SC26 | FUJITSU BOOTH 4619

FUJITSU

Get More Tokens from Every GPU

High-Efficiency Inference Platform

Turn LLMs from private batch jobs into a shared, fair and efficient service for your entire HPC center.

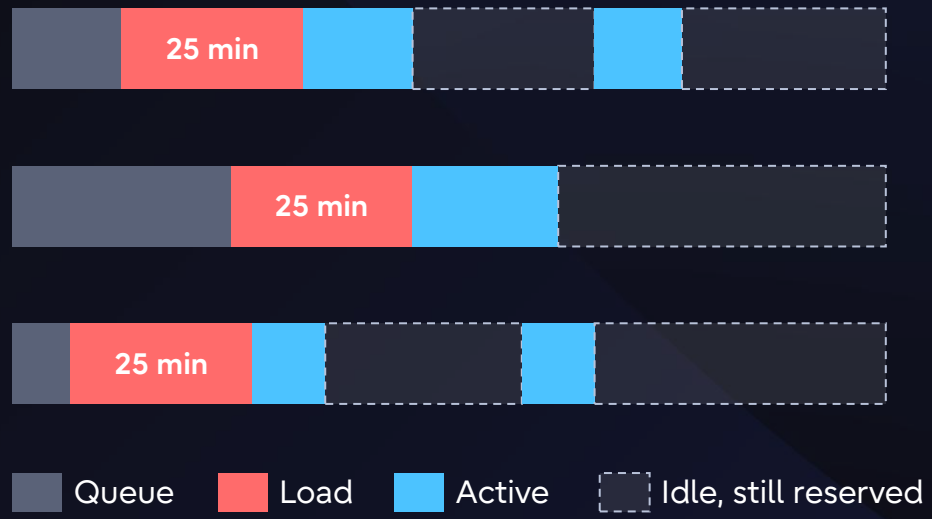
Live demo at Booth 4619

© 2026 Fujitsu Limited

THE PROBLEM

LLMs as batch jobs waste GPUs

One researcher, one LLM job



Each job waits in the queue and spends up to 25 minutes loading a 500B-class model. While researchers think, code or read results, the GPUs sit idle but stay reserved.

The hidden cost

25 min to load a 500B-class model before the first token

Idle GPUs stay reserved while researchers think and code

×3 the same model, loaded separately by three teams

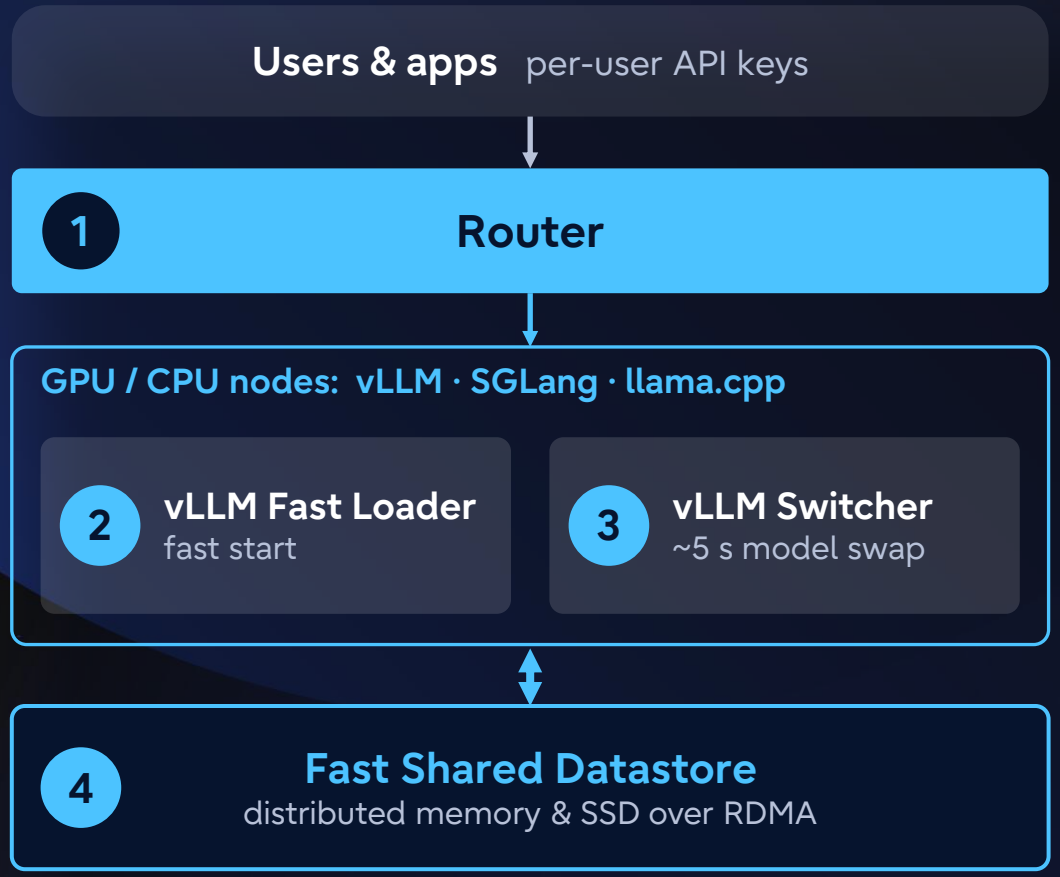
For the center: more GPUs, less science.

Illustrative timeline. Load time based on a measured 1,505 s first start of a 500B-class model.

© 2026 Fujitsu Limited

THE SOLUTION

One shared platform, four technologies



1 Router

One endpoint for every user and app. Fair-share QoS, KV-aware routing, per-group token accounting and non-stop scaling.

2 vLLM Fast Loader

Restores pre-built weights, tuned kernels and compiled graphs. Start serving in minutes, not half an hour.

3 vLLM Switcher

Parks models in distributed remote memory and swaps them onto the same GPUs in about 5 seconds.

4 Fast Shared Datastore

Cluster-wide memory and SSD for models and KV cache. Reuse KV cache, even across vLLM, SGLang and llama.cpp.

Works with common LLM APIs, including OpenAI Chat Completions / Responses and Anthropic Messages.

© 2026 Fujitsu Limited

PROOF

Measured results

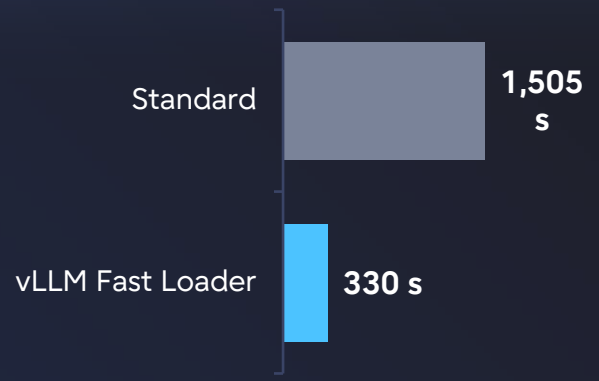
FUJITSU

2 vLLM Fast Loader

1/4.6

startup time

500B-class Model A, 4× B200. Start request to endpoint ready; separate runs.

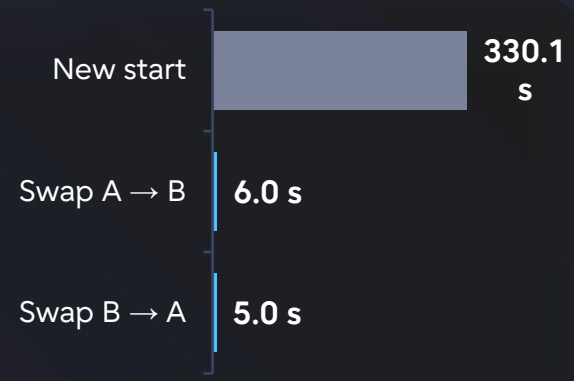


3 vLLM Switcher

~5 s

500B-class model swap

500B-class Model A ↔ Model B, 4× B200. Median of 3 round trips.

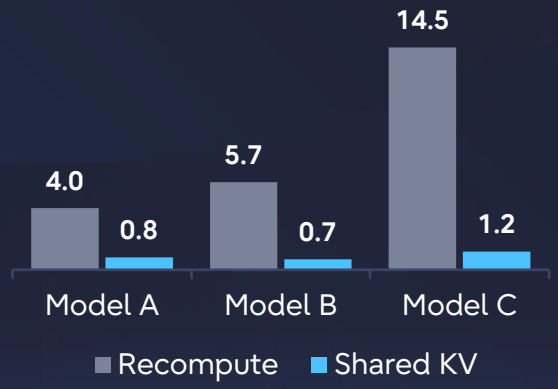


4 Fast Shared Datastore

−91.8%

time to first token

Models A, B and C (31B); long prompt, B200, median of 3. Identical generated text and token IDs.

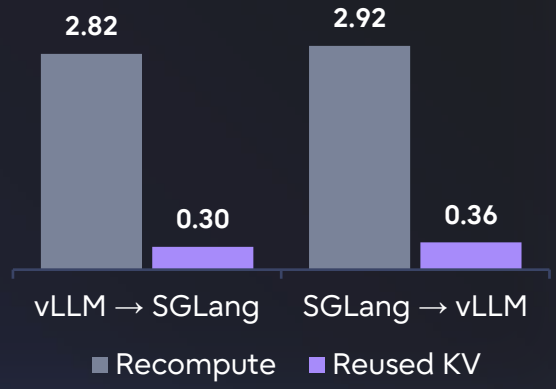


4 Fast Shared Datastore

8–9×

faster first token across engines

31B Model C, 64K-token prompt, TP2, median of 3. KV computed in one engine, reused by the other.



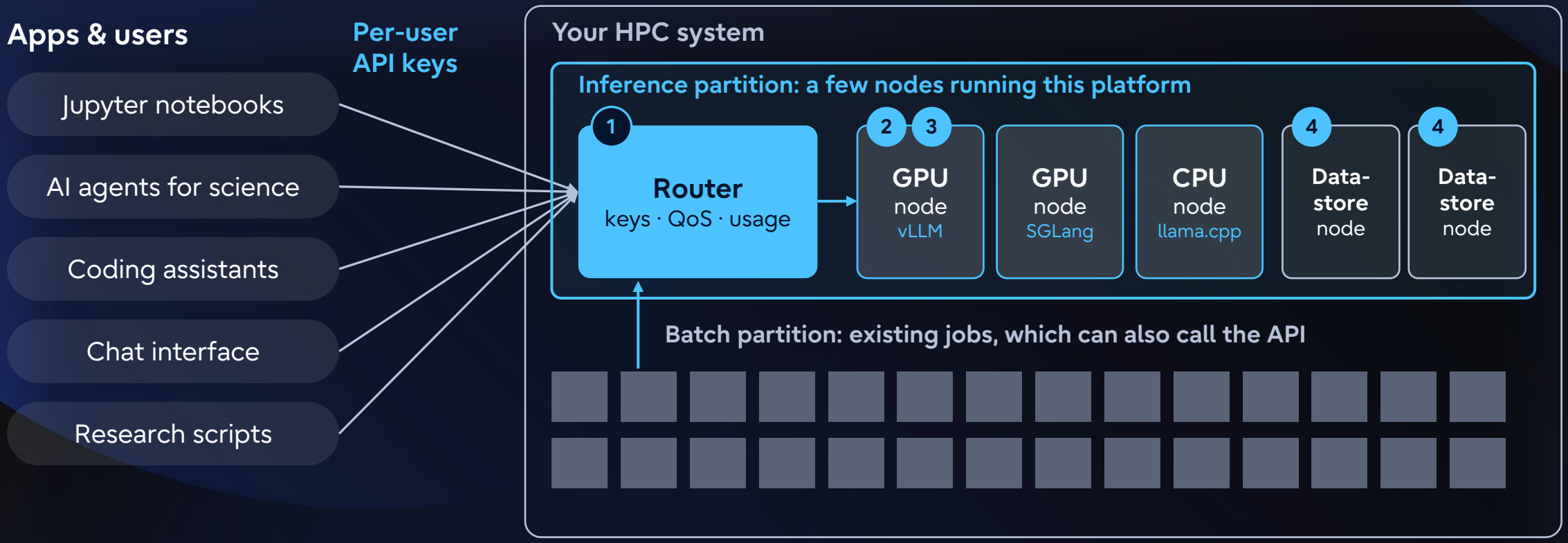
All numbers: internal evaluation on NVIDIA B200 systems, October 2026. TTFT in seconds.

© 2026 Fujitsu Limited

DEPLOYMENT

How it fits your HPC center

FUJITSU



STEP 1 Set up on a few nodes

STEP 2 Issue API keys to users

STEP 3 Connect existing apps

Illustrative deployment. Works with common LLM APIs. Batch partition stays as it is.

© 2026 Fujitsu Limited

WHY IT MATTERS

More science from every GPU

FUJITSU

✓ **More tokens from the same GPUs**

Start in minutes, swap in seconds, and stop recomputing long context.

✓ **Fair, accounted sharing**

One endpoint with per-group priorities, quotas and token accounting, like node-hours.

✓ **Ready for agents and long context**

KV cache shared across the cluster, and across vLLM, SGLang and llama.cpp.

✓ **Fits the HPC system you have**

Runs on a few nodes; the batch partition and your existing apps stay as they are.

Live demo

Booth 4619

Get More Tokens from Every GPU

High-Efficiency Inference Platform

Numbers: internal evaluation on NVIDIA B200 systems, October 2026. Cross-engine KV reuse currently validated with a 31B model.

© 2026 Fujitsu Limited