2

Inference demand drives continued AI buildout

The intense AI infrastructure buildout that has been taking place for years shows no signs of slowing down and the environments that have come online have become a strong foundation for significant software innovation and ecosystem development. Compute architectures will evolve into hybrid classical-quantum workflows and neuromorphic systems, leveraging CPUs, GPUs and specialized accelerators to address diverse computational demands across the infrastructure landscape. The progress in quantum hardware will lead to many scientific quantum advantage demonstrations showing a quantum computer performing a scientifically interesting computation that is impossible classically. These demonstrations will be uncontested by classical techniques and will fuel the race to demonstrate commercially relevant quantum advantage in the years to come.

This report is for informational/discussion purposes only and is not intended to be nor should it be relied on for decision making

AI infrastructure innovation

Innovation across power generation, data center design, silicon architectures and enabling technologies to meet soaring demand for computing resources.

The AI revolution continues to push the limits of existing infrastructure capabilities, and the insatiable demand for computing resources continues to drive the ecosystem to innovate across the entire stack. In addition, heterogenous compute (i.e. different chips for different use cases) continues to gain market traction in an effort to provide best fit for purpose designs to improve overall efficiencies.

The ongoing innovations can be grouped into four main categories:

New power generation methodologies: Gigawatt DC facilities and overall expansion have led to fierce competition for utility lockups and novel energy sources like Small Modular Nuclear Reactor development.

New datacenter designs and cooling strategies: Demand is accelerating facilities trends like modularity, multi-story, high density racks, and AI edge locations. Meanwhile continued power densities are changing how current compute is delivered to each server and how racks of servers are cooled using liquid (including network and storage devices).

New silicon architectures: The chip market continues to expand, with increasing competition from the clouds, traditional silicon companies and the startup ecosystem leading to incredible performance leaps and more cost-efficient alternatives. Custom AI silicon (e.g. chips focused on inference) have continued to gain traction and are being considered for future architectures like AI at the edge.

Enabling technologies: New networking transfer rates, optical technology, Data Processing Units, Disaggregated Storage, Memory Sharing, In-Storage Computing – there are dozens of infrastructure breakthroughs gaining traction as part of this ecosystem's massive capex build out.

The Infrastructure build out is now multiple years in, and the leading question from industry pundits is just how long will it last. Many of these developments are by nature multi-year investments, and 2026 looks poised to continue this trend as many new facilities start to come online.

This report is for informational/discussion purposes only and is not intended to be nor should it be relied on for decision making

Global AI Infrastructure Capex spending in 2025 surpassed $400B, with recent 2026 projections exceeding $600B.[8]

Cloud and colocation providers such as AWS, Microsoft, Google, Meta, and CoreWeave are building next-gen data centers optimized for AI workloads, often co-located near renewable or modular power sources, including nuclear microreactors, geothermal plants and hydrogen fuel cells.

The competition in silicon development is intensifying as organizations work to introduce new types of accelerators. Recent industry activity includes acquisitions of licensing rights and talent from startups, reflecting a trend toward workload-specific compute to complement general-purpose accelerators. Providers are also making significant progress through ongoing improvements in hardware performance and software integration.

Hyperscalers continue to invest heavily in their own silicon, and these chips are now more broadly available, and some can even run in a customers data center. 2026 may likely be a turning point year where these alternatives start to gain more meaningful adoption by customers in the market, having reached significant maturity and demonstrated demand from key AI labs.

Markets beyond silicon are also experiencing significant evolution. Major cloud service providers are advancing their networking capabilities, while startups focused on photonic networking are beginning to gain traction, with recent acquisitions occurring in this area. Companies specializing in storage solutions are seeing substantial growth, with large-scale partnerships being formed to deliver advanced data services to a broad range of customers across the technology stack.

Finally, “AI at the edge” remains an evolving strategy across the industry from CDN companies building out AI compute across their locations, to low-powered custom inference silicon targeting edge use cases and the continued developed of AI chipsets embedded in personal computing.

[8] Big Tech to invest about $650 billion in AI in 2026, Bridgewater says. Reuters. (2026, February 23). https://www.reuters.com/business/big-tech-invest-about-650-billion-ai-2026-bridgewater-says-2026-02-23/.

This report is for informational/discussion purposes only and is not intended to be nor should it be relied on for decision making

Cloud native AI

Accelerating efficient, scalable, and cost-effective AI through open-source tools and advanced optimization techniques.

As more GenAI use cases are scaled into production, there is significant industry focus on efficiently managing infrastructure resources through cloud native software constructs.

The ecosystem is working to develop common and familiar toolsets that abstract hardware complexity and provide workload portability. The result is a competitive landscape across proprietary, specialized software stacks and, increasingly, an open-source approach to managing inference workloads using standard architectures like Linux and Kubernetes.

While training is mostly done offline, inference is done in production and needs to be designed to support things like multi-tenancy, scalability, resiliency and high availability, posing challenges at enterprise scale. New inference engines and distributed serving frameworks are transforming the way LLMs operate in production by optimizing the inference process, rather than the model architecture itself.

Traditional inference was limited by sequential token generation and static memory use, but modern systems leverage “KV cache” optimization – a technique that stores key and value pairs from prior computations – to avoid redundant processing. Emerging frameworks build on this approach by introducing methods such as paged attention and continuous batching, which efficiently share GPU memory and process multiple requests simultaneously.

AI Labs continue to play an important role in driving the overall market to focus on optimized use of infrastructure resources. New innovations from the data science community lead to more efficient and accurate use of resources – from reducing memory requirements to leveraging only the most needed parts of a model for a given inference query and improving the stability and accuracy of large training runs.

Together, these advancements accelerate time-to-inference, lower costs, and enable more frequent model updates – reshaping how AI workloads are built and maintained.

This report is for informational/discussion purposes only and is not intended to be nor should it be relied on for decision making

Market and industry perspectives

The market for AI infrastructure optimization is expanding rapidly as enterprises seek greater efficiency and control over AI costs. According to industry estimates, the AI inference optimization market alone is projected to exceed $100B over the next five years as more startups focus on providing platform technology so reduced overall AI infrastructures costs.[9]

The Cloud Native Ecosystem (i.e. Linux / Kubernetes) is embracing emerging projects like vLLM (inference engine) and llm-d (optimize and manage inference at scale) to promote consistent and powerful inference patterns. Emerging startups are differentiating through software-defined optimization and low-level programming techniques to improve price to performance.

Large web scale companies and research institutions are developing and open sourcing many new projects, including serving engines, disaggregation, and caching techniques, etc.

AI Labs continue to publish new research focused on infrastructure optimization. Key innovators have brought numerous techniques to market, from “Multi-Head Latent Attention” and new “Mixure of Expert” approaches for inference, and recent breakthroughs to improve Training stability which enable larger experiments on less reliable hardware.

Ecosystem convergence of training and inference runtimes will likely occur over time to help abstract the complexity of AI Infrastructure management which may lead to more overall use of AI and the increased deployment of smaller models that can easily be hosted across data center and edge deployments.

[9] PR Newswire. (2025, February 28). Ai inference market worth $254.98 billion by 2030 - exclusive report by MarketsandMarketsTM. Yahoo! Finance. https://finance.yahoo.com/news/ai-inference-market-worth-254-151500286.html.

This report is for informational/discussion purposes only and is not intended to be nor should it be relied on for decision making

Quantum computing: from research breakthroughs to real-world use

Hardware advancements will enable scientific quantum advantage.

Quantum computers (QCs) are transforming specialized computing beyond classical limits. QCs are fundamentally different from classical computers and AI, which handle general-purpose, sequential tasks like logic operations and pattern recognition. By manipulating the probabilities of all possible states simultaneously, QCs leverage quantum mechanics to achieve dramatic speedups for specialized problems – such as simulations, optimization, and cryptography – that classical computers cannot efficiently solve. Ultimately, QCs and classical computers are expected to be complementary, with classical systems handling broad tasks and QCs focusing on specialized, complex challenges.

Recent breakthroughs are overcoming key engineering barriers to fault tolerance. Fault tolerance is the key goal for all QCs, as it enables them to perform long, reliable computations despite errors in physical gates or measurements – a milestone necessary for transformative business impact. Achieving fault tolerance requires encoding logical qubits across many physical qubits and applying continuous error correction. Recent breakthroughs, such as Google's Willow processor surpassing the surface-code threshold[10] and Quantinuum's demonstration of fully fault-tolerant universal gate set with repeatable error correction[11], signal a new phase for the industry. Numerous leading companies estimate to release fault-tolerant quantum computers by 2028–2030, aiming to scale from 30–1,200 qubits today to 1M[12, 13, 14], which experts believe is needed to break current cryptography and unlock world-changing applications.

QCs are playing an expanding role across industries, with progress in sectors such as healthcare, material science, and finance, enabling solutions to complex problems that were previously unsolvable. From an AI perspective, quantum computing will generate data beyond classical simulation, providing valuable input for AI models to advance scientific understanding.

This report is for informational/discussion purposes only and is not intended to be nor should it be relied on for decision making

Market and industry perspectives

The quantum computing industry is highly fragmented, with four to seven main modalities under active research. Each approach uses different setups, quantum particles, and temperatures, facing distinct engineering hurdles that make progress uneven and interdependent, and it remains uncertain which will achieve large-scale real-world impact first. Major tech firms are developing proprietary technologies, forming partnerships, or investing in private companies: focus areas include, superconducting qubits, trapped ions, photonics, neutral atoms.

Political support for quantum computing is accelerating both in US and internationally. The US has made quantum computing a national priority since the National Quantum Initiative Act of 2018, with recent executive orders accelerating research and development (R&D), post-quantum cryptography (PQC) and federal adoption timelines. The administration is considering equity stakes in QC firms and pushing for quantum-resistant upgrades by 2030. Major corporations have aligned with these efforts, including JPMorganChase which named a quantum computing a focus in the firm's $1.5T Security and Resiliency Initiative. Internationally, China leads in government investment and strategic planning, while Europe boasts strong scientific leadership and public funding.

The capital market for quantum computing has experienced strong momentum. Driven by AI's substantial impact on public market capitalizations and private valuations over the past 2–3 years, investors are actively seeking the next major growth opportunity. QC has emerged as a leading area of interest, viewed as a promising new investment frontier.

The long-term revenue potential for quantum computing remains strong. Over the next 2–4 years, industry revenue is projected to reach $1B, primarily driven by early-stage development and testing.[15] Looking ahead to 2040, consulting firms like McKinsey and BCG estimate a $100B market opportunity for quantum computing providers, who are expected to capture 20% of the $500B in total economic value generated by the industry.[16, 17]

[10] Google Quantum AI and Collaborators. Quantum error correction below the surface code threshold. Nature 638, 920–926 (2025). https://doi.org/10.1038/s41586-024-08449-y

[11] Quantinuum, Breaking even with magic: demonstration of a high-fidelity logical non-Clifford gate. As of February 2025.

[12] Craig Gidney. How to factor 2048 bit RSA integers with less than a million noisy qubits. https://doi.org/10.48550/arXiv.2505.15917. As of May 2025.

[13] Ben Bloom, CEO of Atom Computing, Quest for qubits: Quantum computing leaders make their case at Nvidia GTC. As of March 2025.

[14] PsiQuantum Raises $1 Billion to Build Million-Qubit Scale, Fault-Tolerant Quantum Computers. (2023, February). Psiquantum. Retrieved from https://www.psiquantum.com/news-import/psiquantum-1b-fundraise.

[15] Bank of America Institute. Quantum Leaps and Bounds. (2025, October 23). https://institute.bankofamerica.com/content/dam/transformation/quantum-computing.pdf

[16] Jean-Francois Bobier, Matt Langione, Cassia Naudet-Baulieu, Zheng Cui, and Eitoku Watanabe. The Long-Term Forecast for Quantum Computing Still Looks Bright. BCG. (2024, July 18). https://www.bcg.com/publications/2024/long-term-forecast-for-quantum-computing-still-looks-bright.

[17] McKinsey & Company. Quantum Technology Monitor. (2024, April). https://www.mckinsey.com/~/media/mckinsey/business%20functions/mckinsey%20digital/our%20insights/steady%20progress%20in%20approaching%20the%20quantum%20advantage/quantum-technology-monitor-april-2024.pdf?pdf.

This report is for informational/discussion purposes only and is not intended to be nor should it be relied on for decision making

This report is for informational/discussion purposes only and is not intended to be nor should it be relied on for decision making