AI Inference: The Next Step in the AI Evolution

REGIONS
ASSET MANAGEMENT

JANUARY 2026

Since the release of ChatGPT in November 2022, Generative AI has been the main focus for capex spending for hyperscalers. It has spurred large investments in the development and training of generative AI models. Developing and training these models requires a significant amount of computing power and has resulted in development of powerful Graphic Processing Units (GPUs) as well as a build out of AI infrastructure. Capex by the largest technology companies in 2025 is estimated to be over $400B[1] and is expected to grow to over $600B in 2026[2]. Going forward, AI is expected to focus more on application rather than on training AI models.

Table 1: Capex Spending by Big Technology Companies (B)[1,2]

	2025 (E)	2026 (E)
Amazon	$125	$155
Alphabet	$91	$125
Microsoft	$117.5	$160
Meta	$71	$120
Oracle	-	$42
Total	$405	$602

AI INFERENCE

AI inference is the next step in the development of an AI model. AI inference is where a trained model is applied to new / "unseen" data in order to generate outcomes in real time. It is the “doing” phase of the trained model.

Differences between training and inference (Table 2)

1. *Compute:* Training an AI model is the first step in its buildout. It requires developing algorithms based on complex neural networks that can recognize extensive data patterns. This requires large data sets and typically undergoes many iterations in order to minimize errors and improve accuracy. Training also requires optimization of parameters within the model. Therefore, training requires signifcant computational power.
AI inference on the other hand, requires the trained model to be run over many new data points. However, the model does not need to make adjustments to its parameters. Therefore, AI inference requires less computational power.
2. *Latency:* As accuracy is key, training an AI model is time consuming. The time required depends on several factors including model design and optimization techniques. AI inference is generally done in real time; fast responses are critical and low latency is key.
3. *Scalability:* Training models need scalability in order to process high amounts of data. This is vertical scaling and is achieved by using clusters of high powered GPUs. AI inference scalability involves adding more machines/nodes rather than compute power to distribute increased workloads. This is horizontal scaling.
4. *Cost:* AI training requires substantial upfront cost to develop an AI model. However, once a model is developed, it only needs to be periodically modified. AI inference cost per query is very low. However, AI inference is always on, therefore, the overall cost can be much higher than that for training.

Table 2: Differences Between AI Training and AI Inference

	Training	Inference
Objective	Learn patterns, optimize parameters	Make predictions in real time
Data	Existing labeled data sets	New/unseen data
Compute	High computing power	Low per query
Latency	Not critical	Critical, real time
Scalability	Vertical, adding compute capacity	Horizontal, adding nodes/machines
Time	Periodic tuning after initial optimization	Always on
Cost	High initial cost	Low cost per query, high total cost

Examples of AI inference include autonomous systems like self-driving cars, drones, and robots. Chatbots, voice assistants and AI assisted ecommerce are also examples of AI inference.

Authors: Anjali Tagare
Regions Investment Management, Equity Analyst

INVESTMENT THESIS

So far, most of the investment in generative AI has been geared towards developing and training AI models. This has resulted in the development of high power GPUs as well as the buildout of data centers. Returns on these massive investments are expected from the related productivity gains which can be achieved only when these models are used in real world workflows. Therefore, transitioning to AI inference is needed for realizing returns on these investments. Assuming rapid adoption, the AI inference market is expected to grow from an estimated $106B in 2025 to $255B by 2030 for a CAGR of 13.7%[4]. This is significantly greater than the AI training market which is expected to grow from an estimated $3.5B in 2025 to $17B by 2032 for a CAGR of 25% (Table 3)[5].

Table 3: Global Market Size (B)[4,5]

	2025 (E)	2032 (E)	CAGR
AI Inference	$106	$255	13.7%
AI Training	$3.5	$17	25%

Compute: AI inference workloads are expected to dominate over training workloads (Figure 1)[3]. Due to differences in computing, latency requirements and operational cost, AI inference requires specialized chips or accelerators that focus on speed and cost efficiency rather than computing power. These AI accelerators include GPUs, NPUs (neural processing units), FPGAs (field-programable gate arrays), and ASICs (application specific integrated circuits) which are specialized chips which are designed to perform specific tasks. Demand for AI accelerators is expected to grow from about $200B in 2025 to $650B in 2030 for a CAGR of 26% (Table 4)[6]. Unlike GPUs for general compute, where NVDA is the dominant, chips for AI inference are also being designed by companies like Google, Broadcom, Microsoft, Amazon Web Services, Advanced Micro Device, and Intel[7].

Figure 1: Global Data Center Demand by Workload (gigawatts)[3]

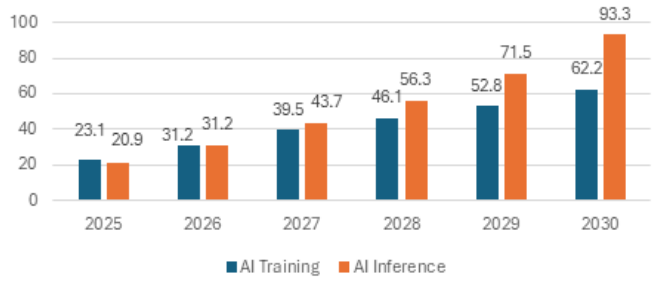

Table 4: Market Size Estimates (B)[6]

	2025	2026	2027	2028	2029	2030	CAGR 2025-2030
AI Accelerator	200.9	308.6	415.5	501.8	577.8	650.2	26%
HBM	33.4	37.9	49.3	63	75.7	87	21%

Storage: Memory is the backbone of AI for training as well as inference. AI training requires large data sets and inference is expected to generate even more data. Data storage/memory is expected to be a bottleneck in the growth of AI. Traditional DRAM memory does not have enough bandwidth and latency that is required for efficient training of AI models. High Bandwidth Memory (HBM) with its low latency and high bandwidth is critical for AL models and is also more energy efficient than DRAM. Demand for HBM is expected to grow from $33B in 2025 to $87B in 2030 for a CAGR of 21% (Table 4)[6]. Micron Technology (MU), Samsung, and SK Hynix are some of the leading manufacturers' of HBM. AI Inference is expected to generate massive amounts of data from various sources including AI generated content, social media user generated content, and ecommerce. This is expected to boost content availability for end user devices like PCs and smart phones and drive a device refresh cycle. This will create incremental demand for companies like Western Digital (WDC) and Seagate (STX) which make hard disk drives (HDD) as well as companies like SanDisk (SNDK) that make flash drives.

SUMMARY

Technology companies have invested heavily in generative AI. So far, these investments have been focused on developing and training AI models and for the buildout of data centers. Going forward, AI is expected to focus more on application or inference rather than training AI models. AI inference is the application of models to real world workflows. AI inference does not need high computing power; however, it does need to be reliable and fast. Therefore, demand for AI accelerators is projected to be high. This will create incremental demand for companies like NVDA, GOOG, AVGO, MSFT, AMD, AWS, and INTC which have developed or are developing these chips. Memory is the backbone of AI for training as well as inference and demand is expected to grow. This will also create incremental demand for companies like MU, Samsung and SK Hynix that make HBM as well as companies like WDC and STX, that make HDD and NSAND and flash drive makers like SDK.

Sources:

1. https://io-fund.com/ai-stocks/ai-platforms/big-techs-405b-bet
2. https://www.mufgamericas.com/sites/default/files/document/2025-12/AI_Chart_Weekly_12_19_Financing_the_AI_Supercycle.pdf
3. https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/the-next-big-shifts-in-ai-workloads-and-hyperscaler-strategies
4. https://www.fortunebusinessinsights.com/ai-inference-market-113705
5. https://www.fortunebusinessinsights.com/ai-training-dataset-market-109241#:~:text=KEY%20MARKET%20INSIGHTS,share%20of%2047.95%25%20in%202024.
6. AI 2030: Building Blocks of the next $1tn Industrial Revolution: Bank of America, 24 June 2025
7. https://www.cnbc.com/2025/11/21/nvidia-gpus-google-tpus-aws-trainium-comparing-the-top-ai-chips.html

---