (Module 13): DeepSeek's V4, previewed on April 24, 2026, was its first model optimized for Chinese chips, and it runs on them for inference, though outside experts think it was still trained mainly on Nvidia's. China's design layer is close to the frontier, and the money is abundant. Beyond the Big Fund's third phase (Module 15) and provincial funds, a wave of listings began in December 2025: Moore Threads and MetaX listed in Shanghai, Biren in Hong Kong in January 2026, and CXMT, the memory maker, in July 2026 (Module 8).

14b. Where It Is Stuck

The gap is the machines and what they make possible: without EUV, SMIC makes each critical layer of an advanced chip by exposing it several times with DUV tools, which is slower and costlier, yields less, and cannot go much below the 5nm class. China's equipment industry competes in etch and deposition, through Naura and AMEC, but trails far behind in lithography. A state program has built a prototype EUV machine, which makes EUV light but no working chips yet (Module 9). Without high-bandwidth memory at scale, China's AI chips are short of memory (Module 8). Without advanced packaging at scale, Chinese accelerators cannot match Nvidia's rack density, and the American controls on 'US persons' removed the servicing, spare parts, and process know-how that keep American tools running, so the installed base slowly degrades. The assessment, then: China is about two generations behind at the leading edge and holding the gap rather than closing it, while it wins the mature-node volume war of Module 11. It can supply its own AI industry with chips that trail Nvidia's (table above). The controls have made China more self-sufficient in everything except the things that matter most, at the price of an American company's largest export market.

Module 14 Key Points

China's Chip Industry

• SMIC's 5nm-class N+3 process, confirmed in Huawei's Kirin 9030, was made without EUV at several times TSMC's cost.

• Huawei's Ascend 950PR entered mass production in April 2026, with about 750,000 units expected this year.

• Chinese labs train strong models; DeepSeek's V4 runs on Chinese chips but was likely trained mainly on Nvidia's.

• Capital is abundant, from the Big Fund to a wave of listings; technology is the limit.

• China is about two generations behind at the leading edge, holding the gap rather than closing it.

Module 14 Quiz

Check Your Understanding

1. How did SMIC reach a 5nm-class process without EUV?

A) Smuggled EUV
B) A Samsung process license
C) Repeated DUV patterning
D) By buying TSMC's old Nanjing tools

2. How many Ascend 950PR units did Huawei expect to ship in 2026, according to Reuters?

A) Roughly 75,000 units
B) About 750,000 of the chips
C) Several million, more than Nvidia sells worldwide in a year
D) About 2.5 million

3. What hardware was DeepSeek's V4 optimized for?

A) Google's TPUs, rented in the cloud
B) Nvidia's newest Rubin GPUs, bought through the H200 licenses and shipped through Singapore
C) Phones and laptops
D) Chinese chips

4. What is the gap that decides China's position, according to the module?

A) Money for new fabs
B) Design talent at home
C) The machines, above all lithography
D) A shortage of engineers

5. What is the module's overall assessment of China's position?

A) At parity with TSMC in 2026
B) About two generations behind at the edge, and winning the mature-node war
C) Out of the leading edge for good
D) Ahead of TSMC in advanced packaging, though behind in lithography and memory

Module 15 — Industrial Policy: The CHIPS Act and the World's Subsidies

The Money Governments Are Spending, and What It Has Built

Recommended Reading / Reference

Book: Chris Miller, *Chip War* (on industrial policy from Sematech to the CHIPS Act)

Reference: The Semiconductor Industry Association and Boston Consulting Group, *Emerging Resilience in the Semiconductor Supply Chain* (2024)

Country or program	Money	Instruments	What it has built (October 2026)	Assessment
United States: the CHIPS and Science Act (2022)	$39B of manufacturing grants, $75B of loan authority, and an investment tax credit of 25%, raised to 35% by the July 2025 tax law	Grants to Intel, TSMC ($6.6B), Samsung, Micron, GlobalFoundries, Texas Instruments, Amkor	TSMC's two Arizona fabs (Module 5); Intel's 18A fab in Arizona; Samsung's Taylor fab, with Tesla as anchor; Micron building in Idaho and New York; Amkor's packaging plant	The fabs are real and mostly late
Taiwan	Tax credits; science parks; power and water	The state as landlord (15b)	The concentration itself	The model everyone imitates and no one matches
Korea	Tax credits under the K-Chips Act; the Samsung-SK Hynix pledge (Module 8)	Credits and infrastructure	Pyeongtaek, with the Yongin cluster and the new fabs of Module 8 still to come	Memory-led (15b)
Japan	Subsidies for TSMC's Kumamoto fabs, Micron's Hiroshima plant, Kioxia, and Rapidus (Module 7)	Grants of up to about half of fab costs	Kumamoto producing since 2024; Rapidus's pilot line (Module 7)	The materials chokepoints (15b)
European Union: the Chips Act (2023)	Credited by the Commission with mobilizing more than €52B of public and private investment	State-aid approvals: TSMC's Dresden fab (with Bosch, Infineon, and NXP); Intel's Magdeburg fab (cancelled July 2025)	Dresden building (Module 5); Magdeburg abandoned; a 'Chips Act 2.0' proposed on June 3, 2026	Money without a company to spend it at the leading edge
China: the Big Fund	Phase I (2014, about $21B), Phase II (2019, about $29B), Phase III (2024, $47B); provincial funds; state procurement	Equity in SMIC, YMTC, CXMT, Hua Hong, and the equipment makers; subsidies; protection	Module 14's industry; the mature-node surplus	Limited by technology (15b)
India	The India Semiconductor Mission (2021): ten approved projects worth about $19B	Grants of about half of project costs	Micron's Sanand assembly-and-test plant, producing since February 2026; three more plants due in 2026; Tata's Dholera fab under construction	Design talent the asset
Others: Singapore, Malaysia, Vietnam, Israel, the Gulf	Various	Grants and tax breaks	Packaging and mature nodes (Malaysia a major packaging center); Intel's Israeli fabs; the Gulf's AI data centers as chip buyers	The supply chain's second tier

15a. The CHIPS Act and What It Became

The CHIPS and Science Act of 2022, the largest American industrial policy in decades, offered the grants, loan authority, and tax credit in the table above, aiming to rebuild leading-edge manufacturing that had fallen to almost nothing in America. By 2026 it had built fabs and changed its form: TSMC's first Arizona fab, announced in 2020 and slowed by labor and permitting troubles, now makes chips for customers including Apple and Nvidia (Module 5 gives both Arizona fabs' status). TSMC's American commitment rose from $12 billion in 2020 to $40 billion in 2022, $65 billion in 2024, and $165 billion in 2025, before the 2026 increase of Module 5. Samsung's Taylor fab found an anchor customer in Tesla, Intel's Arizona fab runs 18A, and Micron, Amkor, and Texas Instruments are building. The form changed under the second Trump administration: Intel's grants became a government stake (Module 7). Other awards were renegotiated for larger commitments, and a tax law raised the credit, as the table shows. The administration's main lever is now the tariff: chipmakers that build in America may import a multiple of their American capacity free of the new chip tariffs, as the US-Taiwan deal of Module 5 sets out, so that where the Act used a carrot, the tariff is a stick. As for what it has achieved, the Semiconductor Industry Association and Boston Consulting Group projected in 2024 that America's share of the world's most advanced logic capacity would rise from zero in 2022 to 28% by 2032, beyond the Commerce Department's goal of 20% by 2030. Critics cite the cost per job, and defenders measure it against the Taiwan risk of Module 5.

15b. The World's Programs

Every advanced economy has a program, and the results depend on the companies available to spend the money. Taiwan's model, with the state as landlord, utility, and educator and the newest node kept at home, built the concentration everyone else is subsidizing against. Korea's is memory-led and among the largest, combining tax credits with the Samsung-SK Hynix pledge for new fabs (Module 8). Japan's money bought TSMC's Kumamoto fabs and the Rapidus bet, and its real asset is the materials chokepoints it never lost. The European Union's Chips Act has been the least successful, though the Commission credits it with mobilizing the sum in the table, much of it national money, and since Intel cancelled its Magdeburg fab, only TSMC's Dresden plant, for automotive chips, is building at scale. Europe still makes under 10% of the world's chips, and the Commission has proposed a 'Chips Act 2.0' that leans more on demand than on new fabs. China's Big Fund is the largest program in the world, and the one limited by technology rather than money (Module 14), while India is entering through assembly, test, and mature nodes, and its first big plant, Micron's in Sanand, is now in commercial production. The lesson across the programs is the industry's structure: subsidies move fabs to where a company willing to run them already is, and they cannot create the company.

Module 15 Key Points

Industrial Policy: The CHIPS Act and the World's Subsidies

• The CHIPS Act offered $39 billion of grants, $75 billion of loan authority, and a tax credit later raised to 35%.

• Under the second Trump administration, Intel's grants became a 10% government stake, and tariffs became the main lever.

• In 2024 the SIA and BCG projected the US share of advanced logic capacity at 28% by 2032.

• Korea's 800 trillion won pledge is memory-led; Japan's real asset is its materials chokepoints.

• Europe still makes under 10% of the world's chips, and Intel cancelled its Magdeburg fab in July 2025.

• Subsidies move fabs to where a company willing to run them already is; they cannot create the company.

Module 15 Quiz

Check Your Understanding

1. What did $8.9 billion of Intel's grants become in August 2025?

A) A low-interest loan
B) A tax credit for 2026
C) A tariff exemption
D) A 10% stake

2. What did the SIA and BCG project in 2024 for America's share of advanced logic capacity?

A) 28% by 2032
B) 5%
C) About 50% by 2030
D) Nearly all of it by 2030, as TSMC moves its leading edge

3. Why does the module call the EU's Chips Act the least successful program?

A) It was too big for Europe's budgets to sustain
B) It banned Chinese chips across Europe and set off a trade war that closed the Dresden plant
C) It built too many fabs at once in Italy
D) Money without a company to spend it

4. What is the lesson the module draws across the world's subsidy programs?

A) Only China's program works, because its state owns the fabs.
B) Tariffs fail everywhere they are tried, so grants are the only tool that works.
C) Subsidies move fabs to where a willing company already is, and they cannot create the company.
D) Money alone builds a chip industry, given enough time.

5. What is the US administration's main lever for building fabs now?

A) Tariff relief for building fabs in America
B) Direct federal ownership of the new American fabs.
C) Larger CHIPS grants.
D) A ban on importing any chip that was made outside the United States or allied countries.

Module 16 — Demand: Where the Chips Go

The End Markets and the AI Capital-Spending Cycle

Recommended Reading / Reference

Book: Carlota Perez, *Technological Revolutions and Financial Capital*

Reference: The cloud companies' quarterly earnings calls, where capital-spending guidance is given (the second quarter of 2026, late July)

16a. The Markets

End market	What drives it in 2026	Evidence (dated)
Data centers and AI	The cloud companies' and AI labs' build-out (below)	Nvidia's data-center revenue (Module 6); server memory prices (Module 8)
Smartphones	Flat unit sales; rising memory content and costs	Memory costs in Samsung's phones rose 211% in the first half of 2026 while its phone prices rose about 7%, as reported; analysts expect its phone unit to post a loss in Q3
PCs	A replacement cycle for older machines, squeezed by memory prices	TrendForce expects PC and phone demand to stay weak through the third quarter as component costs push up device prices
Automotive	More chips per car, electric vehicles, and driver assistance; Chinese carmakers' own chips	Texas Instruments' growth (Module 4) was led by industrial, data-center, and automotive demand; the Nexperia shortage (Module 11)
Industrial, energy, and infrastructure	Power electronics, factory automation, solar and storage inverters, grid equipment, and the data centers' power systems	Analog's modest growth (Module 1)
Communications	Optical and copper networking for AI clusters; 5G is mature	Broadcom's and Marvell's networking growth (Module 12)

Chips go into data centers, phones, computers, cars, machines, and everything else with a plug or a battery. In 2026 the data center dominates, in two ways: directly, the cloud companies and AI labs buy the accelerators, memory, networking, and storage for AI, and indirectly, their demand for memory has raised its price for everyone. Smartphones, PCs, cars, and industrial equipment grow at single digits or not at all, and phone and PC makers now pay far more for memory than they can pass on to buyers, as Samsung's own phone business shows (table above). The AI build-out therefore accounts for nearly all of the industry's growth: through Nvidia's revenue (Module 6), the memory makers' record margins, and the packaging constraint, and all of it traces to a few lines in a few companies' capital budgets. Whether those budgets hold is the question of the next two sections.

16b. The AI Capital-Spending Cycle

Company	Capital spending 2025 (actual)	2026 (guidance after the second quarter)	Cloud business (Q2 2026)	Note
Amazon	$134.7B	About $220B, raised from about $200B, partly for higher memory costs	AWS revenue up 36.7%, at an annual rate of $169B; backlog $496B	The largest spender; Trainium plus Nvidia
Alphabet (Google)	$91.5B	$195-205B, raised from $180-190B	Google Cloud revenue $24.8B, up 82%; backlog $514B	Its own TPUs; Alphabet's debt rose from about $16B to about $100B in a year
Microsoft	$118B (calendar year)	About $175B on a new accounting basis (about $190B on the old)	Azure up 43%; commercial contracted backlog $678B, up 84% (up 25% excluding OpenAI)	The OpenAI partner
Meta	$72.2B	$130-145B	No cloud business; AI in advertising	The largest spender with no cloud revenue
Oracle	—	About $70B (fiscal year to May 2027)	Its OpenAI contract (Module 6)	The most indebted builder
Total	Over $416B (the four)	$720-745B (the four); about $800B with Oracle		Third-party forecasts put 2027 above $1 trillion

Company guidance as compiled by TMT Finance (August 18, 2026) and Platformonomics (February 25, 2026); cloud figures from the companies' second-quarter reports. Microsoft's capital-spending figures include finance leases.

The AI build-out is the largest corporate capital-spending program in history: after their second-quarter results in late July 2026, Amazon, Alphabet, Microsoft, and Meta guided to the 2026 totals in the table above, and Oracle's plan adds to them. The money goes to data centers, chips, networking, and power, and several companies raised their plans partly because memory and other components cost more. The AI labs, which rent much of this capacity, raise money at record scale: after OpenAI's round (Module 6), Anthropic raised $65 billion at a $965 billion valuation in late May and filed confidentially for a public listing on June 1. The spending is the demand for everything in this course: Nvidia's data-center sales (Module 6), TSMC's growth (Module 5), the memory supercycle, and the packaging shortage. It is funded by the cloud companies' cash flows, increasingly by their borrowing, and by the labs' equity rounds, chip-vendor investments, and Gulf money.

16c. Is the Demand Durable?

The case for durability, as its advocates make it, starts with Jensen Huang's 'compute is revenue' (Module 6). The evidence they point to is the buyers' accelerating cloud revenue and enormous contracted backlogs, both in the capital-spending table above. The labs' revenues are compounding: Anthropic's annualized revenue passed $30 billion by April 2026, from about $9 billion at the end of 2025. Inference, the compute used to run models rather than train them, grows with every user and every automated agent, and cheaper AI has so far meant more use. When DeepSeek, a Chinese lab, released a strong reasoning model trained at a reported fraction of the usual cost in January 2025, Nvidia lost 17% of its market value in a day, about $590 billion, but its revenue kept climbing in the fiscal year that followed (Module 6). And most of the spending is paid for by companies with the profits to afford it.

The case against, as its advocates make it, starts with the ratio: the four largest buyers' 2026 capital spending exceeds their combined annual cloud revenue. The labs that rent the capacity mostly lose money

or earn thin margins on it, and their valuations assume decades of growth. The financing is getting looser and more circular (Module 6), as chip sellers invest in chip buyers, Oracle borrows to build for OpenAI, and Alphabet's borrowing has soared (table above). Higher memory prices (Module 8) raise the cost of every server, power is short, and data centers wait years for grid connections. A single lab's failure to raise its next round would cascade through the cloud companies' backlogs and the debt of the GPU-rental firms.

The evidence between them cuts both ways: the cloud growth is real and accelerating, and it is the strongest point for the bulls. The backlogs are concentrated, which is the bears' strongest point: much of Microsoft's backlog growth comes from OpenAI (table above). The spending is still mostly paid from cash flow, but borrowing is rising fast, and the labs' revenues are growing faster than any software business before them, from a base still far below the spending made on their behalf. The course's reading: the build-out is real, and so is the demand behind it, but the demand is concentrated in four buyers and a few labs, so the industry's 2027-28 depends on those income statements more than on any technology. The first sign of a turn will appear in a cloud company's capital-spending guidance or a lab's failed fund-raising, not in a chip company's results. Either way the chips will be needed, and the question is only at what price and who pays.

Module 16 Key Points

Demand: Where the Chips Go

• Four cloud companies guided to $720-745 billion of 2026 capital spending, nearly double the $416 billion of 2025.

• AI drives nearly all the growth, directly in data centers and indirectly by raising memory prices for everyone.

• The bulls cite accelerating cloud revenue and huge backlogs; Google Cloud grew 82% in the second quarter.

• The bears cite spending above cloud revenue, circular financing, rising debt, and memory costs.

• The backlogs are concentrated: Microsoft's grew 84%, but only 25% excluding OpenAI.

• The course's reading: real but concentrated demand, so a turn would show first in a buyer's spending guidance.

Module 16 Quiz

Check Your Understanding

1. How much capital spending did Amazon, Alphabet, Microsoft, and Meta guide to for 2026?

A) About $450 billion
B) $720-745 billion, nearly double 2025
C) $1.5 trillion
D) Roughly $300 billion, about the same as in 2025

2. What is the bears' strongest point, according to the module's weighing of the evidence?

A) Cloud revenue is falling at all four of the largest companies.
B) The chips will be obsolete within a year.
C) The backlogs are concentrated in OpenAI's contracts.
D) Memory prices have started to fall sharply, which will cut spending.

3. Which cloud business grew fastest in the second quarter of 2026?

A) Google Cloud
B) Amazon Web Services, at 36.7%
C) Microsoft's Azure
D) Oracle Cloud, on the OpenAI contract

4. Where does the module expect the first sign of a turn in the cycle?

A) In memory spot prices.
B) In Nvidia's results.
C) In TSMC's monthly sales, which will fall a full year before the cloud companies cut their budgets.
D) In a cloud company's capital-spending guidance.

5. What did Anthropic do on June 1, 2026?

A) Cut prices
B) Filed confidentially for a public listing
C) Bought a chip designer outright
D) Signed a $300 billion Oracle contract

Module 17 — Where Things Stand, and What to Watch

The Argument, the Contradictions, and the Signals Ahead

Recommended Reading / Reference

Book: Chris Miller, *Chip War* (the epilogue)

Reference: The Semiconductor Industry Association's monthly sales releases, published in the first week of each month

17a. The Argument

Two forces shaped the chip industry, and a third now drives it. The first is the learning curve: every node rewards whoever ran the last one, so each link of the chain has settled in one or two companies in one or two places, from design software and lithography to the leading-edge fabs, the memory, and the advanced packaging (Modules 2, 4, and 9-12). The second is the state. Since 2022, Washington, Beijing, and their allies have treated the chain as strategic and have tried to move it with controls, tariffs, subsidies, and, since 2025, equity stakes and shares of revenue (Modules 7, 13, and 15). The third is AI, which made the chip the constraint on the technology every company and government wants, and which in 2026 turned memory, the most cyclical part of the business, into its largest (Modules 6, 8, and 16). The record so far is that the chain bends but does not move. Fabs relocate at a premium and a lag, and the leading edge stays where the yield engineers are (Module 5). Controls slow a rival without stopping it, and the captive market they create pays for the rival's catch-up (Module 14). Subsidies find companies rather than create them (Module 15). And the whole structure rests on the capital budgets of a handful of buyers, whose returns are not yet proven (Module 16).

17b. The Contradictions

Tension	One side	The other
Efficiency and resilience (Modules 2, 5, and 15)	The learning curve makes one company the best and cheapest maker of leading-edge chips	The same concentration puts the world's computing on one island, and governments pay a premium to undo what the economics built
Denial and dependence (Modules 13 and 14)	Controls deny China the machines and the most capable chips	They also create the captive market that funds China's own industry, and Washington now sells licenses for a share of the revenue
The boom and the glut (Modules 8 and 16)	Record memory profits are paying for a new wave of fabs	Every past memory boom ended when the capacity built for it arrived, and today's high prices raise the cost of the servers that drive demand
Customers and competitors (Modules 6 and 12)	Nvidia's largest customers buy most of its output	They are designing chips to replace it, while Nvidia invests in customers that buy its chips
Regulator and shareholder (Modules 7, 13, and 15)	Washington regulates the industry with controls, tariffs, and subsidies	It also owns a stake in Intel and takes a share of Nvidia's licensed China sales, so it profits from the trade it restricts
The leading edge and the trailing edge (Modules 11 and 14)	Policy and money concentrate on the race at 2nm	China is winning the mature nodes by volume, and the cheap chips are the ones that have stopped car plants

Each tension is argued in the modules named in its row.

17c. What to Watch

Date	Signal	Why it matters
October 14, 2026	ASML's third-quarter report (guided at €11-12 billion)	The first read on fab spending for 2027, and on how much of the toolmaker's business China still is
October 15, 2026	TSMC's third-quarter report (guidance in Module 5) and its first word on 2027	The foundry sees its customers' orders before anyone else, so its tone on 2nm and packaging leads the industry's
Late October 2026	The cloud companies' third-quarter results and first signals of 2027 capital spending (Module 16)	The demand behind every other row: a flat plan for 2027 would be the turn Module 16 describes
Late October 2026	Samsung's and SK Hynix's full results (Samsung's guidance in Module 8); Intel's (guided at $15.8-16.8 billion)	Whether memory prices kept rising into the fourth quarter, and whether Intel names an outside customer for 14A
Late November 2026	Nvidia's quarter to October (guidance in Module 6)	Rubin's ramp, and whether higher memory costs begin to show in its gross margin
Early December 2026	World Semiconductor Trade Statistics' autumn forecast, its first for 2027	Whether the forecasters expect memory's prices, and with them the industry's size, to hold
December 2026	The fiscal 2027 defense authorization bill, and whether the MATCH Act rides on it (Module 9)	Whether allied toolmakers lose their remaining sales of older machines to China
Mid-December 2026	Micron's quarter to early December (guided at $61.5 billion)	The first memory maker's results after the fourth quarter's contract prices
Mid-January 2027	TSMC's fourth-quarter report and its 2027 capital budget	TSMC builds years ahead of demand, so its budget is the industry's own forecast for 2028-29
No date	Memory contract prices and the timing of the new Korean and American fabs (Module 8)	The memory cycle turns when capacity arrives, not when demand falls
No date	SMIC's yields, Huawei's Ascend shipments, and China's EUV prototype (Modules 9 and 14)	Whether China's chain closes the gap or only holds it
No date	The Taiwan Strait	The one risk no part of the chain has hedged (Module 5)

Company guidance and the companies' usual reporting calendar; deadlines dated in earlier modules, above all the US-China truce's expiry (Module 13), are not repeated here.

17d. The Last Word

For sixty years the industry's question was technical: whether the transistor could keep shrinking. The question now is who decides where chips are made, and on what terms. Four answers will shape the next five years: whether the AI spending earns its cost, whether a second company can run the leading edge, whether packaging and memory keep pace with the processors, and whether Taiwan stays open. The physics still sets what is possible, and the learning curve still sets who is good at it, but governments now sit at every link, as regulators, customers, and owners. The likeliest course is more of the same: a few companies in a few places, more expensive, more watched, and more political, making the chips that every answer will still need.

Module 17 Key Points

Where Things Stand, and What to Watch

• The learning curve concentrated each link of the chain; since 2022 the state has tried to move it with controls, subsidies, and stakes.

• AI made the chip the constraint on the decade's key technology, and in 2026 turned memory into the industry's largest business.

• The chain bends but does not move: fabs relocate at a premium and a lag, and controls slow rivals without stopping them.

• Its contradictions include customers designing chips to replace their supplier and a government profiting from trade it restricts.

• The next tests are the third-quarter reports, the cloud companies' plans for 2027, and TSMC's capital budget in January.

• The next five years turn on AI's returns, a second leading-edge foundry, packaging and memory, and Taiwan.

Module 17 Quiz

Check Your Understanding

1. According to the module's argument, which two forces shaped the chip industry before AI came to drive it?

A) Moore's law and the price of memory
B) The learning curve, which concentrated each link, and the state, which now tries to move it
C) Consumer demand for phones and PCs
D) Tariffs and the cost of electricity for fabs

2. What does the module mean by saying the chain 'bends but does not move'?

A) The chain has split into two blocs that no longer trade chips or chipmaking tools with each other.
B) Its links have moved, fab by fab, from Asia to America and Japan since the CHIPS Act passed.
C) Fabs relocate slowly, and controls slow rivals without stopping them.
D) Prices bend with each cycle, but the number of chips sold each year has stayed about the same.

3. Which contradiction does the module see in Nvidia's position?

A) Its largest customers are designing chips to replace it
B) Its chips are cheaper than AMD's
C) It owns TSMC's fabs
D) It sells mostly to China, the country whose access to its best chips Washington restricts

4. Why does the module watch TSMC's 2027 capital budget, due in January 2027?

A) It decides how much Taiwan's government will pay toward TSMC's new fabs in Arizona and Japan.
B) It is TSMC's last report before its A16 node enters volume production later in the year.
C) It shows whether TSMC will pay a dividend.
D) TSMC builds years ahead, so its budget forecasts demand.

5. Which contradiction does the module see in Washington's role?

A) It bans all chip exports to China
B) It has no tools to influence where fabs are built, since the industry is private and global, so it relies on persuasion alone
C) It subsidizes only Asian firms
D) It profits from the trade it restricts

Answer Sheet

Answers to every module's Check Your Understanding quiz.

Module	Answers
Module 1 Quiz	1. A 2. B 3. C 4. D 5. A
Module 2 Quiz	1. C 2. B 3. D 4. B 5. D
Module 3 Quiz	1. A 2. C 3. B 4. C 5. D
Module 4 Quiz	1. A 2. C 3. B 4. A 5. D
Module 5 Quiz	1. D 2. A 3. B 4. C 5. B
Module 6 Quiz	1. D 2. A 3. C 4. D 5. B
Module 7 Quiz	1. C 2. A 3. D 4. A 5. B
Module 8 Quiz	1. C 2. B 3. C 4. D 5. A
Module 9 Quiz	1. C 2. D 3. A 4. B 5. C
Module 10 Quiz	1. B 2. D 3. A 4. A 5. D
Module 11 Quiz	1. C 2. B 3. D 4. B 5. A
Module 12 Quiz	1. C 2. C 3. A 4. B 5. D
Module 13 Quiz	1. B 2. C 3. A 4. D 5. A
Module 14 Quiz	1. C 2. B 3. D 4. C 5. B
Module 15 Quiz	1. D 2. A 3. D 4. C 5. A
Module 16 Quiz	1. B 2. C 3. A 4. D 5. B
Module 17 Quiz	1. B 2. C 3. A 4. D 5. D