Futurum RESEARCH

Google Cloud

The Token Cost Reckoning Arrives in the CFO's Office

How to track cost per successful agent loop before your next budget review

SEPTEMBER 2026

Enterprise agents have started calling other agents, and the token volume coming off that traffic looks nothing like the early wave of typing into a chatbot. For most of the past two years, those tokens were priced below what they cost to serve because vendors were buying share. The subsidies are ending now, and the bills have come due at scale.

The wasteful habit of throwing the largest model at every problem, regardless of cost, is losing ground as a result. Industry watchers call it tokenmaxxing. Teams are opening an invoice and asking what they can afford. Finance already tracks cost per transaction on the cloud bill, and the AI line needs that habit applied to tokens. Only about 36% of AI Platform decision makers track cost per request, and barely 15% track token efficiency.[1] Enterprises are still early in AI adoption and need to build discipline now to avoid all-or-nothing usage.

The anatomy of an agent invoice

To measure your token spend, start with what one agent loop consumes in cloud resources:

- **Model calls** bill by input and output token, including reasoning traces.
- **Retrieval** acts as a recurring data tax because the model must re-read the conversation history on every turn, repeatedly billing for the same context.
- **Orchestration** keeps the container alive while the agent waits on a slow third-party API. Reserved accelerator capacity bills straight through the wait.

Consumption-based pricing rewards engineering discipline. The typical engineering response is to evaluate cheaper hardware. Four moves can significantly cut costs before you consider a switch:

- Cache the prompt prefix. The fixed block of standing instructions attached to every request is normally billed every time, but caching lets you pay once and reuse it at roughly a 90% discount.
- Route easy calls to a cheaper model. Send routine work to a cheaper model and reserve the premium one for what matters, which cuts the overall bill roughly 60–80% since the cheapest tier runs about 20% of the top one.
- Batch the non-urgent work. Any job that nobody is waiting on can be submitted for off-peak processing at a flat 50% off, in exchange for not getting it back instantly.
- Cap retries. Each automatic re-attempt on a failed job is billed, so a cap acts as a stop-loss that kills a runaway process on the fourth try instead of the fortieth.

Reserved pricing works the other way: you trade flexibility for a lower rate. That pays when traffic is predictable, and burns cash when it is lumpy. Early agent traffic is as lumpy as anything you will forecast this year.

1 Futurum Research: AI Platforms Decision Maker, 1H2026

The number worth managing isn't on dashboards yet: **cost per successful run, or (CPS): total spend divided by only the runs that produced something a person accepted.** A pipeline that completes 80% of what it starts quietly charges 25% extra for everything it delivers because failed runs are billed the same as good ones. That figure goes alongside the share of reserved accelerator hours spent waiting on tools instead of computing. **Together, those two numbers explain the most surprising invoices**.

$$\textit{Cost per successful run} = \textit{Total spend} \div \textit{Accepted runs}$$

That's the scoreboard. What moves it is four architecture choices that sound technical but wind up on the invoice: which tokens you buy, which silicon runs them, whether the workload can move, and whether you can get the hardware at all. Each is a question finance can put to the CIO.

Are you buying the right tokens for your task?

Frontier intelligence earns its premium on a narrow band of work. Summarizing a ticket or pulling fields from a form can be accomplished by a much cheaper model, and open-weight releases have pushed that floor close to zero. **Match the token to the job** and the blended cost per token comes down without changing anything a customer would notice. Match the task to a small model that passes evaluations and both cost per token and CPS fall in tandem.

Are you wasting budget on the wrong silicon?

In Futurum's 1H2026 AI Chipsets survey, accelerator availability limits cluster expansion for nearly 33% of buyers, ahead of budget and capital at 23%[2]. A team that's waited in a queue once will over-reserve the next time, and that hardware idles while somebody else waits. This creates massive CapEx waste. Siloed environments compound the problem — if you build one stack for core applications and another for AI, you are paying for idle headroom twice

Agent work is mostly orchestration — waiting on tools, moving data, deciding what to call next — which is why the CPU is back, and CPU-to-accelerator ratios keep tightening. Google's Axion, its Arm-based server CPU line, is built for exactly this. Google claims up to **twice the price-performance** of comparable current-generation x86 instances. **Pushing coordination onto cheaper silicon pulls down the blended cost of every agent transaction**.

Excess compute headroom gets bought twice — once for core applications, once for AI. One control plane to coordinate resources can close it.

CFOs should expect greater specialization at the compute layer to run continuous agent loops and complete more tasks per hour. You should start thinking about how you might diversify away from expensive options and create more heterogeneous compute environments among CPUs, GPUs, and XPUs.

2 Futurum Research: AI Chipsets Decision Maker, 1H2026

Can you move the workload if you have to?

Software is what keeps you from using more than one kind of compute. Revalidating a stack on a second accelerator architecture is a real engineering program that most companies fund once, and that's how multi-sourcing stays theoretical. None of that opportunity cost shows up on the cloud invoice. So price the system, not the part. Google Cloud's system is the AI Hypercomputer, the integrated stack from silicon through software that engineers integration costs out upstream. TorchTPU runs PyTorch natively on TPUs, so existing code interoperates with GPUs with minimal changes, and the accelerator becomes a choice. TPU 8i is one target, the inference chip Google claims runs 80% better per dollar than the prior generation. The same infrastructure spend can then serve materially more tokens while performing more reliably due to software compatibility.

Price the system, not the chip. The same infrastructure spend will then serve materially more tokens.

Are you maximizing utilization of committed compute?

When scaling AI, the enterprise reflex is to over-provision a standby buffer to ensure availability. But this leaves depreciating CapEx sitting idle, contributing to an industry-wide server utilization rate of just 12% to 18%. Instead of buying buffer hardware, build flexibility into the architecture.

The CFO has two jobs to maximize uptime: reserve for the demand you can predict, and automate for the demand you can't. For predictable jobs, Google Cloud's Dynamic Workload Scheduler books capacity ahead of known events with calendar mode for a time-bound launch, flex-start for batch and training jobs that can wait.

For everything else, resilience can be automated to minimize impact on the balance sheet:

- Write an ordered list of acceptable hardware.
- Let the platform take the next option when the first is unavailable.
- Migrate back when preferred hardware frees up.

Google Cloud's Google Kubernetes Engine (GKE) runs both workloads from one control plane across the application and AI estates. Custom compute classes store the fallback policy, and Dynamic Resource Allocation allows a job to claim an exact slice of an accelerator rather than an entire card. Commit to a three-year spend baseline, not specific hardware, and save up to 63% off when you sign up for Compute flexible committed use discounts. Availability and price-performance come from software substitution, not standby hardware.

What to ask in the next budget review

The budget data says this tokenomics conversation is overdue[1]:

- Infrastructure and core services take the biggest share of enterprise AI budgets at about **35.8%.**
- **47%** of enterprises expect their AI budget to grow by a quarter or more this year.

Across the four questions above sits **one design target** — **availability, scale, and economics, engineered together** — delivered by two mechanisms: capacity managed dynamically in software, and infrastructure bought as an integrated system, not parts. Each section hands finance a question for the CIO:

- What was your blended cost per token this quarter against last, and which model tiers moved it?
- What share of accelerators you own did useful work, and what would dynamic allocation recover before you buy more?
- If your biggest AI workload had to move to different silicon, what would the move cost, and did your cloud provider price that in?
- When first-choice hardware is unavailable, does software substitute the next acceptable option, or do you carry standby hardware instead?

Put **CPS** next to those answers and judge it all on a five-year horizon. The highest-value use cases are barely out of the gate, and a two-year experiment with token costs shouldn't limit usage of a revolutionary technology.

Sources

1 Futurum Research: AI Platforms Decision Maker, 1H2026 (AI budget allocation and 12-month change; budget by tech-stack layer; inference metrics tracked; AI success metrics; GenAI adoption challenges)

2 Futurum Research: AI Chipsets Decision Maker, 1H2026 (cluster expansion limits; scaling constraints).

Axion price-performance (up to 2x against comparable current-generation x86 VMs) and TPU 8i performance-per-dollar (up to 80% against the prior generation) are Google's published claims, included as vendor context rather than Futurum data

GKE Dynamic Resource Allocation and custom compute class behavior are as described by Google. The cost-per-successful-run illustration uses round numbers for shape rather than a specific deployment.

Important Information About This Report

Authors

Brendan Burke
Research Director, Semiconductors,
Supply Chain & Emerging Tech | The Futurum Group

Daniel Newman
CEO | The Futurum Group

Publisher

Futurum Research

Inquiries

Contact us if you would like to discuss this report and The Futurum Group will respond promptly.

Citations

This paper can be cited by accredited press and analysts, but must be cited in context, displaying author's name, author's title, and "The Futurum Group." Non-press and non-analysts must receive prior written permission by The Futurum Group for any citations.

Licensing

Disclosures

The Futurum Group provides research, analysis, advising, and consulting to many high-tech companies, including those mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

Google Cloud

About Google Cloud

Google Cloud is a full-stack AI platform, combining foundation models, data and analytics services, purpose-built accelerators, and cloud infrastructure to support enterprise AI adoption. Its AI infrastructure strategy centers on AI Hypercomputer and Google's Tensor Processing Units (TPUs), complemented by NVIDIA GPUs and an integrated software and networking stack designed to optimize performance, scalability, and efficiency across training and inference workloads. As AI workloads increasingly shift toward inference and agentic applications, token economics—the cost, speed, and energy efficiency of generating and processing tokens—are becoming increasingly important to the economics of AI deployment, an area where Google Cloud is emphasizing performance-per-dollar and low-latency inference. This combination of vertically integrated AI infrastructure and flexible consumption models positions Google Cloud to compete for organizations seeking to scale AI workloads while maintaining greater control over performance and cost.

Futurum

About The Futurum Group

The Futurum Group is an independent research, analysis, and advisory firm, focused on digital innovation and market-disrupting technologies and trends. Every day our analysts, researchers, and advisors help business leaders from around the world anticipate tectonic shifts in their industries and leverage disruptive innovation to either gain or maintain a competitive advantage in their markets.