

How Nutanix Built a Private AI Factory to Control Token Costs

The Expected Impact

- **1-3 year Payback Window:** Nutanix expects to achieve a payback on its AI factory within one to three years.
- **\$0 Per-Token Pricing:** Targeting an incremental per-token price of \$0 for open-weight models running on owned Nutanix infrastructure.
- **A Clearer Path to Financial Value:** Expected to demonstrate the financial benefits of owning and operating infrastructure compared with ongoing per-token vendor costs.

Unsustainable Agentic Coding Costs

Nutanix is a hybrid cloud leader and AI platform company, offering organizations a unified infrastructure software platform to safely run applications, data, and AI anywhere. Powering modern workloads across thousands of global enterprises, Nutanix relies on an internal engineering powerhouse of over **4,000** developers, QA engineers, and site reliability engineers operating across a rapid software development lifecycle.

When third-party AI coding tools were introduced across the Nutanix engineering team, usage grew rapidly, reaching nearly full adoption among thousands of engineers, including QA and software engineering. Initially, licensing operated on a fixed pricing structure per API. However, as the vendor shifted their commercial terms to a strict per-token pricing model, the cost of using these coding tools powered by frontier models in the public cloud became increasingly difficult to sustain.

This challenge was further amplified by the nature of agentic coding. AI agents can support work across the software development lifecycle, from code generation and QA testing to front-end design and debugging. Unlike traditional chatbot interactions, which often involve a single prompt and response, agents can involve multiple steps, iterations, and tool calls, consuming significantly more tokens per task.

Nutanix engineering leaders realized that using metered cloud frontier models was driving up token costs for engineering workloads. This financial risk was further escalated by agentic AI coding workflows, which could drive substantial monthly token consumption.

Industry

- Software Development

Location

- San Jose, California, USA
- 8,300 employees worldwide

Applications

- Agentic AI Coding, QA Triage

Products

Nutanix Enterprise AI (NAI)

- Nutanix Agent Gateway
- Nutanix Private Inference

Nutanix Kubernetes Platform (NKP)

Nutanix Cloud Infrastructure (NCI)

- Nutanix AHV

Nutanix Unified Storage (NUS)

Nutanix Database Service (NDB)

Solutions

- Agentic AI
- Hybrid Multicloud
- Sovereign Cloud

Investment and Expected Impact

- \$20 million internal AI cluster investment
- Expected payback within 1-3 years
- \$0 per-token pricing using open-weight models, on-prem
- Enabled greater control over data sovereignty by self-hosting open-weight models on-prem

“Developers expect the best models to be available for their software engineering needs,” shared Ganesh Ramesh, Technical Director and AI agent architect at Nutanix. “The token usage for frontier models can quickly balloon into prohibitive costs with the rise of agentic workflows that can be unsustainable in the long run. We needed an infrastructure strategy that would keep frontier model costs constant while enabling the engineering team to maintain their level of innovation. This gives them the freedom to build without having to worry about costs.”

Nutanix needed a way to support its AI adoption at scale, while improving cost predictability, maintaining performance, and reducing dependence on external frontier model providers.

Building an AI Factory to Own Your Intelligence

To address rising AI costs and improve long-term cost predictability, Nutanix leaders saw an opportunity to shift from recurring per-token charges toward an investment in infrastructure it could own and operate.

The company invested **\$20 million** to build a self-operated AI factory running open-weight models, shifting a portion of its AI workloads from external services to infrastructure it owns and operates. Transitioning away from purely external cloud services wasn't as simple as just purchasing hardware. It required sourcing hardware against supply chain constraints, securing funding, demonstrating an expected payback, and coordinating across multiple internal teams across the organization.

“Every enterprise hitting production scale faces the same reality: token bills escalate fast when you send every prompt to a frontier model,” explained Debojyoti Dutta, Chief AI Officer at Nutanix. “We built our strategy around an 80/20 model: offloading bulk context and repetitive tasks to dedicated on-prem, open-weight models, and using frontier APIs surgically. That shift gave us the performance, data privacy, and budget predictability we needed to scale without compromise.”

To eliminate unpredictable cloud token meters and service these AI coding workloads, Nutanix deployed a full-stack, on-premises AI Factory designed to act as an internal token generation engine. Nutanix's strategy combines self-hosted, open-weight models with external frontier models, giving the company flexibility in how it manages AI costs, performance, and data privacy. The approach centers on four key elements:

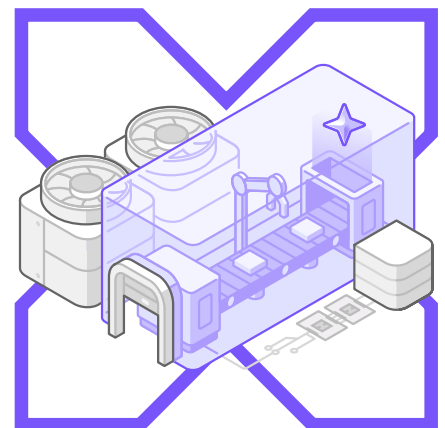
- **Hybrid 80/20 Routing Strategy:** The team allocates **80%** of routine, high-volume AI coding tasks to open-weight models hosted on its own private AI cluster to enable predictable costs with no per-token bill. They allocated the remaining **20%** of complex, high-reasoning tasks to external frontier models in the public cloud.

“Building a private AI factory isn't just about buying GPUs. The AI Factory is a full stack solution.”

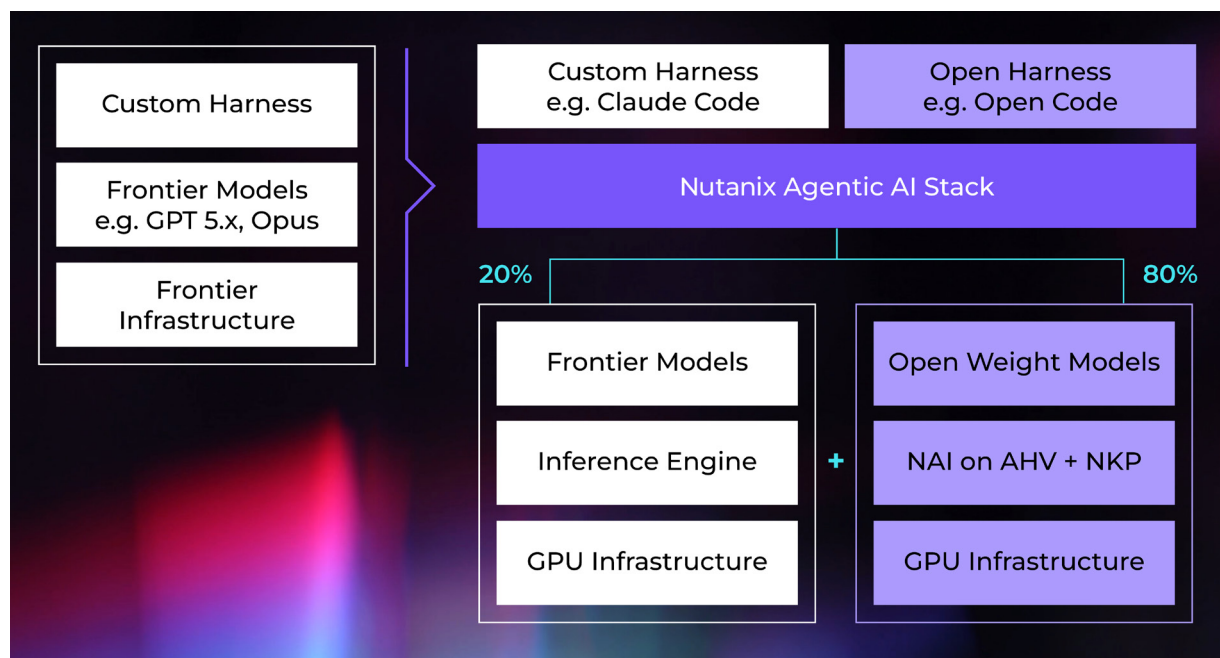
Inbar Stolberg
Senior Director, Engineering
Nutanix

- **AI-Optimized Infrastructure:** Nutanix uses its full-stack Agentic AI solution to run self-hosted AI workloads on its infrastructure.
 - **Nutanix Enterprise AI (NAI):** Enterprise-grade platform for managing and scaling AI workloads across hybrid environments. It hosts and serves on-prem, open-weight models with enterprise-grade inference optimizations through Private Inference, and controls AI traffic through Agent Gateway, giving IT teams a single front door to enforce token rate limits and quotas, tool access, role-based access controls, and audit logging capabilities.
 - **Nutanix Kubernetes Platform (NKP):** Abstracts away the complexity of Kubernetes to give Nutanix AI teams a reliable, production-ready environment. Deep integration with the underlying infrastructure automates node provisioning and GPU resource allocation, allowing engineers to focus on model performance and developer velocity rather than cluster maintenance.
 - **Nutanix Cloud Infrastructure (NCI):** Software-defined platform running directly on self-owned hardware in a self-managed datacenter, integrating compute, storage, and networking into a single unified foundation. This allows a seamless path to integration with existing IT processes and tooling, enabling operation and scale of private AI workloads using the same consistent operations and management workflows they rely on across their existing infrastructure.
 - **Nutanix Unified Storage (NUS):** Provides data services through high performance file and object storage, optimized for hosting and serving on-prem, open-weight AI models leveraged by Nutanix Private Inference, while providing operational metrics and system observability.
 - **Nutanix Database Service (NDB):** Provides a managed PostgreSQL instance in each environment to store Nutanix Enterprise AI data, including application state, model definitions, endpoint configurations, and access control information.
- **Scalable AI Infrastructure Operations:** Enables IT to run, scale, and manage AI workloads with resiliency, reliability, and operational control across the full infrastructure stack.
- **Ensure Data Sovereignty:** Hosting open-weight models on Nutanix owned infrastructure keeps proprietary code, IP, and sensitive data within Nutanix's own environment. Supporting capabilities including role-based access control and governed, audited agent access to tools and data provide control and visibility.

“Building a private AI factory isn’t just about buying GPUs. The AI Factory is a full stack solution,” said Inbar Stolberg, Senior Director, Engineering at Nutanix. “It’s about creating a platform that can give our engineers the tools they need without the public cloud price tag. For us, that means delivering a lower cost per inference than commercial services, with strict governance, resiliency, and reliability on the new AI clusters.”



Nutanix Agentic AI Factory Architecture



Open-Weight Models running on the AI Factory include:

- Laguna S 2.1 (Poolside)
- NVIDIA Nemotron 3 Super (NVIDIA)
- Muse Glimmer (Meta)
- Muse Spark (Meta)
- Beam (Reflection)

Additional models are also planned to be deployed soon, including Thinking Machines' Inkling model. Larger models (e.g. Laguna, Beam) are run across multiple GPUs to optimize performance and throughput, leveraging techniques such as tensor parallelism and KV cache aware routing. The Agent Gateway also routes to frontier models like GPT-5.6-Luna (OpenAI) via Amazon Bedrock.

The Customer Playbook for Enterprise AI

As AI adoption expands across the software development lifecycle, organizations face growing pressure to manage rising token costs while protecting sensitive data. For Nutanix, the AI Factory is both an investment in its own software development productivity and an opportunity to demonstrate how enterprises can manage AI costs and support data sovereignty objectives.



Nutanix's approach provides one example of how an organization can address these challenges by targeting an 80/20 split: running **80%** of everyday agentic coding tasks locally on open-weight models, while the remaining **20%** of high-end reasoning tasks get escalated to frontier cloud models. This operational model can be extended beyond AI coding to other core enterprise workloads, from everyday cowork productivity and internal search and data querying to real-time security threat triage. By combining owned GPU infrastructure, open-weight models, and access to frontier models, Nutanix's internal deployment illustrates one possible framework that gives IT and business leaders control over AI costs and operational scale, while giving engineers and developers the flexibility to innovate.

Ready to Control Your AI Costs and Data?

Learn how Nutanix Agentic AI can help organizations like yours gain greater control over their AI costs and data sovereignty.

[Explore Nutanix Agentic AI today.](#)

NUTANIX

info@nutanix.com | www.nutanix.com | [@nutanix](https://twitter.com/nutanix)

©2026 Nutanix, Inc. All rights reserved. Nutanix, the Nutanix logo and all Nutanix product and service names mentioned are registered trademarks or trademarks of Nutanix, Inc. in the United States and other countries. Kubernetes is a registered trademark of The Linux Foundation in the United States and other countries. All other brand names mentioned are for identification purposes only and may be the trademarks of their respective holder(s).NutanixonNutanix-CaseStudy-10/23/2026