From Tokens to Tokenomics:
Optimizing AI Costs with Open Source



Blake Treves
Managing Director, GlobalLINK Product, State Street
btreves@statestreet.com



Jennifer Bowman
AI Specialist - FSI, Red Hat
jebowman@redhat.com



Aric Rosenbaum
Chief Technologist
Financial Services, Red Hat
aric@redhat.com

Red Hat

Blake Treves

State Street, Managing Director, GlobalLINK Product



e: BTreves@statestreet.com

m: +1-617-462-9562

Background

- Blake is a Managing Director at State Street, leading LINK and GL Pay, with a focus on new product development, innovation, and go-to-market strategy across Global Markets.
- In his decade at, State Street, he has held leadership roles spanning product management, institutional trading, payments and financial technology.
- His work focuses on identifying emerging market opportunities and translating new technologies into scalable, commercially successful products.

Area of Expertise

Capital markets, product innovation, new product development, go-to-market strategy, AI and automation, interoperability, institutional payments, multi-asset sales, trading & analytics

Selected Experience

• Leads LINK and GL Pay, developing next-generation products that connect institutional workflows, technology, and payment capabilities.

• Combines product strategy, technology, and commercial execution to bring innovative financial solutions from concept to market.

Red Hat

Aric Rosenbaum

Red Hat, Chief Technologist, Global FSI



e: aric@redhat.com

m: +1-973-610-4671

Background

- Aric serves as the Chief Technologist in Red Hat's Global FSI team, where he helps clients meet their strategic priorities through the use of open source technology.
- Prior to joining Red Hat, he led large, digital transformation projects at Goldman Sachs' Investment Management Division and was co-founder / CTO of several FinTechs in equity and FX trading.
- In his 20+ years of experience, Aric has designed and built multiple front, middle and back office systems and deployed his first app to the cloud in 2008.

Area of Expertise

Financial services, capital markets, FinTech, microservices, continuous integration / delivery, agile, cloud computing, hybrid cloud

Selected Experience

At Goldman Sachs, Aric co-led a global team of 60+ engineers building the next generation client lifecycle management system.

Aric served on the Governing Board of the Fintech Open Source Foundation (FINOS) from 2022-2024.

Red Hat

Jennifer Bowman

Red Hat, AI Specialist, FSI



e: jebowman@redhat.com

m: +1-412-608-3487

Background

- Jennifer serves as an AI Specialist in Red Hat's AI Specialist team supporting FSI, where she helps clients navigate the fast-paced world of AI and map technology to their strategic priorities.
- Before joining Red Hat, she served on the go-to-market team at Neural Magic prior to its acquisition by Red Hat in January 2025, leading sales GTM strategy and enterprise adoption initiatives.
- Before moving into AI, she built a 12-year career in cybersecurity, helping organizations strengthen their defenses and navigate complex cyber transformations.

Area of Expertise

Financial services, enterprise inference, Agentic AI, compliance, security, governance, hybrid cloud

Selected Experience

Earlier in her career, she successfully established and grew cybersecurity practices at regional managed security services providers, overseeing the full lifecycle from initial service creation and client acquisition to managing engineering resources.

Red Hat

Agenda

Own the Model, Control Your Destiny: Open Source AI for the Enterprise



The Economics of AI



Perspective from State Street

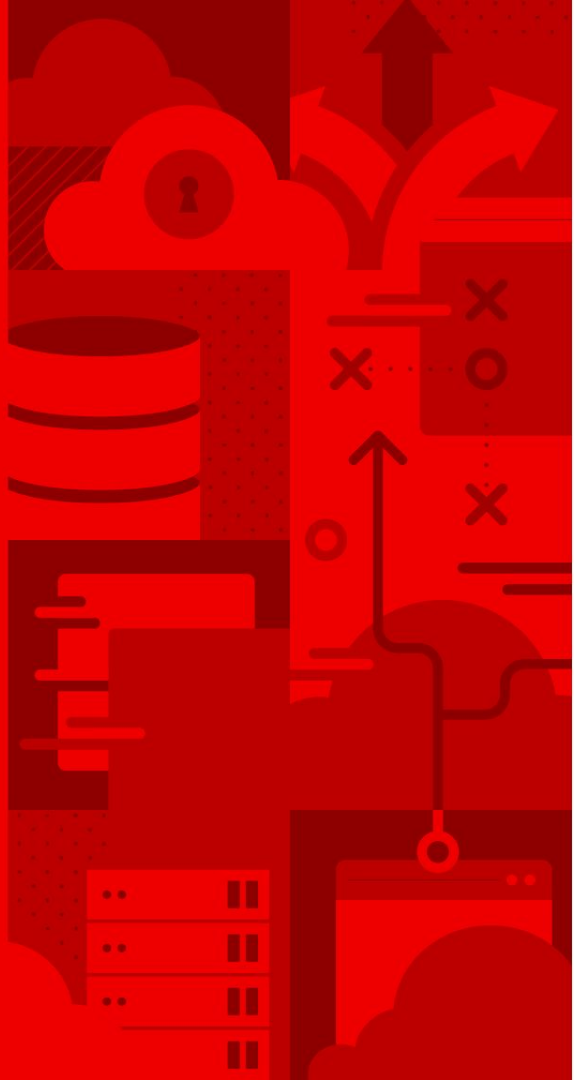


Open or Closed: Who Holds the Levers?



Open Source AI for the Enterprise

Red Hat



The Economics of AI

Red Hat

Google: $811 billion in Purchase Commitments

As of: July 22, 2026

2026 Q2

Purchase Commitments and Other Contractual Obligations

As of June 30, 2026, we had material purchase commitments and other contractual obligations totaling $811.0 billion, of which $200.7 billion was short-term. These purchase commitments primarily relate to costs for technical infrastructure and inventory through long-term supply agreements and open purchase orders. Additional contractual obligations include commitments for content licenses and energy take-or-pay contracts. For additional information related to our long-term supply agreements, energy take-or-pay contracts, see Note 10 of the Notes to Consolidated Financial Statements included in Item 1 of this Quarterly Report on Form 10-Q.

Source: Alphabet 10-Q (Q2 2026)

2026 Q1

Purchase Commitments and Other Cont

As of March 31, 2026, we had m…mitments and other contractual obligations totaling $332.4 billion, of which $138.0 billion …ese purchase commitments primarily relate to costs for technical infrastructure and inventory thr…m supply agreements and open purchase orders. Additional contractual obligations include commitmen… content licenses and energy take-or-pay contracts. For additional information related to our long-term supply agreements, energy take-or-pay contracts, and content licenses, see Note 10 of the Notes to Consolidated Financial Statements included in Item 1 of this Quarterly Report on Form 10-Q.

Source: Alphabet 10-Q (Q1 2026)

+ $479 billion in 3 months

Red Hat

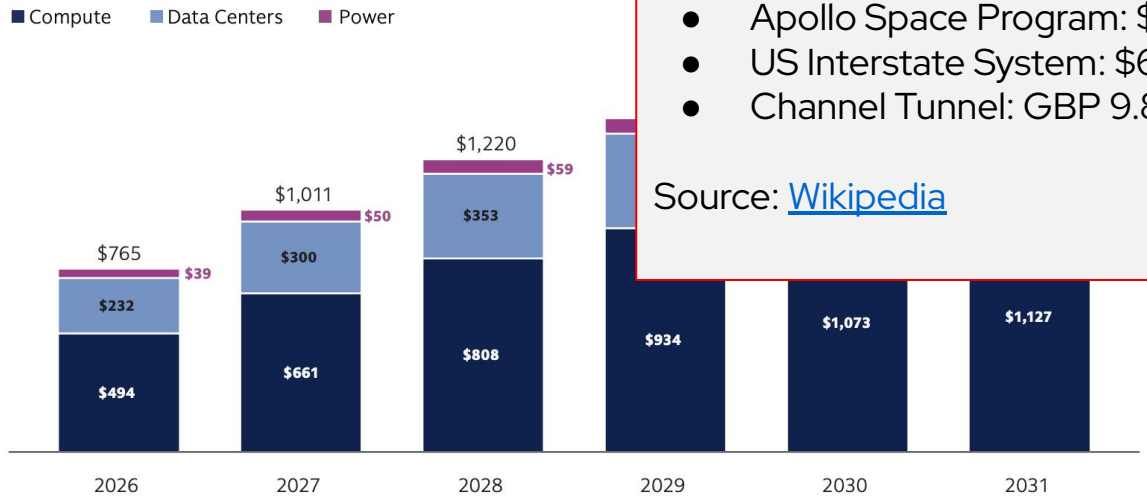
$7.6 trillion (est) Spend over 6 Years (per Goldman Sachs)

Assumptions: $80k per GPU, $15mm per MW for data center spend

Baseline Estimates

Baseline aggregate AI CapEx estimates (bn)

~$7.6tr of capital between 2026 and 2031 across compute, data c



Source: Goldman Sachs Global Institute, Goldman Sachs Global Investment Research NVIDIA projections (as of March 3, 2026)
Note: Forecasts and expectations are based on material assumptions subject to change. Assumes NVIDIA accounts for 75% of total compute spend in each period. Assumes 5% YoY compute growth past the projection period (2031). Uses VR200 (Rubin) chip as baseline spec ($80.5K per GPU [incl. node costs] and 3,000 W per package) across all years. Assumes 1.2 PUE, $15mn per MW for data centers, and $2,500 per kW for new power. Assumes 15% of required data center space is brownfield (i.e., excluded from calculation) in 2026, growing to 30% in 2031. Totals may not sum due to rounding.

To put this in perspective:

- Apollo Space Program: $257 billion (2023 dollars)
- US Interstate System: $634 billion (2024 dollars)
- Channel Tunnel: GBP 9.8 billion (2025 pounds)

Source: Wikipedia

Source: Goldman Sachs (May 1, 2026)

Red Hat

From training debt to token taxes

AI's bottom line remains its ultimate bottleneck

FORTUNE

Subscribe

AI • TECH

Uber burned through its entire 2026 AI budget in four months. Now its COO is questioning whether it's worth it

By Jake Angelo
Former News Fellow

May 26, 2026, 2:03 PM ET



Uber COO Andrew Macdonald
ZED JAMESON/BLOOMBERG VIA GETTY IMAGES

yahoo/finance

Search for news, tickers or comp...

stocktwits

WMT Reportedly Caps Employee AI Usage — Stock Down Nearly 2%



Forbes

Subscribe: Less than $2/wk

MONEY > INVESTING

Microsoft Ends Claude Code Licenses As It Shifts Developers To Copilot

By Jon Markman, Contributor. Analyzing tech stocks through the...

Follow Author

Published Jun 01, 2026, 01:01pm EDT, Updated Jun 01, 2026, 02:44pm EDT

Add Us On Google



Inc.

NEWSLETTERS

SUBSCRIBE

TECHNOLOGY

AI Was Supposed to Save Companies Money. Instead, It's Blowing Up Budgets in a Big Way A new report from KPMG finds that executives are aghast at their AI bills. It's a warning for any business leader planning to go all-in on the tech.

BY CHRIS MORRIS @MORRISATLARGE

JUL 13, 2026

AI AGE

Chamath Palihapitiya says soaring AI token spend will hit companies' earnings

PUBLISHED TUE, JUL 14 2026•10:45 AM EDT UPDATED 16 MIN AGO

Ashley Capoot
@/IN/ASHLEY-CAPOOT/

WATCH LIVE

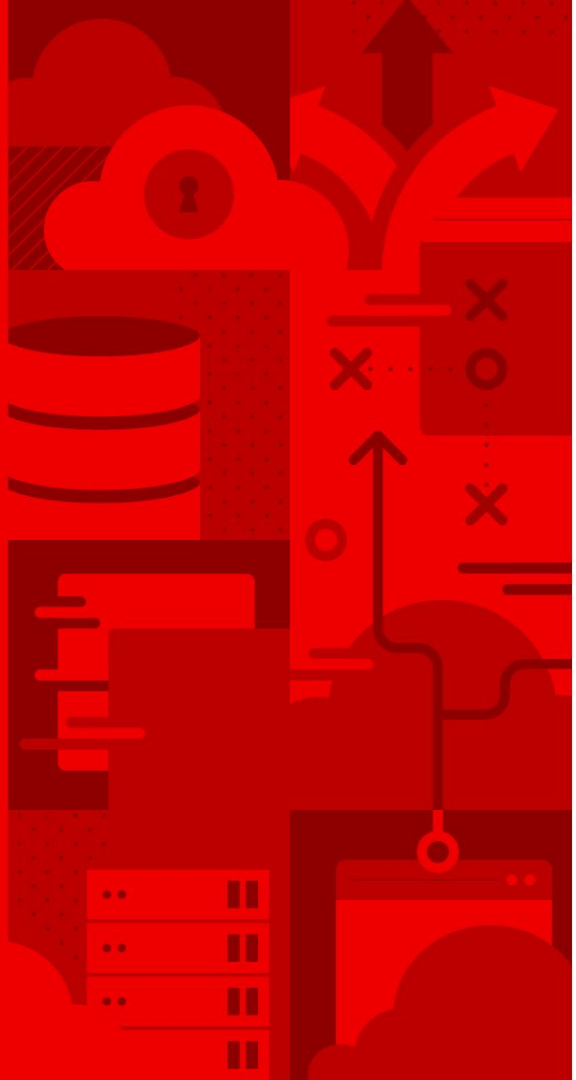
KEY POINTS

- Tech investor Chamath Palihapitiya said AI usage could negatively impact some companies' earnings.
- "CEOs and the CFOs, in my opinion, probably have no idea how much tokenmaxxing is going on inside of their organizations," Palihapitiya told CNBC.
- He is part of a growing chorus of investors and tech executives who are cautioning that the tokenmaxxing era is coming to an end.

CNBC MARKETS VIDEO WATCHLIST MENU

Sources: Fortune article | Forbes article | Yahoo Finance article | Tokenmaxing | Inc Article

Red Hat



Perspective from State Street

Red Hat

AI Value: Measure the Outcome

Assistance was easy to deploy. Better outcomes + solid economics are next.

01 ASSIST

Make knowledge usable

Search approved information; summarize, draft and explain.

UNIT OF VALUE

INDIVIDUAL PRODUCTIVITY

Minutes saved × verified adoption

REQUIRED CONFIDENCE

Grounded answers, data access, quality checks

02 AUGMENT

Improve decisions in context

Interpret signals, explain anomalies and surface options with evidence.

UNIT OF VALUE

DECISION QUALITY

Speed to decision + accepted insights

REQUIRED CONFIDENCE

Provenance, evaluation, human accountability

03 ACT · CONTROLLED

Complete governed steps

Perform or request authorized actions. Clear guardrails and interfaces with existing controls.

UNIT OF VALUE

OPERATIONAL IMPROVEMENT

Revenue (or Cost) – cost to serve

REQUIRED CONFIDENCE

Scoped authority, limits, approval, audit

THE VALUE EQUATION | Measure business benefit vs. integration, oversight and exception costs

Information Classification: Limited Access

Intelligence Across a FinTech Platform

Connect client intent to existing capabilities. Shared controls + differentiated experiences

01 CLIENT EXPERIENCE · LINK

From discovery to next action

Use client context to navigate services, answer workflow questions and reduce repeated handoffs.

TIME-TO-VALUE · CLIENT ENGAGEMENT

02 TRADING & ANALYTICS

From signals to defensible judgment

Explain execution quality, surface outliers and support investigations with traceable evidence.

DECISION SPEED · TRADING OUTCOMES

LINK

03 SUPPORT · CLIENT SERVICE

From fragmented steps to flow

Pull together data across multiple systems to solve client requests faster. Empower support teams to learn.

CLIENT SATISFACTION

04 DATA & PARTNER ECOSYSTEM

From information to embedded value

Deliver permissioned internal and partner insight within the client workflow. Transparency + consent at all times.

INSIGHT-TO-ACTION

SHARED RAILS Identity · entitlements · data provenance · service contracts · evaluation · telemetry · cost controls

Information Classification: Limited Access

LINK AI: Flexible Delivery Model

OUTSIDE-IN

Client-owned AI & agents

Let client tools discover and invoke permitted capabilities through governed interfaces.

INSIDE-OUT

Intelligence embedded in LINK

Use trusted context to guide users from a question or signal to the next valid action.

GOVERNED CAPABILITY LAYER

DISCOVER	AUTHORIZE	EXECUTE	GOVERN
Service catalog & tool semantics	Identity, entitlements, scopes & limits	Deterministic APIs, workflow & approvals	Audit, telemetry & risk management



Information Classification: Limited Access
Information Classification: Limited Access

AI Value: From Fast Assistance to Controlled Automation

ASSIST

- Search enterprise content, summarize, explain, draft
- Communication tools
- Code assistance and troubleshooting
- "Use at your own risk"

Almost instant productivity value

AUGMENT

- Trusted enterprise content (guardrails)
- Decision support + workflow guidance
- Governance + feedback loops
- Some risks managed

Measurable task level value

ACT (CONTROL)

- Authorized actions across systems
- Monitoring, approvals + exception handling
- **Apply to the right use case**

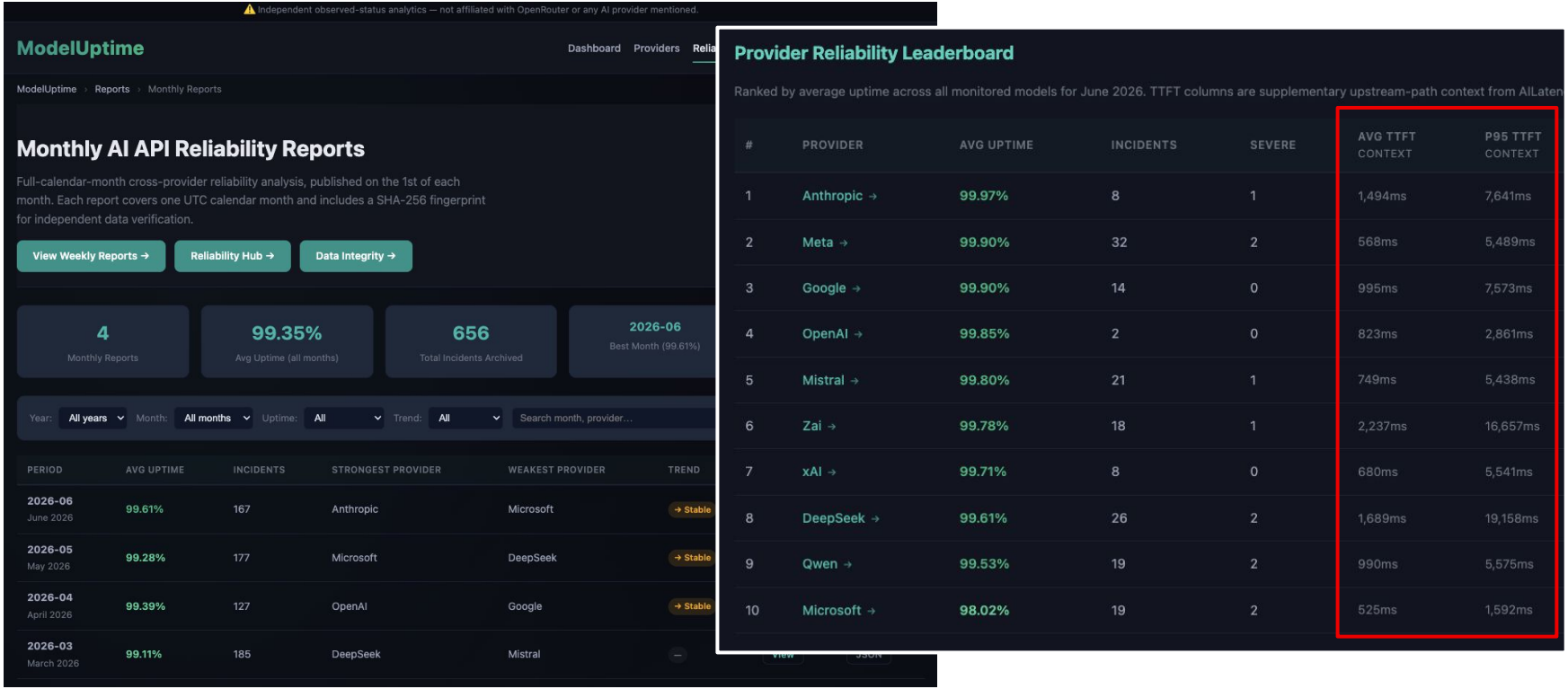
Process level ROI

Demand continues to increase for identity, data governance, permissions, auditability and resiliency in AI projects

Information Classification: Limited Access

Does the Uptime and Latency Match the App's Requirements?

Own the Model, Control Your Destiny: Open Source AI for the Enterprise



Source: ModelUptime.com

Red Hat

Operating Principles

As we engineer our solutions, we need to address core questions

Operating Principles

The operating principles changes from finding raw chips to orchestrating and governing token consumption efficiently.

01. Governance
Who gets access?

02. Gateway
Which model and where is it hosted?

03. Inference
How efficiently is it processed?

04. Token Tracking
How much usage per user, team, or task?

Red Hat

AI Adoption Maturity: From Tokens to Tokenomics

Where are you on the maturity curve and how will you get to the next level?

L1: Initial

Focus

Consuming public SaaS tools (e.g., ChatGPT) with no internal infrastructure.

Architecture

100% Third-party / Public Cloud.

Challenge

High "hidden costs" and zero operational control.

L2: Developing

Focus

Identifying qualified use cases and building foundational AI skills.

Architecture

Experimental exploration of internal "Workbenches."

Challenge

Fragmented experimentation and initial budget planning.

L3: Operational

Focus

Deploying initial POCs and validating performance vs. cost (Inference benchmarking).

Architecture

Hybrid entry point—combining frontier models with local model testing (InstructLab).

Challenge

Managing the "production gap" between lab and scale.

L4: Optimizing

Focus

Implementing MLOps pipelines and formal AI governance.

Architecture

Scaled hybrid architecture—routing workloads between open and closed models.

Challenge

Balancing performance, governance, and predictable "Tokenomics."

L5: Innovator

Focus

Strategic innovation where AI is embedded in core business workflows.

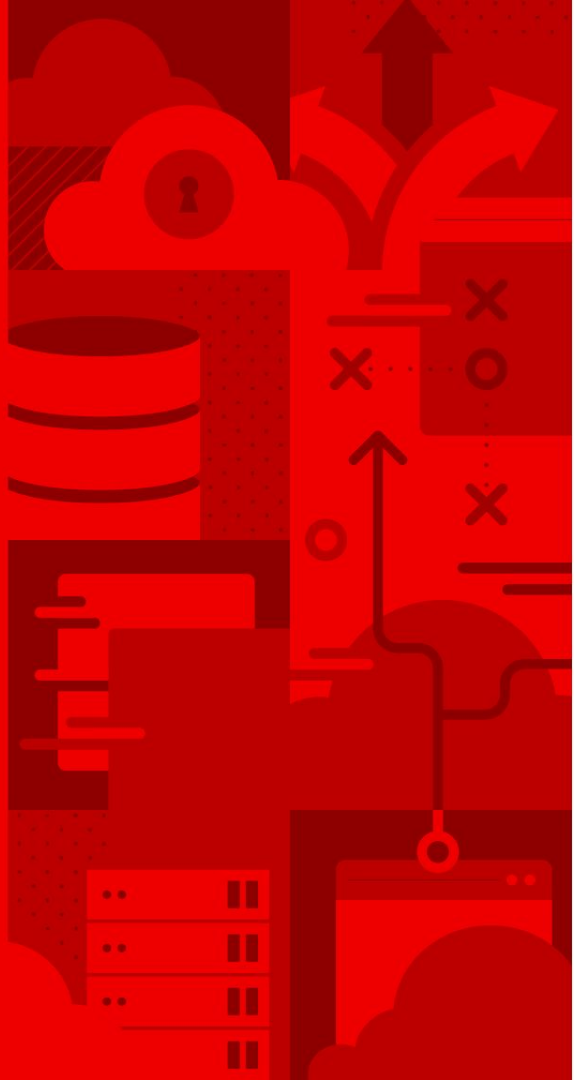
Architecture

Sovereign AI—acting as a "Token Provider" with self-hosted, optimized inference.

Value

Measurable business value through automated, high-performance architectures.

Red Hat



Open or Closed: Who Holds the Levers?

Red Hat

Time Machine: 1989

Dataquest: 82.2% of the workstation market was divided between 5 proprietary vendors











Red Hat

Token Costs Vary Widely

Input Costs: $0.037 - $10.00 / million tokens, Output Costs: $0.10 - $50.00 / million tokens



Source: PricePerToken.com

Red Hat

But the Sticker Price Doesn’t Tell the Whole Story

Some model providers charge more for express lanes



THE ANALOGY

Express Lanes for AI Traffic

Just as physical toll roads charge higher rates to guarantee speed and bypass gridlock, cloud model providers now offer dedicated priority routes to keep enterprise applications responsive.

THE BUSINESS IMPACT

Paying a Premium for Performance

Priority tiers often carry a 1.5x to 2x pricing surcharge. Without this "express" tier, standard serverless APIs can suffer from severe latency spikes or load shedding during peak hours.

Red Hat

But the Sticker Price Doesn't Tell the Whole Story

Some model providers charge more for express lanes

Text and vision models

Per-model pricing for headline models. Fast variants appear as adjacent rows. In each **Standard** or **Priority** cell, prices are **input / cached input / output** (USD per 1M tokens), in that order.

Model	Standard	Priority
Kimi K2.7 Code	$0.95 / $0.19 / $4.00	$1.425 / $0.285 / $6.00
Kimi K2.7 Code Fast	$1.90 / $0.38 / $8.00	—
Kimi K2.6	$0.95 / $0.16 / $4.00	$1.50 / $0.22 / $6.00
Kimi K2.6 Fast	$2.00 / $0.30 / $8.00	—
DeepSeek V4 Pro	$1.74 / $0.145 / $3.48	$2.61 / $0.218 / $5.22
DeepSeek V4 Flash	$0.14 / $0.028 / $0.28	$0.21 / $0.042 / $0.42
GLM 5.2	$1.40 / $0.14 / $4.40	$1.75 / $0.18 / $5.50
GLM 5.2 Fast	$2.10 / $0.21 / $6.60	—
GLM 5.1	$1.40 / $0.26 / $4.40	$2.10 / $0.39 / $6.60
GLM 5.1 Fast	$2.80 / $0.52 / $8.80	—

Source: Fireworks.ai

Fireworks Serverless offers three serving paths:

- **Standard** is the default serving path. No `service_tier` parameter is needed.
- **Priority tier** is for workloads that require higher reliability during peak traffic.
- **Fast** is for workloads that require higher speeds.

Priority tier

Priority tier is for workloads that require higher reliability during peak traffic periods, at a higher price point. Priority tier is prioritized above Standard traffic and is less likely to be load shed (503 server overloaded).

Fast

Fast is a high-speed serving path, useful for interactive applications that require fast response speeds, at a higher price point. Fast variants aim for **100+ tokens per second** of generated throughput. It is not a different model and the quality of the model remains the same.

Red Hat

Does the Model Lifecycle Match the App Lifecycle?

Bank apps may remain in prod for 10+ years

Model status

Claude Mythos Preview (clau
mythos-5), see the migratio

Deprecated models are likely to be less reliable than active models. Move workloads to active models to maintain the highest level of support and reliability.

Current and recently retired models ar

API model name	Current state	Deprecated	Tentative retirement date
claude-fable-5	Active	N/A	Not sooner than June 9, 2027
claude-opus-5	Active	N/A	Not sooner than July 24, 2027
claude-opus-4-8	Active	N/A	Not sooner than May 28, 2027
claude-opus-4-7	Active	N/A	Not sooner than April 16, 2027
claude-opus-4-6	Active	N/A	Not sooner than February 5, 2027
claude-opus-4-5-20251101	Active	N/A	Not sooner than November 24, 2026
claude-opus-4-1-20250805	Deprecated	June 5, 2026	August 5, 2026
claude-opus-4-20250514	Retired	April 14, 2026	June 15, 2026
claude-sonnet-5	Active	N/A	Not sooner than June 30, 2027
claude-sonnet-4-6	Active	N/A	Not sooner than February 17, 2027
claude-sonnet-4-5-20250929	Active	N/A	Not sooner than September 29, 2026
claude-sonnet-4-20250514	Retired	April 14, 2026	June 15, 2026

Source: Anthropic

Red Hat

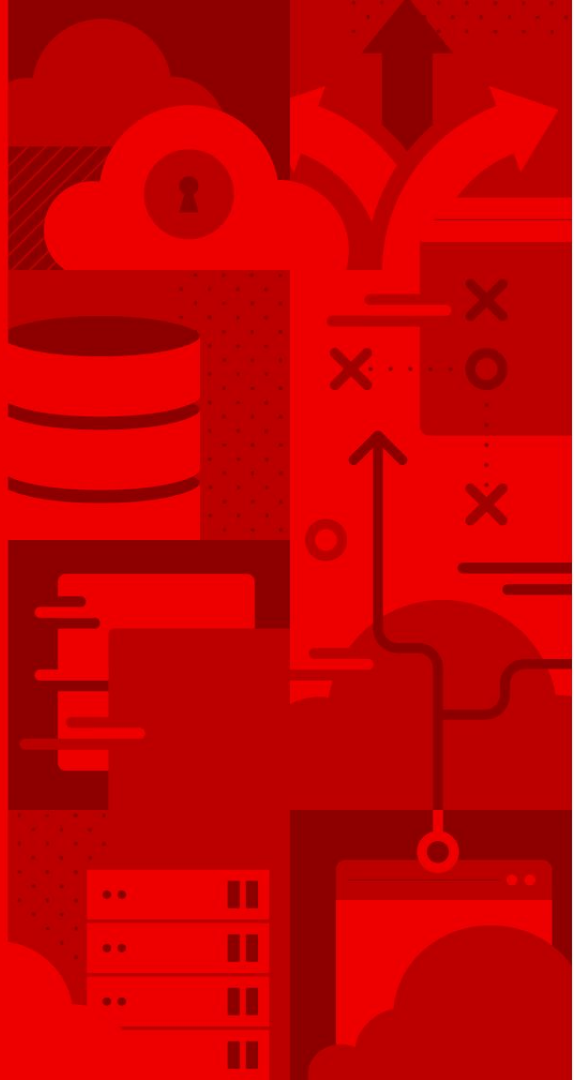
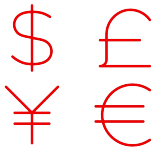


Chart Your Course: Open Source AI for the Enterprise

Red Hat

App Requirements Drive Model / Location Choice

Does the app touch PII?, latency sensitive?, highly variable workload?, unique capabilities?, sovereignty?, data gravity?



Economics

Token cost (in/out), data egress, inference prioritization, model efficiency, CapEx



Data Privacy

PII, MNPI and IP. Risk of data leakage?



Accuracy

Hallucination rates, determinism, and safety



Scaling

Scaling is critical for integrating AI with traditional workloads, as well as training and inferencing multimodal models



Performance

Latency, throughput, rate limits, uptime, ability to customize, time to market



Risk and Governance

Model lifecycle, auditability, 3rd party risk, exit risk, sovereignty

Red Hat

Optimize Inference

Own the Model, Control Your Destiny: Open Source AI for the Enterprise



KIMI deepseek Mistral AI OpenAI LLaMA by Meta NVIDIA IBM Gemma

Qwen cohere

vLLM llm-d

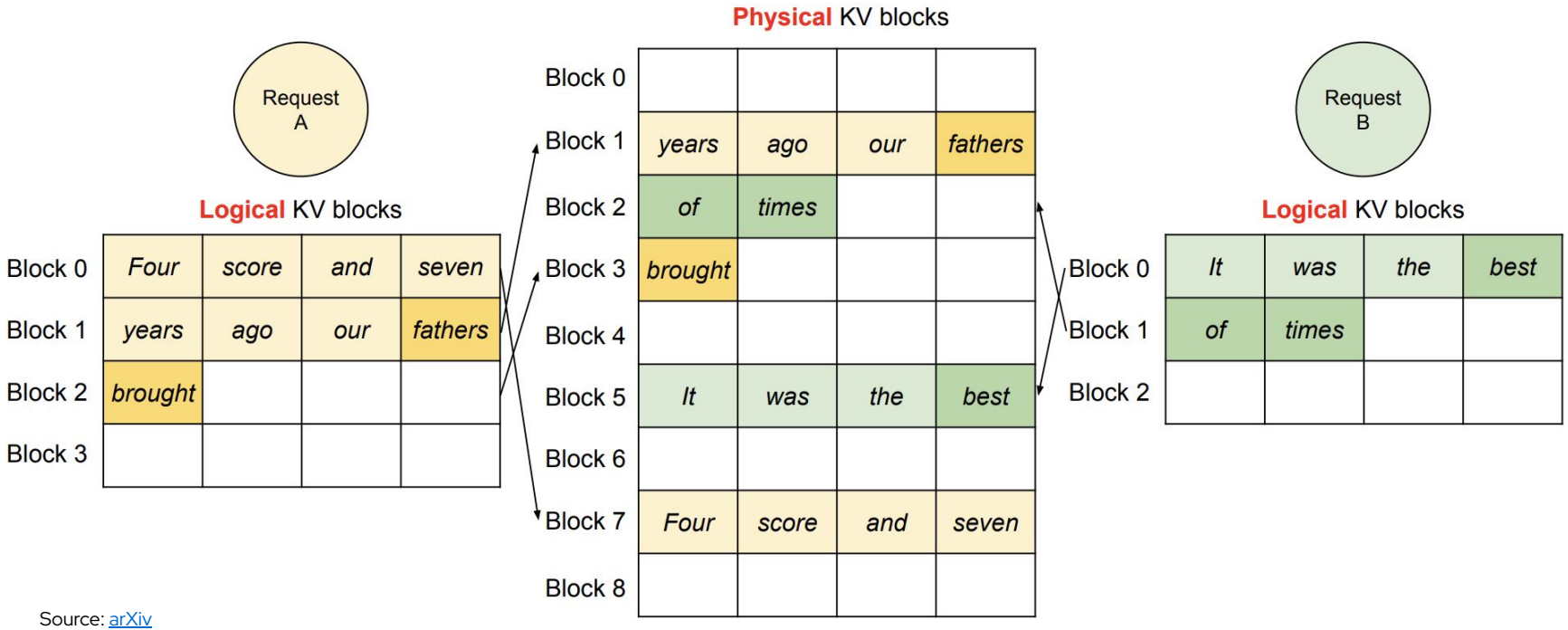


AMD aws IBM intel NVIDIA Google

Red Hat

Optimize Inference: vLLM

Increased throughput and reduced latency through efficient GPU memory management (PagedAttention)

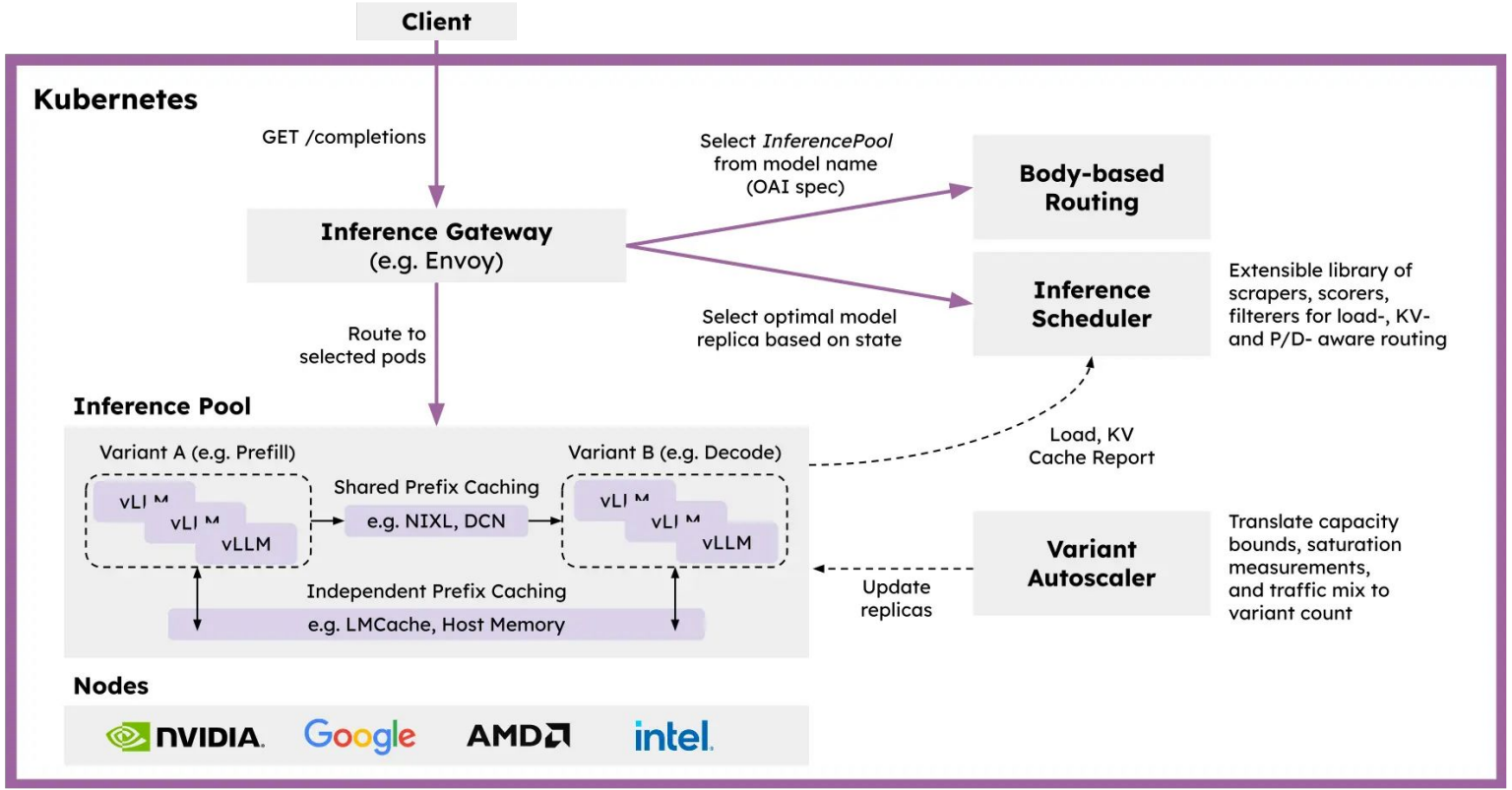


Source: arXiv

Red Hat

Optimize Inference: llm-d

Intelligent routing and disaggregated prefill and decode across a cluster of AI nodes



Red Hat

From Tokens to Tokenomics:
Optimizing AI Costs with Open Source



Blake Treves
Managing Director, GlobalLINK Payments
State Street
BTreves@statestreet.com



Jennifer Bowman
AI Specialist - FSI, Red Hat
jebowman@redhat.com



Aric Rosenbaum
Chief Technologist
Financial Services, Red Hat
aric@redhat.com

Red Hat