

Agentic AI Can Raise Token Use per Task Up to 100 Times, Accelerating the Shift Away From Per-Token Pricing, Futurum Research Finds

2026-09-25

Report finds reserved and owned infrastructure already account for 66% of AI compute consumption, and that a hybrid strategy relying on reserved capacity for predictable, high-volume inference workloads offers greater cost control, performance and flexibility

ATLANTA--(BUSINESS WIRE)-- QumulusAI (Nasdaq: QMLS), a neocloud infrastructure provider purpose-built for the AI computing era, today announced the release of a new Futurum Research report, sponsored by QumulusAI, which finds that agentic AI can increase token consumption per task by 10 to 100 times compared with a simple inference call. That increase can expose organizations using per-token services to unpredictable and rising costs as applications move into production and scale across the enterprise.

The report, *The Off Ramp From Per-Token Pricing: How Enterprises Regain AI Cost Control With Reserved Bare Metal*, examines how enterprises can control rapidly growing AI inference costs by adopting a hybrid strategy that supports moving mature production workloads from per-token pricing to reserved infrastructure. It draws on Futurum Research market forecasts and a survey of 824 AI decision-makers conducted during the first half of 2026.

Key report findings include:

- Reserved and owned infrastructure already account for 66% of AI compute consumption, compared with 19% for on-demand cloud.
- Of AI decision-makers surveyed, 59% primarily run workloads outside hyperscaler public clouds, including their own data centers, colocation facilities and bare-metal or high-performance computing providers.
- Inference spending is forecast to grow from \$120 billion in 2025 to \$885 billion by 2030.
- Agent and reasoning inference is expected to grow 219% in 2026, as multistep AI systems generate far more tokens than traditional prompt-and-response applications.
- AI-first cloud is the fastest-growing infrastructure tier, with spending forecast to increase 107.1% in 2026 and reach \$375.2 billion by 2030.

The report describes a typical progression from serverless APIs for initial experimentation, to on-demand GPUs, reserved infrastructure and ultimately a hybrid AI architecture. In this model, organizations match each workload to the right environment based on utilization, cost, latency, data governance and the level of control required.

“Per-token pricing makes sense when companies are experimenting, but the economics change quickly when AI applications move into sustained production,” said Michael Maniscalco, CEO of QumulusAI. “This research shows why enterprises need a more deliberate infrastructure strategy, using flexible services for experimentation and reserved capacity for predictable, high-volume workloads. The goal is not to replace the public cloud, but to put each workload where it makes the most economic and operational sense.”

Futurum Research recommends considering reserved bare-metal infrastructure for workloads with sustained and predictable utilization, high token volumes, specialized model-serving requirements, or data privacy and residency constraints. Bursty or experimental workloads may remain better suited to serverless or on-demand services.

“Enterprises are not choosing a single infrastructure model for AI,” said Brendan Burke, Research Director, Semiconductors, Supply Chain and Emerging Technology at Futurum Research, and author of the report. “They are building hybrid environments that retain on-demand flexibility while moving predictable production workloads to reserved capacity. As agentic AI raises token volumes, knowing when to make that move will become an important part of controlling costs.”

QumulusAI provides dedicated, reserved bare-metal and virtualized GPU infrastructure for AI inference and training. Its infrastructure gives customers direct control over their compute environment and replaces variable per-token charges with costs based on reserved hardware capacity and utilization.

The Off Ramp From Per-Token Pricing: How Enterprises Regain AI Cost Control With Reserved Bare Metal, is available by clicking to the [full report](#).

About QumulusAI

QumulusAI is a distributed AI cloud platform that delivers accelerated access to high-performance GPU compute. Through an inference-first, demand-led deployment model across a network of data center sites, QumulusAI brings compute closer to customer demand, helping AI teams and enterprises scale production AI workloads with speed, flexibility and control. By combining rapid deployment with flexible private cloud infrastructure, QumulusAI gives customers a faster, more adaptable path beyond the capacity constraints of traditional centralized and hyperscale cloud models. Learn more at [QumulusAI.com](https://qumulusai.com).

Follow us on [LinkedIn](#) and X [@QumulusAI](#).

Forward-Looking Statements

This press release contains forward-looking statements within the meaning of Section 27A of the Securities Act of 1933, as amended, and Section 21E of the Securities Exchange Act of 1934, as amended, including statements regarding the Futurum Research forecasts for growth in AI inference spending, agent and reasoning inference, and AI-first cloud spending, the anticipated increase in token consumption driven by agentic AI and its effect on the costs of per-token services, the expected shift of enterprise production inference workloads from per-token pricing to reserved and hybrid infrastructure, the anticipated cost-control, performance, and flexibility benefits of reserved

infrastructure for predictable, high-volume inference workloads, and the ability of the company's reserved bare-metal and virtualized GPU infrastructure to give customers direct control over their compute environments and to replace variable per-token charges with costs based on reserved capacity and utilization. Words such as "anticipate," "believe," "estimate," "expect," "guidance," "intend," "can," "may," "on track," "plan," "project," "target," "will" and similar expressions are intended to identify forward-looking statements. These statements are based on management's current expectations and assumptions as of the date of this release and are subject to risks and uncertainties that could cause actual results to differ materially, including, among others, the company's dependence on a limited number of large customers; the availability and cost of power, network connectivity and specialized hardware such as graphics processing units; the company's substantial capital requirements and access to financing; competition and rapid technological change in the high-performance computing and AI markets; the company's limited operating history and history of net losses; and those described in the "Risk Factors" section of the company's registration statement on Form S-1, as amended (File No. 333-292514), filed with the U.S. Securities and Exchange Commission (SEC), as such factors may be updated in the company's subsequent filings with the SEC. QumulusAI undertakes no obligation to update or revise any forward-looking statement, whether as a result of new information, future developments or otherwise, except as required by applicable law.

View source version on businesswire.com:

<https://www.businesswire.com/news/home/20260925592810/en/>

Investor Contact

investors@qumulusai.com

Media Contact

media@qumulusai.com

Source: QumulusAI