Analyst Brief, sponsored by:
Penguin Solutions

Optimizing AI Infrastructure for Token Economics

WRITTEN BY:
Mary Johnston Turner
Research Vice President, AI and Workloads, Enterprise Infrastructure, IDC

PUBLISHED:
September 2026

Agentic AI adoption is exploding infrastructure budgets at a time when organizations are racing to innovate and deliver business value. This Analyst Brief shows how tech leaders can use token economics to evaluate AI infrastructure design choices and costs for optimal cost performance.

Introduction

IDC research shows that 78% of enterprise IT leaders agree AI infrastructure represents the single-largest cost element underpinning AI ROI and total cost of ownership (TCO) calculations. After accounting for the cost of AI models, agents, and development activities, infrastructure typically represents approximately one-third of total enterprise AI spending. POCs that started with relatively simple generative AI (GenAI) chatbots running on public cloud infrastructure are rapidly evolving into complex agentic AI workloads supporting hundreds of users and thousands of actions daily and are often subject to strict compliance and security requirements.

Enterprises need to find ways to better predict AI infrastructure performance and therefore align spending. The challenge, however, is that the ongoing runtime cost of AI infrastructure varies widely depending on the types of models, workloads, and infrastructure being used. To be able to conduct fair, consistent budget assessments and trade-off analysis, IT decision-makers are using AI tokens as a standard unit of measurement to compare the economics of running an AI workload using different deployment choices and system designs.

This Analyst Brief highlights how using a token economics–driven planning approach can help IT decision-makers evaluate the cost of AI infrastructure while ensuring each AI workload is deployed on the most cost-effective environment available to support that specific workload's performance, compliance, and security requirements.

Using token economics to frame AI infrastructure choices in the agentic era

Reliably forecasting token economics begins by understanding the operational demands specific to the workload and models being used. Input tokens used to train and fine-tune models are typically less expensive than output tokens used for GenAI inference queries and agentic AI workflows, due to the different ways the tokens are processed. Input tokens are typically processed in parallel using very large-scale, high-performance accelerated computing resources. Because resources are shared, training workloads can maintain very high levels of system utilization and efficiency, often running in hyperscale or neocloud datacenters and GPU-intensive AI factories.

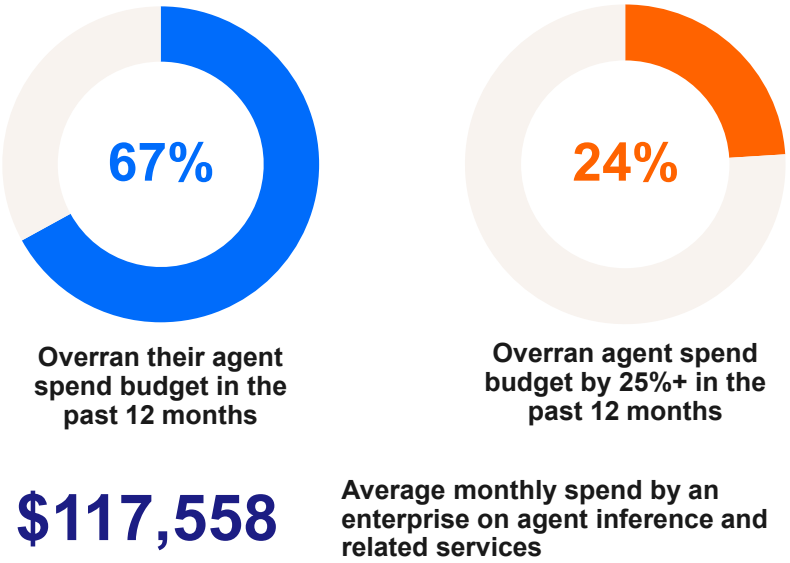
By comparison, output tokens require sequential token processing that cannot be parallelized the same way. They may be generated by smaller, more targeted models that do not require the same level of computational intensity as is needed for larger model training but demand extensive agent-to-agent connectivity and data sharing. Output generation is inherently slower and more memory intensive than input processing, which also makes it harder to predict costs.

IDC's research shows that the expected growth of agentic AI is poised to broadly disrupt IT budgets. Worldwide, 95% of enterprises have at least 1 AI agent in production, with the average

enterprise running 11 agentic workflows today and planning for many more. The rapid rollout of agents has caused 67% of organizations to overrun budgets for agent spending in the past 12 months. Almost a quarter (24%) overran agent spend budgets by 25% or more. IDC estimates the average enterprise agent monthly spend is over $117,000 (see Figure 1). Costs will only increase as agentic AI becomes more widely used, with the potential to fuel agent budget overruns on the order of tens of thousands of dollars a month or more for many organizations that fail to get better insight and control over their business' token requirements.

Figure 1

Rate of agent spend budget overruns



n = 1,835 worldwide IT decision-makers

Source: IDC's *Future Enterprise Resiliency and Spending Survey (FERS Survey)* Wave 4, July 2026

Organizations using token economics planning approaches often start by looking at the trade-offs between usage-based public cloud services and owning dedicated on-premises infrastructure systems. What they quickly find is that, regardless of deployment location, system design choices can often have an even greater impact on overall AI performance and costs. Reasoning chains, tool calls, and interagent messages are almost entirely output tokens and compound across every step of a multiagent task. Extended context windows can also drive up the number of tokens and the cost of creating them. A single user query might look cheap, but the agentic system generates dozens of intermediate output tokens behind the scenes. This can result in large and unpredictable charges appearing unexpectedly.

Critical AI infrastructure bottlenecks and system design choices

Token economics helps simplify the comparison of cost and performance across a range of infrastructure options. The cost per token reflects the full infrastructure stack — compute, memory, network, storage, management, facilities, power, and cooling — needed to generate each AI unit of work (i.e., a token).

The best way to ensure that AI infrastructure systems are optimized for performance and token economics is to carefully match system design choices with the needs of the workload. Important design choices and potential bottlenecks to consider include:

- **Computational resources.** Requirements differ sharply between small models working across a narrow set of data and large-scale fine-tuning or training of frontier or open-weight models.
- **Memory performance.** Bottlenecks appear when the speed of GPUs, CPUs, and accelerators outpaces available memory bandwidth and capacity, limiting the size and complexity of workflows that can run efficiently.
- **Networking and interconnect.** Fabric limits constrain scale-out inference and distributed agentic workloads.
- **Storage I/O.** Contention builds in multitenant inference environments.
- **Configuration and life-cycle management.** The software and practices used to configure, orchestrate, monitor, and tune the cluster determine whether it holds optimal performance across its full life, not just at acceptance testing.

System design choices have a significant impact on system efficiency and cost per token.

Enterprises should anticipate that infrastructure requirements will evolve as context windows expand, concurrency requirements increase, and model architectures become more sophisticated. Decision-makers need to plan how they will deal with growth over time in a modular, scalable, and cost-effective manner.

Shaping your token economics strategy

As business leaders lean into using AI across almost every business process, the volume and cost of inference and agentic token processes threaten to overwhelm budgets and slow implementation. Comparing the cost per token for different infrastructure architecture and deployment options helps decision-makers balance cost control priorities against ensuring access to the right resources to get the job done. Key considerations when designing your organization's token economics strategy should include:

- Understanding the needs of the workloads in terms of output/input token balance, performance SLAs, compliance and security requirements, and scale of users (both human and agentic)

- Evaluating cost per token in the context of business outcomes and SLAs (The cheapest token option may not always be the best for a specific workload.)
- Examining system design, model choices, context, and configuration trade-offs to split up work and optimize integrations across complex workflows
- Planning for scenarios where accelerated adoption and growth drive large, unexpected increases in token volumes over compressed periods of time
- Engaging across business and technology teams to align budgets, SLAs, and outcomes

Token economics represents a new approach to infrastructure planning that aligns engineering, business, and budgetary realities to drive AI-enabled innovation.

About the IDC analyst



Mary Johnston Turner
Research Vice President, AI and Workloads, Enterprise Infrastructure, IDC

Mary Johnston Turner is a research vice president in IDC's Enterprise Infrastructure global research domain and part of the AI and Workloads subdomain. She oversees IDC's research on AI infrastructure strategies, which takes a tech leader–centric view of enterprise AI infrastructure markets and adoption.

More about Mary Johnston Turner

Message from the sponsor

About Penguin Solutions

Penguin Solutions (Nasdaq: PENG) is the AI Factory Platform Company. We design, build, and manage next-generation data centers for enterprises, sovereign AI initiatives, and neocloud providers.

With deep design expertise at the intersection of data center AI infrastructure and memory solutions, our Full-Stack AI Factory Platform combines differentiated infrastructure software, advanced memory, compute systems, end-to-end services, and industry-leading partner technologies to help customers accelerate deployment, optimize token economics, and maximize the return on their AI investments.

Learn more at PenguinSolutions.com.

IDC Custom Solutions

IDC Custom Solutions produced this publication. The opinion, analysis, and research results presented herein are drawn from more detailed research and analysis that IDC independently conducted and published, unless specific vendor sponsorship is noted. IDC Custom Solutions makes IDC content available in a wide range of formats for distribution by various companies.

This IDC material is licensed for external use, and in no way does the use or publication of IDC research indicate IDC's endorsement of the sponsor's or licensee's products or strategies.



IDC Research, Inc.
One Beacon Street, Suite 33100, Boston, MA 02108, USA
T +1 508 872 8200

blogs.idc.com | LinkedIn @IDC | www.idc.com

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications, and consumer technology markets. With more than 1,300 analysts worldwide, IDC offers global, regional, and local expertise on technology and industry opportunities and trends in over 110 countries. IDC's analysis and insight help IT professionals, business executives, and the investment community make fact-based technology decisions and achieve their key business objectives.

©2026 IDC. Reproduction is forbidden unless authorized. All rights reserved. CCPA