Americas Technology: Hardware: Expert Network Series: Former Director of AI Transformation at Microsoft on AI infrastructure demand

Katherine Murphy
+1(212)902-1151 | katherine.a.murphy@gs.com
Goldman Sachs & Co. LLC

Michael Ng, CFA
+1(212)902-8618 | michael.ng@gs.com
Goldman Sachs & Co. LLC

We hosted an investor webinar with William Fong, former Director of AI Transformation at Microsoft on July 10th, 2026 as part of our Expert Network Series to discuss AI infrastructure demand trends.

Bottom line: (1) Agentic AI is driving organizations to invest in on-premise and edge infrastructure because of better token cost, latency, and governance/data sovereignty relative to the cloud; this should be meaningful tailwind for enterprise compute vendors like DELL and HPE; (2) Mr. Fong does not believe that AI infrastructure is at risk of being overbuilt, largely due to the massive demand from agentic AI as well as ongoing demand for net-new training infrastructure; and (3) Organizations are finding budget for hardware investments through infrastructure performance efficiencies (e.g., power cost savings from new generation CPU servers) and reduced headcount.

Key takeaways

Agentic AI is driving workloads on-prem because of better cost, latency, and governance. To optimize across these constraints, Mr. Fong has observed organizations deploying a hybrid compute strategy where they are increasingly running individual AI agents locally on-premise to achieve millisecond-level latency and maintain strict data privacy, while utilizing the cloud for the complex multi-agent orchestration layers. Mr. Fong sees significant opportunities to run these workloads at the edge, spanning personal computers and IoT devices. However, he flagged that supply chain constraints, particularly in memory supply, will serve as a critical bottleneck limiting the shipment volume of these edge platforms.

AI infrastructure likely not at risk of being overbuilt. Addressing market concerns, Mr. Fong emphasized that AI infrastructure is not at risk of being overbuilt, largely due to the massive demand tailwinds from agentic AI. While inference is shifting toward a hybrid model across the cloud, on-premises, and the edge, there remains a robust demand for net-new AI training infrastructure, especially since the vast majority of the world's data (of which the internet represents only ~1/3) resides untrained within enterprise systems. Furthermore, Mr. Fong clarified that recent reports of Microsoft canceling data center MOUs with former bitcoin-mining neoclouds did not signal softening demand for AI training. Instead, these cancellations were driven strictly by the sites' inability to secure guaranteed power

Goldman Sachs does and seeks to do business with companies covered in its research reports. As
should be aware that the firm may have a conflict of interest that could affect the objectivity of th
Investors should consider this report as only a single factor in making their investment decision. Fo
certification and other important disclosures, see the Disclosure Appendix, or go to
www.gs.com/research/hedge.html. Analysts employed by non-US affiliates are not registered/qua
analysts with FINRA in the U.S.



grids and meet the intense energy requirements of modern AI clusters.

Enterprises are finding budget through infrastructure performance efficiencies, operating leverage. First, organizations are driving meaningful cost savings by replacing legacy CPUs with latest-generation, highly efficient servers and upgraded networking; for instance, transitioning to newer platforms like the Arm AGI CPU, which is designed to deliver twice the performance per rack and consume 50 percent less power, allows companies to replace multiple older servers with a single unit. Second, Mr. Fong observed that organizations are realizing strong return on investment from practical use cases such as coding and call center agents. Many of these enterprises are actively reinvesting the operational savings generated from reduced headcount directly back into their AI infrastructure.