TECHDIRECTORY RESEARCH

2026 Semiconductor Industry Outlook

The AI supply chain becomes the market: memory repricing, rack-scale compute, advanced packaging and trade policy now determine the cost of computing far beyond the accelerator aisle.

EVIDENCE CUT-OFF: 25 AUGUST 2026 · SINGAPORE

US$1.51T

WSTS 2026 semiconductor market forecast; 90% year-on-year growth.

~30%

Share of 2026 chip revenue Gartner attributes to AI semiconductors.

US$165.9B

SEMI 2026 manufacturing-equipment sales forecast, up 23.2%.

BOARD BRIEF

Executive judgement

> **The 2026 semiconductor boom is real, but its economics are concentrated in AI accelerators, high-bandwidth memory, advanced packaging and the tools that make them.**

WSTS now forecasts US$1.51 trillion in 2026 semiconductor revenue, 90% above 2025, with memory rising about 250% to more than US$800 billion. Gartner uses a narrower market base and arrives at US$1.32 trillion, but the mechanism is the same: memory revenue nearly triples while AI semiconductors account for roughly 30% of the total. Revenue has outrun physical volume because the scarce unit is not the generic chip; it is a qualified stack of accelerator silicon, HBM, package, interconnect, cooling and power.

The operating evidence is unusually stark. SK hynix reported a 76% operating margin in the second quarter, TSMC reported a 60.3% operating margin, and Broadcom said quarterly AI semiconductor revenue rose 143% year on year to US$10.8 billion. Those disclosures describe pricing power and concentration, not a normal recovery across automotive, analog, industrial and consumer components.

+250%
WSTS forecast growth for memory revenue in 2026.

73%
TSMC share of pure-play foundry revenue in Q1 2026, Counterpoint.

76%
SK hynix Q2 2026 operating margin, company disclosure.

200G
UALink 1.0 per-lane scale-up rate for up to 1,024 accelerators.

Decision frame. The relevant procurement unit is now the delivered workload: tokens per second at a defined latency, availability and power envelope. GPU count alone hides memory capacity, scale-up bandwidth, compiler maturity, data gravity, cluster utilization and the exit cost from a proprietary software stack.

The trillion-dollar headline is primarily a memory-price event

> **Market revenue is rising much faster than chip volume because DRAM, NAND and HBM prices are doing work that unit shipments are not.**

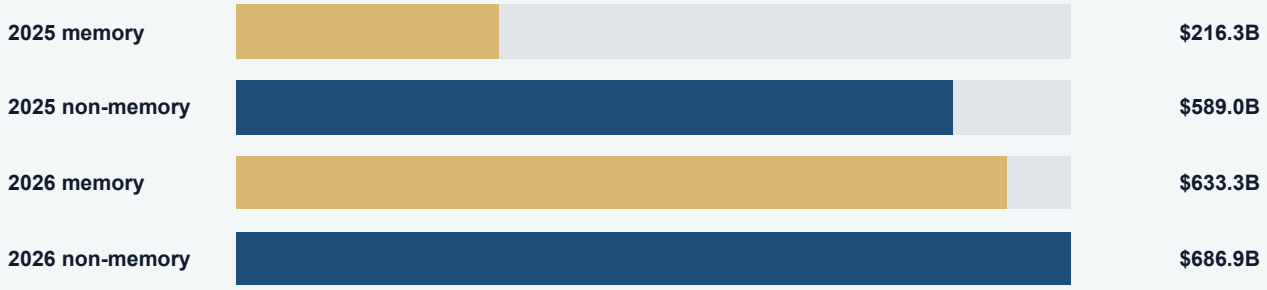
How much of the 2026 revenue surge is durable demand?

The durable part is AI infrastructure utilization and the move toward memory-rich inference; the unstable part is the price paid for each bit. Gartner forecasts DRAM annual prices up 125% and NAND flash up 234% in 2026, with no meaningful relief before late 2027. Its market table puts memory at US$633.3 billion in 2026, up from US$216.3 billion in 2025, while non-memory rises from US$589.0 billion to US$686.9 billion.

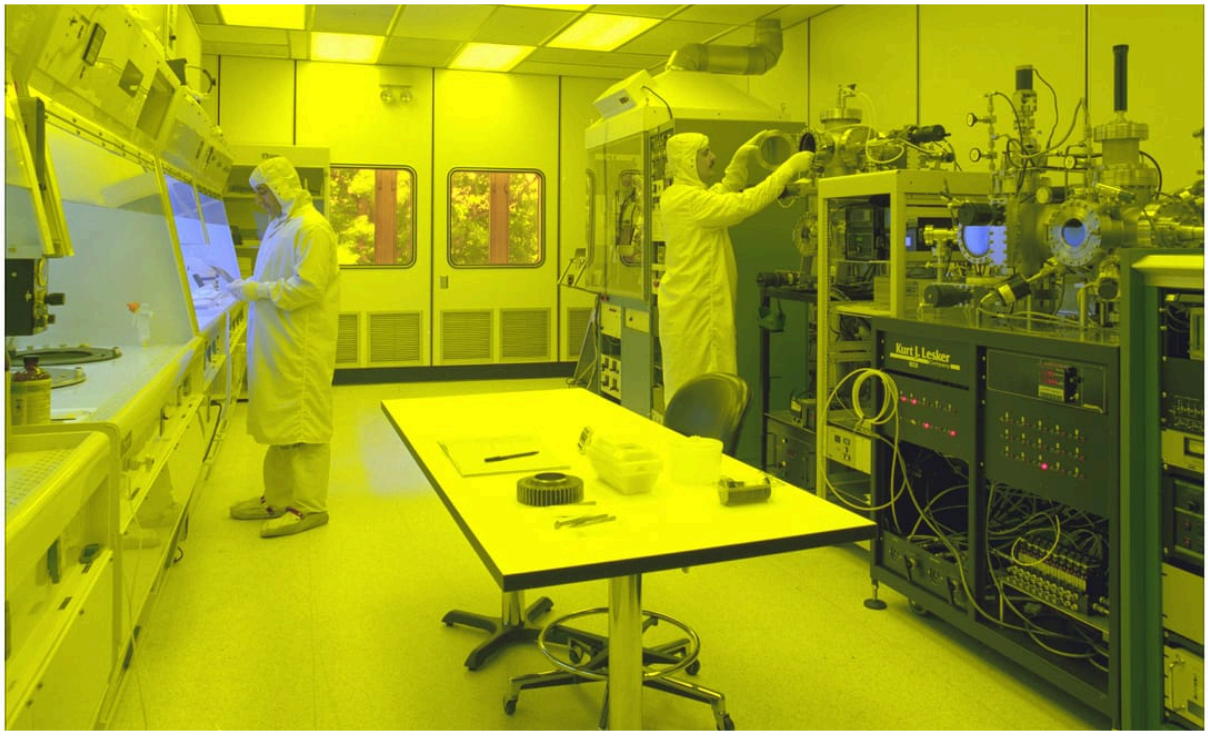
WSTS revised more aggressively after strong late-2025 and early-2026 results, putting the whole market at US$1.51 trillion and memory above US$800 billion. Deloitte’s earlier US$975 billion forecast remains useful as a timestamp, not as a current denominator. Mixing it with the later WSTS or Gartner forecast would overstate AI’s share and conceal how quickly memory pricing changed the market.

Source and release	2026 market	Growth	What the forecast says
WSTS, Spring 2026	US$1.51T	+90%	Memory +~250% to >US$800B; logic +37%; 2027 market ~US$1.9T.
Gartner, 8 Apr 2026	US$1.320T	+64%	Memory US$633.3B; non-memory US$686.9B; AI semiconductors ~30% of total.
Deloitte, late 2025 vintage	US$975B	+26%	Pre-repricing baseline; generative-AI chips approaching US$500B.

Forecasts use different release dates and market definitions. They are shown side by side, not blended.



Gartner worldwide semiconductor revenue forecast, April 2026. Bars share a US$686.9B scale.



Public-domain photograph. NASA Glenn Research Center, image GRC-1998-C-01261; retrieved via NASA Images and documented by Wikimedia Commons.

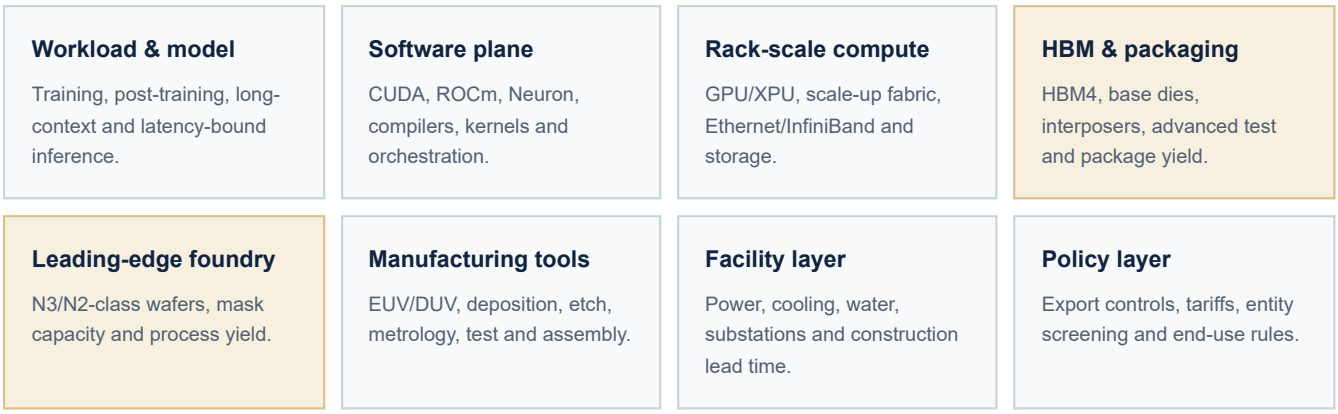
AI has moved from an end market to the organising constraint

> **AI demand now sets design priorities and capacity allocation from leading-edge logic through HBM, packaging, networking and data-centre power.**

Gartner expects AI semiconductors to represent about 30% of 2026 chip revenue and hyperscaler AI infrastructure spending to rise more than 50%. NVIDIA's Vera Rubin platform packages Rubin GPUs, Vera CPUs, NVLink 6, ConnectX-9, BlueField-4 and Spectrum-6 into a rack-scale system; AMD's Helios does the

same around MI450-series accelerators, HBM4, Pensando networking and an open rack design. Broadcom's results show that custom accelerators and AI networking are already a large merchant business, not a laboratory hedge.

The bottleneck has migrated. Compute throughput matters, but memory bandwidth, all-reduce latency, package yield, optical links, coolant distribution and megawatts decide how much of the silicon becomes usable service. A stalled collective operation leaves expensive arithmetic idle; an under-provisioned memory tier turns a nominal accelerator advantage into queue time.



Accelerator choice is now a software, network and contracting choice

> **Peak FLOPS are less decisive than achieved utilization, interconnect behavior, memory headroom and the cost of leaving the platform.**

What does the buyer actually lock in when it chooses an AI accelerator?

NVIDIA's lock-in surface is broad: CUDA libraries, NVLink scale-up, networking, reference racks, management software and a deep systems channel. AMD offers a more open rack posture through Helios and ROCm, but porting and kernel qualification remain operating work, not a checkbox. AWS Trainium moves the commitment into the cloud account, Neuron toolchain, service region and data-egress model; custom ASICs push it into the hyperscaler's own compiler and fleet economics.

Vendor metrics below are architecture disclosures, not normalized benchmarks. NVIDIA states up to 50 PFLOPS of NVFP4 inference compute per Rubin GPU and 260 TB/s of NVLink bandwidth across an NVL72 rack. AMD states 432 GB of HBM4 and 19.6 TB/s of memory bandwidth per MI450-series GPU, with 31 TB of HBM4 and 1.4 PB/s aggregate bandwidth in a 72-GPU Helios rack.

Platform	Published rack or chip data	Lock-in surface	Commercial reading
NVIDIA Vera Rubin NVL72	72 GPUs; 50 PFLOPS NVFP4 per GPU; 3.6 TB/s NVLink per GPU; 260 TB/s rack.	CUDA, NVLink, NVIDIA networking, systems software and certified supply chain.	Lowest migration friction for CUDA estates; concentration remains at silicon, fabric and software layers.
AMD Helios / MI450	72 GPUs; 432 GB HBM4 and 19.6 TB/s per GPU; 31 TB HBM4 and 1.4 PB/s aggregate rack bandwidth.	ROCm maturity, kernel coverage, Infinity Fabric and OEM execution.	Credible second-source leverage where workloads are qualified before contract, not after delivery.
AWS Trainium3 UltraServer	Up to 144 chips; up to 362 FP8 PFLOPS; AWS claims 4.4× compute and 4× energy efficiency versus Trn2 UltraServers.	Neuron SDK, AWS regions, service constructs, commitments and data gravity.	OpEx route with no owned hardware; exit economics depend on code portability and data movement.
Custom XPU / ASIC	No common product metric; Broadcom Q2 AI semiconductor revenue was US$10.8B, +143% YoY.	Internal compiler, model roadmap, volume commitment, foundry and packaging allocation.	Economical at hyperscaler utilization and volume; generally inaccessible as an enterprise diversification path.

All performance figures are vendor-stated and not directly comparable across numerical formats, software versions or workload classes.

Open interconnects are the strongest structural counterweight. UALink 200G 1.0 defines 200 GT/s per lane, up to 800 Gb/s bidirectional per four-lane station, a target request-to-response round trip below 1 microsecond, and scale-up connectivity for as many as 1,024 accelerators. CXL 3.2 addresses a different layer—coherent memory and device connectivity—with recommended transaction targets as low as 40 to 150 nanoseconds for specified cases, but real systems add switch, media and software overhead.

Generated documentary image for TechDirectory Research. The scene depicts a technically plausible direct-to-chip liquid-cooling and scale-out networking installation; it is not a photograph of a named facility.

HBM and advanced packaging hold the pricing power

> **The highest-value constraint is the qualified package that places accelerator logic beside stacked memory, not the bare logic die in isolation.**

SK hynix began mass shipments of HBM4 in the second quarter and reported Q2 revenue of 79.3187 trillion won, operating profit of 60.5426 trillion won and a 76% operating margin. It also disclosed long-term agreements with around ten key customers. Those terms convert capacity allocation into a negotiated asset and reduce the amount of supply available to buyers who arrive with short-duration orders.

Packaging is where wafer economics become system economics. HBM consumes leading DRAM capacity, through-silicon-via processing, known-good-die test and advanced assembly; accelerator packages then compete for interposers, substrate, CoWoS-class capacity and final test. A nominal second source for the GPU does not diversify the memory and packaging layers if both platforms draw from the same three HBM suppliers and the same concentrated advanced-package ecosystem.

Micron’s Singapore facility makes the regional point. The company expects the US$7 billion HBM advanced-packaging site to contribute meaningfully to supply in the first half of calendar 2027, while operations begin in 2026. It is capacity under construction, not immediate relief for a 2026 allocation problem.

Generated documentary image for TechDirectory Research. The packages and equipment are illustrative and carry no vendor identity.

Foundry concentration converts capacity into strategy

> **At the leading edge, supplier diversification is a multi-year design and qualification program rather than a procurement clause.**

Counterpoint measured TSMC at 73% of pure-play foundry revenue in Q1 2026, with Samsung at 7%. TSMC reported US$40.20 billion of Q2 revenue and a 60.3% operating margin, while its N2 process entered high-volume manufacturing in the fourth quarter of 2025. N2P and A16 volume production are scheduled for the second half of 2026, keeping smartphone and HPC roadmaps tied to the same manufacturing organization.

TSMC set a 2026 capital budget of US$52 billion to US$56 billion, up from US$40.9 billion spent in 2025. That spend is both a response to demand and a barrier to entry: new nodes require denser process integration, more expensive tools and years of customer co-qualification. A buyer can dual-source systems; it cannot create a second leading-edge foundry within a contract year.

The equipment cycle confirms where the industry is spending

> **Capital is flowing into leading-edge logic, HBM-oriented DRAM, test and advanced packaging at a pace that reinforces the AI bottleneck before it relieves it.**

SEMI’s July forecast puts 2026 semiconductor manufacturing-equipment sales at US$165.9 billion, up 23.2%. Wafer-fab equipment is expected to reach US$143.9 billion, with foundry and logic tools at US$78.0 billion and DRAM equipment at US$38.8 billion. Test equipment is forecast at US$15.3 billion and assembly and packaging equipment at US$6.7 billion.

ASML's results sit inside that spend. The company reported Q2 net sales of €9.3 billion, raised full-year guidance to €43–45 billion and said it plans to add 30% to its 2026 low-NA EUV capacity of about 65 systems for 2027; it is considering another 30% increase for 2028. Lithography output can expand, but each fab also needs deposition, etch, clean, metrology, facilities, trained staff and production yield.

Equipment segment	2026 forecast	YoY change	AI-related driver
Total manufacturing equipment	US$165.9B	+23.2%	Leading-edge logic, advanced memory, test and packaging.
Wafer-fab equipment	US$143.9B	+23.1%	N3/N2-class logic and HBM-oriented DRAM investment.
Foundry and logic WFE	US$78.0B	+18.9%	AI accelerators, HPC and premium mobile processors.
DRAM equipment	US$38.8B	+39.0%	HBM demand and advanced DRAM node migration.
Test equipment	US$15.3B	+31.0%	More complex devices and stricter AI/HBM reliability requirements.

Trade policy is part of the bill of materials

> **Export eligibility, tariff treatment and end-use documentation now influence architecture availability alongside price and performance.**

U.S. BIS controls introduced in December 2024 cover specified HBM and apply to U.S.-origin products and certain foreign-produced HBM under the advanced-computing Foreign Direct Product rule. In January 2026, BIS moved NVIDIA H200, AMD MI325X and similar exports to China to case-by-case review under stated security conditions. These are different controls with different product scopes; treating "AI chip restrictions" as one rule creates compliance risk.

The January 2026 Section 232 proclamation added a 25% tariff to a narrow category of advanced-computing chips and derivatives, with exemptions that include use in U.S. data centres, domestic R&D, startups, public-sector applications and other qualifying supply-chain uses. The policy makes end use commercially material. A landed-cost model now needs product classification, destination, customer screening, use case and re-export analysis before the quote can be trusted.

The strategic result is geographic duplication with imperfect substitution. New fabs in the United States, Japan and Europe can reduce exposure to one location, but leading-node yield, HBM supply, tool service and advanced packaging remain globally coupled. Regional capacity is not sovereign capacity when its critical inputs still cross controlled borders.

Singapore gains leverage in packaging, equipment and specialty capacity

> **Singapore's strongest AI-semiconductor position is not leading-edge logic; it is the dense middle of the chain where memory, equipment, packaging, specialty wafers and regional operations meet.**

Singapore EDB states that the country accounts for one in ten chips and one-fifth of global semiconductor-equipment production; the industry contributes close to 6% of GDP and employs more than 35,000 people. The footprint spans GlobalFoundries, Micron, Infineon, Applied Materials, KLA, ASM, Kulicke & Soffa, OSAT operations and design activity from firms including Broadcom and Marvell. That breadth is commercially valuable because equipment service, process engineering and package qualification sit close together.

Micron's HBM packaging plant is the direct AI link, but power policy limits the downstream compute that Singapore can host. IMDA's Green Data Centre Roadmap aims to provide at least 300 MW of additional capacity in the near term, with potentially another 200 MW or more through green energy, and supports facilities targeting PUE of 1.3 or lower. The chip and the megawatt therefore belong in the same regional capacity plan.

The local hedge has boundaries. Singapore does not remove dependence on Korean HBM wafer production, Taiwanese leading-edge foundry output, Dutch lithography or U.S.-controlled design software. It does provide an experienced, politically stable node for advanced packaging and equipment operations at a point where the global chain has little spare execution capacity.

Procurement should contract for delivered performance and reversible choices

> **The defensible buying strategy is to separate workload commitments from hardware identity, then price the operational and exit obligations explicitly.**

Which capacity risks can a C-suite buyer contract around—and which remain sovereign?

Buyers can negotiate allocation windows, price holds, acceptance tests, liquidated remedies, portability work, spares, power ceilings and telemetry access. They cannot contract around an export prohibition, a single-source foundry process, a grid connection that does not exist or a package technology for which no qualified alternative has reached yield. Governance improves when those categories are separated before the board approves capital.

CapEx route: owned or leased cluster.
Control over topology, security boundary and depreciation comes with facility readiness, utilization risk, spares, platform operations and residual-value exposure. Acceptance should test real models at target batch size, context length and service-level latency.

OpEx route: cloud accelerator service.
Faster access and elastic capacity trade against region availability, committed-spend terms, egress, managed-service coupling and limited visibility into physical allocation. Contract the service outcome and portability plan, not a nominal accelerator SKU.

- **Benchmark the workload.** Record tokens per second, time to first token, p95/p99 latency, model quality, batch size, context length, power and failed-job rate on production-like data.
- **Expose the dependency tree.** Name the compiler, kernels, scale-up fabric, scale-out network, HBM generation, package, storage path, scheduler, cooling system and export classification.
- **Fund portability before migration.** Maintain a tested second build path, portable checkpoints and a measured requalification interval; "supports ROCm" or "runs on Trainium" is not an exit plan.
- **Separate scarce and routine budgets.** Ring-fence AI compute, high-capacity memory and enterprise storage so memory inflation does not silently cancel ordinary infrastructure refreshes.
- **Review utilization monthly.** Idle accelerators turn a supply shortage into self-inflicted OpEx; report queue time, achieved FLOPS, memory occupancy and tokens per megawatt alongside financial utilization.

Method and source discipline

> **This report uses the supplied editorial draft as a thesis and source map, while every published statistic is rechecked against a current public source.**

Evidence was cut off on 25 August 2026. Company performance and product specifications are primary disclosures and may be preliminary or vendor-stated; architecture figures are not treated as cross-platform benchmarks. Analyst forecasts retain their original release dates and definitions, and incompatible market denominators are not blended. The two generated images are labeled; the NASA lead photograph is public domain and separately credited.

Primary and first-party sources

1. WSTS, Spring 2026 semiconductor market forecast.
2. Gartner, worldwide semiconductor revenue forecast, 8 April 2026.
3. Deloitte, 2026 Global Semiconductor Industry Outlook.
4. SEMI, 2026 mid-year total equipment forecast, 14 July 2026.
5. TSMC, second-quarter 2026 results and 2025 annual report.
6. SK hynix, second-quarter 2026 financial results.
7. ASML, second-quarter 2026 results.
8. Broadcom, fiscal Q2 2026 results.
9. NVIDIA, Vera Rubin platform disclosure.
10. AMD, Helios rack architecture disclosure.
11. AWS, Trainium3 UltraServer disclosure.
12. UALink Consortium, UALink 200G 1.0 white paper.
13. CXL Consortium, CXL 3.2 specification.
14. Counterpoint Research, pure-play foundry share, Q1 2026.
15. U.S. BIS, December 2024 semiconductor and HBM controls and January 2026 licensing policy.
16. White House, January 2026 Section 232 semiconductor proclamation.
17. Singapore EDB, semiconductor industry profile and Micron HBM facility release.
18. IMDA, Green Data Centre Roadmap.
19. Wikimedia Commons record for NASA image GRC-1998-C-01261.

The watchlist is narrower than the headline market

> **Three signals will decide whether 2026’s pricing power persists: HBM contract behavior, advanced-package throughput and paid utilization of deployed AI capacity.**

Memory price indexes will show when new supply begins to matter, but long-term HBM agreements may delay the pass-through. Package lead times and test yields will reveal whether more wafers are becoming shippable systems. Hyperscaler disclosures on custom silicon, committed workloads and infrastructure revenue will show whether capital is meeting paid demand rather than accumulating as underused capacity.

Do not use total semiconductor revenue as the only early-warning signal. By the time the headline turns, purchase orders, contract prices and tool utilization will already have moved. The useful dashboard sits one level down: HBM allocation, CoWoS-class output, AI networking revenue, cloud accelerator occupancy, power-connection dates and export-license status.

Deloitte estimated that fewer than 20 million generative-AI chips sat inside roughly 1.05 trillion semiconductor units sold in 2025—under 0.002% of units, or fewer than one in 50,000.