AI computing broker

Future technologies for AI Infrastructure

November 16-21

Fujitsu Limited

A silent tax on AI infrastructure

70% of organizations are running their GPUs

below **70%** utilization

The State of AI Infrastructure at Scale 2024

How can we drive *peak* GPU utilization for AI?

FUJITSU

Increase throughput on shared memory

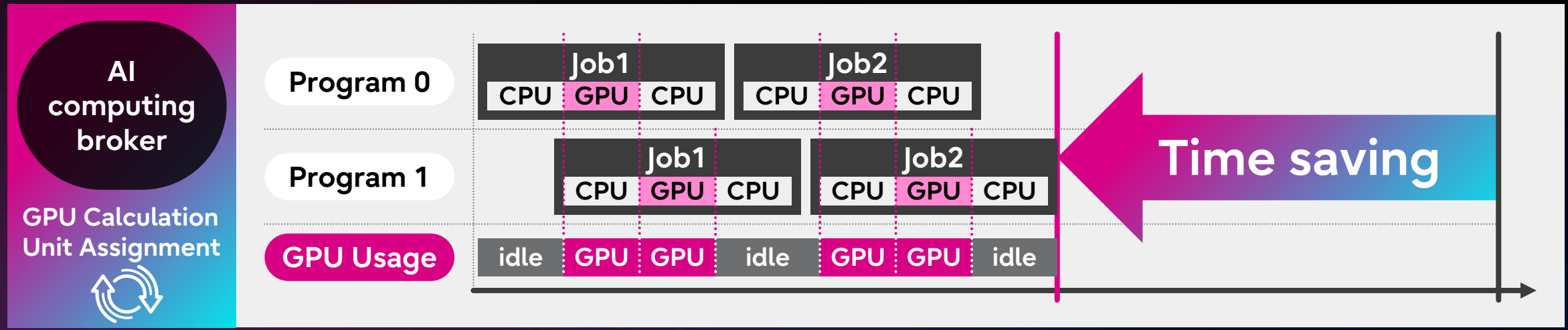
Dynamically allocate resources to application need

Use fewer GPUs across heterogenous workloads

AI computing broker achieves more with your current GPU infrastructure

FUJITSU

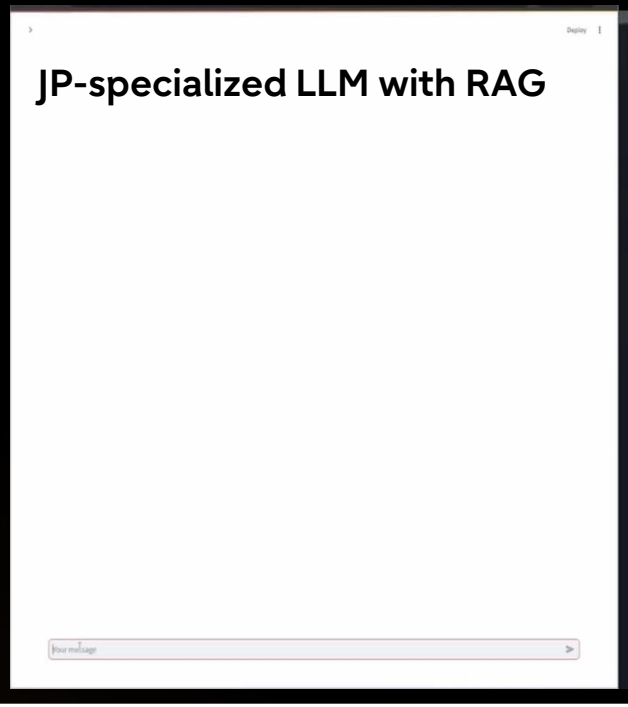
- Runtime-aware GPU allocation
- Full memory access
- Advanced scheduling algorithms
- Docker support
- No code changes
- Cloud or on-prem



Conventional

1 LLM on a server with 4 GPUs

JP-specialized LLM

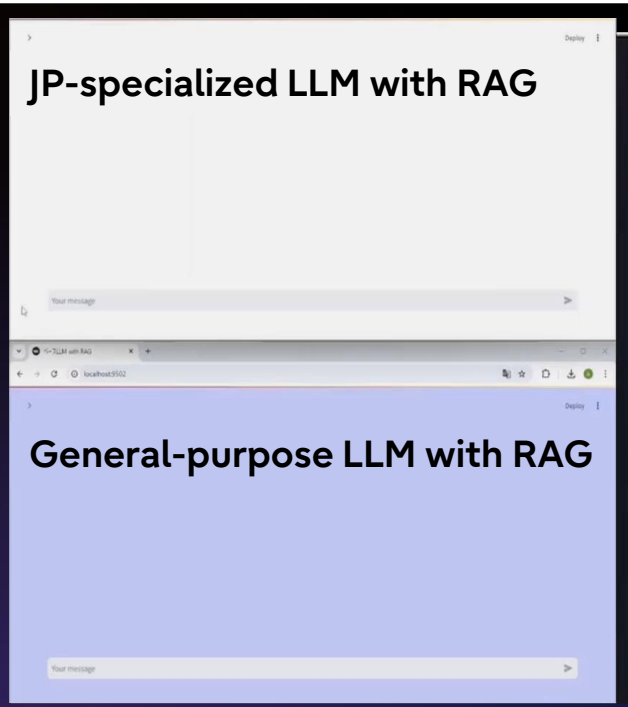


AI computing broker

Multiple LLM can run on a server

JP-specialized LLM

General-purpose LLM



Use case

TRADOM

Streamlining GPU resources for multi-instance AI training

with AI computing broker

+25%

GPU utilization during AI model execution *without code changes*

x 2

Model throughput per unit time

> *ACB proved its ability to significantly streamline GPU resource allocation for AI model generation, enabling the development of substantially* ***more accurate models in significantly less time*** *through AI learning process multiplexing*
>
> *- Junichi Kayamoto, Chief Data Science Officer, TRADOM Inc.*

About TRADOM

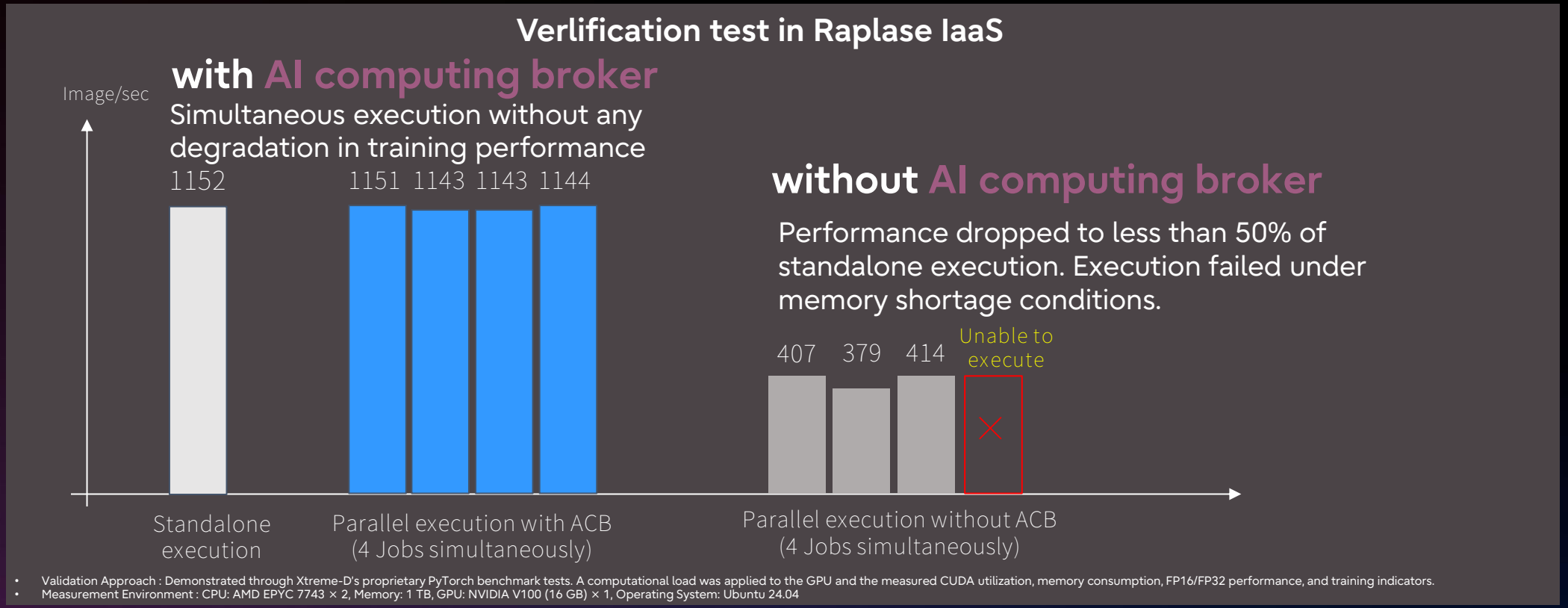
TRADOM delivers cutting-edge AI-powered solutions for managing foreign exchange risk.

TRADOM Inc. : https://www.tradom.jp/company

Use case

Xtreme-D

Improved efficiency and throughput for AI workloads utilizing GPUs, contributing to reduced computational costs



- Validation Approach : Demonstrated through Xtreme-D's proprietary PyTorch benchmark tests. A computational load was applied to the GPU and the measured CUDA utilization, memory consumption, FP16/FP32 performance, and training indicators.
- Measurement Environment : CPU: AMD EPYC 7743 × 2, Memory: 1 TB, GPU: NVIDIA V100 (16 GB) × 1, Operating System: Ubuntu 24.04

About Xtreme-D

Xtreme-D offers a multi-cloud compatible and high-speed AI platform service Raplase(Ra+)

Xtreme-D inc. https://xtreme-d.net/

FUJITSU

50% of large cloud spenders cite ‘waste reduction’ as their top priority.

— State of FinOps 2025

Your GPUs work hard. Could they work smarter?

Ask us about **free trial access** to AI computing broker

Reach out:
awelden@fujitsu.com

FUJITSU

Future proposal

- High-performance Data Store for Efficient LLM Inference
- Interactive HPC - A Scheduler for Parallel Applications

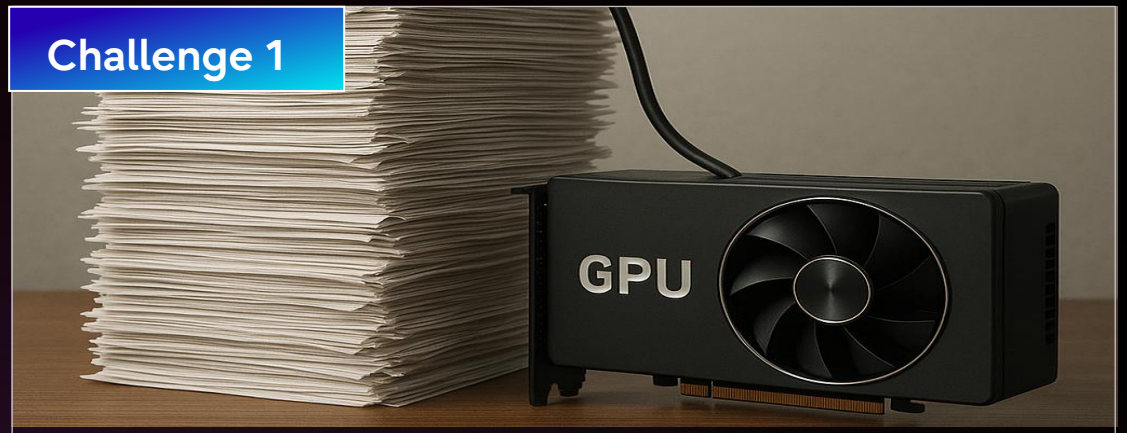
FUJITSU

High-performance Data Store for Efficient LLM Inference

Challenges in LLM inference

- When the number of agents and users increase, performance drops

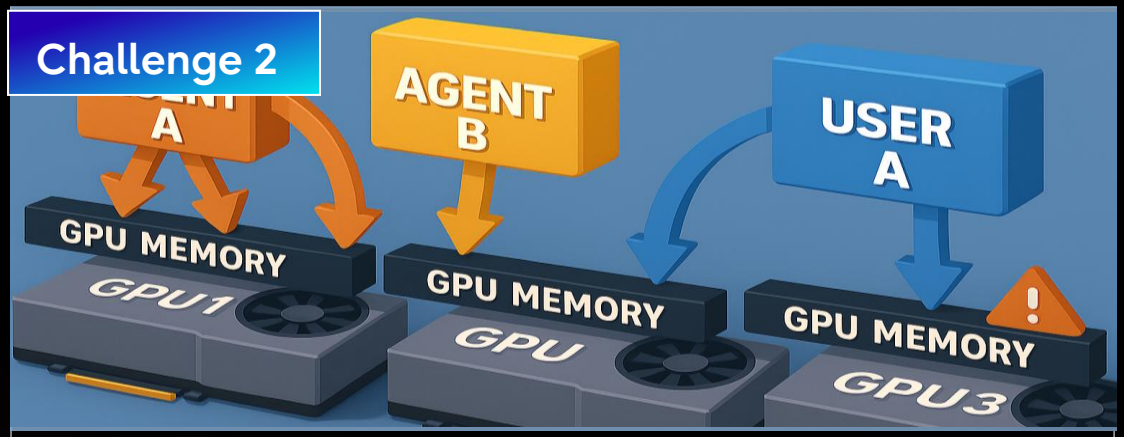
Challenge 1



High GPU load due to large input data

- long inputs → heavy computation
- limited cache → redundant computation

Challenge 2



Cache contention among multiple requests

- multiple requests → memory conflict
- cache eviction → recomputation

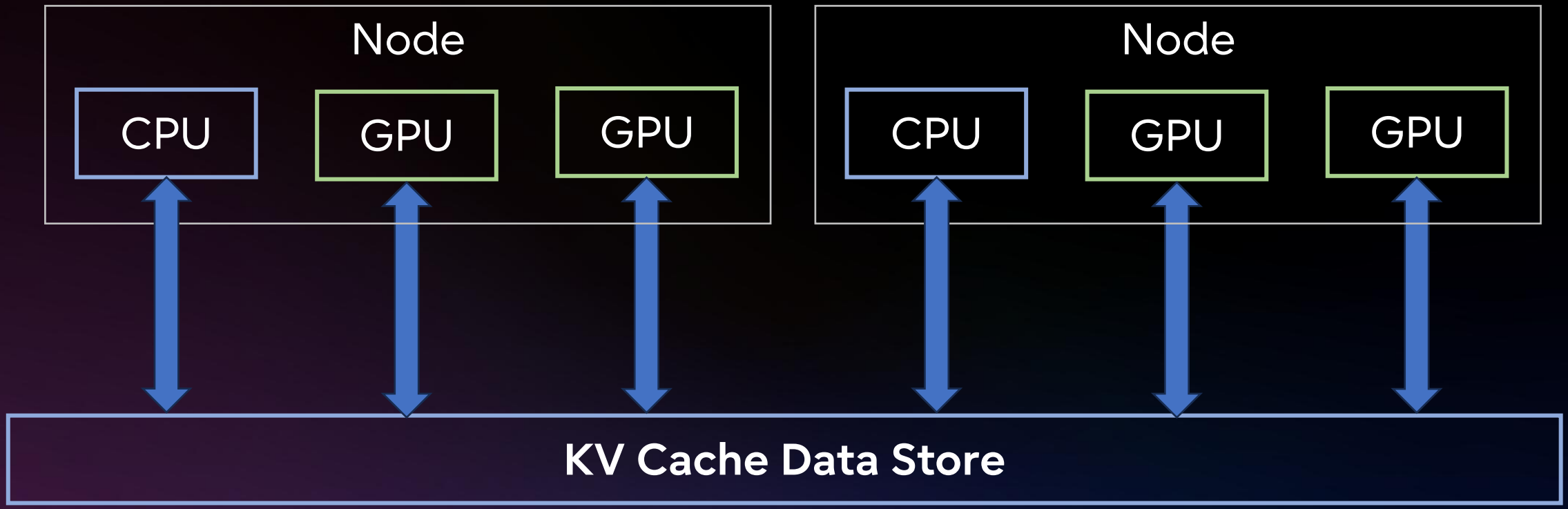
While hierarchical caching is effective, local caches often miss when load balancing sends requests to uncached nodes.

Sharing LLM Context Everywhere

FUJITSU

- Continue long prompts and contexts faster

  without GPU or node constraints – powered by a high-performance data store



Sharing KV cache entries makes the inference cluster more efficient

KV Cache Demo

Legacy Prefill vs L2 Datastore

▶ Start demo | ↺ Reset

R.pdf | O.csv | P.txt

Legacy: Prefill every read

WAIT FOR PREFILL

KV view: SINGLE TIER

Prefill	L1	L2
Every request	GPU KV	Not available
	Entries: 3/16	Entries: 0/0
	HIT:0 MISS:0	L2 HIT:0 Promote:0

...

P.txt

R.pdf | O.csv

auto

Prefill first

Send

Fast datastore + L2 KV

STREAM FROM CACHE

KV view: TWO TIERS

Prefill	L1	L2
Rare	GPU KV	Datastore
	Entries: 3/16	Entries: 7/256
	L1 HIT:0 MISS:0	L2 HIT:0 Promote:0

...

P.txt

R.pdf | O.csv

auto

Serve from L2

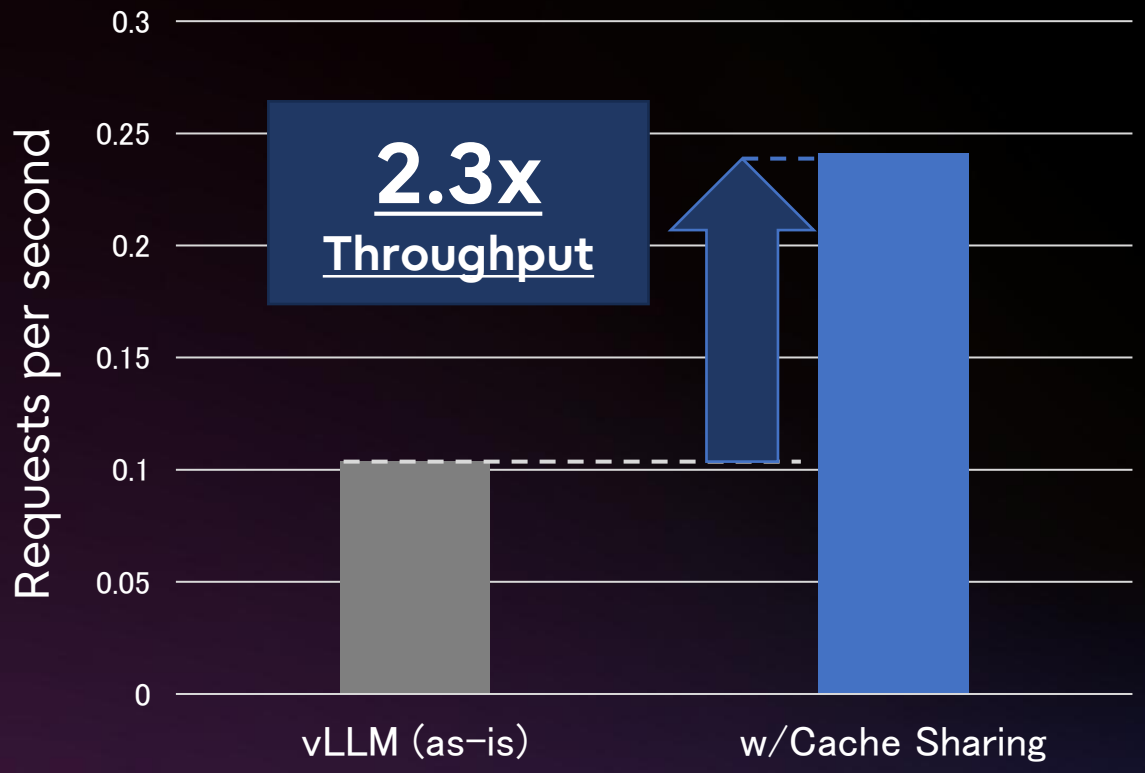
Send

* mock-up

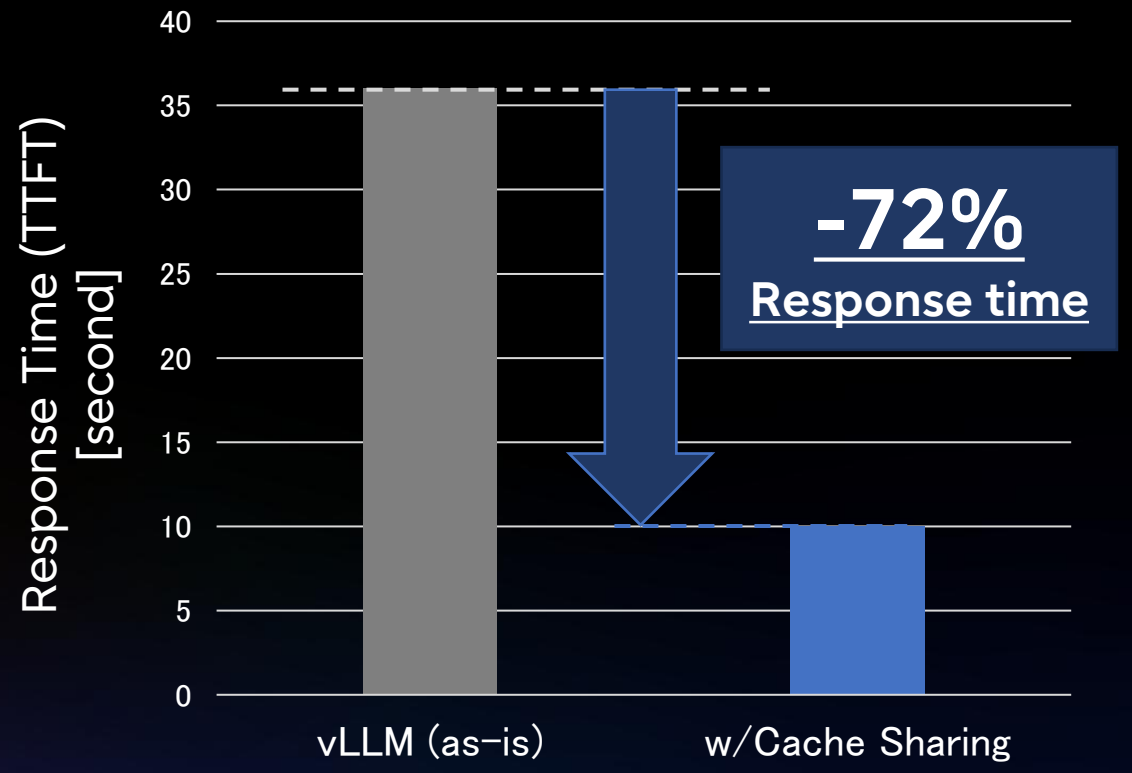
Effect of LLM inference acceleration

FUJITSU

- Inference Throughput



- Inference Response Time



Enhanced performance and reduced GPU usage for agent, personalized AI, and document reference inference.

CxS benchmark: https://github.com/LMCache/LMBenchmark

FUJITSU

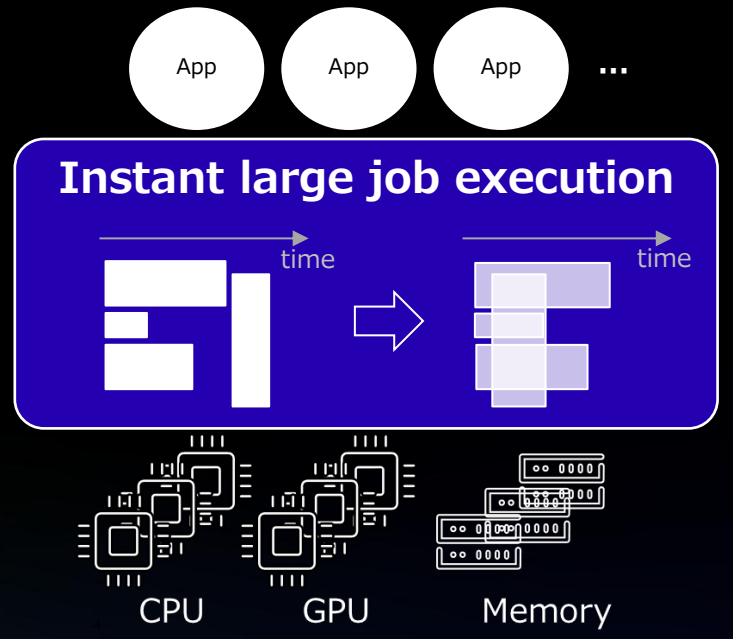
Interactive HPC

Interactive HPC

FUJITSU

A scheduler for parallel applications

- Interactive, large-scale processing
- Real-time data analysis and visualization

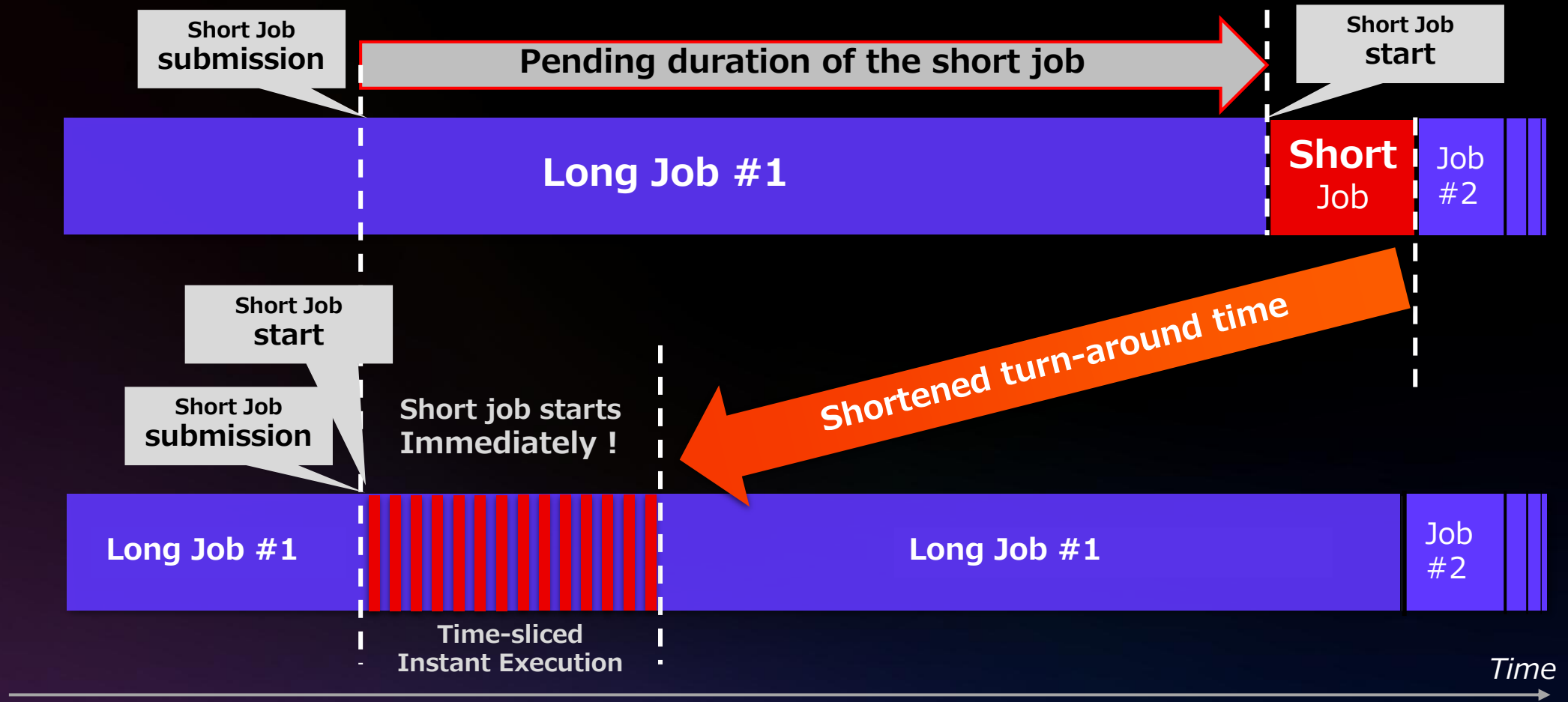


Key Features

- Scalable sub-second global job switching
- CPU/GPU workloads are supported
- Efficient GPU state management
- No user code modification required

Bringing interactivity on HPC

FUJITSU

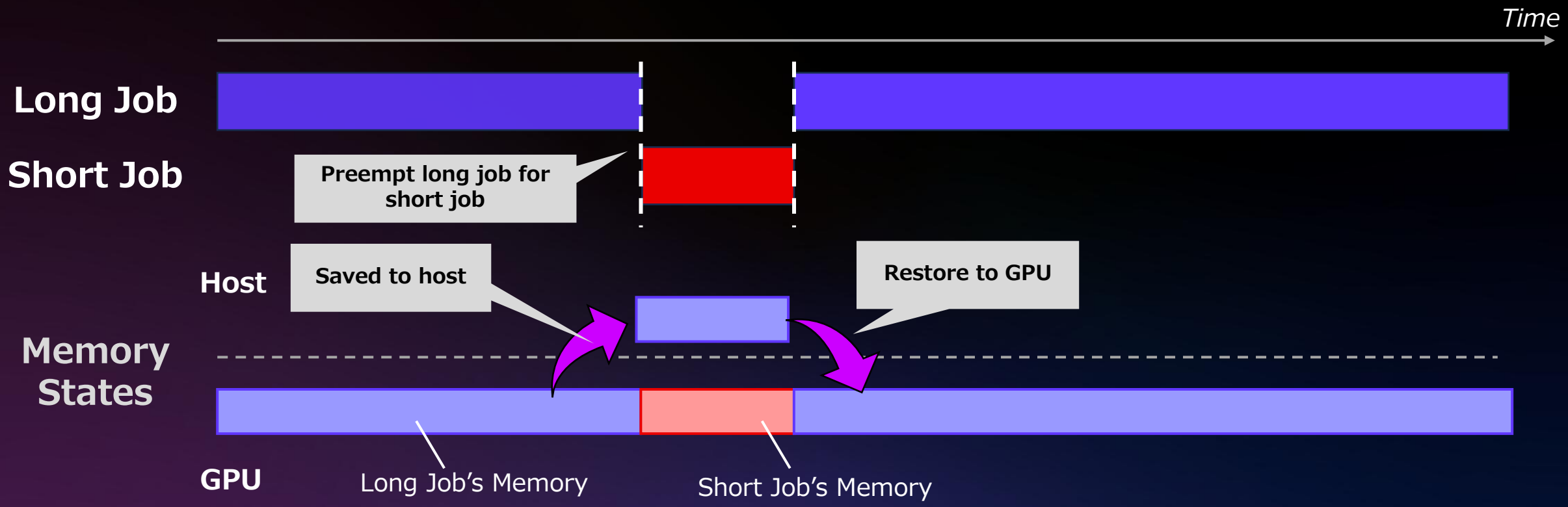


Transparent preemption of GPU Apps.

FUJITSU

[New] An extension to interactive HPC scheduling

- Seamless pause and resume of **any** GPU workload

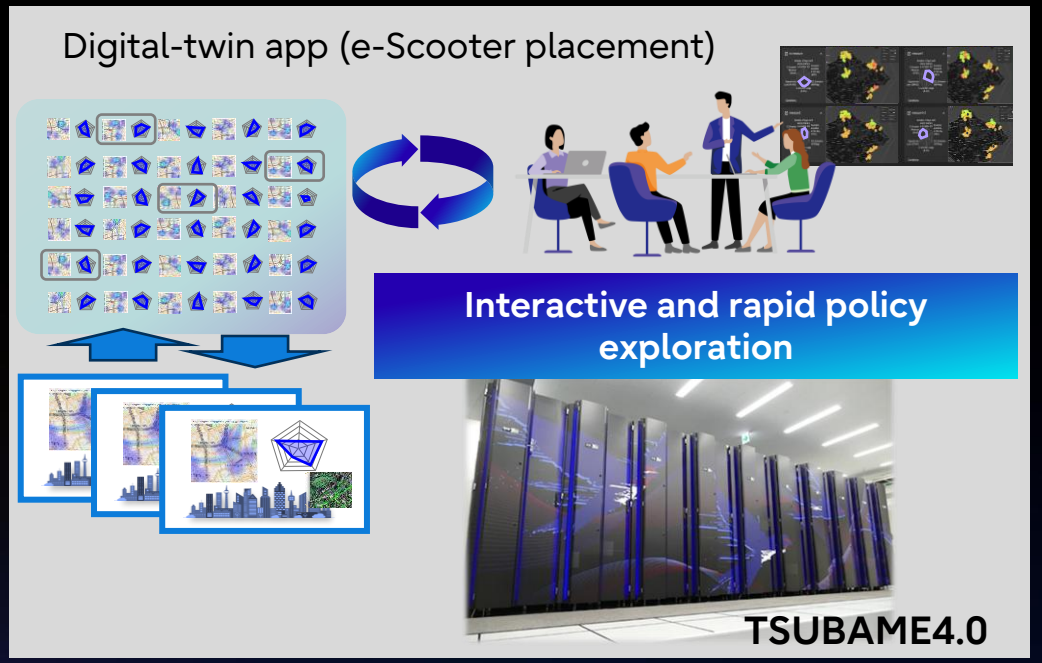


Deployment Case

- 1056-node ARM-based HPC Cluster (System-wide)

Job switch agents deployed on compute nodes

- TSUBAME 3.0/4.0 in Science Tokyo (User space)



☞ In-cooperation with Science Tokyo (Booth #2818)

FUJITSU

Thank you