Bloomberg Intelligence BI

AI Accelerator Chips

2026 Outlook

Rapidly Growing Market To Top $600 Billion by 2033

AI accelerators – GPUs and ASICs – are becoming the backbone of AI infrastructure as computing demand increasingly outpaces Moore's Law of chip capability. The market looks set to log 16% compound annual growth to exceed $600 billion by 2033 on rising hyperscale capital spending, expanding inferencing infrastructure, and broader AI adoption by enterprises and national governments. Nvidia, AMD, Broadcom and Marvell are at the center of the transformation, capturing value as computing shifts from general purpose CPUs to accelerated, vertically integrated systems.

- **Enabling $3.5 Trillion of Capex:** The rapid scaling of Gen AI models is fueling hyperscale and Tier 2 cloud capex, projected at more than $3.5 trillion over the next five years, with GPUs and ASICs as the foundation. Upgrade potential is significant since 65% of global servers still aren't made for AI.
- **GPUs Evolve From Chips:** Surging AI model complexity has turned GPUs from stand-alone chips into systems combining computing, memory, networking and cooling. Clusters of about 100,000 GPUs are being deployed today and could exceed sizes of 1 million by the end of this decade, driving rapid growth in system value.
- **ASIC Market Heads to $120 Billion:** Hyperscalers are deepening their push into custom ASICs to lower costs and differentiate architectures, driving that market toward $120 billion by 2033 at a 27% CAGR, dominated by Broadcom and Marvell. Yet even as ASIC adoption accelerates, sovereign and enterprise AI deployments will sustain GPU demand.

Featured in This Report: Bloomberg Intelligence's **interactive model**, available on the terminal, calculates the AI accelerator market with a bottom-up analysis of GPU and ASIC growth by units and prices.

BI on AI

Jan. 12, 2026

Contents

Section 1.	**Executive Summary**	**2**
Section 2.	**Catalysts to Watch**	**3**
Section 3.	**Market Overview**	**4**
Section 4.	**ASIC Transition**	**16**
Section 5.	**GPU Transition**	**19**
Section 6.	**Competition**	**20**
Section 7.	**Regulations, Geopolitics**	**27**
Section 8.	**Supply Chain**	**30**
Section 9.	**Performance & Valuation**	**31**
Section 10.	**Company Impacts**	**33**
Section 11.	**Methodology**	**39**
Bloomberg Intelligence Research Coverage		**41**
Copyright and Disclaimer		**44**
About Bloomberg Intelligence		**45**

Lead Analyst		
Kunjan Sobhani	Semiconductors, Americas	ksobhani@bloomberg.net
Contributing Analysts		
Woo Jin Ho	Hardware, Networking, Americas	who88@bloomberg.net
Robert Lea	Internet, Application Software, China	rlea19@bloomberg.net
Anurag Rana	IT Services, Americas	arana4@bloomberg.net
Matthew Schettenhelm	Litigation, Americas	mschettenhel@bloomberg.net
Charles Shum	Semiconductors, APAC	cshum2@bloomberg.net
Jake Silverman	Logic ICs, Americas	jsilverma109@bloomberg.net
Mandeep Singh	Technology, Global	msingh15@bloomberg.net
Steven Tseng	Technology, APAC	htseng18@bloomberg.net
Masahiro Wakasugi	Semiconductors, Global	mwakasugi4@bloomberg.net
Oscar Hernandez Tejada	Technology, Americas	ohernandezte@bloomberg.net
Robert Biggar	Technology, Americas	rbiggar3@bloomberg.net
Andrew Girard	Technology, Americas	agirard16@bloomberg.net
Editorial & Visuals		
Tony Robinson, Rik Stevens, Justin DeVoursney, Philippe Tardieu		

More detailed analysis and interactive graphics are available on the Bloomberg Terminal

Section 1. Executive Summary

$604 Billion
Projected AI accelerator market by 2033, from $116 billion in 2024

$3.5 Trillion
Total expected capex by hyperscalers and Tier 2 cloud providers through 2030

25%
Accelerator market's projected growth over next five years

AI Accelerators Power the Future of Computing Innovation

AI accelerator chips – GPUs and ASICs – will be pivotal to the future of computing, driven by escalating demands in AI training and inferencing, with market leaders Nvidia, AMD, Broadcom and Marvell spearheading a rapidly expanding segment that's projected to grow at a 16% compound annual rate to above $600 billion by 2033. Gains will be fueled by expanding model sizes and rapid penetration of AI workloads into server markets previously dominated by central processing units (CPUs), as well as increased system-level integration, which will significantly boost average selling prices.

Key Research Topics

- **Broad Adoption Bolsters Opportunity:** Adoption of AI training and inference workloads cuts across hyperscale cloud, enterprise and sovereign entities, with aggregate capital spending set to reach over $3.5 trillion by decade's end. Expanding AI infrastructure deployment, model complexity and performance demands present a vast market opportunity, as 65% of global servers are non-AI.
- **GPUs Become Integrated Systems:** GPUs are making a transition from discrete accelerator chips to fully integrated systems, incorporating memory, networking, interconnectivity and cooling. Nvidia's NVL72 architecture is the prime example, significantly increasing content per deployment and enhancing vendor monetization.
- **ASICs Address Custom Needs:** Customized ASICs are gaining traction for repetitive or highly optimized tasks for hyperscalers with unique workloads and scale. Broadcom and Marvell continue to expand their foothold in custom ASICs, which are set to grow at a 27% compound annual rate to reach $118 billion by 2033.
- **Regulatory, Geopolitical Risks:** Escalating US export controls, potential tariff disruptions and geopolitical tension affecting critical semiconductor manufacturing hubs like Taiwan create vulnerabilities in supply chains. Robust hyperscaler and enterprise demand, preferential treatment under the US, Mexico, Canada trade pact and possible government incentives for infrastructure investments will soften the blow.
- **Concentrated Competitive Dynamic:** Nvidia's dominant position in GPU-based training and Broadcom's and Marvell's in custom ASICs for inferencing, have created highly concentrated markets where a few players command substantial market shares, limiting opportunities for smaller competitors through the end of the decade, if not longer.

Performance and Valuation

Most AI semiconductor stocks topped non-AI chip shares in 2025, led by AMD's 77% gain and Broadcom's 49% advance. Marvell fell 23%, partly on fears of inconsistent ASIC expansion by Amazon Web Services. Multiples for AI chipmakers also outpaced the broader sector though Nvidia traded at 24x projected earnings, below its five-year average, on expectations that its rapid growth will slow. Custom ASIC stocks continued to fetch premiums, with Broadcom at 30x, exceeding its 18x average, bolstered by hyperscaler and Ethernet demand.

Section 2. Catalysts to Watch

GPUs, ASICs Poised to Maintain Steady Pace of Upgrades

AI accelerator market growth hinges on a steady cadence of new GPU systems, custom ASIC launches, memory upgrades and packaging breakthroughs. Such milestones will determine whether performance, efficiency and cost-per-token gains can be sustained while reshaping competitive dynamics among Nvidia, AMD and ASIC vendors.

Critical Milestones:

- **Early 2026:** Conversion of chip-on-wafer-on-substrate-S packaging to CoWoS-L accelerates, supporting shipments of Nvidia B300 GPU shipments
- **March 2026:** Expected updates at Nvidia GPU Technology Conference on Rubin racks (NVL576 + CPX inference disaggregation) and early disclosures on the next architecture, Feynman
- **2H26:** AMD's first rack-scale system integrates MI450 GPUs with its CPUs and network interface cards
- **June 2026:** AI event could serve as launch for AMD's Helios, aimed at Rubin-class performance using open standards
- **2026:** Adoption broadens for high-bandwidth memory version 4, as 16-high stacks enable at least 384 GB per GPU, raising costs but boosting bandwidth and power efficiency
- **2026-27:** Initial build-outs of OpenAI multiple-gigawatt capacity data centers through deals with AMD, Nvidia and Broadcom
- **2026-27:** Amazon Web Service's Trainium, Google's TPU, Meta's MTIA, Microsoft's Maia ASICs and OpenAI's first in-house ASIC expected to reach large-scale deployment
- **2026-27:** Nvidia Rubin Ultra rollout extends the company's cadence of annual NVLink-scale upgrades; CPX-prefill racks disaggregate inference, lowering cost per token and raising bar for ASIC challengers
- **2028-29:** As AMD MI500 and Nvidia Rubin CPX scale up, rack-scale computing nodes expand with prefill/decoder specialization, redefining inference economics and intensifying ASIC competition
- **2029-30:** Nvidia Feynman next-generation architecture is expected to follow Rubin, potentially paired with HBM5 and new interconnect fabric, marking another system-level leap

Section 3. Market Overview

AI Accelerator Market Heads for $604 Billion by 2033

The AI accelerator market is set to grow at a 16% compound annual rate to 2033, approaching $604 billion, with GPU and custom ASIC chips powering the next wave of AI infrastructure as models become larger, more complex and require more computing capability. GPUs (graphics processing units) will remain the foundation of large-scale training, while ASICs (application-specific integrated circuits) scale rapidly to improve hyperscalers' inferencing efficiency.

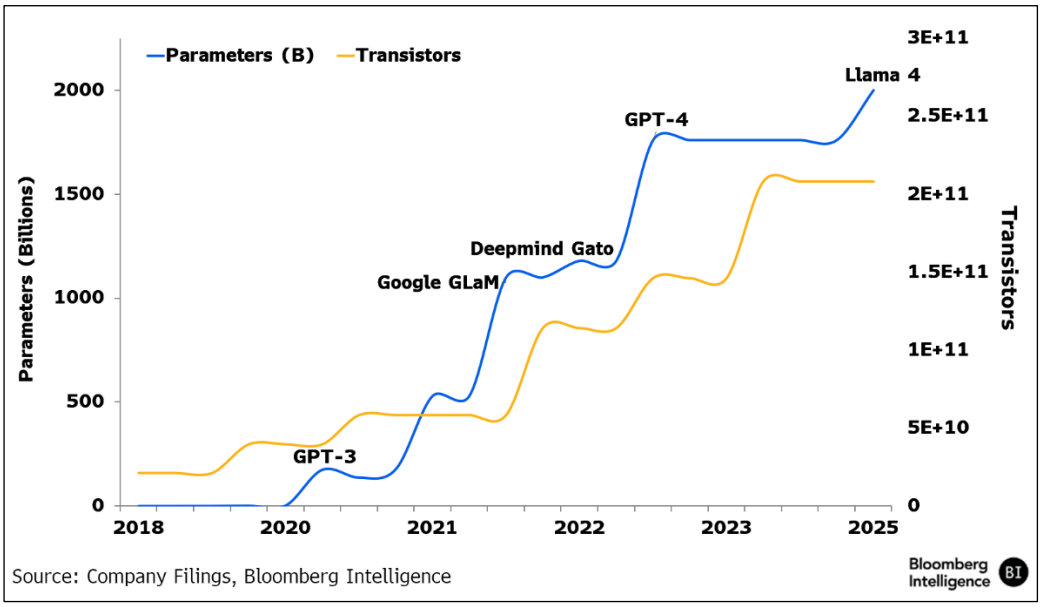
3.1 Traditional CPUs Can't Keep Up With AI

Computing power is surging as chip miniaturization has reached the limit of Moore's Law – that the number of transistors on a chip doubles every two years – due to rising costs, power constraints and physical limitations (see Figure 1). General purpose CPUs (central processing units), which rely on transistor scaling, also are inherently inefficient at handling the massive matrix operations required for AI workloads.

BI

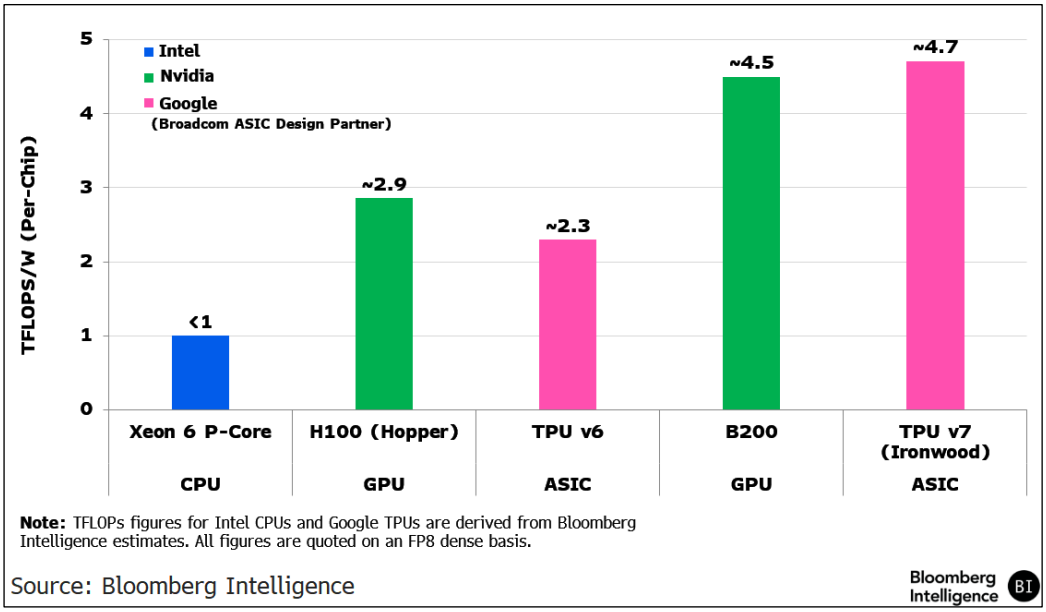
AI performance advances stretch the limits of Moore's Law

Figure 1: Moore's Law vs. AI Model Complexity



In contrast, GPUs and ASICs are designed for parallel execution, enabling significantly higher performance per watt, a critical factor for hyperscale data centers facing power and cooling constraints. As large language models like GPT-4 and Claude 3 now exceed 1 trillion parameters, GPUs such as Nvidia's H200 and Blackwell GB300 deliver parallel matrix operations. Nvidia's Blackwell GPUs, AMD's MI350X and custom AI ASICs have demonstrated five to 10 times the efficiency of CPUs. ASICs from Alphabet's Google, Amazon.com and Meta Platforms improve inferencing efficiency, ensuring exponential AI performance growth beyond the boundaries of Moore's Law. As AI expands, energy efficiency will drive silicon adoption, shifting from general purpose processors.

Figure 2: Performance per Watt



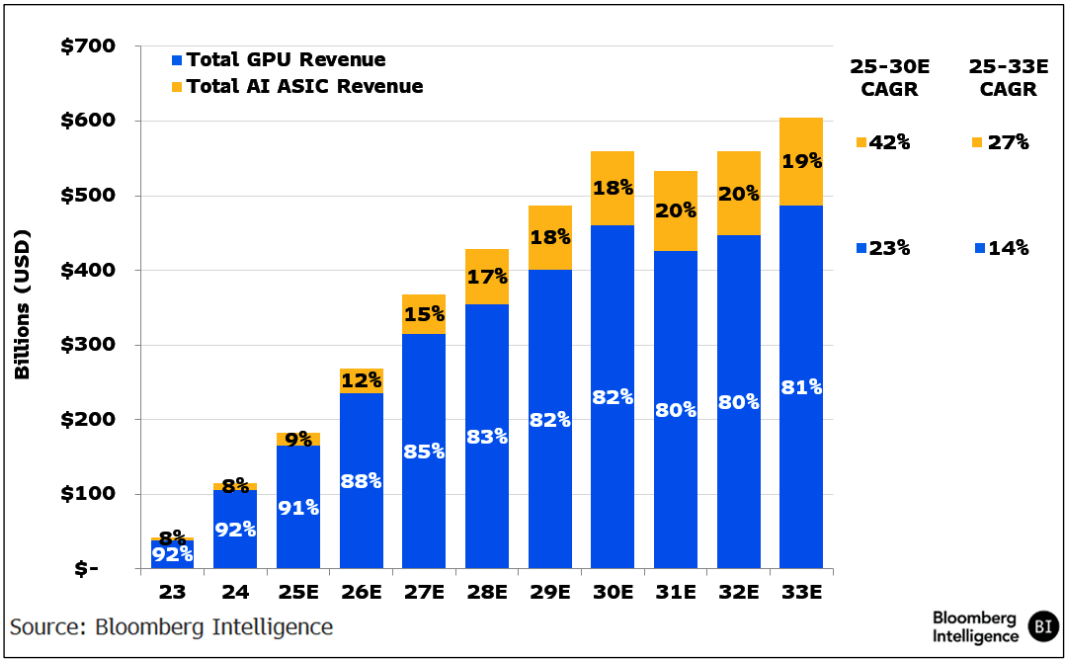
BI **The parallel execution of GPUs and ASICs makes them more efficient than CPUs**

By 2033, the total AI accelerator market, spanning GPUs, ASICs and emerging architectures, will be more than five times what it was in 2024. Though GPUs will retain the dominant share, at 81% of revenue, custom ASICs are growing nearly three times faster, expanding to 19% of the market from 8% in 2024 as hyperscalers ramp up proprietary inferencing silicon. GPU revenue is projected to grow at more than a 14% compound annual rate through 2033, as they remain crucial to AI data centers due to continuing system integration, memory expansion and software lock-in.

Hyperscalers will remain the primary driver of demand through the decade, as Meta, Google, Microsoft and Amazon Web Services lead the adoption of Nvidia's Blackwell architecture. Advanced Micro Devices' MI350 and upcoming MI450 series are also gaining traction, particularly in high-performance and inferencing-focused systems, as seen in the company's 6-gigawatt deal with OpenAI. Rising AI adoption by enterprises, national infrastructure projects and Tier 2 cloud expansions are broadening deployment across regions and verticals, which will reinforce sustained demand for GPU and ASIC systems.

AI data-center construction by hyperscale and neocloud providers, which typically offer GPUs as a service, appears to be well underway. We calculate that cloud and neocloud providers plan to have at least 6 gigawatts of US AI data-center construction in 2025, based on intended corporate capital investment. Multiple 100,000-plus GPU clusters were planned for 2025, including xAI's of at 200,000 GPUs. In 2026, Crusoe Energy and other neocloud providers are bringing 200-megawatt data-center facilities online, which would be enough to house 100,000 Nvidia Blackwell GPU clusters. The trend bodes well for equipment vendors' volume and higher-priced gear.

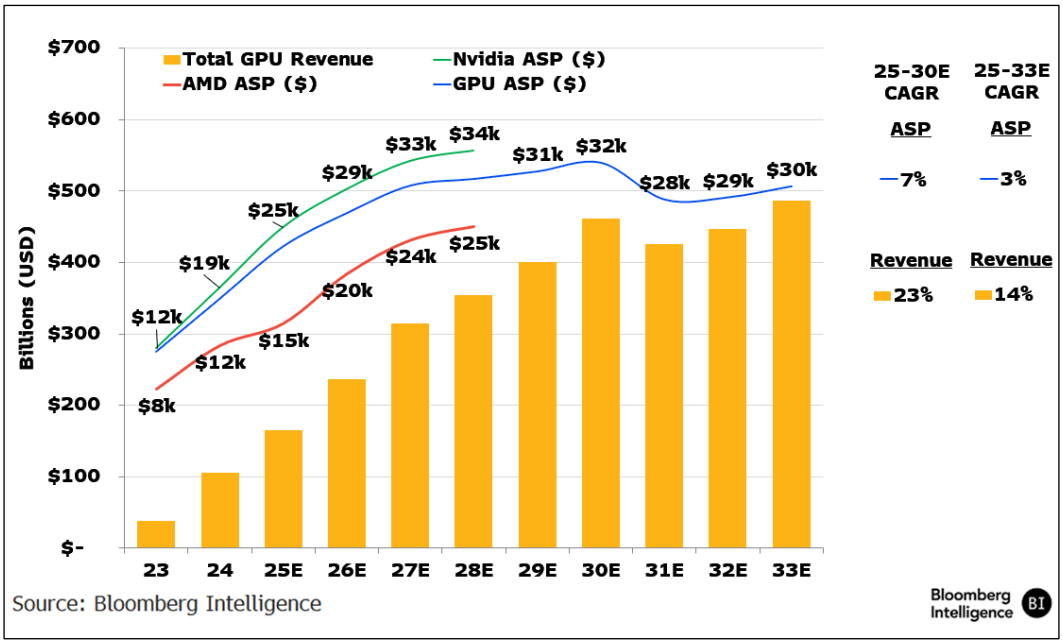
Figure 3: AI Accelerator Total Addressable Market



BI

The AI accelerator market should quintuple by 2033

Figure 4: Nvidia, AMD AI GPU Revenue; Average Selling Prices



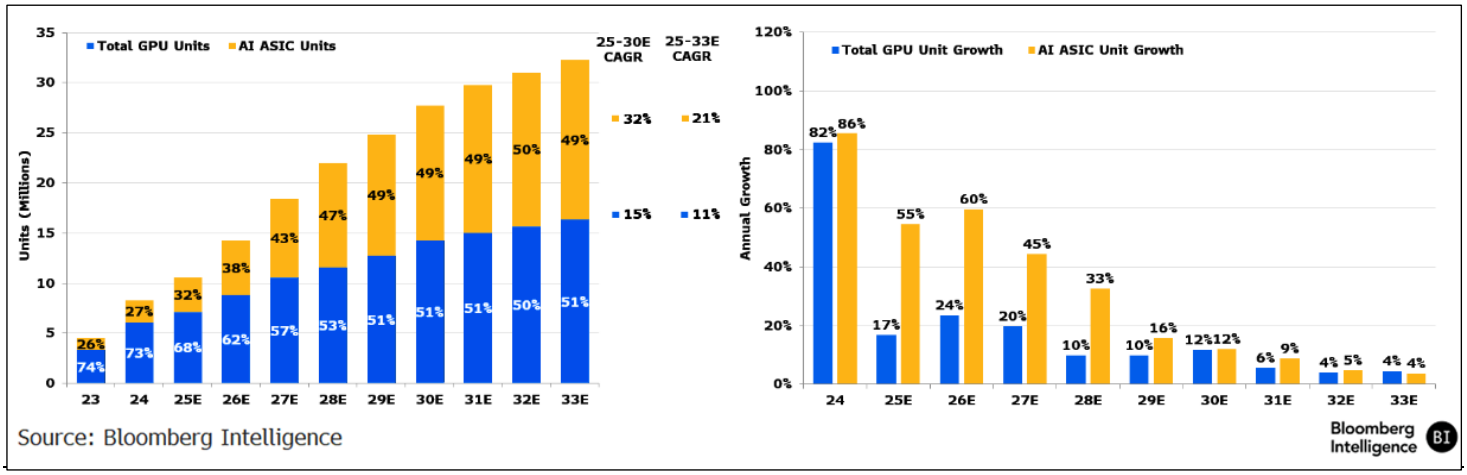
Increased model complexity and larger AI clusters continue to push GPU system values higher, with over 16 million GPUs expected to be deployed annually by 2033. Next generation chips like Nvidia's Blackwell and AMD's MI450 are driving performance and pricing higher, supported by greater high-bandwidth memory content, integrated networking and optimized rack-level designs. Average selling prices for GPUs are projected to rise to $33,000 for Nvidia from $19,000 in 2024, and to $29,000 for AMD from $12,000. AMD is closing the gap, a trend that should

accelerate as its MI450 rack-scale systems launches in 2026. This structural increase in selling prices, combined with cluster expansion, is bolstering GPU growth.

3.2 Custom AI ASIC Chips Not the End for AI GPUs

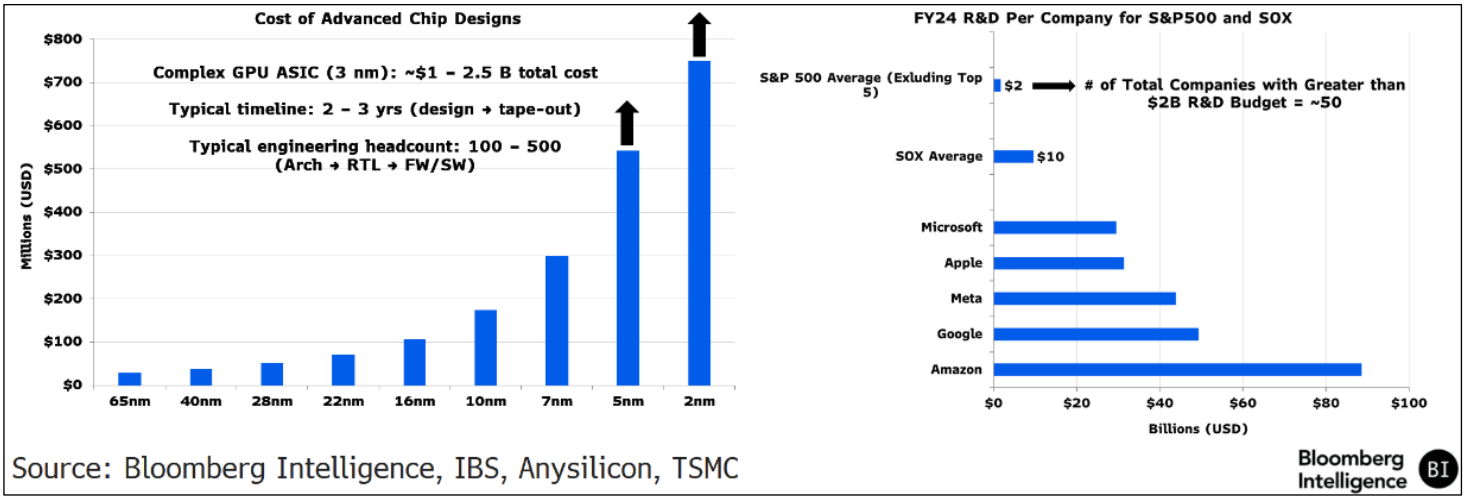
AI ASICs are optimizing hyperscaler computing, reducing reliance on Nvidia for inferencing and repetitive AI tasks. Custom ASICs like Google's TPU and Amazon's Trainium are scaling across inferencing and reasoning workloads, with ASIC unit shipments projected to grow at a 21% compound annual rate through 2033, outpacing GPUs' 11%. ASIC unit volumes should match those for GPUs by 2033 (Figure 5).

Figure 5: AI Computing, Infrastructure Market Share



Despite having a lower cost per chip than GPUs, ASICs' custom silicon programs require significant upfront expenses ($1 billion to $3 billion for advanced nodes), long execution times (two to three years from design to tapeout) and teams with hundreds of hardware and design engineers. Only a few large companies with substantial research-and-development and capital-spending budgets, and the ability to attract top talent and bear upfront sunk costs would be able to set up an ASICs program. More important, even fewer would have enough computing volume for internal use to generate a positive return for such investments. As a result, only the world's top 20 hyperscalers can create ASIC programs over the next five to seven years.

Figure 6: Advanced Chip Cost; R&D Budgets



ASICs' custom silicon programs require significant upfront expenses

Designed to run most efficiently for specific workloads, ASICs are less flexible than GPUs. The current stage of AI development means peak workloads continue to change, aligning well with the flexibility of GPUs and making them suitable for reuse — from training to inferencing, or from pretraining to reasoning — extending their life cycles. Accelerated product releases by leading suppliers like Nvidia and AMD, with large software developer bases, are likely to keep GPUs the main engine for accelerated computing.

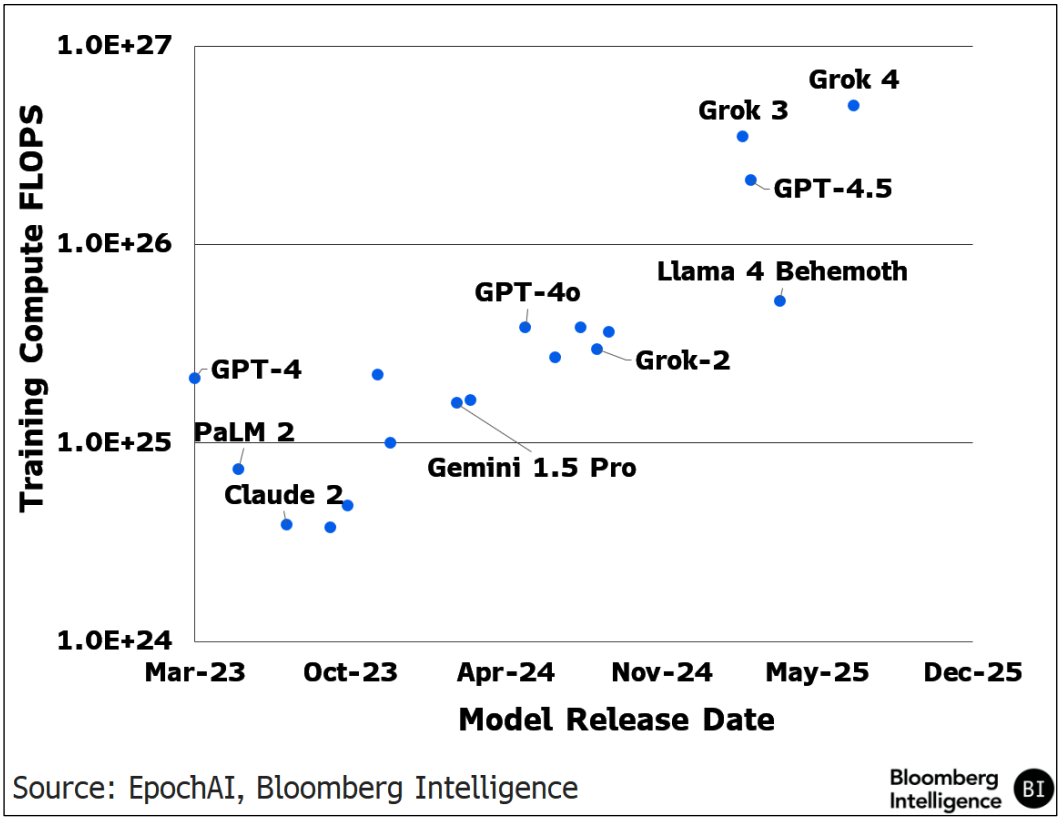
3.3 Token Consumption Driven by Reasoning Models

The industry shift to reasoning models from one-shot answers has increased token throughput at all cloud and large-language-model providers, while cost per token has steadily declined because of improved batching and model efficiency. LLM use has also ballooned: ChatGPT queries increased 2.5-fold from January to June. That supports broader adoption across enterprises, where inferencing economics are critical, while rising cloud revenue offers return on intensifying capital expenditures. With AI agent workflows involving multiple tasks likely to use about 10 times the number of tokens than coding-agent tasks, token consumption will be the biggest driver of inferencing consumption.

Growing computing demand for reasoning models and an increase in the size of GPU clusters remain the biggest drivers of capex increases for hyperscalers. Most companies point to building clusters of 1 million GPUs for their AI infrastructure, which should be a major reason for hyperscale cloud providers to increase capex projections.

BI **Meta, Google, XAI, Anthropic and OpenAI continue to train the latest versions of their models on large clusters**

Figure 7: Frontier Model Training Computing



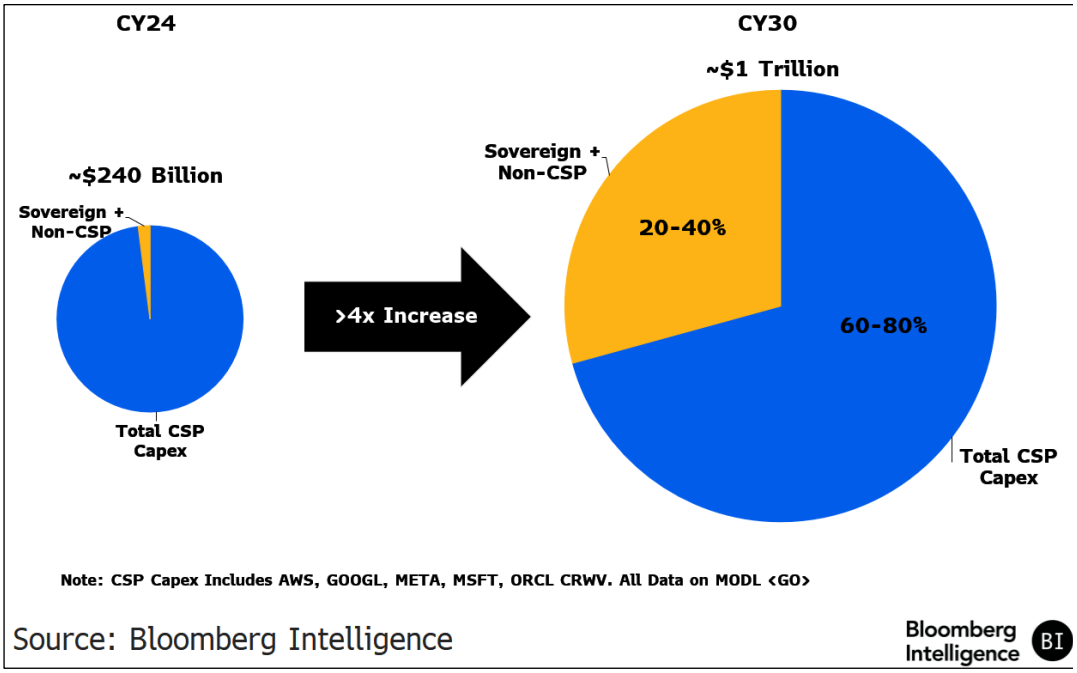
Frontier model companies like Meta, Google, xAI, Anthropic and OpenAI continue to train the latest versions of their models on large clusters, but the market has clearly pivoted toward inferencing amid the widespread adoption of thinking models. OpenAI is generating 2.5 trillion daily tokens from its 1 billion monthly active users, while Google's rollout of AI Overviews to 2 billion users has been the largest driver of token consumption, which has grown about 10-fold in the last year.

3.4 Hyperscalers Continue to Increase Capital Spending

Total AI-related capital expenditures are on track to reach $1 trillion annually, with about $600 billion coming from cloud service providers and the rest from sovereign projects and enterprise customers, diversifying geographic risk (Figure 8). Our calculations project capital spending of $539 billion for the leading hyperscalers – Meta, Oracle, Google, Amazon, CoreWeave and Microsoft – in 2026. Recent earnings calls emphasized the urgency to expand AI infrastructure, driven by tight capacity, growing inferencing workloads and multiyear systems commitments.

Sovereign projects are expected to make up a growing share of AI capex by decade's end

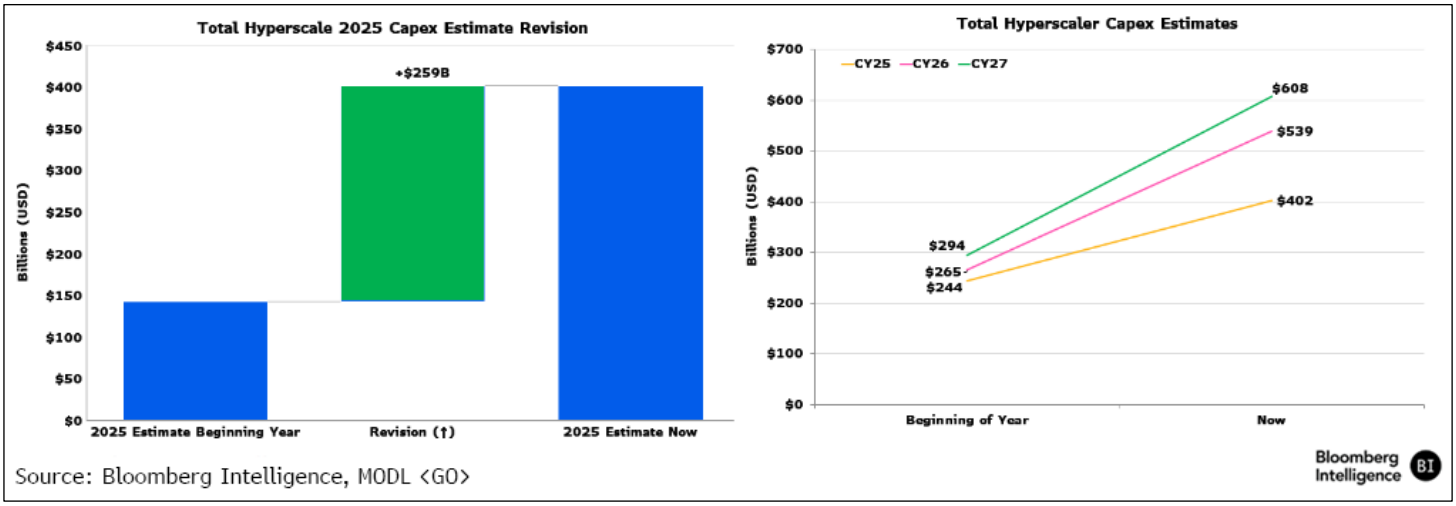
Figure 8: Total Annual AI Capex Breakdown



Our analysis finds that Microsoft is on track to be the largest spender, with projected capital expenditures exceeding $150 billion in 2026, with the bulk of which likely going toward building or leasing AI data centers, driven largely by the continued growth of its partnership with OpenAI. Microsoft will probably allocate most of the investment toward short-lived assets, like chips, networking gear and other AI hardware, rather than long-term infrastructure like buildings. Its focus on providing cloud capacity for AI inference workloads may shift near-term demand for AI model training to other providers. That could lead to increased capital spending by specialized AI infrastructure companies, such as CoreWeave.

OpenAI's latest infrastructure road map alone could be equivalent to about $1 trillion of the projected $3.5 trillion combined capex at Meta, Google, Microsoft, Oracle, CoreWeave and Amazon Web Services for 2025-30 – more than triple their spending for 2023-25. OpenAI's additional supply agreements with AMD, Nvidia and Broadcom build on its previously announced $500 billion Stargate program, lifting the company's total planned computing investment to more than $1 trillion through the end of the decade, establishing OpenAI as the single largest source of expanding AI accelerator demand.

Figure 9: Big Tech Capital Expenditures



BI

The leading hyperscalers are poised to spend $539 billion in 2026

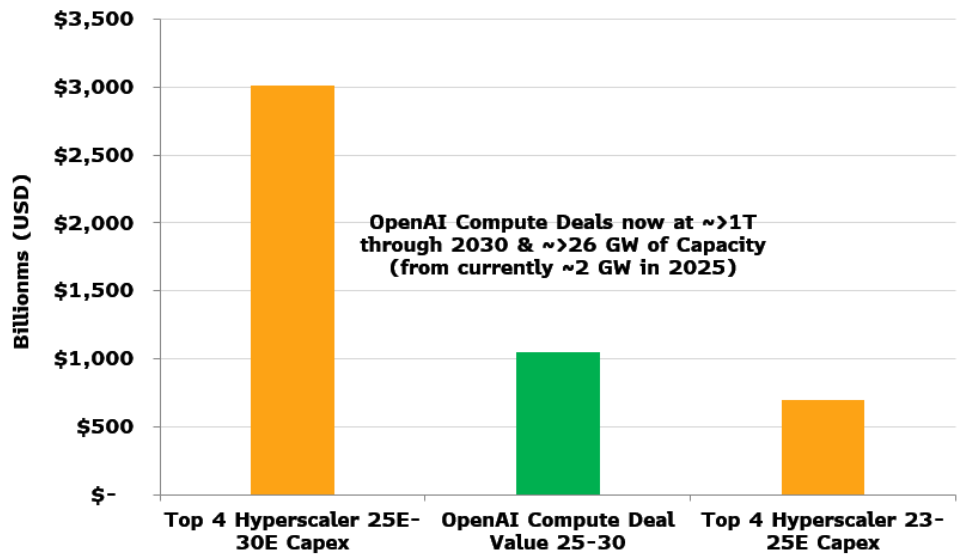
The scale and pace of such commitments are increasing forecasts for earnings and the long-term total addressable market for major semiconductor suppliers, reinforcing investor confidence in sustained AI infrastructure growth through 2030.

OpenAI's partnership with Broadcom for a custom ASIC highlights its push to boost AI accelerator supply for inferencing. OpenAI's efforts to develop its own advanced processor chips will be key to bolstering inferencing efficiency and lowering token pricing to match that of rival Alphabet, which holds a similar advantage from its own tensor processing unit chips.

Broadcom's custom silicon has been used by Google in their TPU partnership and in Meta's MTIA chip. For OpenAI, custom chips are intended to reduce inferencing costs compared with the Nvidia GPUs it uses, which may remain essential for frontier training. OpenAI expects inferencing computing costs to decline to 25% of revenue by 2030 from around 50% in 2024.

Figure 10: OpenAI Computing Deals

Deal / Project	Counterparties	Headline Value ($)	Capacity Planned (GW)	Time Horizon
Stargate (Global Program)	Oracle, SoftBank, MGX (UAE)	$300-350B	10 GW	2025–2029
Stargate US / Oracle Cloud Lease	Oracle	$100-150B	4.5–5 GW	2025–2030
NVIDIA Strategic Partnership	NVIDIA	$300-350B	10 GW	2025–2026 initial; staged thereafter
AMD Multi-Year GPU Supply	AMD	~>$100B	6 GW	Starts 2H 2026; multi-generation
Broadcom Custom Accelerators & Networking	Broadcom (AVGO)	~>$150B	10 GW	2026–2029
CoreWeave Compute Agreements (Cumulative)	CoreWeave	$22.4B	Undisclosed (~1–2 GW est.)	2025–2029



Source: Company Filings, Bloomberg Intelligence

Bloomberg Intelligence BI

BI **OpenAI expects inferencing computing cost to decline to 25% of revenue by 2030 from around 50%**

3.5 Nvidia's Full-Stack AI System Shift Adds Capex Share

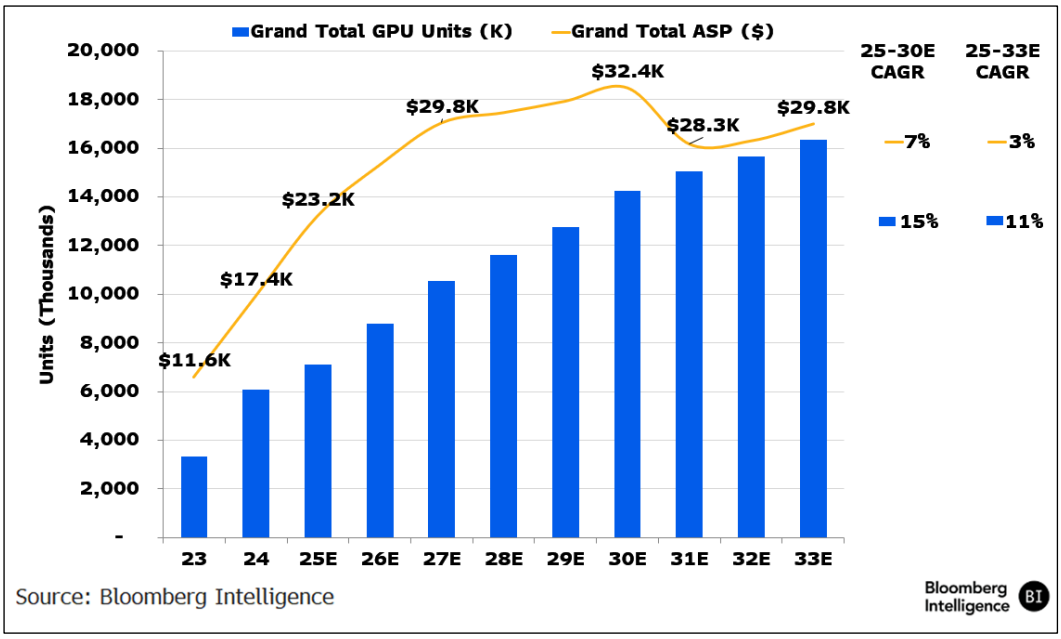
As AI workloads intensify and systems scale, GPUs and other AI accelerators are shifting to full-stack integrated systems from chips. Nvidia's NVL72 and AMD's acquisition of ZT Designs exemplify how each generation expands integration (bundling memory, networking and CPUs), boosts dollar content per deployment with rising average selling prices, deepens platform lock-in and captures a greater share of data-center spending.

Nvidia's Blackwell-based NVL72 moves to 72-GPU rack-scale platforms connected using NVLink from eight-GPU servers. That shifts value to Nvidia, increasing its silicon share to roughly 93% in the GB200 by consolidating components like Grace CPUs, BlueField data-processing units (DPUs), scale-up and scale-out networking and more. Each new generation amplifies the trend. Blackwell boosted average selling prices with more power and memory, while NVSwitches, copper interconnects and direct-to-chip liquid cooling add billing-of-material layers.

That aids monetization per cluster, with average selling prices reaching over $3 million per rack for GB200 from about $250,000 for Hopper systems. Integrated GPU systems drive prices higher by incorporating more system-level components like HBM3e/4 memory, NVLink/Spectrum-X networking and connectivity components. Nvidia further strengthens its control with CUDA, enterprise AI software and NVLink Fusion, ensuring multirack coherence and forcing customers to operate within its fabric, retaining system value.

BI **Shifting to full-stack integrated systems from chips boosts average selling prices**

Figure 11: GPU Systems Average Selling Prices; Unit Growth



That creates a loop in which adoption drives deeper integration, raising dollar content and strengthening customer lock-in, positioning Nvidia and peers like AMD and Broadcom as full-stack data-center vendors rather than simply chip suppliers.

ASICs are the next suite, with each cluster now integrated with much-higher ASIC accelerator counts, connected through integrated networking protocols like Scale Up Ethernet and UALink to operate as a single computing node, increasing average selling prices.

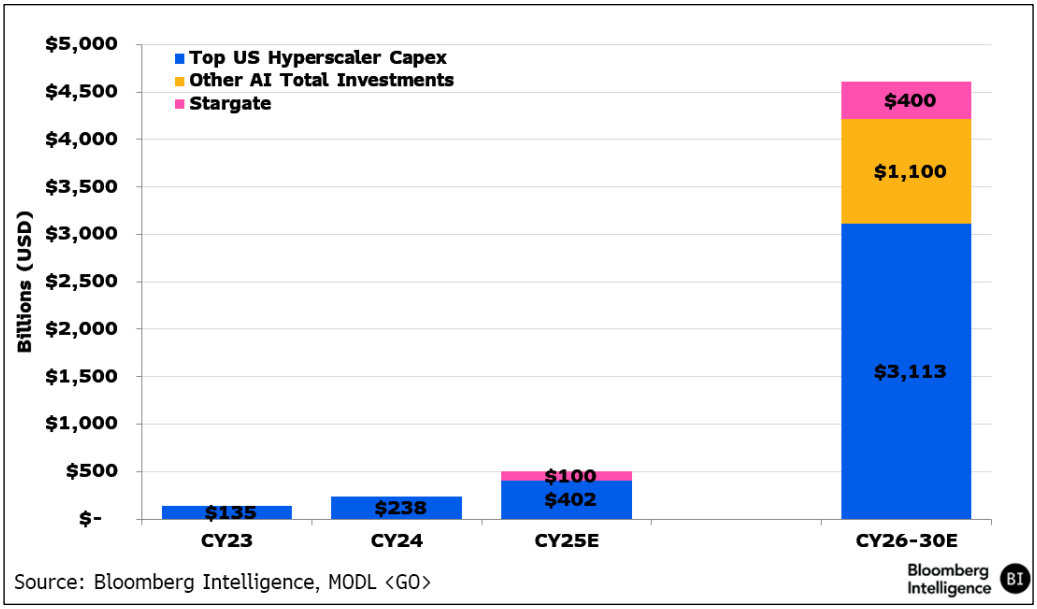
3.6 Governments, Enterprises, Verticals Provide Growth Wave

Governments, enterprises and vertical-specific industries increasingly are deploying AI accelerators to meet their unique needs for data, operational efficiency and specialized AI workloads. That broadens the demand landscape, fueling long-term infrastructure investment and driving the next phase of AI adoption globally.

We calculate that sovereign and enterprise initiatives can contribute as much as $1 trillion cumulatively in demand by the end of the decade (Figure 12). The United Arab Emirates and Saudi Arabia have committed tens of billions of dollars toward national AI infrastructure, the UAE is on track to import more than 500,000 Nvidia GPUs annually through 2027, and Saudi-backed Humain aims to deploy several hundred thousand chips over five years as part of a $10 billion data-center expansion. Such investment can offset the $10 billion to $15 billion of lost China sales in 2025. Nvidia CEO Jensen Huang sees China as a $50 billion total addressable market annually.

BI The US Project Stargate and EU AI funding suggest a long tail for capital spending

Figure 12: Sovereign AI Investments



Though sovereign wealth funds contributed in the mid-single digits to incremental data-center expansion in 2024, Project Stargate in the US and the EU's proposal for a €200 billion fund suggest a long runway for capital spending. Sovereign projects can represent 20-40% of global AI capex by 2030, from around nothing today.

Expansion into verticals such as autos, health care, telecommunications and quantum computing are positioning Nvidia to capture incremental spending from sovereigns and other customers beyond cloud service providers through the end of the decade. The company's auto business is on track to achieve a $5 billion annual run rate across all segments, driven by integrating hardware, software and data and bolstered through partnerships with Uber Technologies and Toyota Motor for self-driving cars, and Aurora Innovation for autonomous trucks.

Most autonomous vehicle technology is built around perception algorithms using data from a suite of sensors. The visual language model supplements that by adding a reasoning component, letting a car interpret and predict its surroundings more deftly. We believe that Nvidia's Thor chip, built on its Blackwell GPU architecture, will encourage original equipment auto manufacturers to integrate visual language models. Tesla, Waymo and Mobileye Global did so in the latest versions of their autonomous driving software.

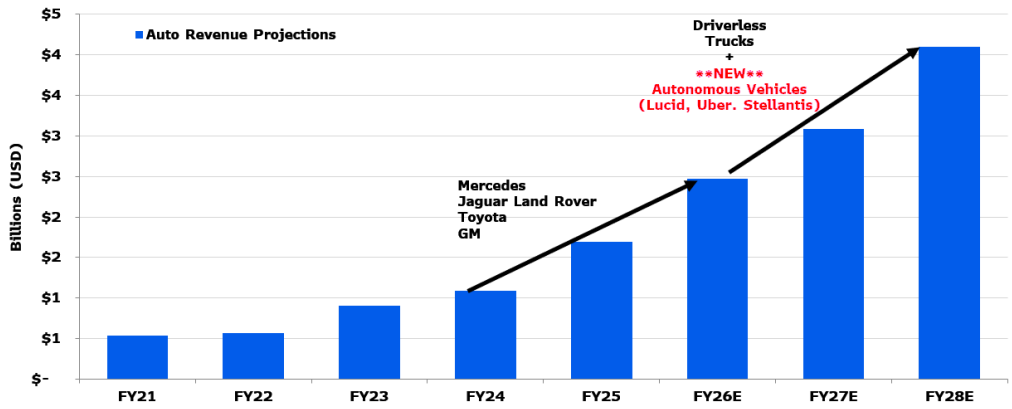
Nvidia has made its Drive AGX Hyperion 10 the common specification for Level 4 autonomous vehicles, those in which the system is fully responsible for driving in limited service areas. Stellantis and Uber use the Nvidia Cosmos platform, harvesting driving data for continuous model training. Lucid plans for its Level 4 consumer cars to deploy Drive AGX Thor, while Mercedes-Benz and others have been testing Hyperion-based programs.

Standardized computing and sensors, a ride-hailing marketplace and a data engine raise Nvidia's per-vehicle content, create recurring software revenue and enlist data-center GPUs for training. Momentum in key markets like China positions Nvidia for sustained growth in autos, diversifying its portfolio through the industry's substantial scale.

Nvidia's auto content generates recurring software revenue

Figure 13: Nvidia Auto Wins, Revenue

Company	Product	**NEW** Country	Ramp/Shipping Dates
Lucid	AGX Thor	USA	Future midsize vehicles (TBD)
Uber	AGX Hyperion 10	USA / Intl	2027 (fleet scaling)
Stellantis	AGX Hyperion 10 (robotaxi)	Netherlands / USA	2028 (start of production)
Mercedes-Benz	AGX Hyperion 10 (L4-ready)	Germany	TBD
Continental	AGX Orin	Germany	2027 (driverless trucks deployment)
Aurora	AGX Orin, DRIVE OS	USA	2027 (driverless trucks deployment)
Toyota	AGX Orin, DRIVE OS	Japan	TBD
General Motors (GM)	AGX Orin, Omniverse, DRIVE OS	USA	TBD
JLR (Jaguar Land Rover)	AGX Orin	UK	2025 (Range Rover, Defender, Discovery, Jaguar)
Rivian	AGX Orin, R1 platform	USA	Shipping now (R1 vehicles)
Hyundai Motor Group	AGX Orin, Omniverse	South Korea	Ongoing
Lucid	AGX Orin, DRIVE Hyperion	USA	2024 (Project Gravity SUV)
Mercedes-Benz	AGX Orin, Omniverse	Germany	Ongoing
Li Auto	AGX Orin, Thor	China	Ongoing (L9 SUV)
Polestar	AGX Orin	Sweden	Ongoing (Polestar 3 SUV)
NIO	AGX Orin, Adam Supercomputer	China	Ongoing (ET5, ES7, and ET7 models)
XPENG	AGX Orin, Thor	China	Ongoing (G6, G9, P7 models)
MediaTek	AGX Orin, RTX Graphics	Taiwan	Ongoing (vehicle segments)
BYD	AGX Orin, DRIVE Hyperion	China	2023 (Hyperion architecture, ongoing)



Source: Source: Bloomberg Intelligence, MODL <GO>, Company Filings

Bloomberg Intelligence BI

Nvidia's RTX Pro servers promise to give Cisco Systems, Dell Technologies, Super Micro Computer and Hewlett Packard Enterprise a kick start into enterprise AI. The product is geared for on-premise enterprise deployment and addresses a challenge for enterprise AI costs, at about $70,000 a server and power consumption at 400-600 watts, compared with $280,000 and 700-1,000 watts for an HGX-based server. RTX Pro's leading use case is digital twins to build simulations for AI, and we expect those systems to complement HGX, MGX and Blackwell AI systems. It may take a couple of years for that segment to add punch to growth until enterprises build a critical mass of AI applications.

Section 4. ASIC Transition

Shift to Custom Chips Improves Control Over Infrastructure

Hyperscalers are speeding the transition to ASICs to optimize power efficiency, cost and performance for big AI workloads. Since ASICs are custom-built for specific tasks, they give Meta, Google, Amazon, Microsoft and others greater control over their AI infrastructures. The shift should boost AI sales for Broadcom and Marvell Technology, with faster growth than GPU units.

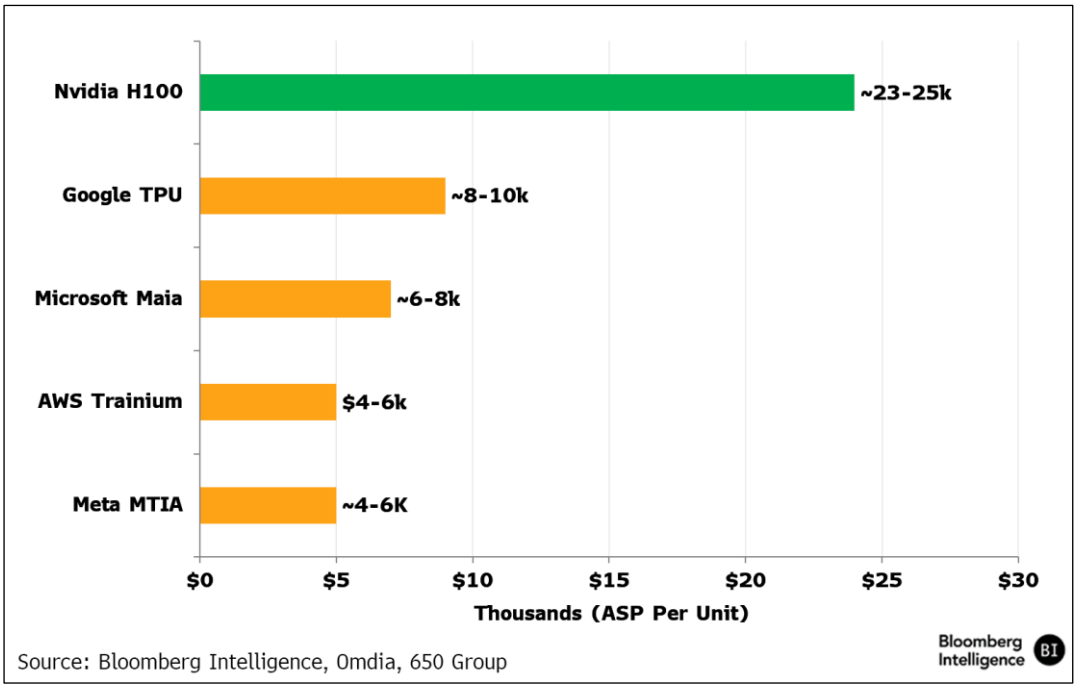
4.1 ASICs Boast Cost, Power Efficiency Advantages Over GPUs

Though GPUs remain dominant in AI training and inferencing, they are the most expensive chips in a server and account for 40-50% of total system hardware spending. With Nvidia newest GPUs priced at $30,000 or more, custom ASICs are an attractive alternative, often cutting per-chip costs by over 50% (Figure 14). That advantage becomes even more compelling in light of hyperscalers' mass volumes, with the top four accounting for 70-80% of total AI hardware capital expenditures and contributing 45-50% of Nvidia's total data-center revenue in 2025. ASICs also reduce operating expenses through their superior power efficiency. Procuring peripheral components directly also reduces costs compared with paying GPU vendors.

BI

GPUs, like those made by Nvidia, can make up 50% of total system hardware spending

Figure 14: ASICs, Nvidia H100 Average Selling Prices



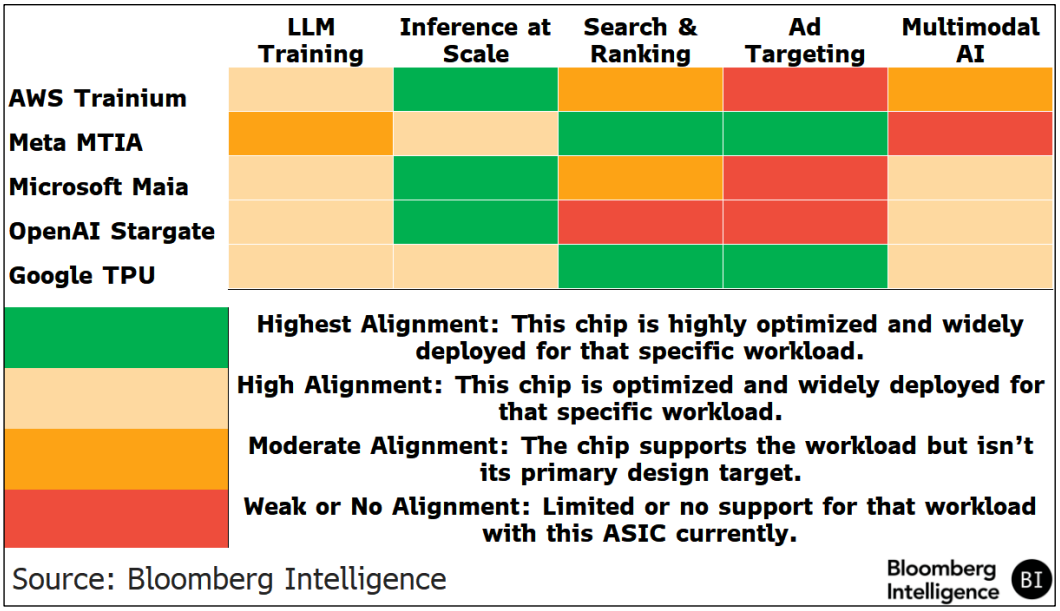
Source: Bloomberg Intelligence, Omdia, 650 Group

General purpose GPUs also can't be significantly customized, capping the ability of hyperscalers to differentiate among their AI offerings. The custom silicon of ASICs allows vertical integration for hyperscalers' hardware, software and networking systems. Such customers have unique, distinguished workloads like ad targeting, ranking, search and inferencing at scale. Meta's MTIA, for example, is tuned for recommendation systems, while Microsoft's Maia targets LLM inferencing and Google's TPU balances training and inferencing. By tailoring hardware to known

workloads, hyperscalers improve performance per watt, reduce latency and streamline their AI infrastructure.

ASICs can be customized for unique workloads like ad targeting, ranking, search and inferencing

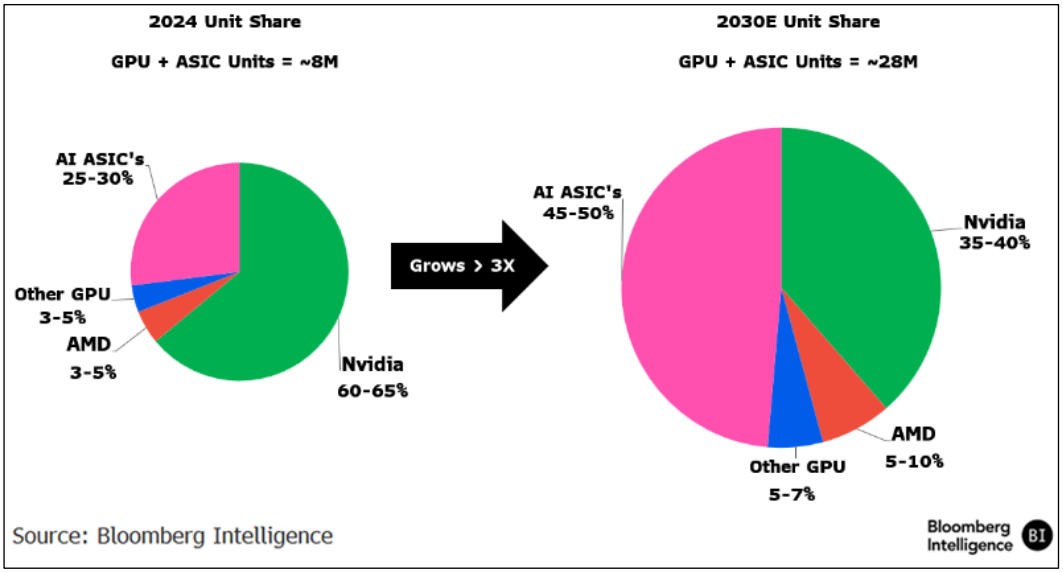
Figure 15: Custom ASIC Workloads



Meta, Google, Amazon, Microsoft, Open AI and other major silicon programs are increasing ASIC deployment for uses spanning inferencing, training and server processing. Broadcom and Marvell dominate as the semiconductor design partners behind such programs, tailoring silicon to their clients' unique demands.

The efficiency breakthroughs of China's DeepSeek could reinforce investments in custom silicon, like Google's TPUs and Amazon Web Services' Trainium, which optimize AI workloads for cost and power efficiency. Given multiyear planning, these custom ASIC programs are likely to remain intact through 2027, particularly in light of advances like DeepSeek. Broadcom and Marvell still are likely to benefit even if training clusters aren't as large as initially expected. Though improved efficiency may temper the need for massive training clusters, hyperscalers' focus on scaling inference ensures continued investment in custom silicon and networking.

Figure 16: AI Accelerator Unit Share



Section 5. GPU Transition

Nvidia Shifts From Providing Chips to Server Systems

Nvidia's evolution from selling individual GPUs to providing fully integrated AI server systems marks a pivotal shift in its business model and mirrors a broader industry trend. The company's DGX and Blackwell architectures deliver complete AI-optimized server systems. The transition increases the content per deployment, capturing more of the data-center value chain. By integrating computing, networking and memory into full-stack solutions, Nvidia is deepening its market penetration, enhancing monetization per system and positioning itself as the dominant force in AI infrastructure.

5.1 AMD Acquisition Is Aimed at Closing Gap With Nvidia

Most peers can't match Nvidia's impressive three-year plan for graphics processing. Blackwell Ultra's ramp-up in the second half of 2025 and expanded memory will strengthen the company's dominance. Its decision to keep the same architecture and chassis design through Rubin Ultra, covering a span of the next three years, cuts execution risk, lowering the chances of the manufacturing hiccups that occurred in the Blackwell transition. Performance improvements at a rapid annual cadence should keep Nvidia's lead intact, making it difficult for merchant and ASIC competitors to catch up.

BI

AMD's M&A strategy is essential to its shift to becoming an integrated-systems firm

Following its acquisition of Silo AI, AMD's purchase of ZT aims to provide the missing piece in its accelerator system-design strategy. It's a timely move as AMD seeks to close the gap with Nvidia, following a similar path toward providing integrated-systems solutions. We see the deal as essential for AMD to avoid falling further behind its rival. Return on investment will depend entirely on execution and realization of revenue synergies as such deals generally have a lower success rate than those focused on cost synergies. ZT's expertise should help AMD shorten its time to market for rack-level solutions, which is currently constrained by the innovation pace of its partners among original-equipment and original-design manufacturers. The transaction also will drive more optimized, nonproprietary solutions, enhancing CPU offerings.

Nvidia's GB300 NVL72 exemplifies the shift toward pre-engineered rack-scale clusters, with 72 GPUs connected by NVLink switches and delivered as a complete, liquid-cooled rack. Scaling out requires additional layers – top-of-rack switches, copper ACC/AEC links and optical transceivers – to stitch racks together into megaclusters.

Spectrum-X Ethernet and InfiniBand NDR/XDR maintain 800-gigabyte to 1.6-terabyte bandwidth, reducing latency and cabling complexity. This system-first model maximizes cost-of-ownership efficiency, shortens deployment cycles and ensures that hyperscalers can scale workloads without interconnect bottlenecks, shifting the unit of computing from the GPU to the rack and fabric itself.

Section 6. Competition

Broadcom Leads ASIC Market as Nvidia Tops in GPUs

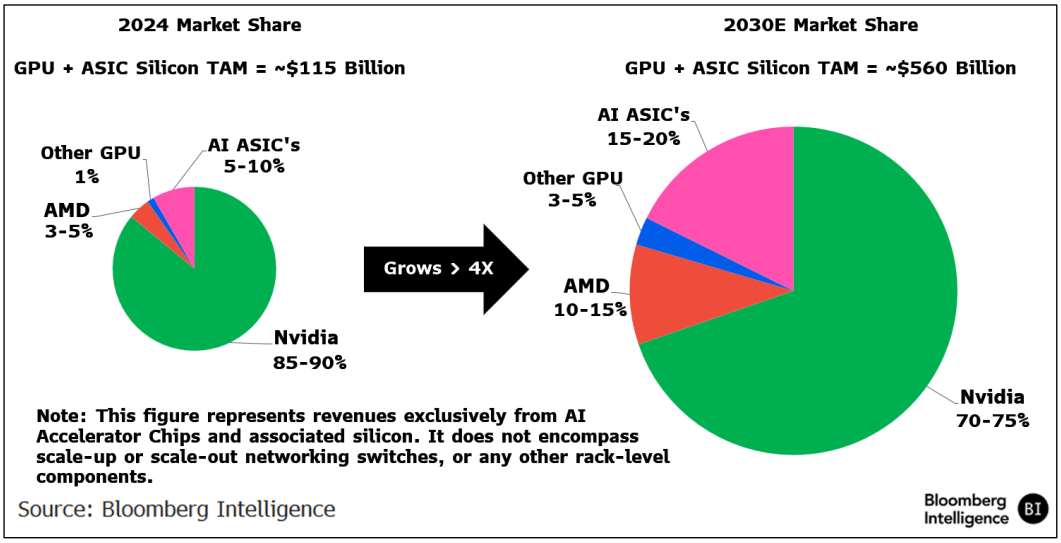
Nvidia's dominant share of the AI training market is unlikely to be challenged meaningfully in the near term, bolstered by its advancements in GPU hardware, software ecosystem and infrastructure integration. Meanwhile, Broadcom's and Marvell's expertise and partnerships with Meta, Google, Amazon and Microsoft will help the pair retain their dominance in ASICs, even as new rivals like MediaTek and Alchip Technologies emerge. In China, tensions with the US may foster a breeding ground for ASIC designers to satisfy growing demand from hyperscalers. Also, Chinese cloud providers like Baidu, Tencent and Alibaba might, like their US counterparts, benefit from custom domestic chips.

6.1 Nvidia Can Keep Grip on at Least 70% of Training Market

Though competition from custom ASICs and alternative accelerators is increasing, Nvidia's CUDA ecosystem, deep software integration, NVLink networking and superior training performance create a wide competitive moat. The transition from Hopper (H100) to Blackwell (B100/GB200) further entrenched Nvidia's lead, with substantial performance and efficiency gains expected. We project that Nvidia will command 70-75% of the AI accelerator market through 2030 (Figure 17). Blackwell is extending the company's dominance with higher computing throughput, expanded memory bandwidth and improved power efficiency. Its full-stack ecosystem, from CUDA software to NVLink and Spectrum-X networking, keeps hyperscalers deeply tied to Nvidia's architecture.

Figure 17: AI Accelerator Market Share

BI **Nvidia can maintain its competitive moat in AI training over the next several years**



AMD is steadily expanding its position with the MI350 and MI400 series, projected to reach a share in the double digits by 2030. Its open-software approach (ROCm) and upcoming Helios rack-scale systems provide a credible second source for training workloads.

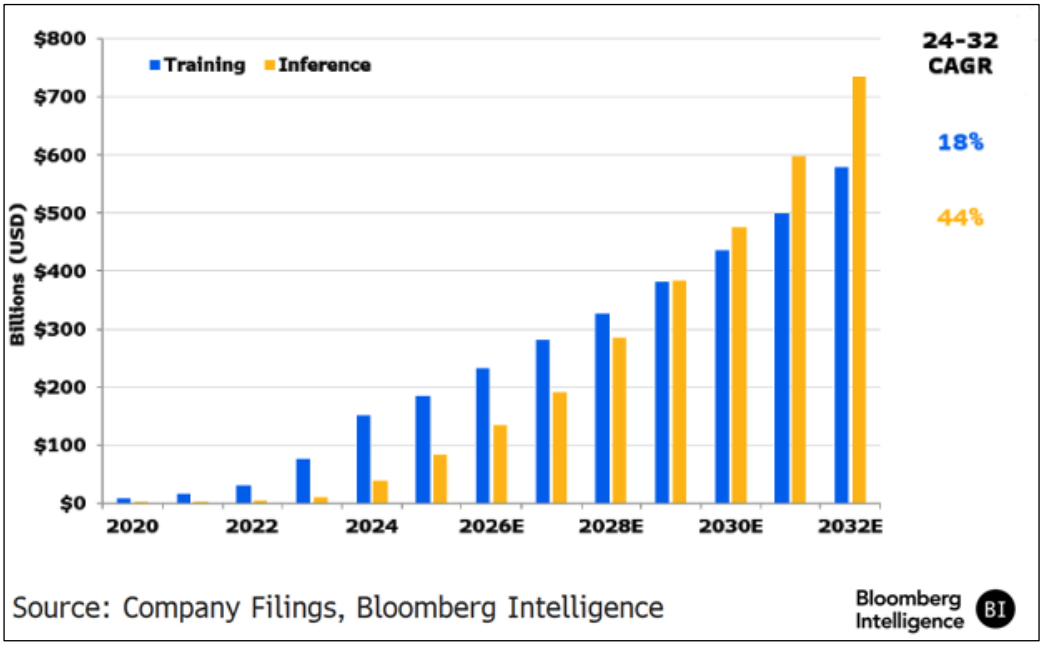
Even as ASICs proliferate for inferencing, training will remain anchored around GPUs, particularly from Nvidia, given their unmatched flexibility, scalability and integration across hardware, networking and software.

Nvidia's software ecosystem continues to bolster its defense against rivals, with innovations like AI Blueprints and NIM microservices and expanding enterprise applications in areas such as fraud detection, call centers and document summarization. Cosmos' foundational models and expanded Omniverse capabilities deepen Nvidia's integration for robotics, simulation and autonomous systems, anchoring enterprise dependence on its platform. Combined with products like DGX Cloud and customized AI agents, Nvidia's software-driven strategy sets the stage for software revenue to outpace its $2 billion annualized run rate.

Nvidia's Blackwell platform is positioned to extend its training dominance into inferencing, making GPU-based inferencing more cost-competitive at scale. Inferencing workloads are becoming the next monetization phase as reasoning and agentic AI models drive higher computing intensity, creating durable demand beyond initial training cycles.

BI **Inferencing is expected to expand faster than training**

Figure 18: Training, Inferencing Markets



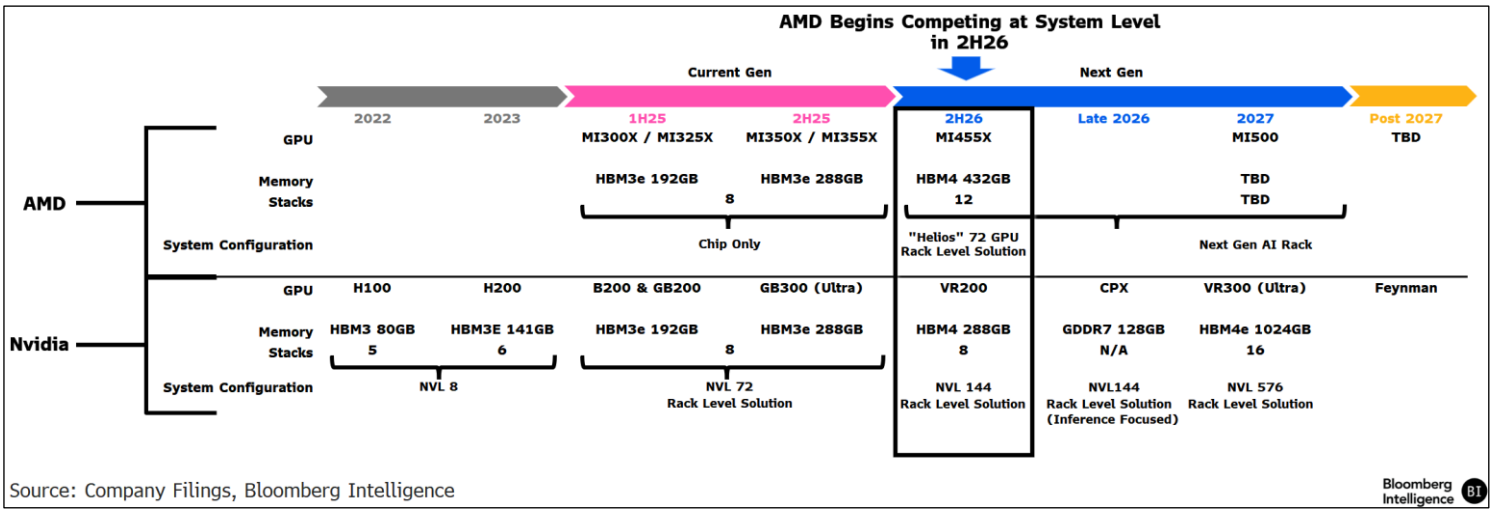
Training remains the larger market for the near term, yet inferencing is expected to expand faster as the decade progresses. Nvidia already attributes a growing share of data-center computing revenue to inferencing, with software tools like TensorRT-LLM, NIM microservices and NVLink-based system optimizations lowering barriers for adoption. The installed CUDA ecosystem and switching costs remain high, reinforcing stickiness as enterprises and national projects scale deployment.

6.2 AMD Helios Rack System Sets Up Next Generation

AMD rollouts are accelerating, with the MI350 series shipping in 2025 and the MI450+Helios rack solution slated for 2026. The timeline already reflects the impact of the March acquisition of ZT

Systems, which has begun to influence system-level design decisions. Helios – a fully integrated rack featuring MI450, EPYC Venice CPUs and Pensando NICs – could be a turning point for AMD to narrow the systems gap with Nvidia. The MI350 series, built on the same modular platform as the MI300, should speed customer adoption and make upgrades easier. AMD now has a credible systems road map, with open standards and scalable integration, giving it a real opportunity to gain market share as AI infrastructure spending accelerates.

Figure 19: AMD, Nvidia System-Level Road Maps



Source: Company Filings, Bloomberg Intelligence

Bloomberg Intelligence BI

AMD is expanding its customer base, including with Nvidia customers

AMD's GPU customer base continues to expand, with inferencing demand growing at around an 80% compound annual rate, driving new deployments. The company now counts seven of the top 10 AI hyperscalers as customers, including Nvidia clients like OpenAI, Microsoft and Meta. AMD's wins extend beyond pilot programs, with its GPUs now supporting inferencing as well as training at scale. It also continues to increase traction with Tier 2 cloud providers and enterprise buyers, aided by a modular product architecture and fast software deployment. As inferencing becomes a larger share of AI spending, AMD's memory-optimized design, value-focused positioning and broad ROCm ecosystem are helping convert customer trials into engagements.

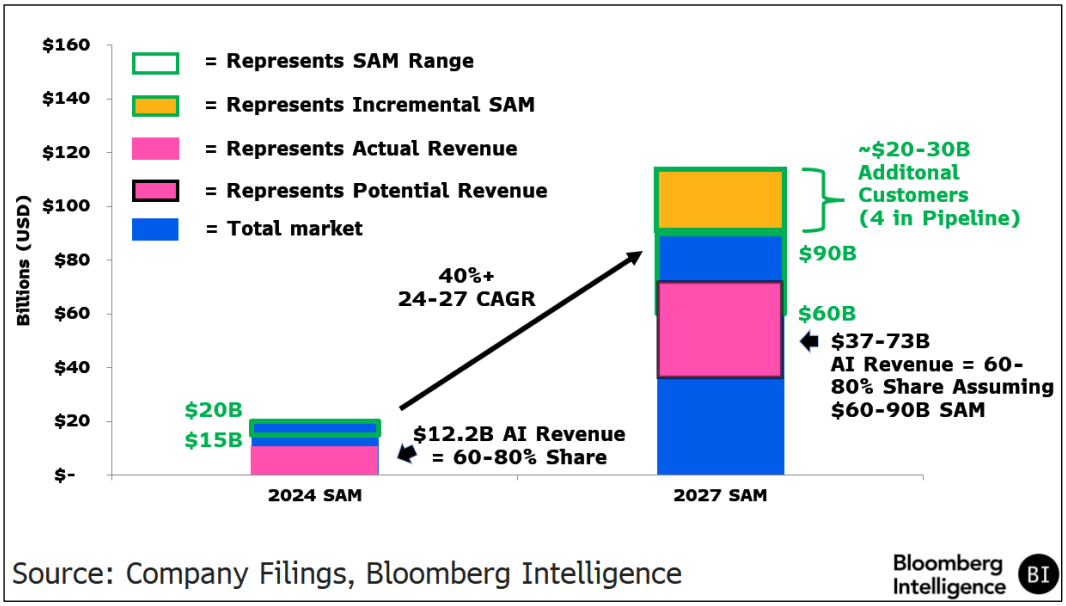
AMD's 6-gigawatt GPU agreement with OpenAI marks a foundational shift in scale, providing visibility into high-volume GPU deployments across multiple generations. The first 1 GW of MI450 chips is scheduled to ship in the second half of 2026, with guidance for incremental data-center AI sales in the double-digit billions once fully ramped up and a clear line to about $10 billion annually by 2027. Management also expects the partnership and associated halo effects to generate up to $100 billion in cumulative ecosystem revenue over the next several years. Consensus following the deal's announcement implied roughly $70 billion in cumulative AI GPU revenue through fiscal 2029, reflecting rising confidence in AMD's long-term competitiveness. The deal could imply additional upside as AMD solidifies its No. 2 spot behind Nvidia in merchant GPUs.

6.3 Broadcom, Marvell to Maintain Hold on ASIC Market

Broadcom has 60-80% of the AI ASIC market, enjoying a first mover advantage with the fifth generation of the Google TPU shipping. The chipmaker is engaged with the top seven

hyperscalers for custom silicon products, including Google, Meta and OpenAI, while playing a crucial role in AI networking with its Tomahawk and Jericho switching platforms. Broadcom targets a $60 billion to $90 billion serviceable addressable market in fiscal 2027, which we think is conservative, as it expands into full-stack AI computing. Its co-packaged optics and networking silicon are becoming the backbone of AI clusters, enabling efficient, high-bandwidth interconnects. As hyperscalers increasingly shift to custom silicon for inferencing, Broadcom remains the top supplier of power-efficient, high-performance ASICs.

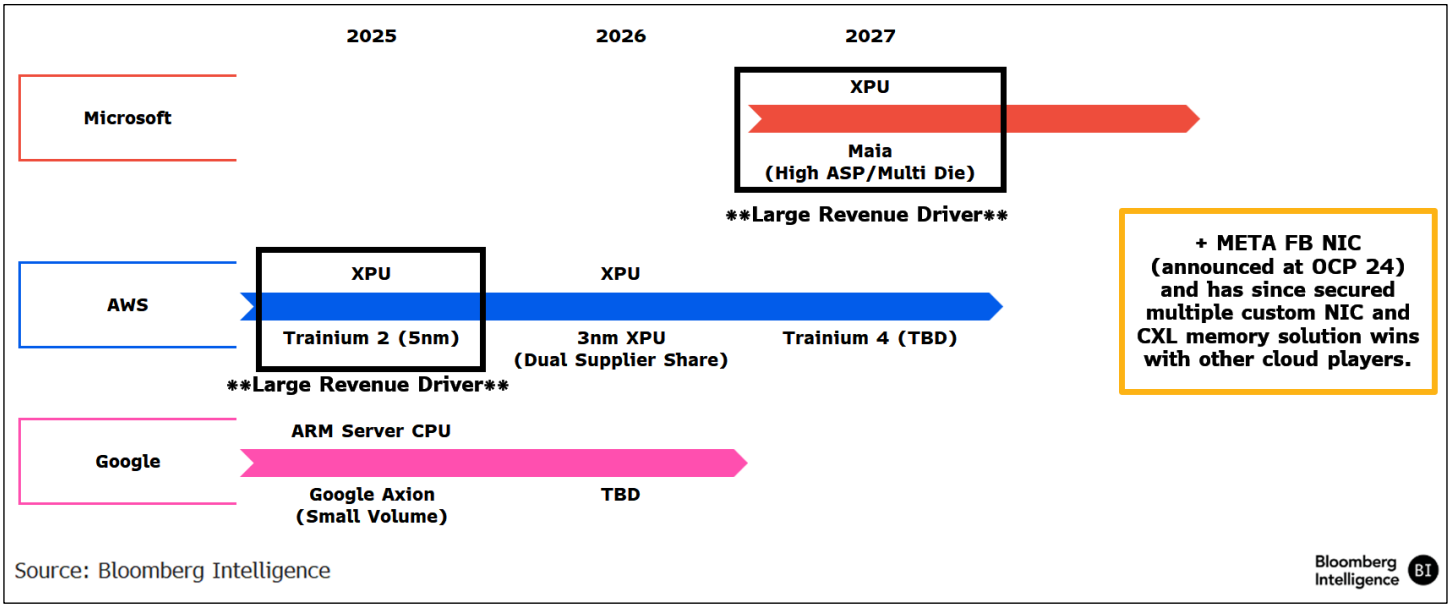
Figure 20: Broadcom AI ASIC Networking Serviceable Addressable Market



Marvell has scored major wins with Meta, Microsoft and Amazon, making it a strong competitor to Broadcom

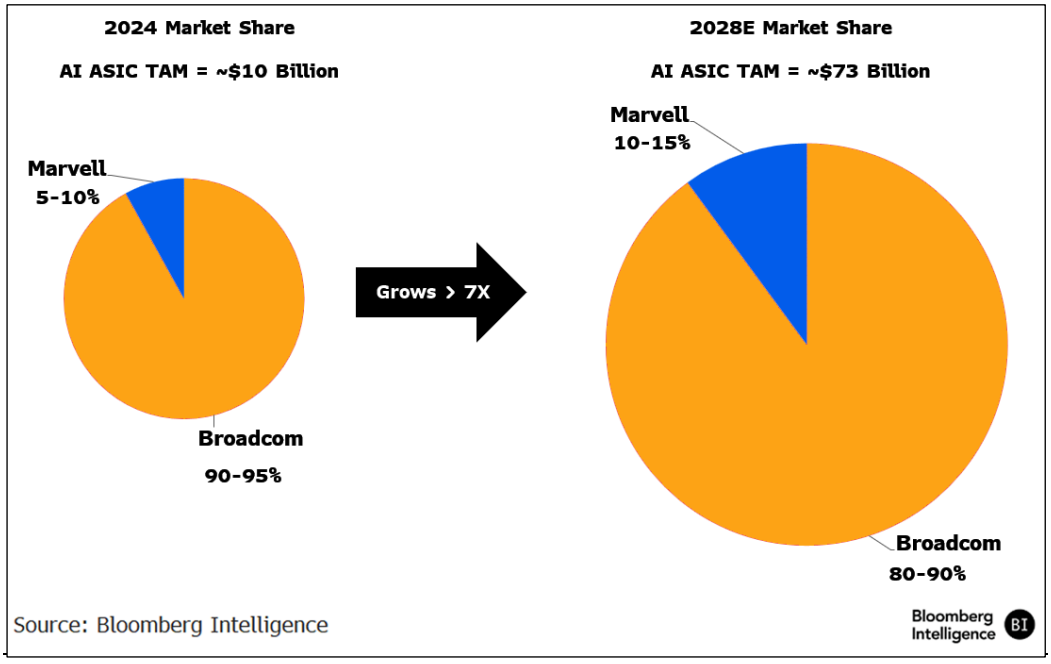
Marvell is a formidable competitor to Broadcom, securing major wins with Meta, Microsoft and Amazon Web Services (Figure 21). Marvell's custom silicon is central to AWS' Trainium chips, which power Amazon's AI clusters, including Project Rainier with 100,000 Trainium2 chips. Marvell also supplies custom silicon for Microsoft's Maia AI accelerator, reinforcing the chipmaker's role in AI infrastructure. Its Teralynx Ethernet switches and Octeon DPUs are gaining traction, particularly as hyperscalers shift toward Ethernet-based AI fabrics.

Figure 21: Marvell's AI ASIC Road Map



Credible rivals continue to mount challenges to the duopoly. Google and AWS reportedly are considering Taiwan ASIC designers like MediaTek and Alchip, which could reduce the incumbents' ability to ramp up volume. Yet we believe Broadcom's and Marvell's wide IP portfolios and advantages as early movers are likely to limit their market-share losses.

Figure 22: AI ASIC Market Share



6.4 As China Localizes Chips, SMIC to Gain Pricing Power

China's Semiconductor Manufacturing International is well-positioned for the country's accelerating AI chip expansion, creating a large captive market and strengthening pricing power. IPOs from clients like Shanghai Biren Technology and Moore Threads Technology will tighten demand for SMIC's 7- to 14-nanometer nodes. That supports our forecast for SMIC's gross margin to recover to 22% in 2026 and for advanced-node revenue to triple by 2028 as yields improve.

BI **Regulatory restrictions create a competitive moat for China's SMIC**

The persistent threat of further US sanctions on AI accelerators has created a powerful geopolitical moat for SMIC. We expect domestically designed AI chips to capture more than 50% of China's cloud and enterprise purchases by 2028 as giants like Tencent and ByteDance prioritize supply-chain security over sourcing from Nvidia. That insulates SMIC from direct competition with TSMC and other leading-edge foundries.

The performance of domestic chips from designers like Huawei Technologies lags behind that of rivals yet is gaining ground through "brute force" system architecture — clustering vast numbers of domestic chips — and leveraging advanced packaging to circumvent manufacturing limitations.

Surging demand for China-designed AI chips and the market's economics will give SMIC substantial room to raise prices for its advanced node wafers (14 nanometers and smaller) over the next three to five years. A single 12-inch silicon wafer dedicated to AI chips can generate over $300,000 in market value, compared with less than $42,000 for mobile processors. As the wafer cost is a small fraction of the AI chip's final selling price, designers can easily absorb significant price hikes from their foundry partner. SMIC, with its China monopoly in advanced node manufacturing, is uniquely positioned to capture a greater share of that value chain.

SMIC's reliance on complex DUV multipatterning for its 7-nm-equivalent nodes results in challenging initial production yields, which we project at a mere 30% for large chip dies. Yet that low starting point means output can grow as the process matures, and SMIC could double production in advanced nodes in several years. The primary catalyst will be gaining experience, accelerated by investment in advanced measurement and testing tools that allow for faster process debugging. As SMIC moves up the learning curve over the next three years, it can reach a yield of at least 60%.

SMIC must scale up advanced-node revenue — about 10% of its total, or $710 million, in 2024 — to increase earnings. We calculate that by boosting yields to 60% for large AI chip dies and raising prices by 12% a year through wafer price hikes and migrations to N+2/N+3 processes from 14 nanometers, revenue could triple to over $2.4 billion by 2028. Such growth will be needed to stabilize margins, after aggressive capacity expansion since 2021 raised depreciation costs by 10% annually. SMIC's gross margin fell to a decade-low of 18% in 2024, even as sales increased 27%.

On-premise enterprise deployment has ignited AI chip demand in China, providing new customers beyond the cloud leaders as state-owned telecom, finance and energy companies have aggressively deployed inference systems to drive business development and ensure data sovereignty. Efficient open-source models like DeepSeek lower the computing bar for AI all-in-one systems that allow enterprises to fine-tune models on private data securely.

China's halt of Nvidia AI chip orders presents a substantial tailwind for SMIC's growth outlook, enabling the company to beat even consensus' accelerated revenue trajectory. Nvidia's 70%

share of China's AI accelerator market, estimated by IDC at 2.7 million units in 2024, opens vast opportunities for local chip designers like Alibaba and Cambricon Technologies. SMIC's advanced 7-nanometer node capacity — set to exceed 20,000s wafer a month by the end of 2025 — should be enough to capture that redirected demand. Assuming a 650-square-millimeter die and a prudent yield estimate of 30%, we calculate that SMIC needs only 11,300 wafers a month to meet domestic AI accelerator needs.

Section 7. Regulations, Geopolitics

US Restrictions Provide Opportunities for China

The latest US export restrictions extend beyond China, limiting the sale of advanced AI chips to key US allies such as Israel, Saudi Arabia and the United Arab Emirates. Microsoft and Amazon, which operate AI data centers in restricted regions, warn that denying allies access to high-end AI accelerators could turn their business toward China. Meanwhile, China's efforts are advancing, with Huawei's Ascend and DeepSeek's AI infrastructure gaining traction. If US allies turn to China's semiconductor ecosystem, Nvidia's global AI accelerator dominance could erode.

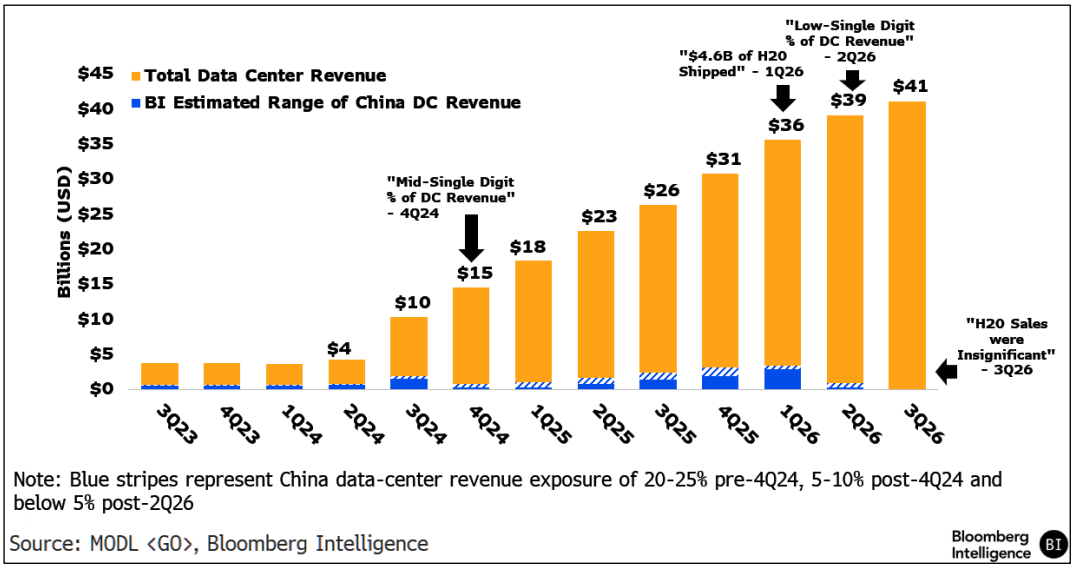
7.1 Nvidia Grapples With Limits in $50 Billion China Market

Expanding US restrictions meanwhile are curbing Nvidia's and AMD's access to China. Changes to export restrictions and additional actions are likely to hasten adoption of Chinese substitutes like Ascend, while reinforcing sovereign AI and custom ASIC investment outside the US.

Nvidia recorded no data-center shipments of its H20 chips in the second quarter though licenses were granted late in the period. Even if the pace picks up, uncertainty around Beijing policy for state entities to avoid US chips clouds visibility. While US hyperscalers and consumer internet companies drove early Blackwell gains, China – framed by Nvidia CEO Jensen Huang as a $50 billion market – remains capped, even with persistent demand from Alibaba, ByteDance and Tencent (Figure 23). Nvidia is lobbying for Blackwell to gain access in China, yet near-term revenue depends on licenses converting to orders.

BI **Nvidia's opportunity in China is limited by restrictions from Beijing and Washington**

Figure 23: China Data-Center Revenue



Note: Blue stripes represent China data-center revenue exposure of 20-25% pre-4Q24, 5-10% post-4Q24 and below 5% post-2Q26

Source: MODL <GO>, Bloomberg Intelligence

Bloomberg Intelligence BI

China's data-center sales, once bolstered by H20 shipments, have drawn the most attention, yet other segments are key to Nvidia's footprint in the region. In fiscal 2025, automotive, professional visualization and "other" accounted for more than a quarter of the company's China revenue, up

from around 11% the previous year. Gaming also remains resilient, supported by adoption of RTX chips and demand for the DIY server.

7.2 Huawei Is Positioned to Benefit from US Chip Embargo

China's Huawei has been a primary beneficiary of the US export controls, which boosted demand for its AI accelerator chips, including the Ascend 910B, marketed as a replacement for Nvidia's H100, which had dominated the market. Though Huawei's chips business is constrained by low production yields on its 7-nanometer fabrication process, rising demand for domestic chips has helped ease the impact of the US controls. The performance of Huawei's AI components lags behind Nvidia's by two to three generations, but industry analysts believe Huawei's Ascend 910C can outperform Nvidia's planned China-specific H20 AI accelerator chip.

Rescinding the Biden-era AI diffusion rule would be a tailwind for GPU makers. The regulation has limited shipments to Tier 2 markets, which accounted for 24% of Nvidia's fiscal 2025 sales, slashing the UAE's allocation by fourfold. We calculate that under a revised framework, the UAE and Saudi Arabia could account for $10 billion to $15 billion in annual sales for Nvidia through 2027 and $1 billion to $2 billion for AMD, cushioning the blow from China curbs. Nvidia is barred from selling Hopper, Blackwell and the H20 in China, implying a $14 billion to $18 billion reduction in revenue in 2025.

7.3 Concentrated Manufacturing Exposes Geopolitical Risks

BI

Over 80% of the world's high-performance AI accelerators rely on TSMC and Samsung

AI chip manufacturing is highly concentrated with geopolitically sensitive Taiwan and South Korea dominating advanced-node production and packaging. Over 80% of the world's high-performance AI accelerators rely on Taiwan's TSMC and South Korea's Samsung Electronics for fabrication, while production of key packaging technologies like high-bandwidth memory and chip on wafer on substrate assembly is similar geographically. With escalating US-China tensions and Taiwan's central role in AI chip production, the risk of supply chain disruption remains a major concern for hyperscalers and AI chip vendors.

A US plan to take nearly a 10% stake in Intel does little to address execution risks. Yet the move, converting CHIPS Act grants into equity, may help the US encourage fabless chipmakers to use Intel's capacity.

Intel is pulling back on greenfield investments, canceling production plans in Germany and Poland and slowing down in Ohio. CEO Lip-Bu Tan said previous expansion was excessive and that capital spending would be deployed only in lockstep with committed customer demand. Arizona and Oregon remain the focus for making its 18A and 14A chips, supported by existing facilities in Israel, Ireland and Malaysia. The company also plans to consolidate Costa Rica test operations in Vietnam and Malaysia. With tariffs and US policy favoring domestic capacity, Intel's longer-term footprint could help customers diversify their supply chain toward the US. Still, a lack of substantial external demand for the 18A or the 14A weighs on the foundry's near-term narrative.

TSMC's disciplined annual technology upgrades reinforce its dominant position, especially as Intel and Samsung struggle to launch compelling alternatives. Samsung is grappling with yield challenges in 3-nanometer gate-all-around architecture and potential conflicts of interest as a foundry as well as a chip designer. Intel's 18A node, while featuring backside power delivery for

better performance efficiency, still lags behing peers in SRAM density and yield. In contrast, TSMC's N2 node features a 35% power saving over current 3-nanometer nodes, which is crucial for power-hungry AI chips. The improved node density also addresses the pressing need for higher computational efficiency.

7.4 Tariff Challenges Appear Manageable for AI Providers

BI **Production in Mexico and the US softens the effect of tariffs**

The impact of tariffs might be more pronounced for networkers than for AI server vendors, but supply-chain flexibility may help the group navigate around the levies. Nvidia's AI systems could be partly insulated. Most GPUs aren't sold as stand-alone semiconductors, shipping in complex systems packed with nonchip components, many assembled abroad and possibly subject to levies. But a large share of those can qualify for exemptions under the US, Mexico, Canada trade agreement if built in Mexico or Canada. We believe that 40-60% of Nvidia's US-bound system imports could fall into that category.

AI-server original equipment manufacturers, which integrate Nvidia GPUs into full racks, also have significant production capacity in Mexico and the US, further blunting the direct impact from duties. That said, as much as half of Nvidia's components, like switches and networking gear, are built in Asia, primarily Taiwan, and may remain exposed even if final assembly occurs in the US.

AI chipmakers — particularly Nvidia, Broadcom, Marvell and, to lesser extent, Intel and AMD — should face limited demand disruption from tariffs. Their enterprise buyers, especially hyperscalers and cloud service providers, have low price sensitivity and are committed to AI infrastructure expansion. Nvidia, AMD, Marvell, Broadcom and Astera Labs have reported no distortion but were modeling outcomes and monitoring trade policy.

Section 8. Supply Chain

Chip Supply Remains Tied to Taiwan as US Expands

AI chip growth is inseparable from advanced-node wafer capacity, which remains concentrated in Taiwan, with its 3- to 5-nanometer production supporting Nvidia's Blackwell and AMD's MI300/MI400 series. The US has a minority of the market, with expansion plans unlikely to ease constraints before 2026. Though China leads in overall wafer volume, it lacks capacity for sub-5-nm chips due to export controls.

8.1 Taiwan Anchors Output of Critical Next Gen Chips

Taiwan is on pace to retain more than 50% of nodes below 5 nanometers through 2026, with the US targeting 18% as Intel's new fabs ramp up. Leading-edge capacity, below 5 nanometers, increasingly is earmarked for GPUs and custom ASICs, with Taiwan and the US controlling nearly all usable wafer starts. The US footprint in leading-edge nodes, at 14-17%, underscores progress in onshoring but remains small next to Taiwan's 87% for 3-5 nanometers.

BI **Most advanced-node capacity is concentrated in Taiwan**

The concentration of advanced-node capacity in Taiwan keeps AI chip supply chains vulnerable. TSMC's N4/N3/N2 plans will support Nvidia's Blackwell and AMD's MI300/MI400 series, with more than 95% of the manufacturer's output still tied to Taiwan despite continued expansion in Arizona. TSMC's installed sub-7-nm capacity could rise 28% by 2026, while its US fabs may grow more than fourfold, albeit from a small base. Intel's plans for its 18A chip remain challenged by delays and limited customer traction, though capacity is poised to expand 18% as the US could encourage more fabless chipmakers to commit to orders. Diversification initiatives aside, Taiwan should retain more than 50% of global supply of nodes below 7 nanometers.

China's wafer expansion is outpacing that of other countries, expanding at 23% — compared with 7-9% for Taiwan and the US — to secure roughly one-third of global capacity. Yet about 98% of China's output is for legacy nodes of greater than 10 nanometers, limiting its relevance for AI accelerators. The country doesn't produce nodes below 5 nanometers due to trade restrictions on extreme ultraviolet lithography machinery, leaving China's AI demand reliant on foreign GPU and ASIC imports and keeping domestic accelerator designers like Huawei years behind Nvidia and AMD. Alibaba's Zhenwu and Baidu's Kunlun P800 chips underscore a shift toward homegrown solutions, but they lag in performance and depend on constrained foundries such as SMIC. Export controls have also forced repeated redesigns of Chinese chips to meet TSMC compliance, slowing development.

While Chinese R&D in chiplet integration and inference-focused designs may ease near-term dependence, the front-end manufacturing gap is likely to persist through the decade, keeping China at a disadvantage in AI infrastructure.

Section 9. Performance & Valuation

AI Chipmakers Outpace Broader Group Despite Volatility

AI chipmakers and networking leaders outperformed the broader semiconductor group in 2025 – just trailing makers of high-bandwidth memory – driven by hyperscaler demand, sovereign customers and system-level gains in average selling prices. While tariff and export risks caused sharp drawdowns early in the year, a rebound since spring lifted most AI-linked stocks to top the Philadelphia Semiconductor Index benchmark. The group also sustained its premium to traditional semiconductor stocks and the broader market, largely supported by Nvidia's and Broadcom's high sales growth trajectories and strong margins.

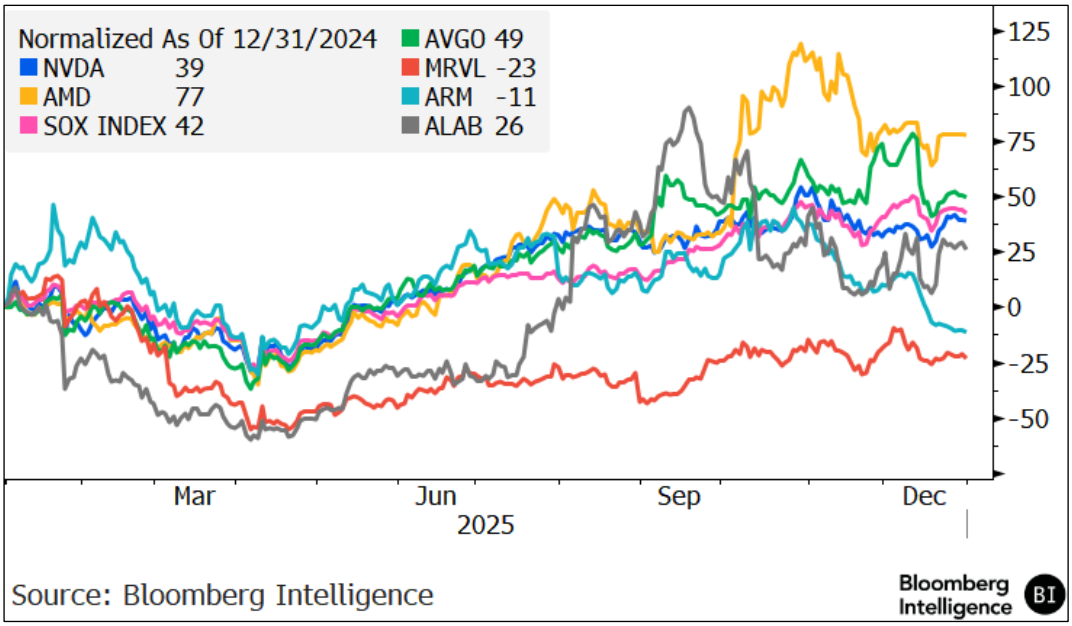
9.1 Performance: AI Chips' Lead Bolstered by AMD, Broadcom

Last year highlighted the resilience of AI-linked chipmakers amid market volatility. While most AI semiconductor stocks bested their non-AI counterparts, companies like AMD, climbing 77% in 2025, and Broadcom, up 49%, led the performance race on expectations that sales are poised to inflect in the second half of 2026 and the first half of 2027. Investors rewarded companies that demonstrated growth above near-term projections.

Marvell was an outlier, dropping 23% on inconsistent ASIC expansion by AWS and a slower-than-expected Microsoft ramp-up. The difference underscores that investors focused on shifts in market share, government action and which players will benefit most from increased AI infrastructure build-outs and system-level revamps.

BI **Investors punished Marvell because of fears of inconsistent ASIC adoption at AWS**

Figure 24: AI Chip Stock Performance



9.2 Valuation: GPUs, ASICs, Networking Command Premiums

AI chipmaker valuations remained above sector averages, though the variation widened. Nvidia's stock traded at 24 times projected calendar 2026 earnings, below its five-year average, largely due to the company's scale and the probability that its high-double-digit growth would slow. AMD, at 34x, was above the stock's long-term average but carried upside possibility from rack-scale adoption of its Helios platform.

Custom ASIC and interconnect companies continued to fetch premiums. Broadcom traded at 30x, compared with an 18x five-year average, supported by hyperscaler ASIC and Ethernet design wins. Astera Labs remained elevated at 72x, indicating investor willingness to pay for early-stage leadership in CXL/PCIe and UALink fabrics. Marvell lagged behind peers, at 27x, amid uncertainty over its share of AWS' business and an intermittent ramp-up for Trainium related designs.

Figure 25: Key Sales, Profit, Margin, Valuation Metrics

		Revenue Growth			EPS Growth			Gross Margin			P/E	
	Companies	2025E	2026E	2027E	2025E	2026E	2027E	2025E	2026E	2027E	2026E	2027E
AI Peers	NVIDIA Corp	64%	53%	28%	57%	65%	28%	71%	75%	74%	24x	19x
	Broadcom Inc	25%	57%	30%	40%	57%	30%	78%	74%	72%	30x	23x
	Advanced Micro Devices Inc	32%	32%	40%	20%	64%	57%	52%	54%	54%	34x	22x
	Marvell Technology Inc	42%	22%	28%	80%	26%	33%	59%	59%	57%	25x	19x
	Arm	26%	21%	19%	-3%	25%	28%	98%	98%	98%	55x	43x
	Astera Labs	110%	45%	38%	107%	34%	35%	76%	73%	72%	72x	53x
	Average	38%	37%	29%	39%	47%	35%	72%	72%	71%	34x	25x
	Median	32%	32%	28%	40%	57%	30%	71%	74%	72%	30x	22x
Broad Based	Average	5%	11%	13%	-4%	32%	36%	53%	54%	56%	28x	21x
	Median	10%	9%	10%	-10%	21%	25%	55%	57%	59%	27x	20x
Handset	Average	5%	1%	8%	5%	-3%	19%	49%	49%	49%	16x	13x
	Median	5%	2%	4%	6%	3%	20%	47%	47%	47%	15x	13x
Mid Cap	Average	5%	10%	12%	38%	30%	33%	47%	48%	51%	29x	21x
	Median	-1%	7%	10%	7%	22%	31%	47%	47%	52%	27x	17x
Small Cap Peers	Average	4%	15%	13%	4%	38%	28%	51%	52%	58%	34x	23x
	Median	5%	17%	16%	-1%	33%	18%	48%	51%	53%	33x	20x
SOX	SOX Average	5%	9%	11%	23%	25%	27%	46%	47%	49%	27x	19x
	SOX Median	0%	8%	9%	3%	18%	26%	47%	47%	48%	23x	16x

Note: Periods represent calendar years, not fiscal years. Valuation Data as of January 5th, 2026
Note: Negative EPS and P/E above 100X are considered N.M.

Source: Bloomberg Intelligence

Bloomberg Intelligence BI

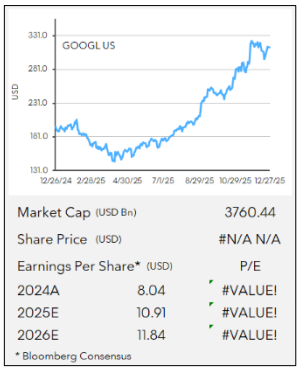
The Philadelphia Semiconductor Index rebounded to a high of 26x from its 2025 low of 17x as AI demand proved resilient and broadened beyond Nvidia. Compression early in the year tied to tariff concerns proved temporary, with computing revenue forecasts revised to a 20% CAGR through 2028 from 9.5% from third-party analysts such as IDC.

Section 10. Companies to Watch

Sales Growth Is Broad, Yet Most Profits Flow to Select Few

The AI accelerator market will be large enough to increase revenue for all major vendors, but the concentrated leadership of Nvidia, Broadcom and AMD means they're poised to capture the lion's share of profits. Smaller companies face greater execution risks as hyperscaler customers consolidate around rack-scale systems and custom ASICs.

10.1 Alphabet Leverages First-Mover Advantage in ASICs



Market Cap (USD Bn)		3760.44
Share Price (USD)		#N/A N/A
Earnings Per Share* (USD)		P/E
2024A	8.04	#VALUE!
2025E	10.91	#VALUE!
2026E	11.84	#VALUE!

* Bloomberg Consensus

1,300 Trillion
Tokens processed monthly

$155 Billion
Google Cloud backlog

Outlook: Alphabet's strong growth for its Gen AI products and cloud segment shows a higher capex return than hyperscale peers like Meta. Traction with AI Overviews and AI Mode has reduced competitive risks from LLM search peers, such as ChatGPT. We believe the company's TPU chips and vertical integration support its cost efficiency compared with other foundational models. Guidance for capex in 2025, up about $10 billion from its previous target, bodes well for the cloud segment, where mid-30% growth seems sustainable. A 30-40% increase in 2026 capex is in line with Meta and could help sales growth acceleration for Gemini, Cloud and Waymo expansion as it expands to new cities.

AI Accelerator Impact: Alphabet has a first-mover advantage in AI ASICs, with its TPUs in their sixth generation. The dynamic is likely to further drive growth in units and average selling prices in our model, as Alphabet's newer LLMs with higher parameters, rising inference tokens due to Gemini and internal AI monetization, and its efforts to sell TPU-based computing capacity to external customers like Anthropic gain steam. In addition, its cloud service, GCP, will likely continue to adopt Nvidia's GPU solutions.

10.2 Amazon Web Services Is No. 2 Contributor to ASIC Volume

Market Cap (USD Bln)		2,417.16
Share Price (USD)		230.82
Earnings Per Share* (USD)		P/E
2024A	5.53	41.74
2025E	7.09	32.54
2026E	7.93	29.12

* Bloomberg Consensus

>$100 Billion
Annual capex beginning in 2026

$128 Billion
AWS sales, 2025

Outlook: Amazon's long-term growth potential is intact as it gains market share for each vertical. Margin can expand despite near-term hurdles from tariff-related costs and capital spending to support AI investment. In retail, convenience, automation and value will be pivotal to maintaining its lead and driving continued double-digit gains in ad revenue. The company's ability to build cloud capacity and hardware for AI will likely determine AWS' near-term growth and could accelerate customer shift to the public cloud. Amazon's push for pharmacy and grocery share are large undertakings that we'll monitor closely.

AI Accelerator Impact: AWS is now the second-largest contributor to global ASIC unit volume, with Trainium deployments directly reinforcing Marvell's custom-silicon revenue pipeline and driving a significant share of the ASIC total addressable market. The operation's dual-track architecture — mixing Trainium-based inferencing with Blackwell GPU-based training — raises total accelerator demand across ASICs and GPUs, making AWS a structural driver of modelwide unit assumptions. As AWS pushes Trainium for external customer adoption with lower inferencing cost-per-token than GPUs, ASIC shipments will continue rising, even as Nvidia GPUs remain indispensable for large-scale training.

10.3 AMD Set to Enter Vertically Integrated Systems

Market Cap (USD Bln)		359.72
Share Price (USD)		214.16
Earnings Per Share* (USD)		P/E
2024A	1.00	214.16
2025E	2.45	87.56
2026E	4.51	47.51

* Bloomberg Consensus

14%
Projected share of GPUs for training and inferencing, 2030

>$28,000
GPU expected average selling price, more than doubling

Outlook: AMD's near-term growth remains rooted in strong CPU demand for cloud, enterprise and premium PCs, with EPYC and Ryzen average selling prices rising as the company gains market share. Hyperscalers' planned CPU build-outs into 2026 will support continued data-center strength, while the MI355X ramp-up stays on track. Major GPU upside likely awaits the MI450 and Helios rollouts in the second half of 2026. With expanding CPU share and rising AI adoption, AMD remains positioned for steady growth this year and an inflection as Helios deployments increase.

AI Accelerator Impact: AMD has gained share with its MI300/350 series and is preparing to scale up with Helios racks this year, marking its entry into vertically integrated systems. Acquisitions like ZT Systems expand its design capabilities, while ROCm gains traction as an open alternative to CUDA. AMD's outlook is improving as it makes inroads with Tier 2 cloud and enterprise buyers, though execution risk remains. The company's shift to rack-scale accelerators positions it as the main credible second source behind Nvidia of GPUs for training and inferencing, with its unit share modeled to reach about 14% by 2030. Even with modest market share gains, average AI GPU average selling prices rising to at least $28,000 by decade's end from $12,000 in 2024 should increase margins as memory, interconnect and rack integration lift dollar content.

10.4 Broadcom Dominance of ASIC Design Has Staying Power

Market Cap (USD Bln)		1,649.78
Share Price (USD)		346.10
Earnings Per Share* (USD)		P/E
2025A	4.77	72.56
2026E	7.68	45.06
2027E	11.42	30.31

* Bloomberg Consensus

Outlook: Broadcom's AI visibility widened meaningfully with a $73 billion, six-quarter backlog powered by Anthropic's additional $11 billion order, the third quarter's $10 billion in TPU racks and a newly converted fifth XPU customer. The setup lifted fiscal 2026-27 AI revenue potential above previous expectations, even as OpenAI shifted toward a 2027-29 contribution. System-level rack delivery boosts dollar content and entrenches Ethernet leadership. Though mix pressure weighs on gross margin, VMware and broader infrastructure software provide operating-margin stability.

60-80%
AI ASIC market share

>$60 Billion
Guidance for AI serviceable available market by fiscal 2027

AI Accelerator Impact: Broadcom dominates hyperscale ASIC design, with leadership for Google's TPU, Meta's MTIA and OpenAI's XPU. Broadcom's Ethernet switching and copackaged optics portfolio further anchor its role in AI infrastructure. It targets a $60 billion to $90 billion AI serviceable available market by fiscal 2027, supported by XPU and networking attach rates. The company is the clearest beneficiary of the shift to custom silicon, controlling 60–80% of AI ASIC share and expanding networking bill-of-materials content. We model its AI revenue growing around a 60% compounded annual rate through 2028, supported by TPUs, MTIAs and XPUs, as well as Ethernet/CPO attach.

10.5 Marvell Follows Rapidly in ASICs, Networking

Market Cap (USD Bln)		75.80
Share Price (USD)		84.98
Earnings Per Share* (USD)		P/E
2025A	-1.02	-
2026E	2.97	28.57
2027E	1.92	44.27

* Bloomberg Consensus

Outlook: Marvell raised its outlook for fiscal 2027-28 revenue on sharper data-center visibility, with optical, switching and expanding XPU-attach content driving growth above capital spending. Strong purchase orders from AWS and next-generation custom programs support confidence into 2027, while the following year should benefit from layered-in orders for Microsoft's Maia and more attach ramp-ups. The Celestial AI deal adds longer-dated upside to the scale-up strategy, strengthening Marvell against Broadcom. Networking breadth and accelerating 2028 custom contributions increasingly suggest that its multiyear AI trajectory can be maintained.

20-25%
Potential ASIC market share

>40%
Expected CAGR for broad AI portfolio through 2028

AI Accelerator Impact: Marvell has secured key design wins for AWS' Trainium2 and Microsoft's Maia while expanding its XPU and connectivity attach rates. Yet execution has been intermittent and pressure from Broadcom and Alchip remains intense. Marvell's modular IP portfolio, including HBM customization, could set it apart in future generations. The company's exposure to hyperscale ASIC programs cements Broadcom as the No. 2 custom silicon provider, with a potential for 20–25% of the market. And its path to scale up is clear: AWS's Project Rainier calls for about 100,000 Trainium2 chips, and we model Marvell's broader AI portfolio at a CAGR above 40% through 2028, with Teralynx Ethernet and Octeon DPUs providing a second growth lever even as margins trail those of Broadcom.

10.6 Microsoft Poised to Remain Biggest Driver of GPU Growth

Market Cap (USD Bln)		3,515.07
Share Price (USD)		483.62
Earnings Per Share* (USD)		P/E
2025A	13.64	35.46
2026E	15.96	30.30
2027E	18.74	25.81

* Bloomberg Consensus

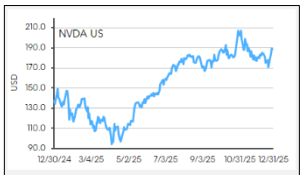
$162 Billion
Capital expenditure, 2026

No. 1
GPU customer

Outlook: Microsoft's aggressive AI investments are starting to yield results, positioning the company to expand its share of the cloud infrastructure market over the next 12-24 months. Though OpenAI has been instrumental in powering some of its software products, Microsoft's vast distribution arm will likely play a more critical role, fueling mid- to high-double-digit sales growth in fiscal 2026. Lower-margin AI workloads may squeeze gross margin in the near term, but we expect management to control expenses to limit any impact on adjusted operating margin. Capital spending, including finance leases, could climb 61% this year to $142 billion.

AI Accelerator Impact: Microsoft is Nvidia's largest GPU customer and is set to remain the biggest single contributor to GPU unit growth as the primary cloud provider for OpenAI. With AI infrastructure capex projected to exceed $100 billion annually starting this year, Microsoft will be a major driver of demand for Blackwell and Rubin rack-scale systems. Beginning in 2027, the company's Maia road map should also lift ASIC volumes, benefiting Marvell as hyperscalers diversify computing for large-scale inferencing.

10.7 Nvidia Maintains Firm Grip on GPU Market for Training



Market Cap (USD Bln)		4,589.06
Share Price (USD)		186.50
Earnings Per Share* (USD)		P/E
2025A	2.94	63.44
2026E	4.61	40.48
2027E	7.34	25.42

* Bloomberg Consensus

70%
Projected share of training market through decade

>$2.8 Million
Rack-scale platform anchor price

Outlook: Nvidia's fiscal third-quarter results reinforced that there's strong demand for AI infrastructure as its GB300 drives most data-center revenue and momentum builds beyond hyperscalers. The company's $500 billion Blackwell and Rubin pipeline could expand, given national deals in the Middle East and Anthropic's first large-scale use of Nvidia architecture. With Rubin set to ramp up in the second half of this year at higher average selling prices, and additional sovereign, enterprise and AI lab projects, Nvidia's growth should continue through 2027.

AI Accelerator Impact: Nvidia remains the dominant supplier of GPUs for AI training and, increasingly, for inferencing, with Blackwell and the coming Rubin racks extending its leadership. The transition from selling chips to full systems through the NVL72, DGX and GB300 has raised its average selling prices and gross margin, while CUDA and NVLink Fusion deepen its software and ecosystem lock-in. Sovereign AI and enterprise expansion diversify demand beyond hyperscalers. The company's control of the full AI stack — GPUs, memory, networking, software, and systems — secures its position in training, where we model it having about 70% of the market through 2030. That reinforces its bulwark against ASICs and merchant peers. Nvidia's NVL72 platforms set the industry's anchor price at $2.8 million to $3.1 million a rack, shaping the mix for bill of materials and valuations across AI semiconductors.

10.8 OpenAI Seen Driving Over $1 Trillion in Accelerator Capex

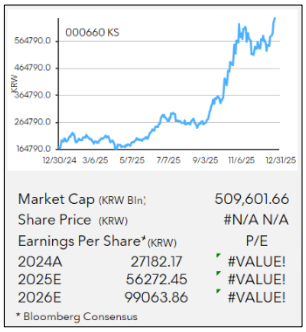
6 Billion
Tokens per minute for application programming interface

$200 Billion
Expected 2030 revenue

Outlook: OpenAI's expectation for 70% revenue growth through 2030 from its $13 billion run rate hinges on its LLMs becoming the intelligence layer for applications including supply chain, engineering, customer relationship management and enterprise resource planning. Recent alliances with Oracle, Nvidia and Broadcom highlight its focus on cloud capacity expansion and GPU supply, while reducing reliance on Microsoft. Given OpenAI's aggressive infrastructure expansion, we believe it will leverage competitive pricing to boost consumption and sustain its market-share lead over Anthropic, Google Gemini and Meta.

AI Accelerator Impact: As a leading frontier LLM provider, OpenAI continues to drive significant AI accelerator capex spent through cloud service providers like Microsoft and Oracle. OpenAI itself could drive more than $1 trillion of capex spent on the back of Stargate commitments and its large deals with Nvidia, AMD and Broadcom to implement more than 30 gigawatts of capacity through 2030. Its plan to ramp up its own ASIC with Broadcom will drive higher unit volume alongside sustained momentum for growth in GPU unit volume.

10.9 SK Hynix Sales Can Expand on AI Memory Chips



Market Cap (KRW Bln)		509,601.66
Share Price (KRW)		#N/A N/A
Earnings Per Share* (KRW)		P/E
2024A	27182.17	#VALUE!
2025E	56272.45	#VALUE!
2026E	99063.86	#VALUE!

* Bloomberg Consensus

55-60%
Global market share for high-bandwidth memory

50%
Sales CAGR over next three years

Outlook: SK Hynix's sales growth could continue into this year, powered by demand for HBM and NAND chips for AI servers. HBM chips' higher operating profit margins support solid gains in operating profit. High-performance DRAM for AI can boost sales further, especially if it garners wide adoption from hyperscale data centers. AI servers could require three times more NAND chips for data storage than general servers, boding well for the company's NAND sales. Demand for its solid-state-drive solutions for enterprise customers could increase due to their high quality and advanced technology. A potential demand recovery for smartphones and PCs could accelerate SK Hynix's sales growth further in 2026.

AI Accelerator Impact: SK Hynix's annual sales growth can average about 50% over the next three years, as HBM4 adoption begins in 2026, aided by higher average selling prices tied to 12- and 16-high stacks and increased interface bandwidth. After leading development of HBM3 and HBM3E, the company should maintain its top market share in HBM4 as well. Backed by its strong HBM technological capabilities, SK Hynix has established close relationships with GPU suppliers and AI ASIC customers, positioning itself favorably for next-generation product development. HBM4 and HBM4E will introduce larger die sizes and higher through-silicon via counts, raising wafer and packaging intensity while increasing AI GPU bill-of-material costs. With GPUs and AI ASICs absorbing more HBM per device, SK Hynix's execution will be central to maintaining the pace of accelerator performance scaling.

10.10 TSMC Advanced Packaging Key to Next-Gen Accelerators

Market Cap (TWD Bln		43,307.32
Share Price (TWD)		1,550.00
Earnings Per Share* (TWD)		P/E
2024A	45.25	34.25
2025E	64.37	24.08
2026E	76.76	20.19

* Bloomberg Consensus

20%
Projected sales CAGR, 2026-27

No. 1
Advanced foundry status

Outlook: TSMC is a key enabler of the AI-chip cycle, reinforcing its leadership in advanced nodes (2-5 nm) and 2.5D/3D packaging. Despite headwinds like new US tariffs on global semiconductor demand, TSMC is likely to sustain 2026-27 sales growth of around 20% and outpace foundry peers. The 2-nm process node is the company's next big catalyst, getting broad first-wave adoption not just from smartphones, but also from AI GPUs and networking chips.

AI Accelerator Impact: Driven by the dual mandates of computing speed and energy efficiency, we expect that AI silicon architects will join mobile chips as the lead adopters of TSMC's leading-edge nodes starting in 2026. Further, TSMC's advanced packaging road map, specifically for CoWoS and COUPE, remains the linchpin for scaling die size and maximizing memory bandwidth in next-generation accelerators.

Section 11. Methodology

Our market-sizing model forecasts the AI accelerator market through 2033, covering GPUs and ASICs used for accelerated workloads in servers. The model incorporates bottom-up approaches to reach our projected drivers and top-down approaches to validate our assumptions for demand and supply. It will be updated regularly as vendor disclosures, hyperscaler capex plans and pricing dynamics evolve. The analysis projects revenue based on expectations for AI chip unit growth, system-level content expansion and average selling prices, along with capacity constraints in memory, packaging and networking that shape real-world deployment limits. Accelerator demand is modeled primarily around data-center use cases, where chips are deployed for training, inferencing and, increasingly, for agentic and reasoning workloads.

AI Chip Forecast: We forecast AI accelerator unit growth based on rising adoption of compute-intensive AI workloads that require high-bandwidth memory, advanced packaging and integrated networking. Historical unit demand is derived from data-center compute revenue where disclosed, paired with assumed chip average selling prices calibrated against vendor price ranges, system original-equipment-manufacturer pricing and reported hyperscaler spending. We use hyperscaler commentary, cloud capex disclosures and third-party infrastructure estimates as top-down validation checks. We assume that all accelerator design cycles are on an annual cadence, reflecting aggressive road-map compression and synchronized generational refresh cycles across GPUs and hyperscaler ASICs.

Content per Accelerator: Our scenario forecasts system-level content per accelerator using known configurations and projected increases in functional blocks that drive total system value. These include memory, board level components and networking, which are priced into GPU average selling prices and future possible pass-through content revenues. We don't include explicitly switching related revenues, which are typically captured in networking segments of GPU and ASIC vendors. Our assumptions include all compute-related revenue components like advanced packaging and liquid cooling, which typically get bundled in GPU/ASIC sales. We use historical high-bandwidth memory and interconnect trends as a starting point, noting that memory bandwidth and interconnect density remain among the largest constraints for model size and inference cost.

Average Selling Prices: We project average selling prices using blended pricing across GPU and ASIC architectures, informed by vendor disclosures, original-equipment-manufacturer system pricing, hyperscaler procurement data and industry cost trends. GPU pricing reflects rising material intensity from HBM, interposers and liquid cooling, as well as the shift to rack-scale systems where the accelerator vendor captures a larger share of the bill of materials. ASIC pricing is modeled using historical cost structures of TPU, Trainium and MTIA-class devices, adjusted for process-node transitions and HBM attach.

Supply and Capacity Forecast: Our supply modeling incorporates constraints in advanced-foundry capacity, high-bandwidth memory availability and advanced packaging. These factors determine practical ceilings on accelerator shipments, similar to how capacity and yields shape supply in memory and packaging markets. From 2027–30, we assume that packaging and memory suppliers expand capacity sufficiently to ease but not eliminate supply tension, enabling midcycle pricing deceleration as yields improve.

Market Share Forecast: Market shares are validated against hyperscaler capex allocation patterns, vendor guidance and historic internal silicon adoption curves.

Scenario Analysis and Validation: We use base, slow-growth and fast-growth scenarios varying unit growth, system content and average selling price assumptions across accelerator types. Sensitivities are anchored in model-size trajectories, token consumption growth, power availability, rack-scale deployments and shifts in spending for training versus inferencing. Model outputs are validated against hyperscaler forward capex disclosures, AI data-center build-out data, OEM backlog commentary, GPU procurement notes, industry forecasts from IDC/Omdia and long-term computing demand assumptions. We incorporate bottom-up revenue checks using vendor-reported data-center sales where available, ensuring alignment between units, pricing curves and real-world spending.

Bloomberg Intelligence Research Coverage

Bloomberg Editorial and Research:
John Micklethwait, Editor-in-Chief; **Reto Gregori**, Deputy Editor-in-Chief

Research Management

David Dwyer, Global Director of Research
Drew Jones, Deputy Global Director of Research
Sam Fazeli, Director of Global Industry Research
Alison Williams, Director of Global Strategy Research
Paul Gulberg, Director of Americas Industry Research
Catherine Lim, Director of APAC Industry Research
Sue Munden, Director of EMEA Industry Research
Alexandra Davidov, Associate Director of Global Industry Research
Lindsey Houghton, Associate Director of Global Strategy Research
Frank Jacobs, Global Chief Operating Officer

Content Management

Tim Craighead, Global Chief Content Manager
Karima Fenaoui, Research Content Manager, Communities & EM
John Lee, Research Content Manager, APAC
Renato Prieto, Research Content Manager, FICC
Justin Spitzer, Cross-Asset Strategy and Credit
Roger Thomson, Research Content Manager, Americas
Rod Turnbull, Research Content Manager, EMEA
Mariam Traore, Research Digital Content Specialist
Matthew Bloxham, Global Head of Alternative Data & Analytics
Brian Egger, Global Head of Financial Modeling
Donna Weston, Co-head Global Research Editorial
Douglas Zehr, Co-head Global Research Editorial

Equity Strategy

Markets

Christopher Cain, Quantitative Analysis, US
Michael Casper, Small Caps and Sectors, US
Nitin Chanduka, Emerging Markets, India
Marvin Chen, China and North Asia
Laurent Douillet, Europe
Anthony Feld, Technical Analysis, Global
Kumar Gautam, Emerging Markets, Global
Jennie Li, Equity Strategy, Americas
Andrew Martynov, Equity Strategy, EMEA
Kaidi Meng, Equity Strategy, Europe
Wendy Soong, Americas
Sufianti, Emerging Markets, ASEAN
Gillian Wolff, Global

Funds

Eric Balchunas, Exchange Traded Funds, Global
David Cohne, Mutual Funds, Global
Henry Jim, Exchange Traded Funds, Europe
Athanasios Psarofagis, Exchange Traded Funds, Americas
James Seyffart, Exchange Traded Funds, Americas
Rebecca Sin, Exchange Traded Funds, APAC

Thematic

Breanne Dougherty, Global
Andrew Silverman, Tax Policy and Corporate Actions, Global
Shirley Wong, APAC

FICC Strategy

Noel Hebert, Director of FICC Strategy Research

Markets

Erica Adelberg, MBS, Americas
Basel Al-Waqayan, Fixed Income Strategist, MENA
Reto Bachmann, Chief Securitization Strategist, Global
Negisa Balluku, Litigation-Bankruptcy, Americas
Mahesh Bhimalingam, Credit Strategy, EMEA
Philip Brendel, Distressed Debt, Americas
Rod Chadehumbe, ABS, Americas
Sam Geier, Credit Strategy, Americas
Noel Hebert, Credit Strategy, Americas
Jason Lee, Credit Strategist, Asia
Mike McGlone, Commodity Strategy, Global
Tanvir Sandhu, Derivatives, Global
Damian Sassower, Emerging Markets, Global
Timothy Tan, Credit Strategy, APAC

FX and Rates

Audrey Childe-Freeman, FX, G-10
Stephen Chiu, FX and Rates, APAC
Ira Jersey, Rates, US
Davison Santana, FX and Rates, LatAm
Sergei Voloboev, FX, Emerging Markets, Global
Huw Worthington, Rates, EMEA

ESG Research

Eric Kane, Director of ESG Research

Shaheen Contractor, Global
Rob Du Boff, Governance, Global
Gail Glazerman, Integration, Americas
Yasutake Homma, Japan
Eric Kane, Global
Grace Osborne, Integration, EMEA
Christopher Ratti, Fixed Income, Global
Andrew John Stevenson, Climate, Global
Conrad Tan, Integration, APAC

Market Structure Research

Larry Tabb, Director of Market Structure Research

Jackson Gutenplan, Equities, Americas
Brian Meehan, Fixed Income, Americas
Nicholas Phillips, Equities, EMEA
Larry Tabb, Equities and Fixed Income, Global

Credit Research

Americas

Himanshu Bakshi, Consumer Finance, Banking, Global
Mike Campellone, Specialty Apparel, Consumer Hardlines, Global
Cecilia Chan, Gaming Lodging & Restaurants, Internet Media, China
Jean-Yves Coupin, Health Care, Corporate Bonds, Americas
Spencer Cutter, Oil & Gas, Global
Stephen Flynn, Entertainment, Cable & Satellite, Americas
Matthew Geudtner, Aerospace, Global, Machinery, Americas
David Havens, Consumer Finance, Investment Mgmt, Americas
Mike Holland, Hospitals, Specialty-Generic Pharma, Americas
Julie Hung, Packaged Food, Beverages, Americas
Arnold Kakuda, Investment Banking, Americas
Jody Lurie, Gaming Lodging & Restaurants, Americas
Robert Schiffman, Hardware & Storage, Internet Media, Global

Asia

Andrew Chan, Real Estate, Infrastructure, China
Sharon Chen, Telecom Carriers, Infrastructure, India
Pri De Silva, Banking, Aerospace & Defense, APAC
Daniel Fan, Real Estate, China
Rena Kwok, Banking, APAC
Mary Ellen Olson, Metals & Mining, Oil & Gas, APAC

EMEA

Tolu Alamutu, Real Estate, Banking, EMEA
Ruben Benavides, Banking, Europe
Aidan Cheslin, Telecom Carriers, EMEA
Jeroen Julius, Banking, EMEA
Stephane Kovatchev, Industrials, Construction, EMEA, Americas
Timothy Riminton, Basic Materials, Europe
Paul Vickars, Electric Utilities, EMEA, Oil & Gas, Global

Industry Research

Consumer

Consumer Products & Services

Rami Abi-Samra, Consumer and Utilities, Middle East
Deborah Aitken, Luxury, Personal Care Products, Global
Brian Egger, Gaming & Lodging, Americas
Conroy Gaynor, Travel & Leisure, EMEA
Michael Halen, Restaurants, Americas
Drew Reading, Homebuilders, Americas

Retail & Wholesale

Charles Allen, Retail Staples & Wholesale, Specialty Apparel, EMEA
Lindsay Dutch, Consumer Hardlines, Retail REIT, Americas
Mary Ross Gilbert, Specialty Apparel Stores, Americas
Abigail Gilmartin, Athleisure & Footwear, Americas
Poonam Goyal, E-Commerce, Specialty Apparel Stores, Americas
Catherine Lim, Consumer Goods, E-Commerce, APAC
Tatiana Lisitsina, Consumer Hardlines, Online Apparel, EMEA

Consumer Staples

Jennifer Bartashus, Packaged Food and Retail Staples, Americas
Ignacio Canals Polo, Consumer Staples, EMEA
Lea El-Hage, Retail Staples, Australia
Duncan Fox, Beverages, Packaged Food, EMEA
Diana Gomes, Consumer Health, Household Products, Global
Diana Rosero-Pena, Packaged Food, Retail Staples, Americas
Kenneth Shea, Beverages, Tobacco & Cannabis, Americas

Energy

Will Hares, Energy Sector Head

Rob Barnett, Solar Energy Equipment, Americas, EMEA
Talon Custer, Oil & Gas, Americas
Vladimir Do Nascimento Pinto, Energy & Utilities, Americas
Henik Fung, Oil & Gas, Gas Utilities, APAC
Brett Gibbs, Biofuels, EMEA
Will Hares, Oil & Gas, EMEA
Scott J. Levine, Oil & Gas, Industrials, Americas
Vincent G. Piazza, Oil & Gas, Americas
Salih Yilmaz, Oil & Gas, EMEA

Utilities

Patricio Alvarez, Electric Utilities, Gas Utilities, EMEA
Nikki Hsu, Electric Utilities, Americas
Gabriela Privetera, Electric Utilities, Americas

Industrials

Steve Man, Director of Industrial Research

Automotive

Joanna Chen, Automobiles, APAC
Gillian Davis, Automobiles, EMEA
Michael Dean, Automobiles, EMEA
Steve Man, Automobiles, Americas
Tatsuo Yoshida, Automobiles, Japan

Industrial & Industrial Services

Christopher Ciolino, Machinery, Industrials, Americas
Christina Feehery, Industrials, Americas
George Ferguson, Aerospace & Defense, Global
Stuart Gordon, Business Services, Europe
Takeshi Kitaura, Industrials, Japan
Will Lee, Aerospace & Defense, Americas
Mustafa Okur, Electrical Equipment, Industrials, Americas
Wayne Sanders, Defense, Americas
Bhawin Thakker, Industrials, EMEA
Karen Ubelhart, Industrials, Machinery, Americas
Omid Vaziri, Industrials, EMEA
Denise Wong, Infrastructure, APAC
Eric Zhu, Aerospace & Defense, APAC

Transportation

Francois Duflot, Airlines, Americas
George Ferguson, Airlines, Global
Conroy Gaynor, Airlines, EMEA
Lee Klaskow, Freight Transportation & Logistics, Global
Kenneth Loh, Marine Shipping, Logistics Services, APAC

Materials

Jason Miner, Agriculture Sector Head
Grant Sporre, Metals and Mining Sector Head

Chemicals

Daniel Cole, Agricultural Chemicals, Americas
Sean Gilmartin, Specialty Chemicals, Americas
Alexis Maxwell, Agricultural Chemicals, Canada
Jason Miner, Agriculture and Chemicals, Americas
Alvin Tai, Agriculture, Malaysia, EMEA

Containers & Packaging

Ryan Fox, Packaging, Americas

Financials

Financial Services

Hideyasu Ban, Financial Services, Japan
Edmond Christou, Financials, Middle East
Ben Elliott, Consumer Finance, Americas
Diksha Gera, Global Payments and Fintech, Americas
Paul Gulberg, Investment Management, Exchanges, Banks Americas
Gabriel Gusan, Financial Services, Americas
Matt Ingram, Financial Services, Australia and Korea
Sarah Jane Mahmud, Banking, Market Structure, ASEAN and India
Neil Sipes, Investment Management, Investment Banking, Americas
Salome Skhirtladze, Financials, Middle East
Alison Williams, Investment Banking, Global
Sharnie Wong, Investment Banking, Exchanges, APAC

Banking

Francis Chan, Banking & Fintech, China & Hong Kong
Herman Chan, Banking, Americas
Tomasz Noetzel, Banking, EMEA
Philip Richards, Banking, EMEA
Lento Tang, Banking, EMEA
Mar'yana Vartsaba, Banking, EMEA

Insurance

Jeffrey Flynn, Life Insurance, Americas
Charles Graham, P&C Insurance, Life Insurance, EMEA
Steven Lam, Insurance, APAC
Matthew Palazola, P&C Insurance, Americas
Kevin Ryan, Life Insurance, P&C Insurance, EMEA

Real Estate

Ken Foong, Real Estate, Singapore
Iwona Hovenko, Real Estate, Business Services, EMEA
Kristy Hung, Real Estate, China
Jeffrey Langbaum, Residential REIT, Office REIT, Americas
Sue Munden, Real Estate, EMEA
Patrick Wong, Real Estate, APAC

Health Care

Aude Gerspacher, Director of Health Care Research

Biotech & Pharma

Sam Fazeli, Biotech, Global
Andrew Galler, Biotech, Americas
Aude Gerspacher, Pharma, Biotech, Americas
Grace Guo, Biotech, Pharma, Americas
Matt Henriksson, Medical Equipment & Devices, Americas
Glen Losev, Hospitals, Managed Care, Americas
Jamie Maarten, Biotech, China
John Murphy, Large Pharma, Biotech, Americas, EMEA
Max Nisen, Biotech, Americas
Jonathan Palmer, Medical Devices, Supply Chain, Americas
Jean Rivera Irizarry, Biotech, Pharma, Americas
Michael Shah, Biotech, Specialty-Generic Pharma, EMEA
Ann-Hunter van Kirk, Biopharmaceuticals, Americas

Metals & Mining

Richard Bourke, Basic Materials, Americas
Michelle Leung, Metals & Mining, China, Japan
Emmanuel Munjeri, Metals & Mining, South Africa
Alon Olsha, Metals & Mining, EMEA
Grant Sporre, Metals & Mining, EMEA

Construction Materials

Sonia Baldeira, Infrastructure & Building Materials, Global
Kevin Kouam, Infrastructure & Building Materials, Global

Technology

Mandeep Singh, Director of Technology Research

Hardware

Woo Jin Ho, Hardware & Networking, Americas
Ken Hui, Semiconductors, Europe
Charles Shum, Semiconductors, APAC
Jake Silverman, Logic ICs, Americas
Kunjan Sobhani, Logic ICs, Americas
Steven Tseng, EMS/ODM, Consumer Electronics, APAC
Masahiro Wakasugi, Semiconductors, EMS/ODM, Global

Software

Tamlin Bason, IT Services, Americas, EMEA
Nicole D'Souza, Internet Services & Software, Americas
Robert Lea, Internet Media, Application Software, China
Nathan Naidu, Entertainment Content, Internet Media, Japan
Niraj Patel, Application Software, Americas
Sunil Rajgopal, Application Infrastructure Software, Americas
Anurag Rana, Application Software, IT Services, Americas
Mandeep Singh, Software, Internet, Hardware/Semis, Americas

Communications

Media

Matthew Bloxham, Media, Advertising, Telecom, EMEA
Geetha Ranganathan, Entertainment, Cable, Advertising, Americas
Tom Ward, Media, Advertising, Telecom Carriers, EMEA

Telecommunications

John Butler, Telecom & Towers, Infrastructure Software, Americas
John Davies, Telecom Carriers and Media, EMEA
Erhan Gurses, Telecom Carriers, EMEA
Chris Muckensturm, Telecom Carriers, ASEAN

Litigation and Policy

Nathan Dean, Financials Policy, Americas
Holly Froum, Consumer, Industrials Litigation and Policy, Americas
Josephine Garban, Health Care Patent Litigation, Americas
Jennifer Rie, Antitrust Litigation and Policy, Americas
Matthew Schettenhelm, TMT Litigation and Policy, Americas
Elliott Stein, Financials Litigation, Americas
Justin Teresi, Antitrust Litigation and Policy, Americas
Duane Wright, Health Care Policy, Americas

Copyright and Disclaimer

Copyright

Disclaimer

The data included in these materials are for illustrative purposes only. The BLOOMBERG TERMINAL service and Bloomberg data products (the "Services") are owned and distributed by Bloomberg Finance L.P. ("BFLP") except (i) in Argentina, Australia and certain jurisdictions in the Pacific islands, Bermuda, China, India, Japan, Korea and New Zealand, where Bloomberg L.P. and its subsidiaries ("BLP") distribute these products and (ii) in Singapore and the jurisdictions serviced by Bloomberg's Singapore office, where a subsidiary of BFLP distributes these products. BLP provides BFLP and its subsidiaries with global marketing and operational support and service. Certain features, functions, products and services are available only to sophisticated investors and only where permitted. BFLP, BLP and their affiliates do not guarantee the accuracy of prices or other information in the Services. Nothing in the Services shall constitute or be construed as an offering of financial instruments by BFLP, BLP or their affiliates, or as investment advice or recommendations by BFLP, BLP or their affiliates of an investment strategy or whether or not to "buy," "sell," or "hold" an investment. Information available via the Services should not be considered as information sufficient upon which to base an investment decision. The following are trademarks and service marks of BFLP, a Delaware limited partnership, or its subsidiaries: BLOOMBERG, BLOOMBERG ANYWHERE, BLOOMBERG MARKETS, BLOOMBERG NEWS, BLOOMBERG PROFESSIONAL, BLOOMBERG TERMINAL and BLOOMBERG.COM. Absence of any trademark or service mark from this list does not waive Bloomberg's intellectual property rights in that name, mark or logo. All rights reserved. © 2025 Bloomberg.

Bloomberg Intelligence is a service provided by Bloomberg Finance L.P. and its affiliates. ("Bloomberg"). Bloomberg is not an officially recognized credit rating agency in any jurisdiction and customers should not use or rely on Bloomberg Intelligence to comply with applicable laws or regulations that prescribe the use of ratings issued by accredited or otherwise recognized credit rating agencies. Bloomberg Intelligence Credit and Company research may not be available in certain jurisdictions.

Bloomberg Intelligence shall not constitute, nor be construed as, investment advice or investment recommendations (i.e., recommendations as to whether or not to "buy," "sell," "hold," or to enter or not to enter into any other transaction involving any specific interest) or a recommendation as to an investment or other strategy. No aspect of the Bloomberg Intelligence function is based on the consideration of a customer's individual circumstances. Bloomberg Intelligence should not be considered as information sufficient upon which to base an investment decision. Customers should determine on their own whether they agree with Bloomberg Intelligence. Bloomberg Intelligence should not be construed as tax or accounting advice or as a service designed to facilitate any Bloomberg Intelligence customer's compliance with its tax, accounting, or other legal obligations. Bloomberg believes that the information it uses in Bloomberg Intelligence comes from reliable sources, but does not guarantee the accuracy of information contained in Bloomberg Intelligence. Employees involved in Bloomberg Intelligence may hold positions in the securities analyzed or discussed on Bloomberg Intelligence.

Bloomberg makes no claims or representations, or provides any assurances, about the sustainability characteristics, profile or data points of any underlying issuers, products or services and users should make their own determination on such issues.

About Bloomberg Intelligence

Your go-to resource for making better investment decisions, faster.

Bloomberg Intelligence (BI) research delivers an independent perspective providing interactive data and research across industries and global markets, plus insights into company fundamentals. The BI team of 500 research professionals is here to help clients make more informed decisions in the rapidly moving investment landscape.

BI's coverage spans all major global markets, more than 135 industries and 2,000 companies, while considering multiple strategic, equity and credit perspectives. In addition, BI has dedicated teams focused on analyzing the impact of government policy, litigation and ESG.

BI is also a leading Terminal resource for interactive data. Aggregated, from proprietary Bloomberg sources and 600 independent data contributors, the unique combination of data and research is organized to allow clients to more quickly understand trends impacting the markets and the underlying securities.

Bloomberg Intelligence is available exclusively for Bloomberg Terminal® subscribers, available on the Terminal and the Bloomberg Professional App.

Take the next step.

For additional information,

press the <HELP> key twice on the Bloomberg Terminal®.

Beijing
+86 10 6649 7500

Hong Kong
+852 2977 6000

New York
+1 212 318 2000

Singapore
+65 6212 1000

Dubai
+971 4 3641000

London
+44 20 7330 7500

San Francisco
+1 415 912 2960

Sydney
+61 2 9777 86 00

Frankfurt
+49 69 92041210

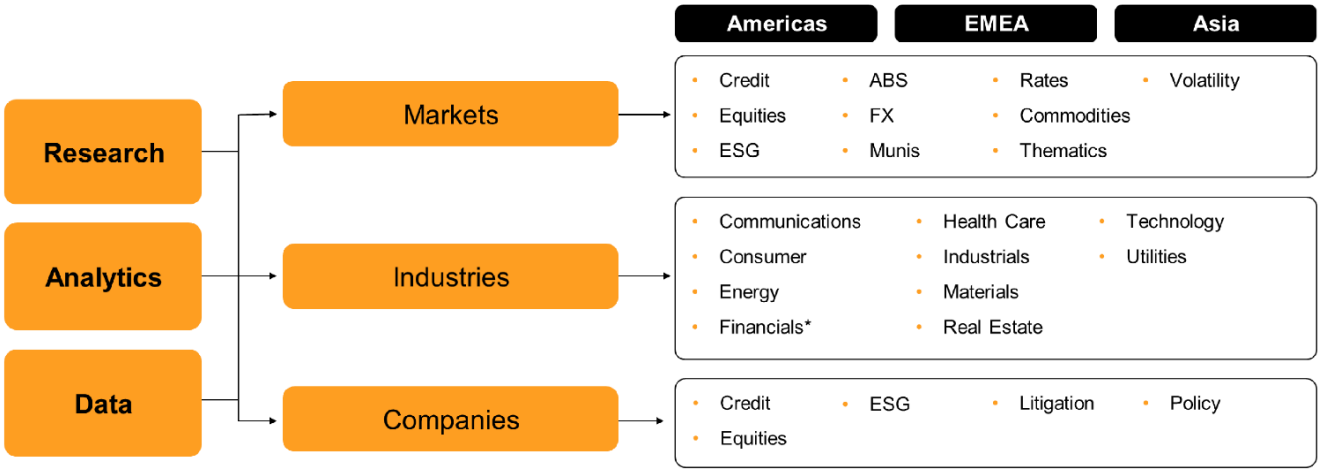
Mumbai
+91 22 6120 3600

Sao Paulo
+55 11 2395 9000

Tokyo
+81 3 4565 8900

Bloomberg Intelligence

Research, analytics and data tools to help you make informed investment decisions



Bloomberg Intelligence by the Numbers.

500

research professionals

135+

industries

600+

data contributors

2,000+

companies

21

markets covered