The Semiconductor Revolution: A First Principles Analysis of the AI Chip Era

Aldrich K Wooden, Sr

Graduate Student
Southern New Hampshire University
aldrich.wooden@snhu.edu

October 13, 2025

Abstract

The semiconductor industry stands at an inflection point where physics, economics, and geopolitics converge to reshape global technology. This analysis deconstructs the entire value chain from atomic-scale fabrication to trillion-dollar market implications, revealing why three companies control humanity's computational future and what this means for AI development through 2030. Using first principles reasoning—starting from transistor physics, lithography wavelength limits, fabrication process complexity, and capital intensity economics—this paper demonstrates why the industry's oligopolistic trajectory is inevitable, not coincidental. The analysis covers quantum-mechanical transistor operation, extreme ultraviolet lithography physics, advanced packaging bottlenecks, high-bandwidth memory constraints, and market dynamics across foundries, equipment manufacturers, and AI accelerator producers.

1 Introduction

Modern semiconductor manufacturing represents humanity's most sophisticated manipulation of matter, operating at the intersection of quantum mechanics, plasma physics, and surface chemistry. Understanding these fundamentals reveals why the industry has consolidated into an oligopoly and why breaking into leading-edge manufacturing now requires over \$100 billion [46].

The \$500 billion semiconductor industry enables the \$5 trillion AI economy—a 10× multiplier where atomic-scale precision determines algorithm-scale possibilities. This paper provides a comprehensive first principles analysis of chip fabrication, materials, manufacturing infrastructure, supply chain architecture, and market dynamics, then reconstructs these elements to analyze market impact across the AI ecosystem over a 3-5 year period.

2 Part 1: Deconstruction Using First Principles

2.1 Quantum Mechanics Meets Industrial Scale

2.1.1 The Transistor at Quantum Scale

A modern Metal-Oxide-Semiconductor Field-Effect Transistor (MOSFET) controls electron flow through quantum-mechanical band engineering [1, 2]. When voltage exceeds the threshold (∼0.5V), electrons accumulate at the oxide-semiconductor interface via band bending, creating an inversion layer merely 1-2 nanometers thick—roughly 5 silicon atoms. This channel enables current flow between source and drain, switching trillions of times per second.

The core challenge: gate oxides now measure just 1.0-1.2nm thick. At this dimension, quantum tunneling becomes severe—electrons possess sufficient wave function overlap to tunnel directly through the dielectric. Tunneling current increases exponentially as thickness decreases according to $J \propto \exp(-2\sqrt{2m\Phi}t/\hbar)$, where t is thickness and Φ is barrier height. This is a hard physical limit.

The solution came through materials science innovation: high-κ dielectrics [4]. Since gate capacitance $C = \kappa\epsilon_0 A/t$, using hafnium oxide (HfO_2) with dielectric constant $\kappa \approx 25$ instead of silicon dioxide ($\kappa \approx 3.9$) allows physically thicker gates (reducing tunneling) while maintaining equivalent electrical thickness.

2.1.2 Short-Channel Effects

When channel length approaches the depletion width (∼10-20nm), source and drain electric fields interfere, causing drain-induced barrier lowering (DIBL) [1]. The drain voltage now modulates the source-channel barrier, reducing threshold voltage and increasing subthreshold leakage. The subthreshold swing—voltage change needed per decade of current—has a classical limit of 60 mV/decade at room temperature, set by thermal energy kT/q.

2.2 EUV Lithography: Taming 200,000°C Plasma

Lithography resolution determines what features can be printed, following $R = k_1\lambda/(n \cdot \sin\theta)$. Reducing wavelength λ or increasing numerical aperture NA improves resolution [6, 8].

2.2.1 Creating 13.5nm Photons

ASML's EUV systems fire a CO_2 laser at 30-micrometer tin droplets dispensed at 50,000 drops per second [6, 7]. A pre-pulse flattens each droplet into a pancake shape, then the main pulse (tens of kilowatts) vaporizes the tin into plasma reaching 200,000°C—40 times the sun's surface temperature. Highly ionized tin atoms (Sn IX to Sn XIV) emit characteristic 13.5nm radiation via electronic transitions between 4d and 4f orbitals.

This wavelength represents 91.8 eV photon energy—so energetic that all materials absorb it strongly. No transmissive optics can exist. The entire optical system must use mirrors [8]. Multilayer mirrors with 40-50 alternating layers of silicon and molybdenum, each 2-4nm thick, achieve only ∼70% reflectivity per bounce through Bragg reflection. With 10 mirrors in the optical path, total light throughput drops to 2-3% of source power.

Current NXE systems with NA = 0.33 achieve 13nm resolution. High-NA EXE systems (NA = 0.55, costing $380 million each) reach 8nm resolution—enabling 1.7× smaller features and 2.9× transistor density [9,10]. Each EUV machine weighs 180 tons, requires 40 freight containers to ship, takes 250 engineers six months to install, and consumes over 1 megawatt of power.

2.2.2 DUV Immersion Lithography

Argon fluoride excimer lasers produce 193nm photons with 6.4 eV energy [11]. By immersing the space between lens and wafer in ultra-pure water (refractive index $n = 1.44$ at 193nm), numerical aperture increases from 0.93 to ∼1.35, improving resolution to ∼38nm. Multiple patterning techniques then multiply this: self-aligned quadruple patterning (SAQP) achieves 19nm pitch from 76nm single exposure [12,13].

2.3 Technology Nodes: Marketing Names Masking Density Warfare

The "nm" in node names stopped representing physical dimensions years ago. What defines a node today is transistor density, measured in millions of transistors per square millimeter (MTr/mm^2) [14,15]:

- 7nm: 91 MTr/mm^2 with gate pitch 54-57nm
- 5nm: 138 MTr/mm^2 (1.8× denser), gate pitch 45-51nm
- 3nm: 292 MTr/mm^2 (2.1× denser), gate pitch 48nm for FinFET, 45nm for GAA

2.3.1 FinFET to Gate-All-Around Transition

FinFETs wrap the gate around three sides of a vertical silicon fin (5-7nm wide, 30-50nm tall), providing superior electrostatic control over planar transistors [16]. But at 3nm, fin width cannot shrink further without quantum confinement effects dominating.

Gate-All-Around (GAA) transistors use horizontally stacked nanosheets (2-4 sheets, 5-8nm thick, 5-40nm wide) with gates surrounding all four sides [16,17]. This provides better subthreshold swing (∼63-65 mV/dec vs 70+ for scaled FinFETs) and tunable performance. Samsung achieved 23% performance improvement, 45% power reduction, and 16% area reduction moving from 5nm FinFET to 3nm GAA [17].

TSMC's N3 uses FinFET, targeting N2 GAA in 2025 [15, 18]. Samsung shipped the industry's first GAA transistors in mid-2022 but struggles with yields around 50% compared to TSMC's 85%+ [19, 20]. Intel's 18A (1.8nm-class) combines RibbonFET (GAA) with PowerVia backside power delivery [5,22].

2.4 Manufacturing Processes

2.4.1 Ion Implantation

Ions accelerated through 0.5 keV to 3 MeV potentials penetrate silicon substrates, losing energy through nuclear stopping (direct collisions) and electronic stopping (interaction with electron clouds) [23,24]. The implant profile follows an approximate Gaussian distribution with projected range R_p and straggle ΔR_p determined by energy and ion mass [25]. Post-implant annealing at 1000-1100°C repairs crystal damage and activates dopants [23].

2.4.2 Reactive Ion Etching

RF power at 13.56 MHz ionizes process gases (SF_6 for silicon, CF_4 for oxide, Cl_2 for metals) [26,27]. DC bias accelerates ions perpendicular to the wafer. Chemical reactions form volatile compounds ($F\cdot + Si \rightarrow SiF_4$) while ion bombardment provides directionality [28]. Achieving high anisotropy (vertical/lateral etch ratio >10:1) requires sidewall passivation through polymer deposition [3,27].

2.4.3 Atomic Layer Deposition

ALD uses self-limiting surface reactions with alternating precursor exposures [29]. For Al_2O_3, trimethylaluminum reacts with surface hydroxyl groups until all sites are consumed, then water exposure regenerates hydroxyl groups. Each cycle deposits exactly one monolayer (∼0.1nm) [29,30]. This process provides unmatched conformality—100% step coverage even in 100:1 aspect ratio trenches.

Leading-edge chips require 1,500-2,000 individual process steps across 80+ mask layers [31,32].

2.5 Materials Architecture

2.5.1 Silicon Wafers

Monocrystalline silicon production via the Czochralski process pulls single crystals from molten silicon at 1410°C, achieving 99.999999999% purity (11 nines) [33,35]. Japan dominates wafer production: Shin-Etsu and SUMCO together control ∼60% of the global market [34].

2.5.2 Critical Materials

Hafnium enables high-κ dielectrics. Germanium improves carrier mobility. Gallium nitride (GaN) and silicon carbide (SiC) power next-generation RF and power electronics [36,37]. Advanced materials push physics boundaries with interconnects using cobalt and ruthenium instead of copper [4,38].

2.6 The Equipment Oligopoly

2.6.1 ASML's EUV Monopoly

ASML spent decades developing EUV with government support and collaboration from Intel and TSMC [10,39]. The company sources 85% of components globally—Zeiss provides multilayer mirrors (monopoly), Trumpf supplies lasers [39,40]. ASML ships 55-90 EUV systems per year at $183-200 million each ($380 million for High-NA) [10]. Lead times reach 12-18 months [41,42].

2.6.2 Deposition and Etch Equipment

Applied Materials leads in CVD, PVD, and ion implantation with ∼20% overall equipment market share [43]. Lam Research dominates etching at 45% market share overall and 80%+ for advanced nodes. Tokyo Electron holds 92% of coater/developer equipment [44]. KLA Corporation owns metrology with 56% market share [45]. Equipment lead times extend to 14 months for critical tools [41].

2.6.3 Fab Construction Costs

A leading-edge 3nm/5nm fab costs \$15-25 billion total: 70-80% equipment (\$10-20B), 20-30% facility construction (\$3-6B) [46]. Clean room requirements reach ISO Class 3 (1,000 particles/m^3 $\geq$0.5μm)—100,000× cleaner than a surgical operating room [46].

2.7 Supply Chain Geography

2.7.1 TSMC Concentration

Taiwan Semiconductor Manufacturing Company fabricates 62-70% of all foundry revenue and over 90% of the world's most advanced logic chips [47]. TSMC invests 50-60% of revenue in capex (\$29B in 2024, \$38-42B projected 2025) while maintaining 53-55% gross margins [48, 49].

2.7.2 Geographic Fragility

Taiwan hosts 46% of global semiconductor foundry capacity [34]. Research estimates \$10 trillion in economic losses from a full-scale China-Taiwan conflict [50]. Taiwan faces natural disaster risks, water scarcity, and energy dependency [51].

Diversification proceeds slowly. TSMC's Arizona fabs face construction delays (19 months in Taiwan, 38+ months in the US) [52, 53]. Intel's Ohio fabs won't produce until 2027-2028 [54]. Lead times from equipment order to chip delivery span 24-36 months minimum [41, 55, 56].

2.8 Packaging: The New Bottleneck

2.8.1 CoWoS Technology

TSMC's 2.5D packaging technology uses silicon interposers to connect logic dies with HBM memory stacks [57, 58]. TSMC Chairman Mark Liu stated in 2023: "It's not the shortage of AI chips, it's the shortage of our CoWoS capacity" [59]. NVIDIA consumed 60% of available CoWoS capacity [60].

CoWoS production grew from ~12,000 units/month (2023) to 15-20K (2024) targeting 25-30K by year-end [53, 60, 61]. TSMC invested \$3.6B in advanced packaging (2022), Samsung \$2B [5, 59, 62, 63].

3 Part 2: Reconstruction and Market Impact

3.1 The AI Accelerator Landscape

3.1.1 NVIDIA's H100 Hopper

Built on TSMC 4N (custom 5nm variant), the H100 packs 80 billion transistors across an 814mm^2 die [64]. It delivers 1,979 TFLOPS FP8 sparse compute and 990 TFLOPS FP16 with 80GB HBM3 memory at 3.35 TB/s bandwidth.

3.1.2 Blackwell B200/GB200

Using dual-die packaging on TSMC 4NP with 208 billion transistors total, Blackwell delivers 2,250-2,500 TFLOPS FP16 and unprecedented 20 PFLOPS FP4 [65,66]. Memory expands to 192GB HBM3E at 8 TB/s bandwidth [67]. NVLink 5 provides 1.8 TB/s inter-GPU bandwidth [68, 69]. Pricing reaches $60-70K for GB200 systems [70].

3.1.3 AMD MI300X

With 153 billion transistors across chiplets, MI300X delivers 163 TFLOPS FP32 and industry-leading 192GB HBM3 at 5.3 TB/s bandwidth [71, 72]. Training performance reaches ∼75% of H100 when optimized [73, 74].

3.1.4 Hyperscaler Custom Silicon

Google's TPU v6e Trillium delivers 4.7× v5e performance (∼925 TFLOPS) [75,76]. Amazon's Trainium2 achieves 667 TFLOPS BF16 per chip with 96GB HBM3e at 2.9-3.2 TB/s bandwidth [77–81].

3.1.5 Alternative Architectures

Cerebras WSE-3 uses a 46,225mm^2 wafer-scale die—57× larger than H100—containing 4 trillion transistors [82–86]. Groq's LPU uses deterministic tensor streaming architecture [87, 88]. Intel's Gaudi 3 targets price competition [89].

3.2 Memory Wall: HBM3E and the Bandwidth Crisis

AI is bandwidth-limited, not compute-limited. HBM3E delivers 1.15 TB/s per stack at 9.6 Gbps/pin across 1024-bit interfaces [90]. SK hynix dominates HBM with 52.5% market share, shipping first 8-Hi HBM3E (24GB stacks) in March 2024 and 12-Hi (36GB) in Q4 2024 [91, 92].

Samsung follows with 42.4% share but faced delays [93, 94]. Micron trails with 8-Hi sampling September 2024 [95]. HBM capacity grew from 12,000 wafers/month (2023) to 25-30,000 wpm (2024) [90, 93].

HBM4 arrives 2026 with 2+ TB/s per stack using 2048-bit interfaces [90, 96].

3.3 Networking: The Hidden Enabler

Training at 10,000+ GPUs demands perfect synchronization. NVLink 4 (H100) delivers 900 GB/s bidirectional bandwidth [67,68]. NVLink 5 (Blackwell) doubles to 1.8 TB/s [68].

NVIDIA's Quantum-X800 InfiniBand achieves 400-800 Gbps with <500ns latency [97,98]. Ethernet at 400-800 Gbps using RoCEv2 achieves 95% throughput via NVIDIA's Spectrum-X [97, 99]. Co-packaged optics arrive 2025-2026 using silicon photonics [100, 100].

3.4 Market Dynamics

3.4.1 TSMC's 2025 Record Capex

At \$38-42B, TSMC plans building 8-9 fabs plus one packaging facility [18,48,49,101]. N3 wafers cost \$18-20K each, N4/N5 \$15-17K, N7 \$10-12K [18].

3.4.2 Samsung and Intel Challenges

Samsung's market share eroding to 9.3%, with 3nm GAA yields of only 30-40% versus TSMC's 60%+ [18, 19, 21, 47]. Intel bets foundry future on 18A execution combining RibbonFET with PowerVia [18, 22, 102].

3.4.3 SMIC and China

Using DUV multi-patterning, SMIC produces 7nm chips at ∼50% yields and 50% higher costs [12].

3.5 Hyperscalers: \$315 Billion Capex

The Big Four hyperscalers deploy \$315B in 2025 capex [103, 104]. Amazon leads with \$100B+, Microsoft ∼\$80B, Google ∼\$75B [105–107].

AWS Trainium2 claims 50% lower TCO versus GPUs [77]. Custom ASICs deliver 30-50% cost savings at sufficient scale [79, 81].

3.6 Technology Roadmap

TSMC N2 launches 2H 2025 with GAA transistors, delivering 15% density increase, 10-15% performance gain, and 25-30% power reduction [18, 48, 49, 101]. A16 (1.6nm) in 2026-2027 adds backside power [18, 19, 21].

High-NA EUV enables future scaling with \$380M systems achieving NA = 0.55 and 8nm resolution [22, 102].

3.7 Bottleneck Migration

Current bottleneck (2024-2025): CoWoS and HBM capacity constraints. Next bottleneck (2026-2028): power and cooling infrastructure [103]. AI datacenters face 120kW per rack power density requiring liquid cooling.

3.8 Market Structure Evolution

TSMC extends lead toward 70% market share by 2028 in base-case scenarios [18,19,21,47]. Custom silicon reaches 30-40% by 2030 versus 15% today [77, 79].

4 Conclusion

The semiconductor industry's evolution represents humanity's mastery over matter at near-atomic scales. Yet this technical achievement creates strategic fragility: over 90% of advanced chips originate from Taiwan, requiring EUV machines only ASML produces.

The AI revolution amplifies these dependencies while driving unprecedented capital deployment of $315 billion in hyperscaler investments and $38-42 billion in TSMC manufacturing capacity.

First principles analysis reveals the industry's oligopolistic trajectory is inevitable, not coincidental. Leading-edge manufacturing exhibits natural monopoly characteristics—$20-30 billion fab costs, 100,000+ wafer/month minimum efficient scale, and decades of accumulated process knowledge. These barriers ensure only TSMC, Samsung, and possibly Intel can compete at the frontier through 2030.

The bottleneck migrates from silicon manufacturing (easing by 2027 as capacity expansions mature) to power delivery infrastructure by 2028, requiring AI datacenters to consume 11-12% of US electricity. Understanding first principles—from transistor physics to fabrication economics—illuminates why this outcome flows inevitably from fundamental constraints in physics, chemistry, economics, and geopolitics.

References

[1] Wikipedia Contributors, "MOSFET," 2025. [Online]. Available: https://en.wikipedia.org/wiki/MOSFET. [Accessed: Oct. 12, 2025].

[2] WikiChip, "MOSFET - Metal-Oxide-Semiconductor Field-Effect-Transistor," 2025. [Online]. Available: https://en.wikichip.org/wiki/mosfet. [Accessed: Oct. 12, 2025].

[3] Tantec, "The difference between Reactive ion etching & Plasma etching," 2025. [Online]. Available: https://tantec.com/what-is-the-difference-between-reactive-ion-etching-and-plasma-etching/. [Accessed: Oct. 12, 2025].

[4] Wikipedia Contributors, "High-κ dielectric," 2025. [Online]. Available: https://en.wikipedia.org/wiki/High-_dielectric. [Accessed: Oct. 12, 2025].

[5] Igor's Lab, "Intel's Fab 52 - New production facility for the 18A process in Arizona," 2025. [Online]. Available: https://www.igorslab.de/en/intels-fab-52-new-production-facility-for-the-18a-process-in-arizona/. [Accessed: Oct. 12, 2025].

[6] IEEE Spectrum, "How Tiny Star Explosions Drive Moore's Law," 2025. [Online]. Available: https://spectrum.ieee.org/euv-light-source. [Accessed: Oct. 12, 2025].

[7] CNBC, "Inside ASML, the company advanced chipmakers use for EUV lithography," 2022. [Online]. Available: https://www.cnbc.com/2022/03/23/inside-asml-the-company-advanced-chipmakers-use-for-euv-lithography.html. [Accessed: Oct. 12, 2025].

[8] Wikipedia Contributors, "Extreme ultraviolet lithography," 2025. [Online]. Available: https://en.wikipedia.org/wiki/Extreme_ultraviolet_lithography. [Accessed: Oct. 12, 2025].

[9] ASML, “5 things you should know about High NA EUV lithography,” 2024. [Online]. Available: https://www.asml.com/en/news/stories/2024/5-things-high-na-euv. [Accessed: Oct. 12, 2025].

[10] Wikipedia Contributors, “ASML Holding,” 2025. [Online]. Available: https://en.wikipedia.org/wiki/ASML_Holding. [Accessed: Oct. 12, 2025].

[11] SPIE, “193nm immersion lithography: Status and challenges,” 2025. [Online]. Available: https://www.spie.org/news/immersionlitho-intro. [Accessed: Oct. 12, 2025].

[12] Tom’s Hardware, “SMIC and Huawei could use quadruple patterning for Chinese 5nm chips: Report,” 2025. [Online]. Available: https://www.tomshardware.com/tech-industry/semiconductors/smic-and-huawei-could-use-quadruple-patterning-for-chinese-5nm-chips-report. [Accessed: Oct. 12, 2025].

[13] Wikipedia Contributors, “Multiple patterning,” 2025. [Online]. Available: https://en.wikipedia.org/wiki/Multiple_patterning. [Accessed: Oct. 12, 2025].

[14] Semiconductor Engineering, “5nm Vs. 3nm,” 2025. [Online]. Available: https://semiengineering.com/5nm-vs-3nm/. [Accessed: Oct. 12, 2025].

[15] Wikipedia Contributors, “3 nm process,” 2025. [Online]. Available: https://en.wikipedia.org/wiki/3_nm_process. [Accessed: Oct. 12, 2025].

[16] Aminext, “TSMC Process Node Deep Dive: N7 to N2, FinFET & GAA Evolution,” 2025. [Online]. Available: https://www.aminext.blog/en/post/tsmc-process-node-evolution-finfet-gaa-1. [Accessed: Oct. 12, 2025].

[17] Samsung, “Samsung Begins Chip Production Using 3nm Process Technology With GAA Architecture,” 2022. [Online]. Available: https://news.samsung.com/global/samsung-begins-chip-production-using-3nm-process-technology-with-gaa-architecture
[Accessed: Oct. 12, 2025].

[18] NextBigFuture, “Samsung Versus TSMC Versus Intel,” 2025. [Online]. Available: https://www.nextbigfuture.com/2025/07/samsung-versus-tsmc-versus-intel.html. [Accessed: Oct. 12, 2025].

[19] DIGITIMES, “Samsung squeezed: TSMC scales 3nm heights, SMIC cracks 5nm,” 2025. [Online]. Available: https://www.digitimes.com/news/a20250602PD219/3nm-samsung-samsung-foundry-smic-5nm.html. [Accessed: Oct. 12, 2025].

[20] Semiconductor Engineering, “Transistors Reach Tipping Point At 3nm,” 2025. [Online]. Available: https://semiengineering.com/transistors-reach-tipping-point-at-3nm/. [Accessed: Oct. 12, 2025].

[21] 3D InCites, “IFTLE 624: TSMC widens lead on Samsung, Leads in 2nm Chip Production,” 2025. [Online]. Available: https://www.3dincites.com/2025/04/iftle-624-tsmc-widens-lead-on-samsung-leads-in-2nm-chip-production/. [Accessed: Oct. 12, 2025].

[22] Wikipedia Contributors, “2 nm process,” 2025. [Online]. Available: https://en.wikipedia.org/wiki/2_nm_process. [Accessed: Oct. 12, 2025].

[23] Semicorex, “Ion Implant and Diffusion Process,” 2025. [Online]. Available: https://www.semicorex.com/news-show-5315.html. [Accessed: Oct. 12, 2025].

[24] ScienceDirect, “Ion Implantation - an overview,” 2025. [Online]. Available: https://www.sciencedirect.com/topics/chemistry/ion-implantation. [Accessed: Oct. 12, 2025].

[25] A. Doolittle, “ECE 6450 Ion Implantation Lecture,” Georgia Institute of Technology, 2025. [Online]. Available: https://alan.ece.gatech.edu/ECE6450/Lectures/ECE6450L5-Ion%20Implantation.pdf. [Accessed: Oct. 12, 2025].

[26] KeyLink, “What is Reactive Ion Etching (RIE)?,” 2025. [Online]. Available: https://www.keylinktech.com/plasma-surface-technology/process/plasma-etching/reactive-ion-etching/. [Accessed: Oct. 12, 2025].

[27] Halbleiter.org, “Dry etch processes,” 2025. [Online]. Available: https://www.halbleiter.org/en/dryetching/etchprocesses/. [Accessed: Oct. 12, 2025].

[28] ScienceDirect, “Reactive Ion Etch - an overview,” 2025. [Online]. Available: https://www.sciencedirect.com/topics/engineering/reactive-ion-etch. [Accessed: Oct. 12, 2025].

[29] ScienceDirect, “Atomic Layer Deposition - an overview,” 2025. [Online]. Available: https://www.sciencedirect.com/topics/engineering/atomic-layer-deposition. [Accessed: Oct. 12, 2025].

[30] IEEE Spectrum, “The High-k Solution,” 2025. [Online]. Available: https://spectrum.ieee.org/the-highk-solution. [Accessed: Oct. 12, 2025].

[31] Wikipedia Contributors, “Semiconductor device fabrication,” 2025. [Online]. Available: https://en.wikipedia.org/wiki/Semiconductor_device_fabrication. [Accessed: Oct. 12, 2025].

[32] Semiconductor Industry Association, “Chipmakers Are Ramping Up Production to Address Semiconductor Shortage,” 2025. [Online]. Available: https://www.semiconductors.org/chipmakers-are-ramping-up-production-to-address-semiconductor-shortage-heres-why-
[Accessed: Oct. 12, 2025].

[33] Shin-Etsu Chemical, “Silicon Wafers,” 2025. [Online]. Available: https://www.shinetsu.co.jp/en/products/electronics-materials/silicon-wafers/. [Accessed: Oct. 12, 2025].

[34] Center for Strategic and International Studies, “Mapping the Semiconductor Supply Chain: The Critical Role of the Indo-Pacific Region,” 2025. [Online]. Available: https://www.csis.org/analysis/mapping-semiconductor-supply-chain-critical-role-indo-pacific-region. [Accessed: Oct. 12, 2025].

[35] SUMCO, “Monocrystalline pulling process,” 2025. [Online]. Available: https://www.sumcosi.com/english/products/process/step_01.html. [Accessed: Oct. 12, 2025].

[36] Center for Strategic and International Studies, “From Mine to Microchip,” 2025. [Online]. Available: https://www.csis.org/analysis/mine-microchip. [Accessed: Oct. 12, 2025].

[37] Japan External Trade Organization, “Japan’s Sublime Semiconductor Supply Chain,” 2025. [Online]. Available: https://www.jetro.go.jp/en/invest/insights/japan-insight/japan-sublime-semiconductor.html. [Accessed: Oct. 12, 2025].

[38] Semiconductor Engineering, “Big Changes In Tiny Interconnects,” 2025. [Online]. Available: https://semiengineering.com/big-changes-in-tiny-interconnects/. [Accessed: Oct. 12, 2025].

[39] The Generalist, “ASML: A Monopoly on Magic,” 2025. [Online]. Available: https://www.generalist.com/briefing/asml. [Accessed: Oct. 12, 2025].

[40] WisdomTree, “How ASML Is Redefining Technology, One Nanometer at a Time,” 2025. [Online]. Available: https://www.wisdomtree.com/investments/blog/2025/01/16/how-asml-is-redefining-technology-one-nanometer-at-a-time. [Accessed: Oct. 12, 2025].

[41] THE ELEC, “Fab equipment lead time delayed to up to 2 years,” 2025. [Online]. Available: https://www.thelec.net/news/articleView.html?idxno=3179. [Accessed: Oct. 12, 2025].

[42] Tom’s Hardware, “ASML: Only 60% of Chipmaking Tool Orders Can Be Met This Year,” 2025. [Online]. Available: https://www.tomshardware.com/news/asml-only-60-percent-of-chipmaking-tool-orders-can-be-fulfilled. [Accessed: Oct. 12, 2025].

[43] TradingView, “ASML vs. AMAT: Which Semiconductor Equipment Leader Is a Better Buy?,” 2025. [Online]. Available: https://www.tradingview.com/news/zacks:c216de105094b:0-asml-vs-amat-which-semiconductor-equipment-leader-is-a-better-buy/. [Accessed: Oct. 12, 2025].

[44] BALD Engineering, “Tokyo Electron Delivers Record FY2025 Results Amid AI Boom,” 2025. [Online]. Available: https://www.blog.baldengineering.com/2025/05/tokyo-electron-delivers-record-fy2025.html. [Accessed: Oct. 12, 2025].

[45] SemiWiki, “KLA Blows Away Competition in the Semiconductor Metrology/Inspection Market,” 2025. [Online]. Available: https://semiwiki.com/semiconductor-services/282881-kla-blows-away-competition-in-the-semiconductor-metrology-inspection-marke
[Accessed: Oct. 12, 2025].

[46] Construction Physics, “How to Build a $20 Billion Semiconductor Fab,” 2025. [Online]. Available: https://www.construction-physics.com/p/how-to-build-a-20-billion-semiconductor. [Accessed: Oct. 12, 2025].

[47] PatentPC, “Samsung vs. TSMC vs. Intel: Who’s Winning the Foundry Market?,” 2025. [Online]. Available: https://patentpc.com/blog/samsung-vs-tsmc-vs-intel-whos-winning-the-foundry-market-latest-numbers. [Accessed: Oct. 12, 2025].

[48] Tom’s Hardware, “TSMC to spend $42 billion on expansion in 2025,” 2025. [Online]. Available: https://www.tomshardware.com/tech-industry/semiconductors/tsmc-to-spend-usd42-billion-on-expansion-in-2025-ambitious-plans-detail-nine-proc
[Accessed: Oct. 12, 2025].

[49] Notebookcheck, “TSMC plans nine 2nm fabs and record $38-42 billion capex in 2025,” 2025. [Online]. Available: https://www.notebookcheck.net/TSMC-plans-nine-2nm-fabs-and-record-38-42-billion-capex-in-2025.1018499.0.html. [Accessed: Oct. 12, 2025].

[50] Center for Strategic and International Studies, “A World of Chips Acts: The Future of U.S.-EU Semiconductor Collaboration,” 2025. [Online]. Available: https://www.csis.org/analysis/world-chips-acts-future-us-eu-semiconductor-collaboration. [Accessed: Oct. 12, 2025].

[51] Discovery Alert, “Critical Minerals Powering US Semiconductor Manufacturing: Supply Chain Challenges,” 2025. [Online]. Available: https://discoveryalert.com.au/news/critical-minerals-semiconductors-2025-coalition/. [Accessed: Oct. 12, 2025].

[52] National Institute of Standards and Technology, “Intel Corporation (Arizona),” 2025. [Online]. Available: https://www.nist.gov/chips/intel-corporation-arizona-chandler. [Accessed: Oct. 12, 2025].

[53] Supply & Demand Chain Executive, “What to Expect in the 2025 Semiconductor Supply Chain,” 2025. [Online]. Available: https://www.sdcexec.com/sourcing-procurement/manufacturing/article/22918774/a2-global-electronics-what-to-expect-in-the-2025-semiconductor-supply-chain. [Accessed: Oct. 12, 2025].

[54] Intel Newsroom, “Updates: Intel’s 10 Largest Construction Projects,” 2025. [Online]. Available: https://newsroom.intel.com/intel-foundry/updates-intel-10-largest-construction-projects. [Accessed: Oct. 12, 2025].

[55] Businesskorea, “Delay in Chip Production Equipment Delivery Puts Samsung and TSMC on Alert,” 2025. [Online]. Available: https://www.businesskorea.co.kr/news/articleView.html?idxno=90573. [Accessed: Oct. 12, 2025].

[56] Deloitte Insights, “Europe’s semiconductor chip shortage,” 2025. [Online]. Available: https://www.deloitte.com/us/en/insights/industry/technology/semiconductor-chip-shortage-supply-chain.html. [Accessed: Oct. 12, 2025].

[57] AnySilicon, “Understanding CoWoS Packaging Technology,” 2025. [Online]. Available: https://anysilicon.com/cowos-package/. [Accessed: Oct. 12, 2025].

[58] TSMC, “CoWoS,” 2025. [Online]. Available: https://3dfabric.tsmc.com/english/dedicatedFoundry/technology/cowos.htm. [Accessed: Oct. 12, 2025].

[59] SemiAnalysis, “AI Capacity Constraints - CoWoS and HBM Supply Chain,” 2023. [Online]. Available: https://semianalysis.com/2023/07/05/ai-capacity-constraints-cowos-and/. [Accessed: Oct. 12, 2025].

[60] TechNode, “TSMC’s advanced packaging capacity under strain for AI chips,” 2023. [Online]. Available: https://technode.com/2023/09/25/tsmcs-advanced-packaging-capacity-under-strain-as-nvidia-amd-and-amazon-increase-
[Accessed: Oct. 12, 2025].

[61] AnandTech, “TSMC to Expand CoWoS Capacity by 60% Yearly Through 2026,” 2025. [Online]. Available: https://www.anandtech.com/show/21405/tsmc-to-expand-cowos-capacity-by-60-every-year-through-2026. [Accessed: Oct. 12, 2025].

[62] Samsung, “Samsung Electronics Unveils Plans for 1.4nm Process Technology,” 2022. [Online]. Available: https://news.samsung.com/global/samsung-electronics-unveils-plans-for-1-4nm-process-technology-and-investment-for
[Accessed: Oct. 12, 2025].

[63] TrendForce, “CoWoS Capacity Shortage Challenges AI Chip Demand,” 2024. [Online]. Available: https://www.trendforce.com/news/2024/02/19/insights-cowos-capacity-shortage-challenges-ai-chip-demand-while-taiwanese-manufa
[Accessed: Oct. 12, 2025].

[64] Civo, “Comparing NVIDIA’s B200 and H100: What’s the difference?,” 2025. [Online]. Available: https://www.civo.com/blog/comparing-nvidia-b200-and-h100. [Accessed: Oct. 12, 2025].

[65] Wikipedia Contributors, “Blackwell (microarchitecture),” 2025. [Online]. Available: https://en.wikipedia.org/wiki/Blackwell_(microarchitecture). [Accessed: Oct. 12, 2025].

[66] NVIDIA, “The Engine Behind AI Factories - NVIDIA Blackwell Architecture,” 2025. [Online]. Available: https://www.nvidia.com/en-us/data-center/technologies/blackwell-architecture/. [Accessed: Oct. 12, 2025].

[67] CUDO Compute, “NVIDIA introduces Blackwell GPU lineup,” 2025. [Online]. Available: https://www.cudocompute.com/blog/nvidias-blackwell-architecture-breaking-down-the-b100-b200-and-gb200. [Accessed: Oct. 12, 2025].

[68] Network World, “Nvidia networking roadmap: Ethernet, InfiniBand, co-packaged optics,” 2025. [Online]. Available: https://www.networkworld.com/article/4050881/nvidia-networking-roadmap-ethernet-infiniband-co-packaged-optics-will-shape-data-
html. [Accessed: Oct. 12, 2025].

[69] P. Goll, “Nvidia’s Blackwell GPUs: B100, B200, and GB200,” 2025. [Online]. Available: https://medium.com/@paulgoll/nvidias-blackwell-gpus-b100-b200-and-gb200-2441119b6941. [Accessed: Oct. 12, 2025].

[70] AnandTech, “NVIDIA Blackwell Architecture and B200/B100 Accelerators Announced,” 2025. [Online]. Available: https://www.anandtech.com/show/21310/nvidia-blackwell-architecture-and-b200b100-accelerators-announced-going-bigger-wi
[Accessed: Oct. 12, 2025].

[71] HPCwire, “AMD’s Horsepower-packed MI300X GPU Beats Nvidia’s Upcoming H200,” 2023. [Online]. Available: https://www.hpcwire.com/2023/12/07/amds-horsepower-packed-mi300x-gpu-beats-nvidias-upcoming-h200/. [Accessed: Oct. 12, 2025].

[72] TRG Datacenters, “AMD Instinct MI300X vs. NVIDIA H100,” 2025. [Online]. Available: https://www.trgdatacenters.com/resource/mi300x-vs-h100/. [Accessed: Oct. 12, 2025].

[73] SemiAnalysis, “MI300X vs H100 vs H200 Benchmark Part 1: Training,” 2024. [Online]. Available: https://semianalysis.com/2024/12/22/mi300x-vs-h100-vs-h200-benchmark-part-1-training/. [Accessed: Oct. 12, 2025].

[74] Tom’s Hardware, “AMD MI300X performance compared with Nvidia H100,” 2025. [Online]. Available: https://www.tomshardware.com/pc-components/gpus/amd-mi300x-performance-compared-with-nvidia-h100. [Accessed: Oct. 12, 2025].

[75] Next Platform, “Lots Of Questions On Google’s ”Trillium” TPU v6, A Few Answers,” 2024. [Online]. Available: https://www.nextplatform.com/2024/06/10/lots-of-questions-on-googles-trillium-tpu-v6-a-few-answers/. [Accessed: Oct. 12, 2025].

[76] Google Cloud, “Introducing Trillium, sixth-generation TPUs,” 2025. [Online]. Available: https://cloud.google.com/blog/products/compute/introducing-trillium-6th-gen-tpus. [Accessed: Oct. 12, 2025].

[77] SemiAnalysis, “Amazon’s AI Self Sufficiency - Trainium2 Architecture & Networking,” 2024. [Online]. Available: https://semianalysis.com/2024/12/03/amazons-ai-self-sufficiency-trainium2-architecture-networking/. [Accessed: Oct. 12, 2025].

[78] AWS, “Navigating GPU Challenges: Cost Optimizing AI Workloads on AWS,” 2025. [Online]. Available: https://aws.amazon.com/blogs/aws-cloud-financial-management/navigating-gpu-challenges-cost-optimizing-ai-workloads-on-aws/. [Accessed: Oct. 12, 2025].

[79] CloudOptimo, “Amazon’s Custom ML Accelerators: AWS Trainium and Inferentia,” 2025. [Online]. Available: https://www.cloudoptimo.com/blog/

amazons-custom-ml-accelerators-aws-trainium-and-inferentia/. [Accessed: Oct. 12, 2025].

[80] Amazon Web Services, “AI Accelerator - AWS Trainium,” 2025. [Online]. Available: https://aws.amazon.com/ai/machine-learning/trainium/. [Accessed: Oct. 12, 2025].

[81] SemiAnalysis, “Amazon’s AI Resurgence: AWS & Anthropic’s Multi-Gigawatt Trainium Expansion,” 2025. [Online]. Available: https://semianalysis.com/2025/09/03/amazons-ai-resurgence-aws-anthropics-multi-gigawatt-trainium-expansion/. [Accessed: Oct. 12, 2025].

[82] Cerebras, “Cerebras Systems Unveils World’s Fastest AI Chip with Whopping 4 Trillion Transistors,” 2025. [Online]. Available: https://www.cerebras.ai/press-release/cerebras-announces-third-generation-wafer-scale-engine. [Accessed: Oct. 12, 2025].

[83] Cerebras, “Exa-scale performance, single device simplicity Cerebras Wafer-Scale Cluster,” 2025. [Online]. Available: https://8968533.fs1.hubspotusercontent-na1.net/hubfs/8968533/CerebrasWaferScaleClusterdatasheet-final.pdf. [Accessed: Oct. 12, 2025].

[84] ServeTheHome, “Cerebras WSE-3 AI Chip Launched 56x Larger than NVIDIA H100,” 2025. [Online]. Available: https://www.servethehome.com/cerebras-wse-3-ai-chip-launched-56x-larger-than-nvidia-h100-vertiv-supermicro-hp
[Accessed: Oct. 12, 2025].

[85] TweakTown, “Cerebras WSE-3 wafer-scale AI chip: 57x bigger than largest GPU,” 2025. [Online]. Available: https://www.tweaktown.com/news/96843/cerebras-wse-3-wafer-scale-ai-chip-57x-bigger-than-largest-gpu-with-4-trillion-t
index.html. [Accessed: Oct. 12, 2025].

[86] GPUnet, “Understanding Wafer Scale Processors - Cerebras CS-3,” 2025. [Online]. Available: https://medium.com/@GPUnet/understanding-wafer-scale-processors-cerebras-cs-3-c040f3d599eb. [Accessed: Oct. 12, 2025].

[87] A. Upadhyay, “The Architecture of Groq’s LPU,” 2025. [Online]. Available: https://blog.codingconfessions.com/p/groq-lpu-design. [Accessed: Oct. 12, 2025].

[88] Voiceflow, “What’s Groq AI and Everything About LPU [2025],” 2025. [Online]. Available: https://www.voiceflow.com/blog/groq. [Accessed: Oct. 12, 2025].

[89] AMAX, “Intel Gaudi 3 vs Gaudi 2: AI Accelerator Comparison,” 2025. [Online]. Available: https://www.amax.com/the-next-step-for-intel-accelerators-a-look-at-intel-gaudi-3/. [Accessed: Oct. 12, 2025].

[90] SemiAnalysis, “Scaling the Memory Wall: The Rise and Roadmap of HBM,” 2025. [Online]. Available: https://semianalysis.com/2025/08/12/scaling-the-memory-wall-the-rise-and-roadmap-of-hbm/. [Accessed: Oct. 12, 2025].

[91] SK hynix, "SK hynix Develops World's Best Performing HBM3E," 2025. [Online]. Available: https://news.skhynix.com/sk-hynix-develops-worlds-best-performing-hbm3e/. [Accessed: Oct. 12, 2025].

[92] KED Global, "SK Hynix chief says no delay in 12-layer HBM3E supply," 2025. [Online]. Available: https://www.kedglobal.com/korean-chipmakers/newsView/ked202410230012. [Accessed: Oct. 12, 2025].

[93] TrendForce, "HBM3 Initially Exclusively Supplied by SK Hynix," 2024. [Online]. Available: https://www.trendforce.com/presscenter/news/20240313-12075.html. [Accessed: Oct. 12, 2025].

[94] TrendForce, "Samsung Reportedly Unable to Supply HBM3E to NVIDIA in 2024," 2024. [Online]. Available: https://www.trendforce.com/news/2024/12/13/news-samsung-reportedly-unable-to-supply-hbm3e-to-nvidia-in-2024-as-gap-with-sk- [Accessed: Oct. 12, 2025].

[95] Tom's Hardware, "HBM roadmaps for Micron, Samsung, and SK hynix," 2025. [Online]. Available: https://www.tomshardware.com/tech-industry/semiconductors/hbm-roadmaps-for-micron-samsung-and-sk-hynix-to-hbm4-and-beyond. [Accessed: Oct. 12, 2025].

[96] KED Global, "Samsung supplies HBM3E to AMD's new accelerators," 2025. [Online]. Available: https://www.kedglobal.com/korean-chipmakers/newsView/ked202506130004. [Accessed: Oct. 12, 2025].

[97] World Wide Technology, "Introduction to NVIDIA's AI/ML GPU networking solutions," 2025. [Online]. Available: https://www.wwt.com/article/introduction-to-nvidias-aiml-gpu-networking-solutions. [Accessed: Oct. 12, 2025].

[98] NVIDIA, "Accelerated Scientific Innovation with InfiniBand," 2025. [Online]. Available: https://www.nvidia.com/en-us/networking/products/infiniband/. [Accessed: Oct. 12, 2025].

[99] World Wide Technology, "The Battle of AI Networking: Ethernet vs InfiniBand," 2025. [Online]. Available: https://www.wwt.com/blog/the-battle-of-ai-networking-ethernet-vs-infiniband. [Accessed: Oct. 12, 2025].

[100] Next Platform, "Nvidia Weaves Silicon Photonics Into InfiniBand And Ethernet," 2025. [Online]. Available: https://www.nextplatform.com/2025/03/18/nvidia-weaves-silicon-photonics-into-infiniband-and-ethernet/. [Accessed: Oct. 12, 2025].

[101] Embedded, "TSMC's 2nm Technology Almost Ready for Mass Production," 2025. [Online]. Available: https://www.embedded.com/tsmcs-2nm-technology-almost-ready-for-mass-production/. [Accessed: Oct. 12, 2025].

[102] Tom's Hardware, "Amid Intel's deals, Intel Foundry remains notably absent," 2025. [Online]. Available: https://www.tomshardware.com/tech-industry/semiconductors/intel-foundry-services-is-mia. [Accessed: Oct. 12, 2025].

[103] Goldman Sachs, "AI to drive 165% increase in data center power demand by 2030," 2025. [Online]. Available: https://www.goldmansachs.com/insights/articles/ai-to-drive-165-increase-in-data-center-power-demand-by-2030. [Accessed: Oct. 12, 2025].

[104] Dcpulse, "The Great AI Infrastructure Race: Hyperscaler CapEx to Hit $315B by 2025," 2025. [Online]. Available: https://dcpulse.com/statistic/the-great-ai-infrastructure-race-hyperscaler-capex. [Accessed: Oct. 12, 2025].

[105] Visual Capitalist, "Charted: The Rise of AI Hyperscaler Spending," 2025. [Online]. Available: https://www.visualcapitalist.com/the-rise-of-ai-hyperscaler-spending/. [Accessed: Oct. 12, 2025].

[106] Network World, "AWS plans to outspend Microsoft and Google on AI infrastructure," 2025. [Online]. Available: https://www.networkworld.com/article/3819826/aws-plans-to-outspend-microsoft-and-google-on-on-ai-infrastructure.html. [Accessed: Oct. 12, 2025].

[107] Revolgy, "Who's winning the Q2 2025 AI cloud race: AWS, Microsoft, or Google Cloud?," 2025. [Online]. Available: https://www.revolgy.com/insights/blog/q2-2025-ai-cloud-race-aws-microsoft-google-cloud. [Accessed: Oct. 12, 2025].