Red Hat Summit Connect

From Tokens to Tasks.

Transforming raw LLM telemetry into accountable governance.

Adnan Drina
Specialist Solution Architect

Ben Taljaard
Senior Solution Architect

Red Hat



Adnan Drina

Specialist Solution Architect
Red Hat



Ben Taljaard

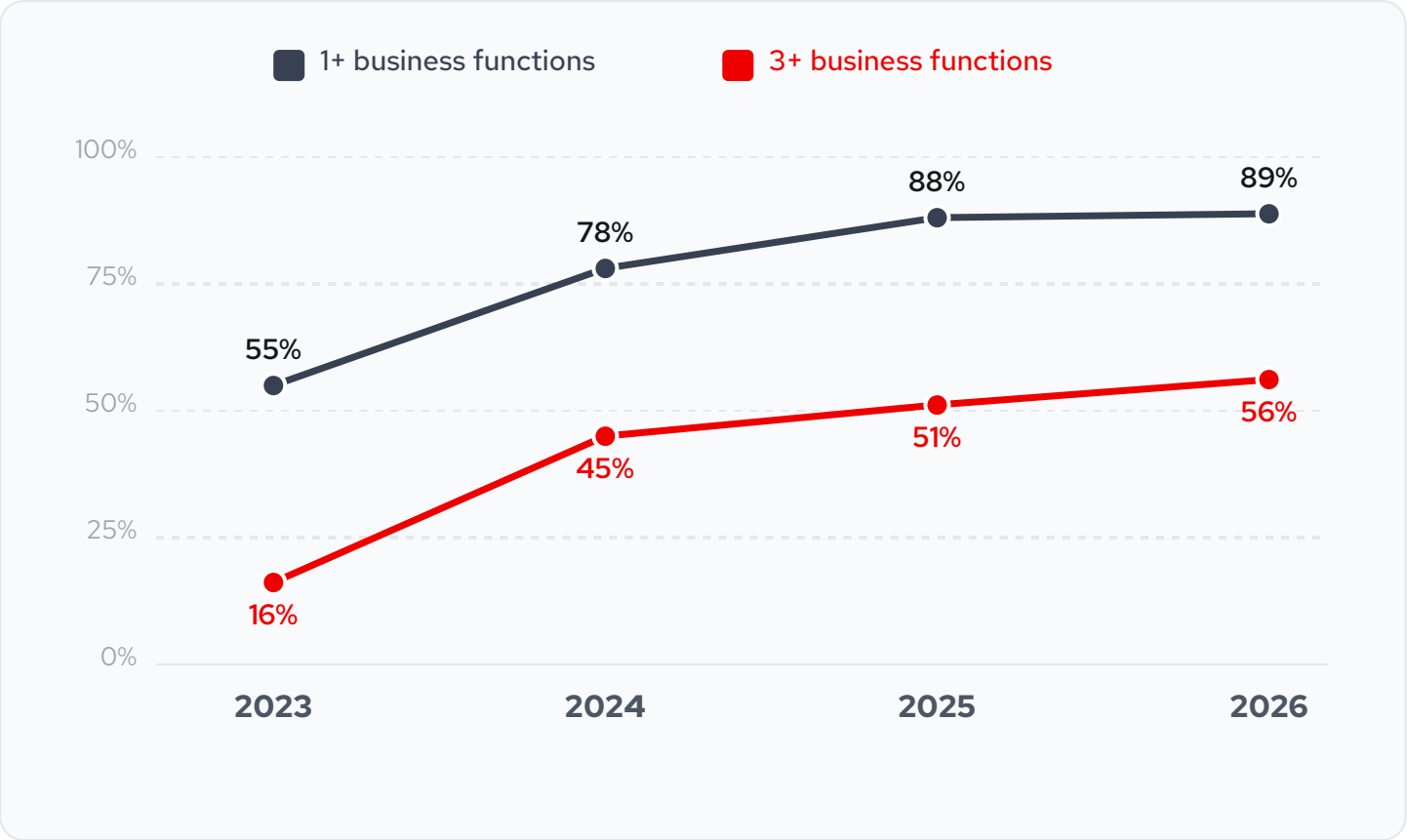
Senior Solution Architect
Red Hat

First some data...



AI adoption is growing

AI use across multiple business functions is growing with it



COST BARRIER

1 in 5

Respondents say operating costs limit AI use

in their organization

(operating costs include tokens)

Source: McKinsey, The state of AI: 2026 report

Managing AI spend is a top priority for FinOps

Key focus areas for enterprise cloud financial management

#1 PRIORITY

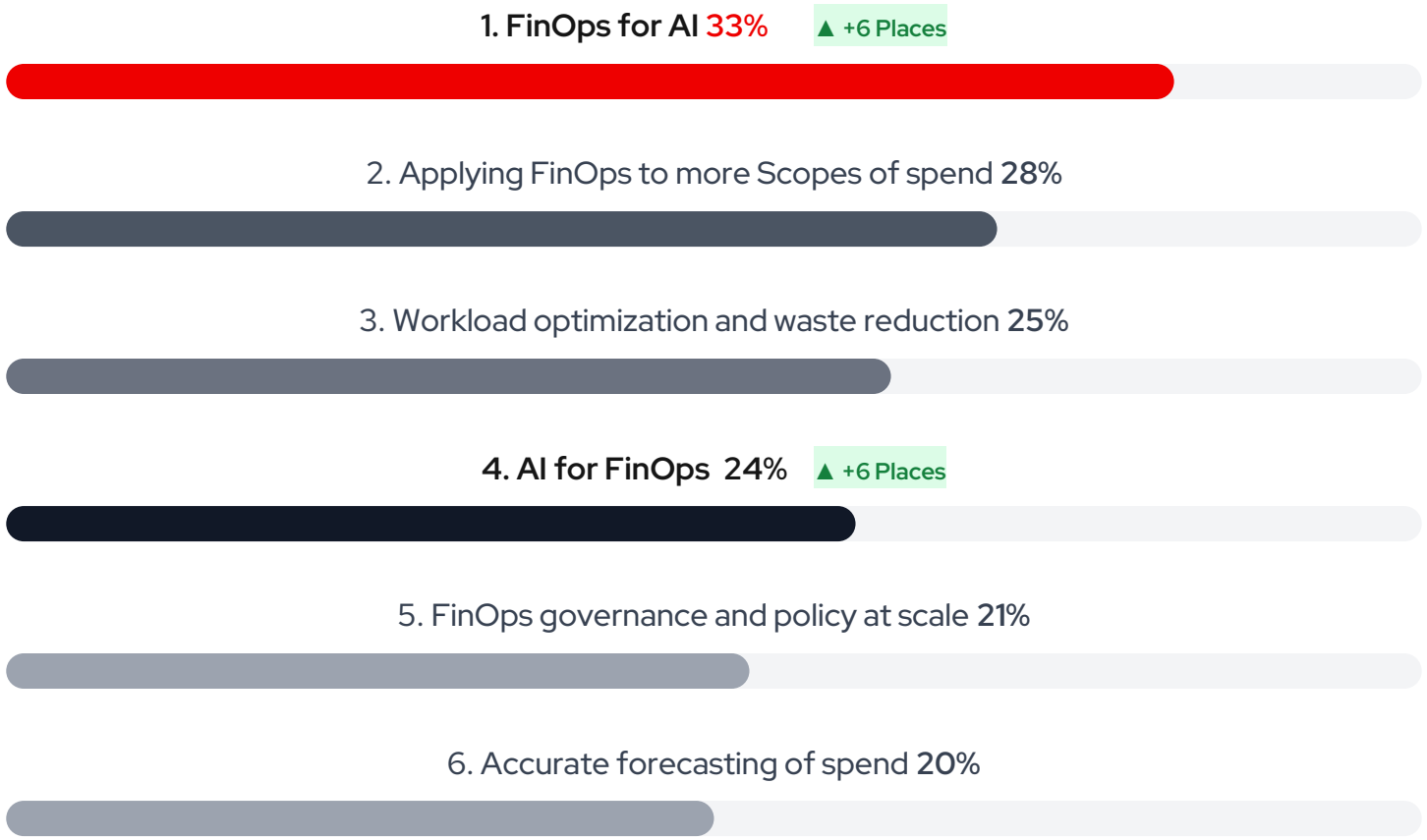
FinOps for AI

33%

Managing and optimizing AI costs surged **+6 places** to rank as the top overall focus for cloud financial teams.

AI for FinOps

24% of respondents are actively deploying AI tools within FinOps (+6 places).



Source: State of FinOps report 2026 (finops.org)

Shadow AI in the enterprise

AI use extends beyond approved services

ENTERPRISE RISK

Unapproved AI Usage

70 - 80%

of employees using AI at work use tools their employer **has not approved**.

AI Tool Adoption in the Workplace

Prevalence of unsanctioned software vs. approved platforms

■ Unapproved Tools (75% avg) ■ Approved Corporate Services (25%)

Key Takeaways for Enterprise IT

- **Widespread Exposure:** Employees actively seek AI solutions independently to streamline daily tasks.
- **Governance & Security Gap:** Unsanctioned tools introduce data privacy, compliance, and IP protection risks.
- **Control Need:** Organizations require centralized visibility, token tracking, and unified AI Gateway access.

Source: Stanford Digital Economy Lab, The Enterprise AI Playbook (2026)

How do I get some control over AI use?



Enterprise AI Inference Platform

Red Hat OpenShift AI

Operationalize **Model as a Service** across the Hybrid Cloud

AI Gateway

API Management & Governance Layer
API Management, Token Tracking, Authentication for any model

Distributed Inference

Orchestration & Serving Control
Intelligent request routing, multi-node scaling, and optimized inference dispatch

llm-d

KServe

VALIDATED, OPTIMIZED MODELS & TECHNIQUES

Nemo | GPT OSS | Gemma | Llama | Granite | Mistral | DeepSeek | Qwen

Inference Server

High-Performance Runtime Engine
Optimized model execution, vLLM acceleration, and low-latency throughput

vLLM

HARDWARE & ACCELERATOR INFRASTRUCTURE

NVIDIA GPUs | AMD GPUs & CPUs | Google TPUs | AWS Inferentia | Intel Xeon & CPUs | IBM Spyre

Over to Adnan...
(MaaS demo)

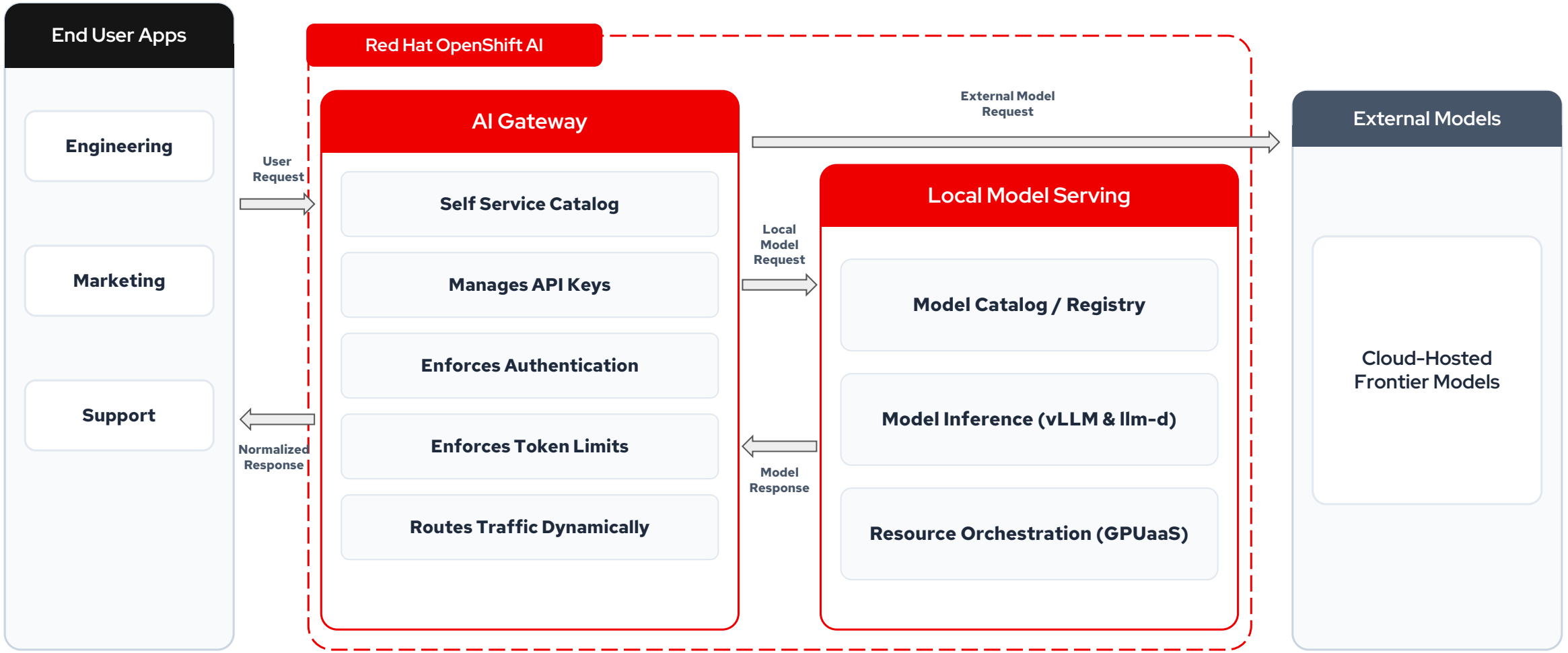




Enterprise AI Inference Platform on OpenShift AI

Model as a Service Demo Recap

Red Hat OpenShift AI



How much does my AI cost?



What is the Public AI Pricing model?

Based on Tokens and Tiers

Pay per token

APIs bill per million tokens.

Output costs **3–5× more** than input.

Larger models cost **10–50× more** than small ones.

Flat-fee Subscriptions

Chat apps charge a **flat monthly fee** per tier.

Usage limits reset over time instead of per-token billing.

Enterprise deals

Volume discounts & reserved capacity.

Per-seat license negotiated for large enterprise customers.

What drives the bill

Context length (history resent each turn)

Reasoning tokens, billed as output

Prompt caching: steep discount

Batch jobs: ~50% cheaper

Multimodal: images, audio, search, tools

PUBLIC AI ECONOMICS

More Tokens Usage Mean More Cost

Unlike flat-fee SaaS, API usage is billed per token of input and output.



Asymmetric Output Premium

Generated tokens cost 3x to 5x more than input prompts.

Context Window Bloat

Repetitive conversational history re-bills on every round trip query.

Cost scales with usage: every additional token is billed; caching, batch and volume discounts only soften the curve.

THE STRATEGIC TRANSITION

The Logical Next Move: Self-Host

Shift from variable per-token pricing to predictable, high-throughput compute.

cloud

Dedicated Cloud GPUs

Deploy open-weight models on reserved AWS, GCP, or Azure GPU instances.

bolt **Fast elasticity with capped hourly rates**

dns

Owned Infrastructure

Run localized clusters on-premises for predictable, sustained high-volume workloads.

verified_user **Maximum cost ceiling and full data privacy**

Core Advantage: No per-token bill: a fixed infrastructure cost, up to cluster capacity

Fixed Cost Model



Infrastructure breakdown

Key Cost Drivers & Effective Cost Per Token

Hardware runs continuously: **cluster utilization is the single make-or-break metric** for token ROI.

GPU Hardware & VRAM

Hardware capex (3–4 year depreciation) or reserved cloud GPUs.

Primary CapEx / OpEx

Power & Hosting

High-draw electrical feeds, cooling, and EU datacenter rack space.

Utility & Facilities

CORE VALUE DRIVER

Utilization is Key

Fixed costs tick 24/7. Idle GPUs burn money; fuller batches cut the cost of every token.

Sets the cost per token

MLOps & Engineering

vLLM serving, continuous batching, quantization, and failover monitoring.

Operational Floor

EFFECTIVE COST PER TOKEN:

Fixed monthly cost ÷ tokens actually served.

Double the utilization and the cost per token halves.

How do I get visibility into my AI spend when self-hosting?



Example Showback Model

Develop a Cluster Showback Model

Define an operational/showback model together with FinOps

Pricing Logic

Base Rate = Shared Pool Cost ÷ Target Capacity

Team Charge = Workload Volume × Tier Rate

TIER 1 - Dedicated Lease

Reserved GPU Capacity

Customer-facing production & heavy 70B+ model serving

Commercial Model: Fixed flat lease per dedicated GPU

SLA / Priority: Non-preemptible with zero noisy neighbors

Idle Capacity: Paid directly by tenant; zero platform risk

Charge = Assigned GPUs × Cost per GPU

Pricing: **Fixed Monthly Lease**

TIER 2 - Shared Pool

Dynamic Multi-Tenant

Internal copilots, agent tools, interactive RAG APIs

Commercial Model: Pay-as-you-go per million tokens

SLA / Priority: Shared scaling with prefix-cache affinity

Idle Capacity: Headroom reserve baked into base rate

Charge = Tokens × Weight × Base Rate

Pricing: **Standard Rate Card**

TIER 3 - Off-Peak/Batch

Opportunistic Compute

Offline vectorization, evaluation, synthetic generation

Commercial Model: Deep discount for off-peak consumption

SLA / Priority: Sheddable traffic during off-peak hours

Idle Capacity: Soaks spare capacity to lower deficit

Charge = Tokens × Weight × Discount

Pricing: **Discounted Metered**

TIER 4 - External API

External SaaS Models

Proprietary third-party LLM APIs (e.g. OpenAI, Anthropic)

Commercial Model: Direct API vendor cost pass-through

SLA / Priority: Provider-managed SLAs & rate limits

Idle Capacity: Zero idle cost; pure variable spend

Charge = External API Pass-Through Cost

Pricing: **Pass-Through Cost**

CLUSTER BALANCE LOGIC:

Total Cluster Cost = Dedicated Leases + Shared Usage + Off-Peak Usage + Unallocated Balance

Transparent FinOps

Telemetry & Observability

Collect Key Metrics

Essential telemetry signals to audit team spend, monitor GPU efficiency, and verify SLAs across MaaS, Serving, GPU, and Model Quality layers.

ACCESS & CONSUMPTION

Layer: MaaS

Tokens per subscription & user
Every request attributed to a team, use case, and model.

Access & limits
Which subscription reaches which model; hourly token limits per user.

Rate-limited requests
Requests rejected over a subscription's threshold limit.

Who used what?

SERVING

Layer: vLLM ON llm-d

Time to First Token (p95)
Initial response latency per model instance.

Prompt vs. completion tokens
The prefill and decode compute mix per serving model.

Serving & busy time
Minutes each worker served, and minutes actively busy.

How did the model serve it?

UTILIZATION

Layer: GPU & CAPACITY

GPU utilization
How busy each model's hardware accelerator is.

Capacity utilization
Tokens served vs. measured capacity

Paid vs. used capacity
Dedicated leases used/unused; shared commitments pace.

How well is the GPU used?

MODEL QUALITY

Layer: MLFLOW

Request traces
Every turn with team, subscription, tokens, cost & reason.

Judge pass rates
Relevance, completeness, and team guidelines per tier.

Model comparison
Same judges on external vs. private model before a move.

Was the answer good?

SHOWBACK LINK:
MaaS tokens × offer-catalog prices, against GPU capacity and judged quality: cost per team, utilisation per worker, cost per passing answer.

Showback Analytics



How can OpenShift AI, help us collect and display metrics?

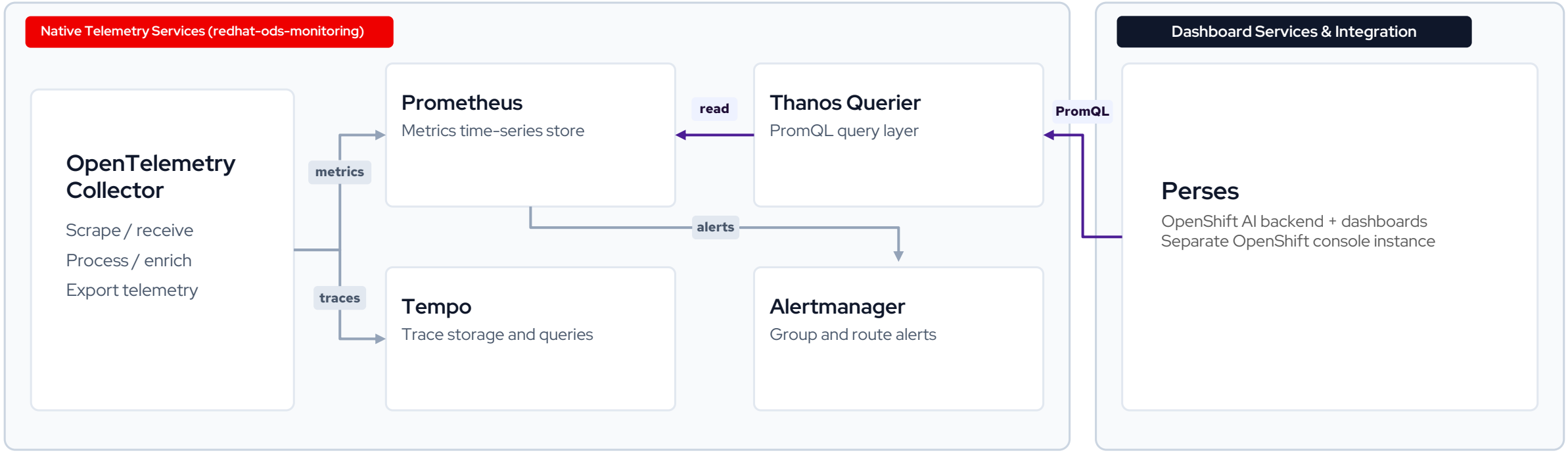


The OpenShift AI observability stack

Installed Prerequisite Operators

Cluster Observability Operator (Prometheus, Thanos, Alertmanager, Perses), **OpenTelemetry Operator** (Red Hat build of OpenTelemetry), **Tempo Operator** (Red Hat distributed tracing)

OpenShift AI Operator: DSCInitialization → spec.monitoring.managementState: Managed



Metrics sources: vLLM · Limitador · DCGM | ServiceMonitor / PodMonitor | **Traces:** OTLP instrumentation

Over to Adnan...
(Cost per team)
(Efficiency and utilization)

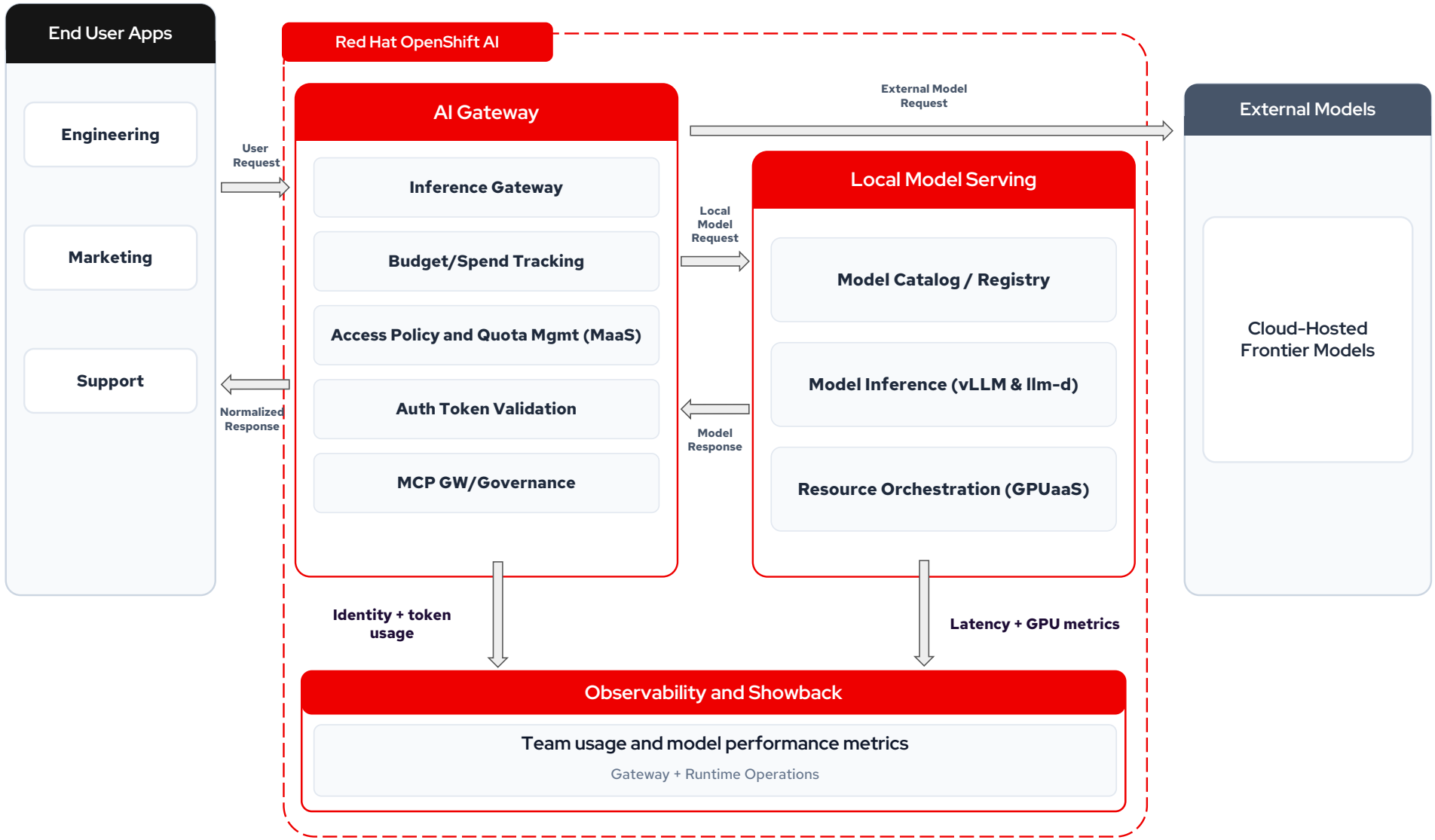




Enterprise AI Inference Platform on OpenShift AI

The Observability and Showback Stack Demo Recap

Red Hat OpenShift AI



Where to next?



Tokens Connect the GPU Bill to Business Value

Bridging infrastructure expenditure, usage metering, and functional work output.

GPU Cost (Fixed €) → **Cost per Token** → Rate Card → **Tokens per Task** → **Cost per Outcome**

PLATFORM TEAM

Lower Cost per Token

- **Raise utilisation:** Continuous batching & off-peak jobs
- **Quantise models:** FP8 & AWQ memory footprint cuts
- **Prefix caching:** Reuse attention across prompt templates
- **Fast serving:** vLLM & optimized execution runtimes

Infrastructure Efficiency

PRODUCT TEAMS

Lower Tokens per Task

- **Smart routing:** Send simple requests to small models
- **Trim prompts:** Strip bloated context & zero-shot fluff
- **Cap generation:** Bound max tokens & reasoning depth
- **Constrain agents:** Multi-agent loops only when needed

Application Efficiency

RATE CARD

Not All Tokens Are Equal

- **Output > Input:** Decode phase is memory-bandwidth bound
- **Cached hits:** KV cache hits are nearly cost-free
- **Long context tax:** Massive contexts consume GPU slots
- **Model weight:** 70B+ models command higher tier rates

True Cost Pricing

CORE PRINCIPLE: **Meter in tokens, judge in outcomes.** Cheaper tokens spur wider usage and multi-agent loops, meaning total spend often grows even as unit cost falls.

Key Takeaways: Connecting GPU Spend to Value

Connecting infrastructure metering to enterprise value and cost predictability.

01

Unified AI Platform Architecture

Control & observe end-to-end model consumption on OpenShift AI:

- **Central AI Gateway:** Standardize access, authentication, and policy enforcement.
- **Local & External Models:** Dynamically route workloads between internal runtimes and cloud models.
- **Full Observability:** Capture identity, token usage, latency, and GPU telemetry for showback.

02

Tokens as the Unit of Economics

Bridge infrastructure spend to business output via token metering:

- **Infrastructure Efficiency:** Lower unit cost with vLLM, continuous batching, and FP8/AWQ cuts.
- **Application Efficiency:** Reduce tokens per task using smart routing and prompt trimming.
- **True Cost Pricing:** Differentiate rates based on model weights and context length.

03

Governed Scale & FinOps Alignment

Manage the unit-cost paradox and scale responsibly:

- **Meter Tokens, Judge Outcomes:** Unit costs drop, but usage expands overall spend.
- **Cross-Team FinOps Alignment:** Unify platform efficiency with application value creation.



Red Hat Summit

Connect

Thank you

linkedin.com/company/red-hat

facebook.com/redhatinc

youtube.com/user/RedHatVideos

x.com/RedHat

Red Hat