Forbes

INNOVATION / AI

Dell COO Says Agentic AI Is Breaking Cloud Economics, Forcing Data Center Rebuild

VICTOR DEY, CONTRIBUTOR

May 19, 2026

When token costs fall 80% in a single year but token consumption for reasoning alone surges 320 times, something fundamental breaks – not just in the economics of cloud AI, but in the physical architecture of every data center built to run it. That arithmetic explains everything Dell Technologies unveiled at its annual conference in Las Vegas this week, and why the company is placing one of the largest strategic bets in its history.

The computing infrastructure giant asserts that agentic AI will continue to generate such continuous demand for inference that cloud-only enterprise strategies will become both economically indefensible and thermodynamically impossible to sustain at scale.

"Last year I said token use was undercalled," Jeff Clarke, vice chairman and chief operating officer of Dell Technologies, tells me. "I was too conservative." He notes that the token request volume now being generated comes from autonomous systems that do not sleep, pause between tasks, or wait for business hours. "The risk isn't overbuilding, but being caught flat-footed as the demand curve bends and never comes back," he says.

In agentic environments, data workloads continuously trigger additional reasoning cycles and machine-to-machine interactions, causing infrastructure demand to compound instead of scaling predictably.

"Eighty-three percent of the world's data sits on-prem," Clarke says. "Not in the cloud, at the edge, inside existing infrastructure." Routing that data to the cloud to meet the agents creates latency, governance exposure, and bandwidth costs that compound alongside the token bills. Dell's answer is its "deskside to data center" strategy – a continuum stretching from local AI workstations running autonomous agents on-premises all the way to liquid-cooled rack-scale systems engineered for continuous enterprise reasoning workloads that would bankrupt a cloud-only budget inside eighteen months.

DELL TECHNOLOGIES

Dell's Jeff Clarke says token consumption for AI reasoning is up 320x as agentic systems break cloud economics, forcing a complete rethink of enterprise data center architecture.

The company aims to help run agents where the context lives and let the economics follow.

In Q4 FY2026, Dell generated $9 billion in AI-optimized server revenue, up 342% year-over-year, and now projects roughly $50 billion in AI server revenue

for FY2027. Its Infrastructure Solutions Group's operating margins improved 530 basis points across the fiscal year, weakening one of the primary bear arguments that AI server growth would permanently erode profitability.

Why Enterprise AI Projects Stall Before Reaching Production

Clarke believes that most enterprises don't have an AI ambition problem, but rather an AI execution problem rooted in data, and it runs deeper than most organizations want to admit.

"Ninety percent of enterprise data is unstructured. None of it is connected in a way that agentic AI can actually use." Every week an organization runs agents on top of a broken data foundation is a week during which those agents produce outputs that cannot be trusted, audited, or scaled.

"The right architectural call is to move AI to the data, not the data to AI. Agents need direct access to data files, which requires a context and orchestration layer, plus rigorous tracking of data lineage and updates. Get it wrong and your AI projects stall before they reach production."

Dell's AI Factory with Nvidia is addressing this through orchestration layers, semantic search systems, and AI-ready storage platforms. However, the most strategically important announcement was the company's Deskside Agentic AI, a platform allowing enterprises to run autonomous agents locally rather than routing everything through cloud inference. The system combines Dell workstations, Nvidia NemoClaw software, OpenShell runtime security, and Dell services into a stack purpose-built for persistent multi-agent workflows.

Clarke notes that most enterprises have not yet felt the full cost of routing agentic workloads entirely through cloud inference – but they will. The cumulative bill for multi-step reasoning systems eventually becomes one of the largest line items in an enterprise technology budget. By the time most finance teams see it clearly enough to act, the window for making the architectural decisions has already closed.

Thermodynamics Are Constraining Enterprise AI

The physical constraints on AI infrastructure are simultaneously arriving faster than most facilities teams anticipated, and in ways that have less to do with software procurement than with watts, coolant flow and the structural load limits of raised floors.

Dell unveiled eleven new PowerEdge servers spanning air-cooled and liquid-cooled environments – anchored by the liquid-cooled PowerEdge M9825 with AMD EPYC 6th Gen processors, purpose-built for ultra-dense AI and HPC workloads in factory-integrated IR7000 racks, and the new XE5845 and XE7845 servers designed for PCIe-based AI deployments at scale, supporting next-generation GPUs. Likewise, the company also unveiled the PowerCool CDU C7000, a cooling distribution unit designed to handle the heat output of GPU-dense AI clusters that conventional air cooling cannot dissipate at scale.

Alongside compute and cooling, Dell introduced PowerStore Elite – an intelligent storage platform that triples performance and density versus prior generations, packs up to 5.8 petabytes into a single 3U appliance, with every component modular and field-upgradable.

While companies want autonomous systems capable of continuous reasoning, they simultaneously demand perfect auditability and operational explainability. In finance, healthcare, and other regulated industries, an agent acting incorrectly is an operational and compliance risk. Clarke explained that all agentic work should follow three parts: intent, action and validation. Humans define the objective, the agent executes, and humans verify the outcome.

"What changes is that the middle part moves out of direct human control at speeds and scales that make micromanagement impossible," Clarke says. The boardroom question, he argues, has already shifted. "It can no longer just be 'are you secure?' rather 'do you understand what your systems are doing on your behalf?' Every action an agent takes needs a receipt."

Increasingly, Wall Street appears to believe that the companies building that governance layer may capture the next major wave of enterprise AI value.

Record Revenue, Competition and a $43 Billion Moat

Dell shares reached an all-time high of $263.99 earlier this month – on numbers increasingly difficult for Wall Street to dismiss. Revenue reached $33.4 billion in Q4 FY2026, beating consensus estimates by roughly $1.8 billion. But the figure that fundamentally changed how investors view Dell was the $43 billion AI server backlog entering FY2027 – the largest disclosed AI infrastructure pipeline of any OEM globally.

According to IDC, the company also led the global OEM server market in Q4 2025 with $12.5 billion in quarterly server revenue and 10% market share. The competitive landscape, however, is tightening – though each major competitor still carries structural constraints.

HPE's AI backlog sits near $5 billion, though only a fraction of Dell's $43 billion pipeline. HPE is building credible sovereign AI and hybrid infrastructure offerings, particularly through its Cray HPC business and Juniper Networks acquisition, but it is monetizing the AI capex cycle at a structurally slower pace.

Lenovo, however, is the most underestimated challenger. Its Neptune liquid-cooling business grew 300% year-over-year, while its gigawatt-scale AI factory initiative with Nvidia signals serious long-term ambition. Lenovo's TruScale consumption-pricing model also resonates with enterprises hesitant to commit large upfront capital expenditures, and its Asia-Pacific footprint gives it stronger access to markets where Dell remains comparatively thinner.

Moreover, Cisco looms differently. The company increasingly sits beneath the AI infrastructure war itself, positioning around the networking, routing, and policy layers connecting distributed AI systems. Low-latency AI fabrics, intelligent traffic orchestration between agents, and network-level governance increasingly fall inside Cisco's territory – and directly underneath the broader infrastructure stack Dell is building and the company has recognized the threat.

Its AI Ecosystem Program, spanning Nvidia, Google, Hugging Face, Palantir, ServiceNow and SpaceXAI, further anchored by the Dell Automation Platform and Automation Studio, is partly a hedge against commodity hardware status in a market where value migrates toward orchestration.

"Validation at scale is the hard part," Clarke says. "The answer to fragmentation is a coherent architecture underneath all of it. Breadth without coherence is just noise."

The enterprise AI divide is already forming. On one side are companies rebuilding their infrastructure, data architecture, and operating culture around autonomous systems. On the other are enterprises dropping AI agents into environments designed for human workflows and hoping the system holds together.

"Inside almost every company deploying AI today, about 5% of the people drive 95% of the value," Clarke notes. "That gap compounds week after week. It's no longer 'where do I find AI engineers?' It's 'how do I get everyone to think and work like one?' And that requires genuine change management and the willingness to break down the cross-functional silos that keep AI value concentrated in pockets. No amount of infrastructure solves that problem for you."

The model was never going to be the hard part. It was always everything underneath it.