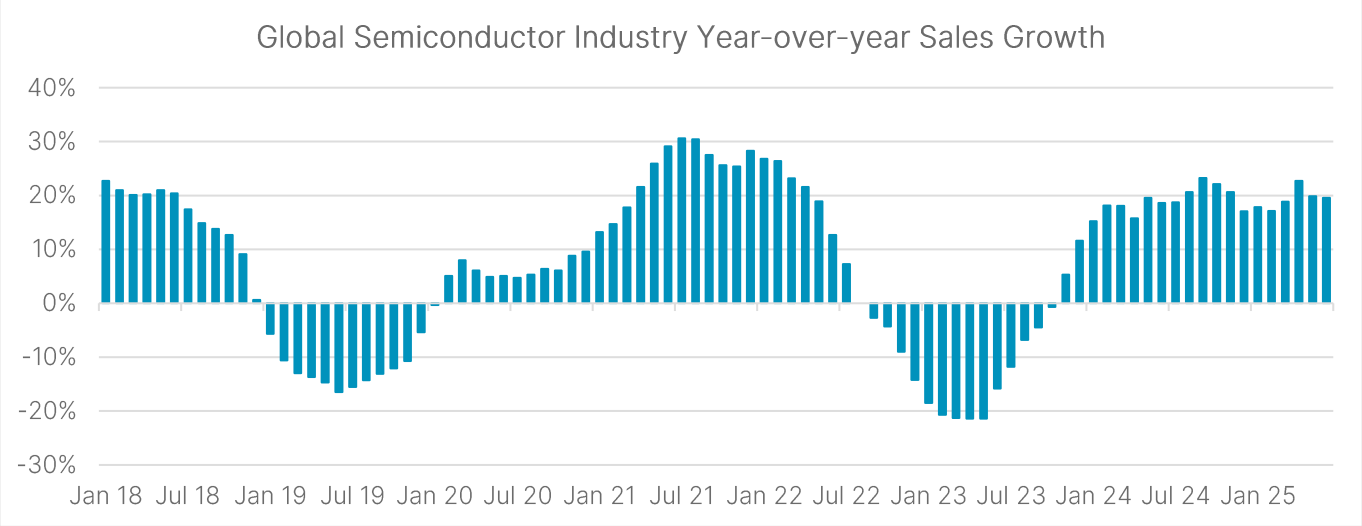
Semiconductors in Focus: Trends Shaping the Next Wave of Innovation

David Tsoi, CFA, CAIA, FRM, CESGA, CAMS, *Lead Index Research Strategist*

Artificial intelligence (AI) continues to be the most transformative technology of our era, with semiconductor companies leading the charge and powering groundbreaking advancements. After a robust recovery in 2024 driven by demand for logic and memory chips, the global semiconductor market is forecast to grow by 15% this year, reaching a total value of $728 billion, with the Americas and Asia Pacific expected to lead the growth.[1] The expansion of data centers continues to drive significant growth, especially for companies specializing in AI and semiconductor innovations. Global sales in June 2025 were US$60 billion, representing a year-over-year increase of 20%.[2]



Source: Semiconductor Industry Association. As of August 4, 2025.

AI growth remains intact

As AI-driven monetization opportunities begin to take shape, hyperscaler capital spending remains on the rise despite mounting tariff and economic headwinds. Global data center capex soared by 53% year-over-year in Q1 2025, marking the sixth straight quarter of double-digit annual expansion.[3] Microsoft, Amazon and Google reported that demand persistently exceeds available infrastructure capacity for AI workloads, with projections indicating that additional capacity will continue to expand throughout the year. Amazon is set to invest at least US$20 billion in Pennsylvania[4] and US$13 billion in Australia[5] to expand its data center infrastructure for AI and cloud services. Meta's capital spending could increase further in 2026, as it is building multiple multi-gigawatt data center clusters to fuel its AI ambitions, with the first facility slated to go live next year. The company has made AI central to its advertising strategy and plans to enable brands to fully design and target campaigns using AI tools by the end of next year. Based on the client's budget, these

[1] https://www.wsts.org/esraCMS/extension/media/f/WST/7175/WSTS-Q2-Release-2025-08-04.pdf/
[2] https://www.semiconductors.org/global-semiconductor-sales-increase-27-0-year-to-year-in-may/
[3] https://www.delloro.com/news/hyperscaler-blackwell-and-custom-accelerator-rollouts-drive-53-percent-capex-growth-in-1q-2025/
[4] https://www.aboutamazon.com/news/aws/amazon-pennsylvania-investment-cloud-infrastructure-ai-innovation/
[5] https://www.aboutamazon.com/news/aws/amazon-data-center-investment-in-australia/

new tools would generate the entire advertisement, including images, videos and text, and deliver it to the targeted audience.[6]

In the past, AI demand has primarily focused on training workloads, particularly for frontier models. While leading tech companies continue pouring resources into building ever-larger AI models, they are also reallocating more investment toward inference. Inference is the stage where trained AI models process new data to generate insights, make predictions or support decision-making. While training a model is essentially a one-time expense, prompting a model (inference) produces tokens, each of which carries a cost. During the Google I/O 2025 keynote, Alphabet CEO Sundar Pichai shared that the firm processed 480 trillion tokens across its products and APIs in April 2025, 50 times more than the same month a year earlier.[7] The rapid surge in token volume reflects growing usage and adoption of AI models, signaling a greater need for computing power and driving higher demand for chips.

The age of AI reasoning

The shift in investment toward inference has also gained momentum with the launch of new reasoning models. While traditional AI models respond swiftly and excel at pattern recognition, they often fail to understand broader contexts and struggle with complex reasoning. Reasoning models are built to deconstruct complex problems into smaller, manageable steps and solve them through explicit logical reasoning. They are specifically trained to show their work and follow a more structured thought process, which results in longer computation times for user queries. These models demand significantly more compute during inference to reason through intricate problems. This evolution from basic pattern recognition to structured reasoning is pivotal to AI, unlocking its potential to tackle complex real-world challenges effectively. As AI adoption rapidly expands, demand for inference will correspondingly intensify.

The ascending wave of AI agents

AI agents are positioned to revolutionize how organizations function, delivering breakthroughs in productivity and operational efficiency. They are intelligent systems designed to execute tasks independently by comprehending objectives, formulating decisions and taking actions to achieve predetermined goals. While humans define the desired outcomes, AI agents autonomously select optimal actions required to accomplish those goals. These agents boast a broad spectrum of uses, from supporting academic research and streamlining online purchases to planning leisurely vacations. Customer service, sales and marketing, and IT and cybersecurity are the three business functions where AI agents are most frequently deployed or planned for implementation in the next six months.[8] As enterprises progressively integrate AI agents across diverse operational applications, demand for computational infrastructure is escalating dramatically.

The rise of custom AI chips

Hyperscalers are increasingly focused on ASIC (application-specific integrated circuits) infrastructure to meet surging AI demand. ASICs are custom-built for specific workloads and can execute those tasks far more efficiently and at a substantially lower cost than high-performance GPUs. Although the initial investment to develop ASIC infrastructure is considerable, the long-term cost of running GenAI workloads on these chips is expected to be lower once the upfront expense is absorbed. For example, in April 2025, Google unveiled Ironwood, its seventh-generation Tensor Processing Unit (TPU), specifically designed for inference workloads.[9] While Google's in-house TPUs were once limited to internal use, the company is expanding external access to drive faster growth of its cloud business. Marvell Technology projects the custom computing device market will surge to US$55.4 billion by 2028, more than eight times its size in 2023.[10]

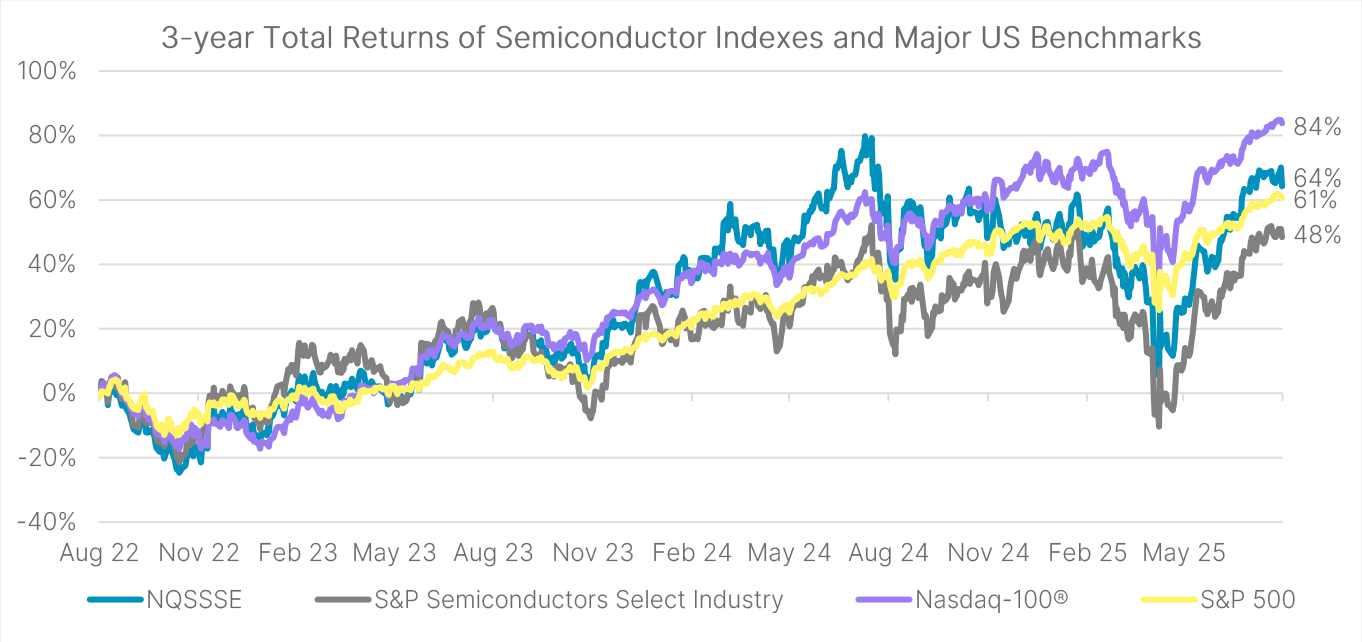
[6] https://www.wsj.com/tech/ai/meta-aims-to-fully-automate-ad-creation-using-ai-7d82e249/
[7] https://blog.google/technology/ai/io-2025-keynote/
[8] Source: PwC's AI Agent Survey (May 2025)
[9] https://blog.google/products/google-cloud/ironwood-tpu-age-of-inference/
[10] https://www.marvell.com/content/dam/marvell/en/company/assets/marvell-custom-ai-investor-event-2025.pdf

AI drives robust demand for high-bandwidth memory (HBM) technology

HBM represents a cutting-edge memory technology engineered to deliver faster data access while reducing energy consumption, which is critical for the performance of AI processing. HBM's market share in the dynamic random access memory (DRAM) segment is forecast to leap from 18% in 2024 to more than 50% by 2030.[11] Starting with the next-generation HBM4, the base die will be produced using logic processes, enabling lower power consumption and customizable features tailored to client requirements. Driven by escalating computational demands from AI training and inference workloads, HBM's market outlook remains strong. As the primary HBM supplier for Nvidia, holding a 62% share of global HBM shipments in Q2 2025[12], SK Hynix projects the global HBM market to expand by 30% annually through 2030.[13]

NQSSSE™ – the leading multifactor index for the semiconductor industry

With 30 constituents as of July 31, 2025, the Nasdaq US Smart Semiconductor™ Index (NQSSSE) is a modified factor-weighted index that aims to provide exposure to US companies within the semiconductor sector. The index universe is screened by four factor scores – (i) profit score (gross income), (ii) ROA score (return on asset), (iii) momentum score (the average of 3-month, 6-month, 9-month and 12-month USD price returns), and (iv) value score (cash flow[14]). NQSSSE delivered a total return of 64% over the past three years, outperforming the S&P Semiconductors Select Industry Index by 16 percentage points.

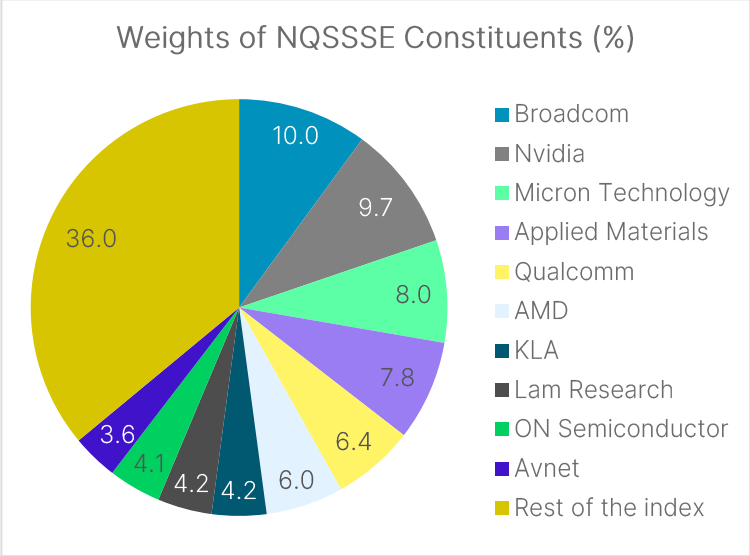


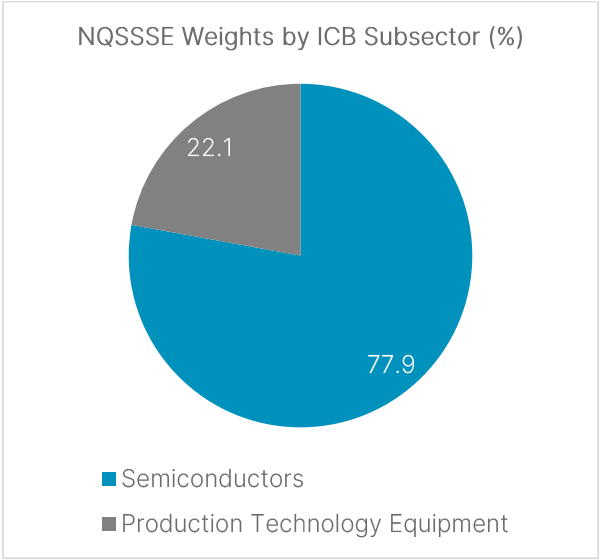
Source: Nasdaq Global Indexes, Bloomberg. As of July 31, 2025.

The constituents of NQSSSE are weighted based on their cash flows, with the top five constituents capped at 8%, the rest capped at 4%, and all constituents having a minimum weight of 0.5% during semi-annual rebalancing. For the full index methodology, please visit our website.

As of the end of July 2025, the ten largest constituents accounted for 64.0% of the index weight. 77.9% of the index weight is in the Semiconductor Subsector, with the rest in the Production Technology Equipment Subsector, according to the Industry Classification Benchmark (ICB) classification system.

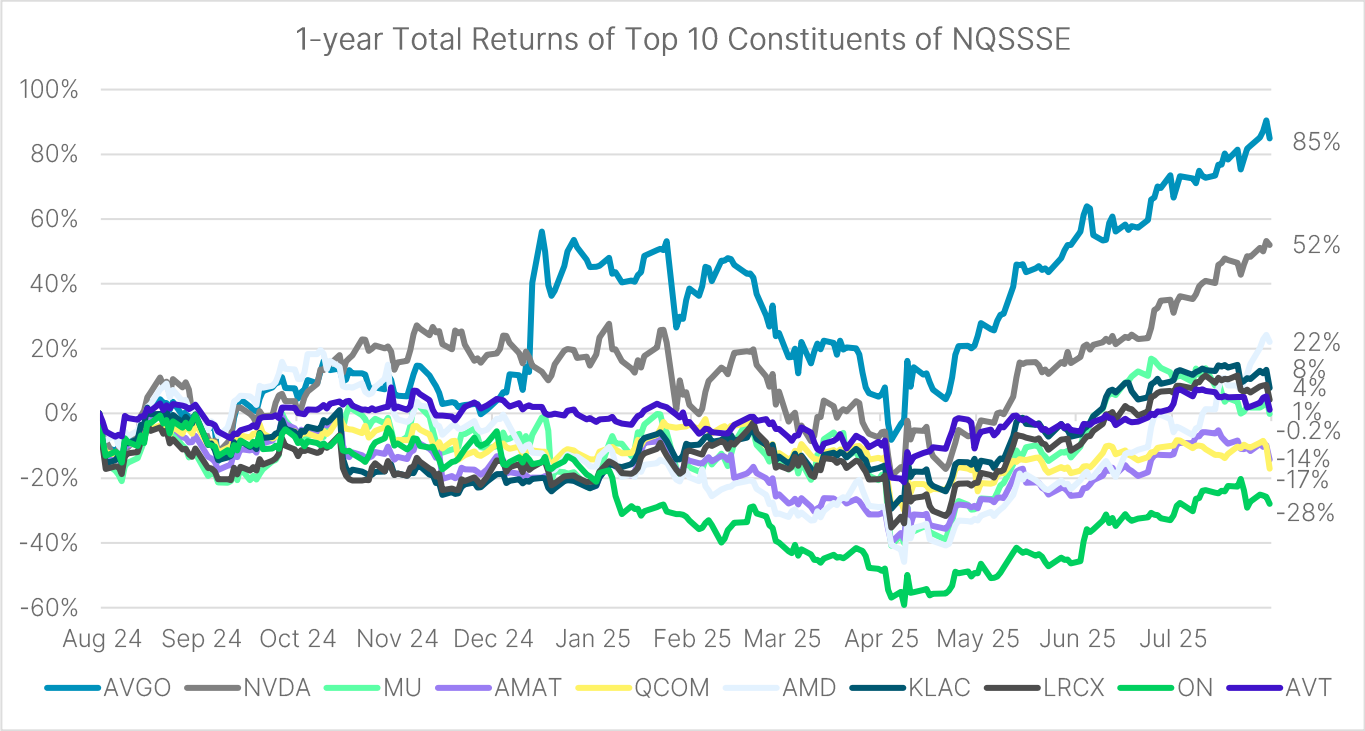
[11] https://www.yolegroup.com/strategy-insights/memory-industry-at-a-crossroads-why-2025-marks-a-defining-year/
[12] https://www.counterpointresearch.com/en/insights/samsungs-q2-2025-memory-performance-disappoints-but-signals-h2-recovery/
[13] https://www.reuters.com/world/asia-pacific/sk-hynix-expects-ai-memory-market-grow-30-year-2030-2025-08-11/
[14] Defined as operating cash flow minus stock-option expense. The cash flow score is multiplied by 0.01. This factor score is given limited weight as it serves as a tiebreaker where other factors are inconclusive.





Source: Nasdaq Global Indexes, FactSet. As of July 31, 2025.

Six of the ten largest holdings posted positive total returns over the past 12 months. On average, the top ten firms achieved a one-year total return of 11%. Within the same industry, stock performances varied significantly. The difference in one-year total returns between the best- and worst-performing stocks among the top ten constituents (Broadcom and ON Semiconductor) was a staggering 113 percentage points. This demonstrates the importance of diversification, even when investing in a single sector or theme.



Source: Nasdaq Global Indexes, FactSet. As of July 31, 2025.

Broadcom (weight: 10.0%)

As the largest constituent in NQSSSE, Broadcom emerged as the second-best performer among all index components, generating a one-year total return of 85% through July 2025. The company continues to dominate the AI ASIC and AI networking semiconductor markets. The chipmaker is engaged with the top seven hyperscalers for custom silicon products, including accelerators for Google, Meta and ByteDance. Moreover, Broadcom has significant exposure to enterprise software following its acquisition of VMware in 2023. Its infrastructure software division posted a 76% operating margin last quarter, up from 60% a year ago.[15]

Nvidia (weight: 9.7%)

As the second-largest constituent in NQSSSE, Nvidia was the second-best performer among the top ten constituents, gaining 52% over the past 12 months. It became the first publicly traded company in history to achieve a US$4 trillion market valuation in July 2025, after joining the trillion-dollar club in May 2023.[16] Nvidia's technological edge remains pronounced, as Blackwell shipments accelerate, driven by soaring AI reasoning demand and realized economies of scale. Beyond AI, the company identifies robotics as its most substantial addressable growth opportunity, with autonomous vehicles representing the first major commercial deployment. The firm recently unveiled new NVIDIA Omniverse libraries and NVIDIA Cosmos world foundation models that empower developers to build next-generation robots and autonomous vehicles by integrating AI reasoning with scalable, physically accurate simulations.[17]

Micron Technology (weight: 8.0%)

With a quarter of the market share in DRAM revenue in Q1 2025[18], Micron Technology stands as the third-largest NQSSSE constituent, delivering a 30% total return in the first seven months of this year. The company lifted its August-quarter gross margin guidance from 42% to 44.5%.[19] Management highlighted strong momentum in customer engagements for 2026 HBM volumes and expressed confidence in selling out next year's planned supply, including HBM3E and HBM4. HBM4 samples have been shipped to leading customers, who will begin qualification closer to the launch of their next-generation GPU or ASIC platforms. Micron is also capitalizing on an improved pricing environment, particularly in DRAM across data center, smartphone and PC markets.

Conclusion

Beyond the escalating demand for advanced training capabilities to support increasingly complex AI model architectures, AI inference has surfaced as a pivotal growth driver. The rapid increase in token volume signifies expanding usage and adoption of AI models. The proliferation of AI agents is poised to transform various industries and substantially increase compute demand. This dynamic landscape highlights the pivotal role of the semiconductor sector in driving the next wave of AI innovation.

The Nasdaq US Smart Semiconductor Index (NQSSSE) delivered a total return of 64% over the three years through July 2025. The First Trust Nasdaq Semiconductor ETF (US: FTXL) tracks the index.

[15] https://investors.broadcom.com/static-files/a5d6db22-6861-47e5-901b-13961fbc5321/
[16] https://www.reuters.com/technology/nvidia-sets-eye-1-trillion-market-value-2023-05-30/
[17] https://nvidianews.nvidia.com/news/nvidia-opens-portals-to-world-of-robotics-with-new-omniverse-libraries-cosmos-physical-ai-models-and-ai-computing-infrastructure/
[18] https://www.counterpointresearch.com/en/insights/post-insight-research-notes-blogs-sk-hynix-takes-top-spot-for-first-time-on-continued-hbm-demand
[19] https://investors.micron.com/news-releases/news-release-details/micron-updates-fourth-quarter-fiscal-2025-guidance/