Allianz Global Investors

JUNE 2026

From cloud to edge: the next wave of AI opportunity

As the focus in artificial intelligence (AI) moves beyond rapid model training toward real-world deployment, attention is increasingly shifting to inference, and how AI can deliver fast, efficient, and scalable outcomes.

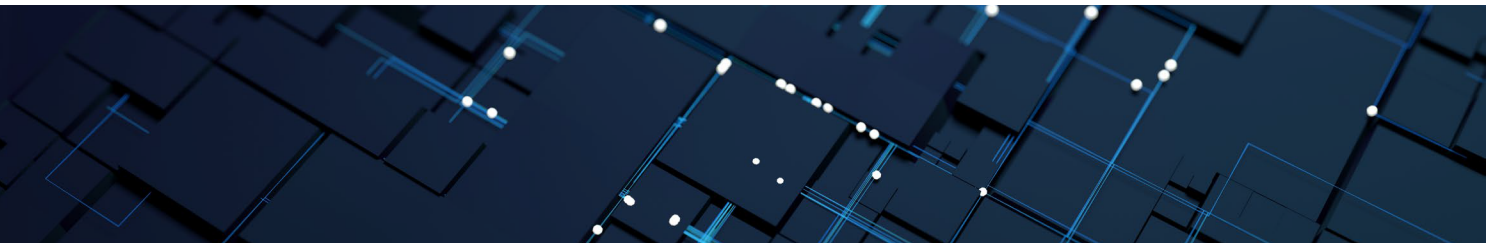
The training of large language models (LLMs) requires huge clusters of AI chips and is almost exclusively done on the cloud, where hyperscalers rent out compute capacity consisting of graphical processing units (GPUs) and application-specific integrated circuits (ASICs). Inference, on the other hand, is the process of applying knowledge to real world use cases. This uses a different mix of chips with the aim of low latency (responding to requests quickly) and high throughput (processing a large number of requests).

While the training of models has progressed quickly over recent years, prioritising speed over cost efficiency, we are reaching an inflection point for AI as the advancement of LLMs has enabled the next phase of inference, and AI's commercial application, to take centre stage. What is stark is how quickly we've reached this inflection point. Andrej Karpathy (founding member of OpenAI and the individual who coined "vibe coding") stated in

Jeremy Gleeson
CIO Global Tech Equity

Sunil George
Senior Portfolio Manager
Global Tech Equity

Brad Reynolds
Portfolio Manager
Global Tech Equity



October 2025 that the hype around coding agents was "exaggerated". Then, just 3 months later, he stated that:

"LLM agent capabilities (Claude & Codex especially) have crossed some kind of threshold of coherence around December 2025 and caused a phase shift in software engineering... This is easily the biggest change to my basic coding workflow in 2 decades of programming and it happened over the course of a few weeks."

This is not the first time a frontier technology has been underestimated, but the speed of change here has been remarkable. Gartner, the research and advisory firm, has predicted that, in 2026, 55% of AI-optimised infrastructure as a service (IaaS) spending will go to inference workloads, rising to over 65% in 2029. This raises the question of what the most suitable architecture will be to optimise this next phase of inference, with edge AI –running AI models locally, or close to use cases, rather than on the cloud – being a leading candidate.

From content delivery to distributed AI infrastructure

Compared to centralised cloud computing, edge AI makes use of end devices and/or a network of points of presence that are located geographically closer to the query. In this respect, it is not a new architecture but consists of many elements previously known as the "internet of things". The customer base of these distributed networks has traditionally been entertainment services such as gaming, and media streaming, and yet with the ongoing rapid growth in AI, using all available channels to distribute compute is becoming more important than ever. A sub-division of this architecture that is seeing a great deal of focus is the network of small datacentres, known as content delivery networks (CDNs). Due to the distributed nature of these CDNs, they now have a role to play in AI inference.

This shift from CDNs to edge AI providers can be seen in the recent USD 1.8 billion deal between Akamai and Anthropic. The legacy CDN has over 4,300 points of presence across 130 countries, and while initially created from content-delivery for entertainment services, this network also maps cleanly onto distributed inference at the edge. This example highlights how edge AI will grow to complement cloud AI, becoming a vital tool in the efficient deployment of AI inference. The benefits of shifting inference to the edge include lower latency, increased throughput, availability, privacy, fewer bandwidth bottlenecks, and reduced cost per query due to the lower marginal cost of inference on local devices.

While AI models with a significant degree of complexity must be processed on the cloud due to compute limitations on the edge, inference for more focused applications, where speed of inference and high throughput are key determinants of efficacy, could be optimised on the edge. Examples include, but are not limited to:

- Autonomous vehicles – benefit through speed of inference and high throughput
- Wearables/AR/VR – benefit through constant availability
- Industrial Robotics – benefit through availability and cost per query
- Healthcare – benefit through privacy and fewer bandwidth bottlenecks

A bifurcated, tiered inference architecture will likely emerge where low-cost, latency-sensitive tasks are executed on the edge, while more complex workloads remain in the cloud, with routing determined by compute intensity. As efficiency gains expand capabilities across both layers, competitive advantage is likely to shift toward workload routing and cost optimisation rather than model scale alone.

New opportunities across the AI value chain

The shift to inference on the edge is creating new investment opportunities, alongside expanding existing investment cases. And many of these opportunities are aligned with the themes in global tech that **we have already identified.**

The clearest new opportunity is with the **transition of CDNs into edge AI providers. Such providers** will get a boost from increased revenue growth for their network services, going from commoditised bandwidth providers to critical infrastructure for real-time inference. This also pairs well with these providers' existing services, as many also offer cybersecurity and networking functionality. The shift would also likely come with a re-rating, something which has occurred across many other sub-sectors heavily exposed to AI.

Another existing investment opportunity and Theme of 2025 for the Global Technology team that will likely get renewed impetus is the **optical networking industry** which facilitates the **high-speed interconnection** of distribution networks. This industry will benefit from expansion of the AI network from the central hyperscaler cloud architecture to the vastly expanded edge network, encompassing thousands of points of presence. Optical networking will enable the increased throughput and speed of traffic which accompanies inference, and in particular inference from distributed edge networks.

Inference on the edge will also further drive the **revival of analog semiconductors** as AI begins to interact with the real world. The growth of analog semiconductor content per device, along with a greater number of distributed endpoints from inference on the edge, will further propel this revival. This was one of the Global Technology team's Themes for 2026 and we remain confident in this theme playing out.

Finally, where cloud architecture has caused a huge supercycle in memory with a focus on HBM and DRAM which optimises for throughput, inference on the edge could see a **broadening in the demand for different types of memory**. The adoption of KV Cache (Key-Value Cache) as AI's **short-term working memory** during a conversation highlights this as without it, the AI would act like it has amnesia, forgetting the beginning of your prompt every time it types a new word. The KV Cache must be saved in the computer's fastest memory (DRAM) so the AI can look at it instantly, reducing latency. Regardless of the optimal memory for inference at the edge, the likely outcome is additional fuel for the **memory supercycle** and a potential broadening of memory demand. This is another Theme for 2026 within the Global Technology team.

Together, these developments highlight how inference on the edge is not just a technological shift, but a structural expansion of the AI ecosystem, unlocking new efficiencies, applications, and investment opportunities across the value chain.

MACS-P-20811269

ADM-W5661105