SOLUTIONS OVERVIEW · BUSINESS EDITION

More tokens from every GPU server with TuringData Cache Fabric and Red Hat AI Enterprise

Serve more AI users on the GPU servers you already own, at a lower cost per token.

Highlights

- Up to 11.7× the tokens from the same GPU server
- First token in 2.1 s instead of 36.7 s under load
- Lower cost per token on hardware you already own
- No change to your models or applications
- A measured result on your own GPUs in days

Technology used

- **Red Hat AI Enterprise**
- **TuringData Cache Fabric**

Inference is becoming the bigger AI bill

As Vibe Coding, AI assistants, agents and document analysis move into production, inference becomes the main consumer of GPU capacity. Gartner forecasts that inference takes 55% of AI-optimized IaaS spending in 2026 and 59% in 2027.[1]

Much of that capacity is spent repeating tasks. Large language models re-read the full context on every request: the system prompt, the documents, the conversation history. Most of it was processed moments earlier, yet the GPU computes it again.

Optimize GPU investments and reduce inferencing costs

TuringData Cache Fabric with Red Hat AI Enterprise optimizing your investments, while significantly reducing your token costs. It integrates with the vLLM inference engine provided by Red Hat, and manages and accelerates the KV cache: the GPU's saved work on context it has already processed.

The GPU reuses what it has done and computes only what is new. The same servers produce more tokens, serve more users and cost less per token. Achieving more with less!

Key benefits

- **Higher GPU efficiency.** GPUs stop recomputing context they have already processed.
- **Lower cost per token.** The same servers, power and space produce far more tokens.
- **More tokens per server.** Up to 11.7× token throughput in the joint benchmark with Red Hat.
- **Higher concurrency.** Serve more workloads and users within the same response time.

[1]Gartner press release, “Gartner Forecasts Worldwide AI-Optimized IaaS Spending to Grow 96% Through 2026”, 10 August 2026. Market context only.

Inference today vs. TuringData Cache Fabric on Red Hat AI Enterprise

	Inference today	With the joint solution
GPU work	Every request recomputes its full context	**Context that was already processed is reused**
User experience	Users wait longer for the first token as usage grows	**The first token stays fast as usage grows**
Capacity	More users means more GPU servers	**The same GPU servers serve more users and workloads**
Cost per token	Stays high	**Falls, because the same hardware produces far more tokens**
Change required	Not applicable	**None to your models or applications**

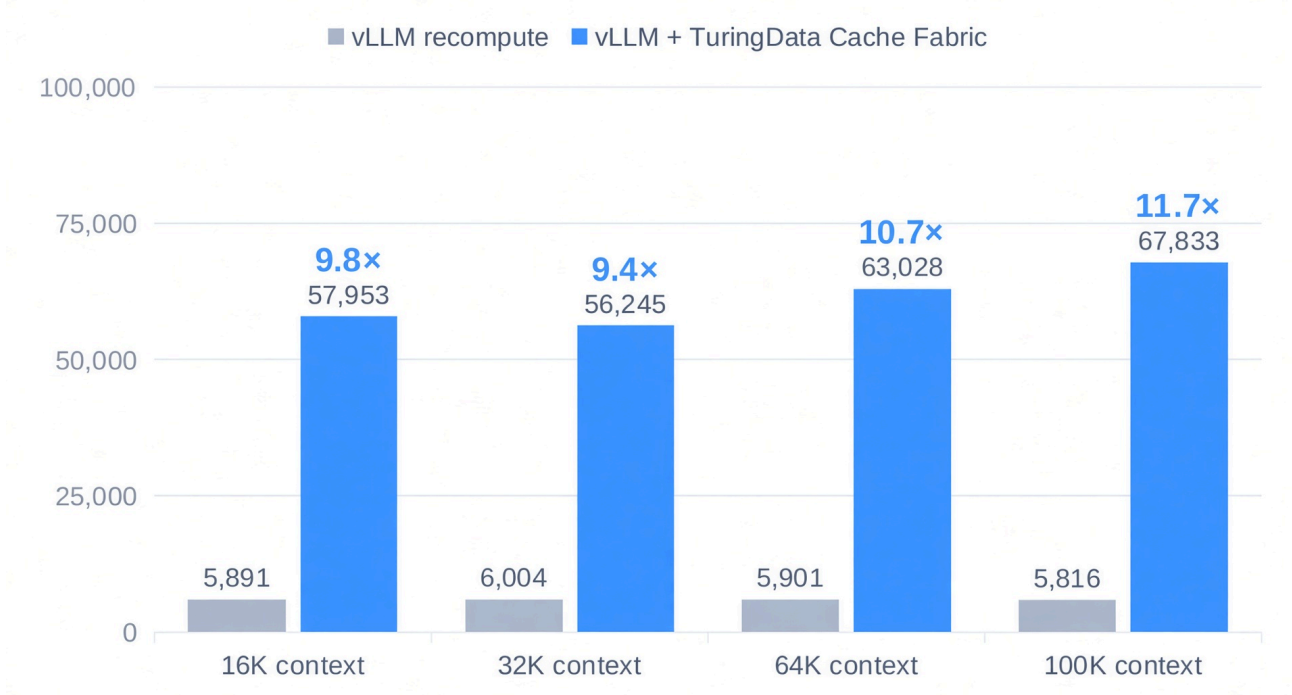
Exceptional performance benchmark results together with Red Hat

Same GPU server, same model. The only change is turning on TuringData Cache Fabric.

11.7×
the tokens per second from the same GPU server: 5,816 → 67,833

2.1 s
to the first token, down from 36.7 s

9.4× to 11.7×
token throughput gain at every context length tested, 16K to 100K tokens



Total token throughput, tokens per second. One server with 8× NVIDIA H200, GLM-5.2-FP8, vLLM 0.23.0, Red Hat GuideLLM 0.7.3. Each pair is one measured point at the highest concurrency tested for that context length. The 11.7× and 2.1 s figures are at 100K-token context.

Where it pays off most

Any workload where long context repeats from one request to the next.

Use case	What repeats
Coding assistants	The same repository and files go in with every request.
Long-context analysis	Many questions on the same long contract, report or log.
AI agents	Instructions, tools and task history repeat at every step.
RAG	Shared instructions and popular documents keep coming back.
Multi-round conversation	The full chat history is sent again on every turn.

The longer the context and the more it repeats, the bigger the gain.

Time to value and implementation approach

TuringData Cache Fabric is an enterprise-grade solution that runs on the GPU servers you already own, so there is no hardware lead time. The first measurable results can be noticed within days, with production readiness in about four weeks.

Phase	Timing	What happens	Outcome
1. Discover and scope	Week 1	Pick one workload and agree what success looks like.	Agreed workload and targets
2. Prove it on your GPUs	Weeks 1–2	Side-by-side test on your own GPU server, without and with Cache Fabric. 1 day on one server, 2 to 4 days on site.	Measurable results on your own workload
3. Pilot	Weeks 2–3	Switch it on for the first production workload.	First workload live
4. Production rollout	Week 4	Extend across GPU servers and workloads.	Solution in production

Indicative timeline for a software-only rollout. Actual timing depends on the model and on how ready the environment is.

Interested in finding out more?

Start with a proof of concept on your own GPU servers: your vLLM baseline against TuringData Cache Fabric, side by side.

- **Customers' Pre-requisition:** a technical owner, one dedicated GPU server with access.
- **TuringData Offers:** engineers to execute Proof of Concept with a performance comparison report

Contact TuringData at marketing@turingdata.io or visit www.turingdata.io.

About TuringData

TuringData Pte. Ltd. is an AI storage company. It helps enterprises and AI cloud providers get more from every GPU.

STORAGE FOR AI TRAINING
TuringData Platform
TuringData Flash

KV CACHE FOR AI INFERENCE
TuringData Cache Fabric
TuringData ContextCube

WEBSITE
www.turingdata.io

EMAIL
marketing@turingdata.io

ADDRESS
18 Cross Street, #02-101
Singapore 048423