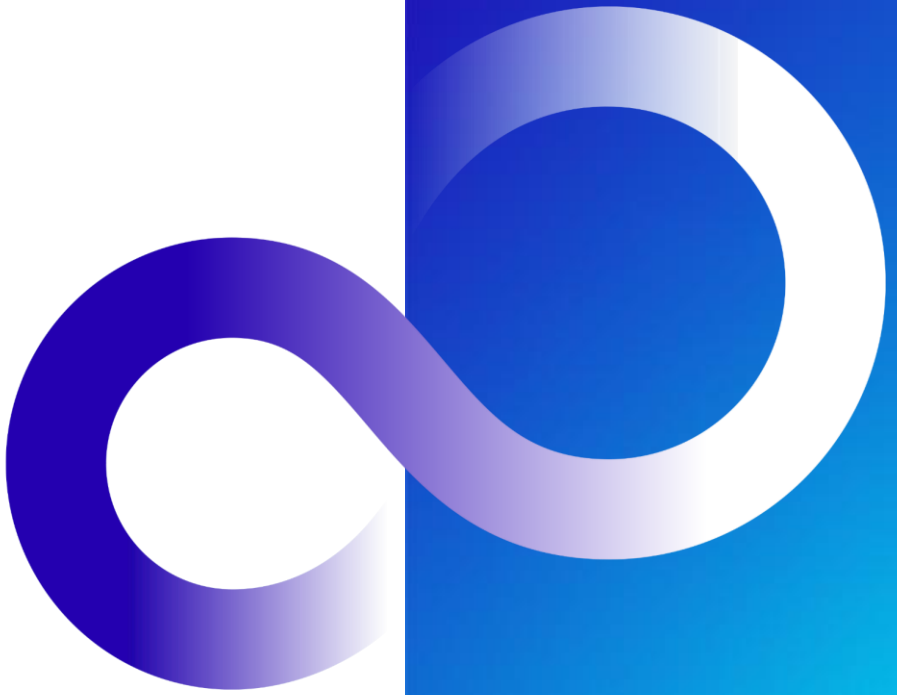
AI Computing Broker Walking Deck

Fujitsu Research of America

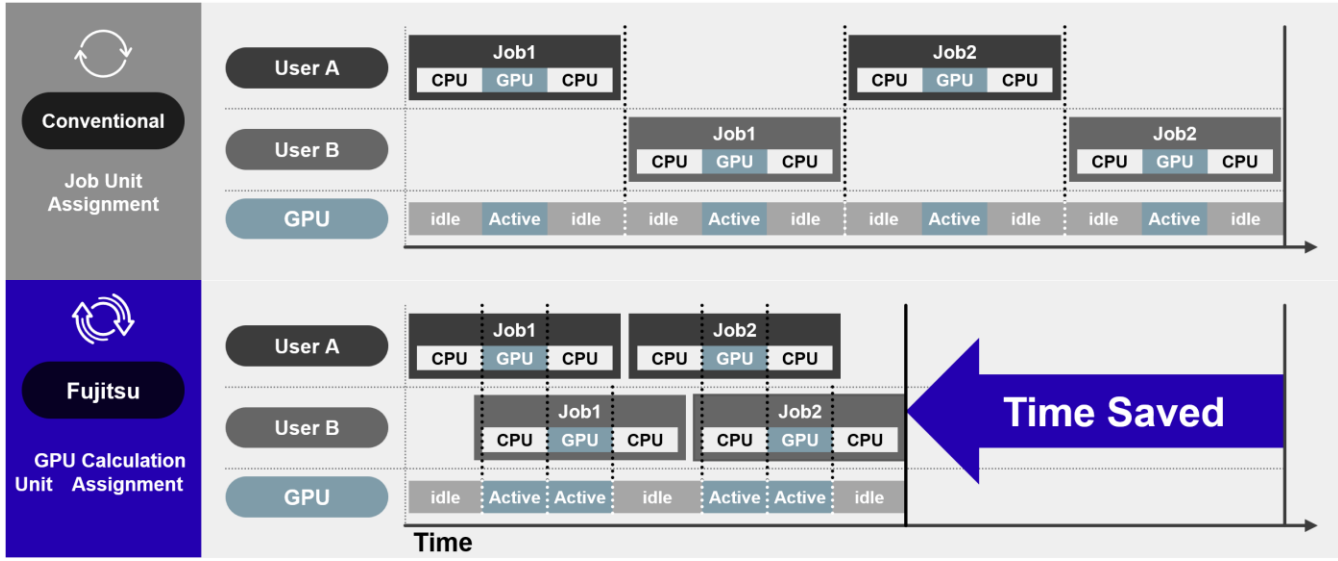


Executive Summary

Challenge

GPU idle time is significantly **decreasing ROI** on GPU infrastructure spending and **slowing down development speed**.

FUJITSU
AI Computing Broker
(ACB)



AI computing broker

dynamically allocates GPUs based on workload activity

Plug-and-Play integration

With no application-layer code changes needed

Efficiency gains

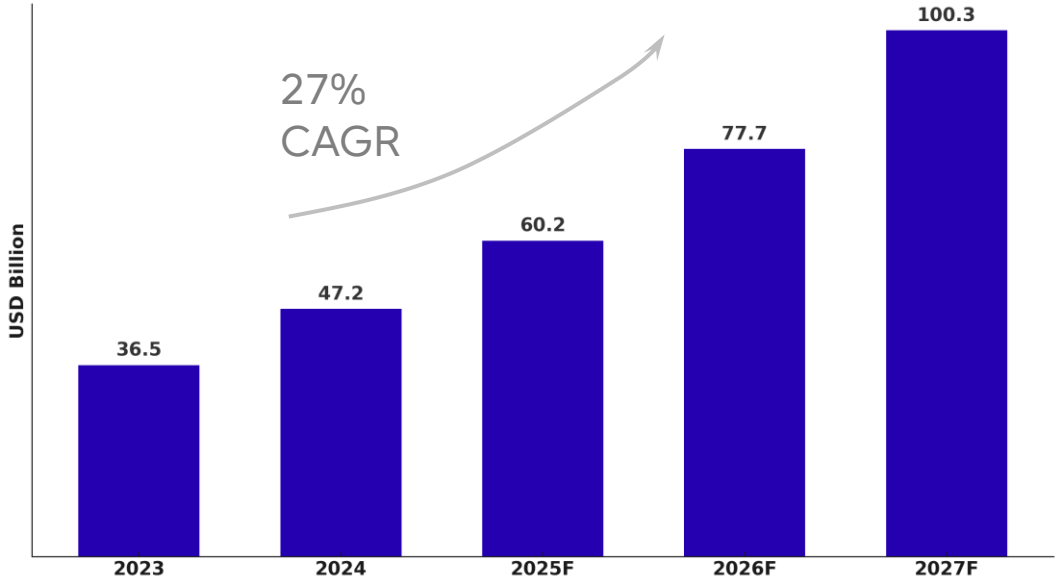
of up to **45%** for complex AI workloads

Maximize Your AI Infrastructure ROI Today!

Low GPU Utilization Compromising Enterprise AI ROI

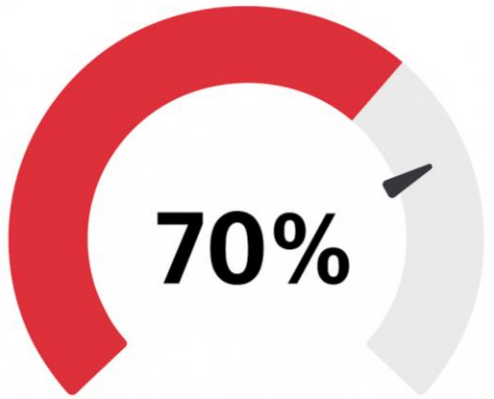
FUJITSU

The AI infrastructure Market poised for Explosive Growth: Expanding at 27% CAGR, nearing $100B in 2027



Key Insight:
Explosive infrastructure growth is driven by complex AI models, broader enterprise adoption, and rising GPU costs

Underutilized GPUs Are Undermining ROI on Infrastructure Spend

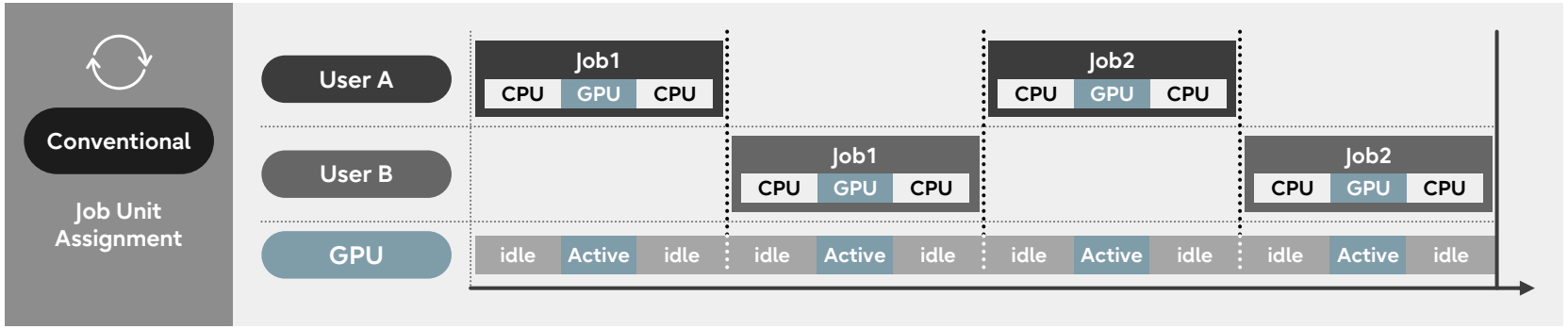


- ❌ **High Infrastructure Cost**
  Wasted spend on idle GPUs.
- ❌ **Slow AI Development**
  Underutilized compute delays model deployment.
- ❌ **Limited Cutting Edge AI**
  Infrastructure prevents effectively running cutting-edge models.

Average GPU utilization remains under 70% even during peak periods across enterprises.
(The State of AI Infrastructure at Scale 2024 Report)

Why Optimization Is Hard: Complexity of AI Workloads

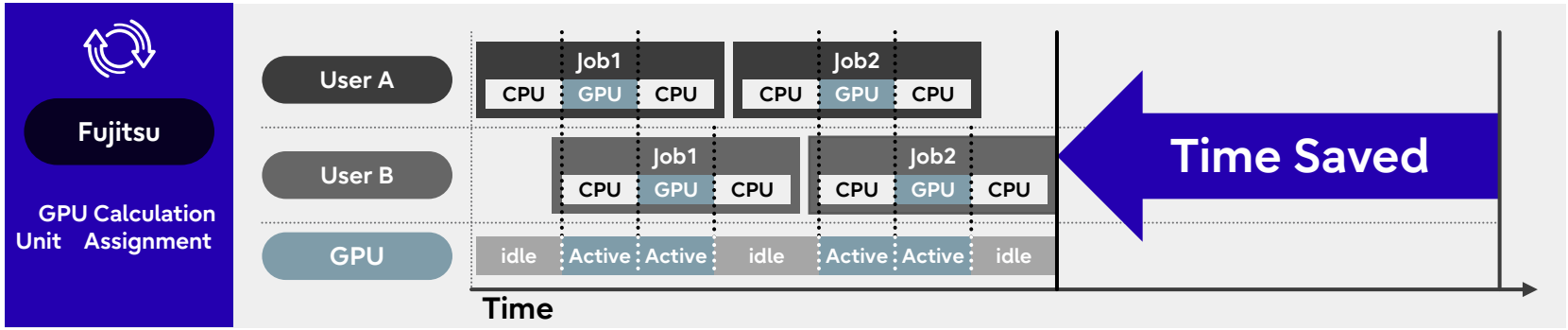
Key issues	Why
Static Allocation	GPUs stay occupied for entire jobs, causing idle time during CPU-heavy phases
Heterogeneous Compute Profiles	AI tasks need different CPU/GPU ratios at each stage (e.g., AlphaFold2, risk prediction)
Inefficient Scheduling	Simple schedulers fail to fully utilize shared GPUs



Source: https://en-documents.research.global.fujitsu.com/ai-computing-broker/documents/ACB_WhitePaper_en.pdf

Solving GPU Underutilization: AI Computing Broker Maximizes GPU Utilization Across Complex AI Workloads

Key benefits	How ACB Works
Higher GPU Utilization	**Runtime-aware GPU allocation:** Monitors AI framework to allocate GPUs needed
Reduced Infra Cost	**Full Memory Access:** Active program has access to the full GPU memory
Accelerated AI Dev.	**Advanced Scheduling Algorithms:** Employs techniques like backfill to optimize job placement and maximize aggregate utilization



Source: https://en-documents.research.global.fujitsu.com/ai-computing-broker/documents/ACB_WhitePaper_en.pdf

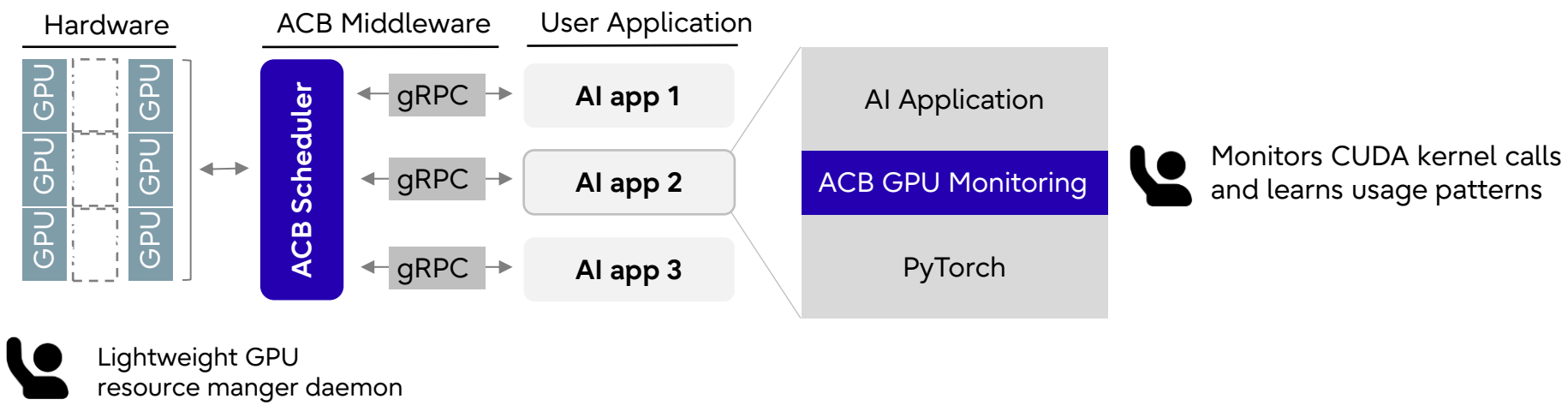
AI Computing Broker: Unlocking GPU Efficiency

FUJITSU

AI computing broker (ACB) Core Features

- Runtime aware GPU scheduler middleware
- No code changes in user program required
- Docker support
- Multi server

ACB is a runtime aware GPU scheduling middleware

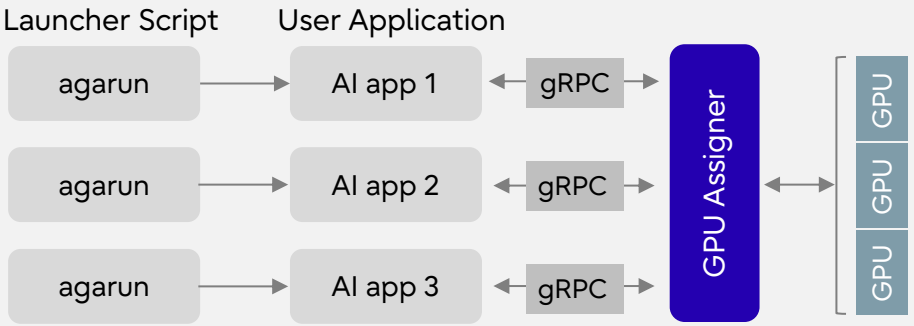


A Peak Under the Hood: Architecture of ACB

FUJITSU

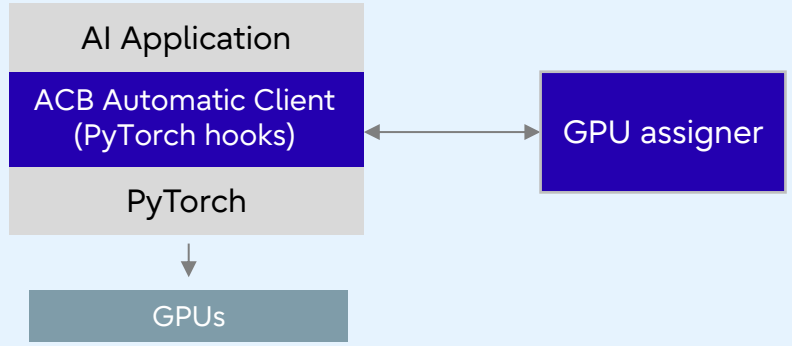
ACB enables seamless, flexible GPU sharing and efficient utilization across diverse AI applications.

Server-side (lightweight app)



The GPU assigner daemon manages the GPU pool and allocates resources to applications via gRPC protocols

Client-side



The ACB client intercepts PyTorch calls to monitor and analyze GPU usage patterns

Source: https://en-documents.research.global.fujitsu.com/ai-computing-broker/documents/ACB_WhitePaper_en.pdf

Proven Efficiency Gains: Where ACB Delivers Value

FUJITSU

Proven Customer Impacts:

- **2× Training Throughput** for AI workloads
- **5× Memory Oversubscription** for LLM inference
- Delivered across **multiple AI architectures, datasets, and industries**

AI Models:

Transformers, YOLO, GAN, LLM, hybrid workflows

Industry Verticals:

GPU cloud provider, Finance, AI Tech

GPU Type:

H100, A100, L40S, RTX

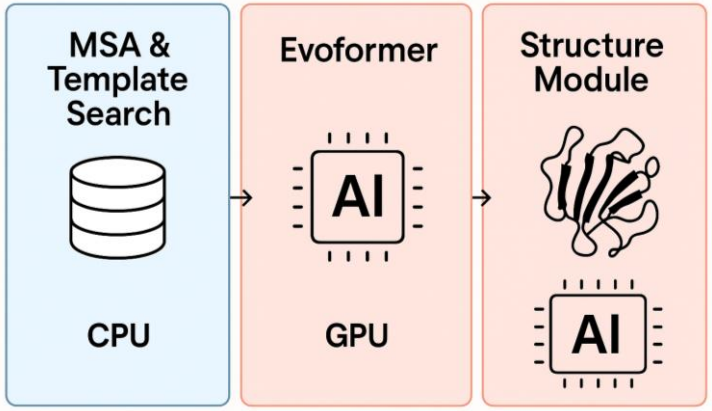
Pain point:

Model training throughput, GPU requirement, GPU scheduling

ACB in Action: Increasing GPU Utilization by 45% for AlphaFold2

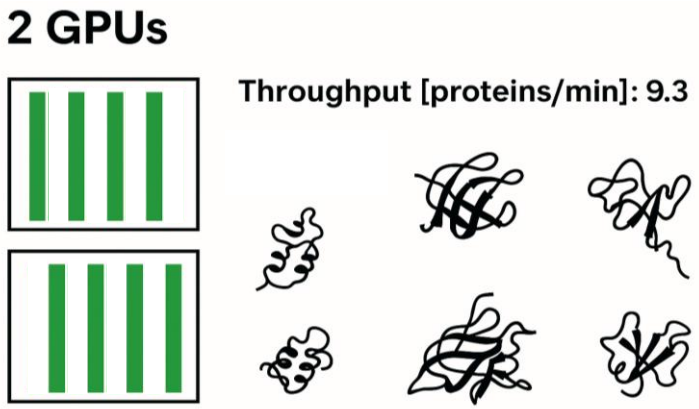
FUJITSU

AlphaFold2 solving a Grand Challenge in Structural Biology with AI



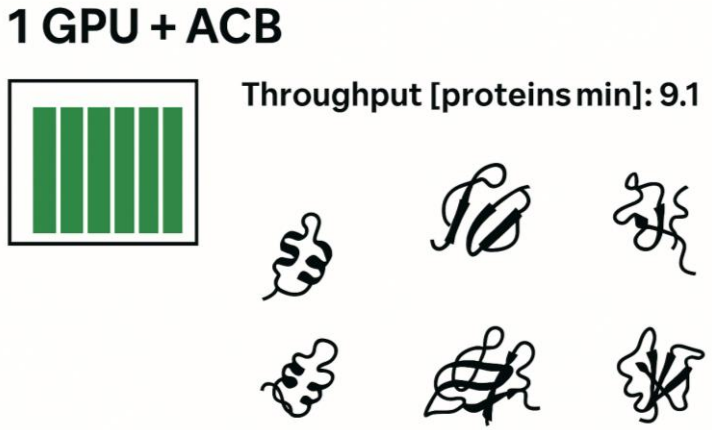
- **Revolutionized structural biology**: Protein shape prediction from sequences
- **Drug discovery Impact**: Screening large combinatorial libraries in protein engineering
- **Complex Architecture**: Multi-stage process with diverse CPU/GPU demands

GPU Usage Is Not Consistently High in AlphaFold2 Inference



- **Baseline Performance**: 2 GPUs yield 9 proteins/min
- **Template Search**: Significant GPU idle time
- **Static Allocation**: GPUs reserved for full job duration

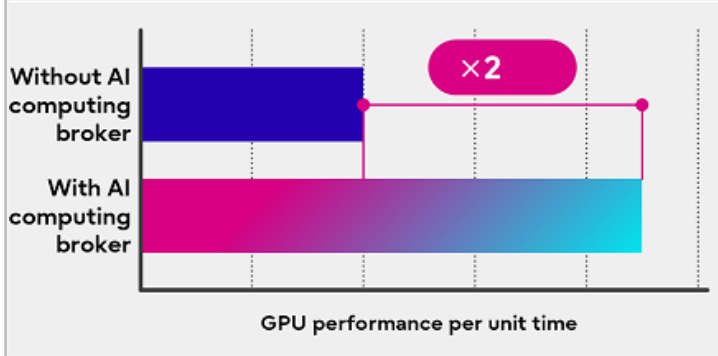
AI Compute Broker Achieves Same Throughput With a *Single* GPU



- ACB reclaims idle GPU time for a second worker
- Increases GPU Utilization by **45%***
  * Tested with T4 GPUs
- Watch live demo here

AI Compute Broker is Delivering Value for Customers in Finance, Cloud Computing, and IT

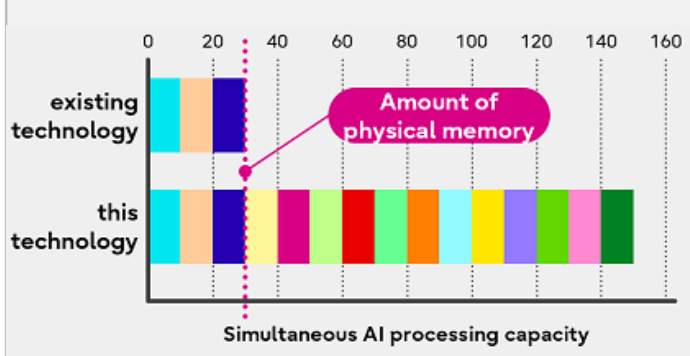
Japanese FinTech



Use case: AI model prototyping for foreign exchange risk prediction

Key Result: Increased throughput 2-fold

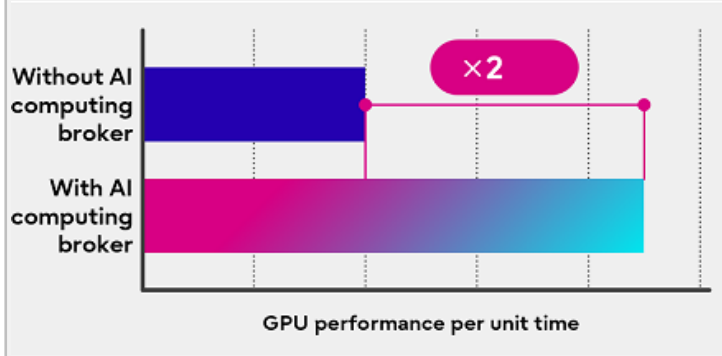
Japanese cloud computing



Use case: Host several AI models on a single GPU node

Key Result: Manage five times the physical GPU memory capacity

Japanese AI Tech



Use case: Object Recognition Model Training for IaaS business

Key Result: Increased throughput 2-fold

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

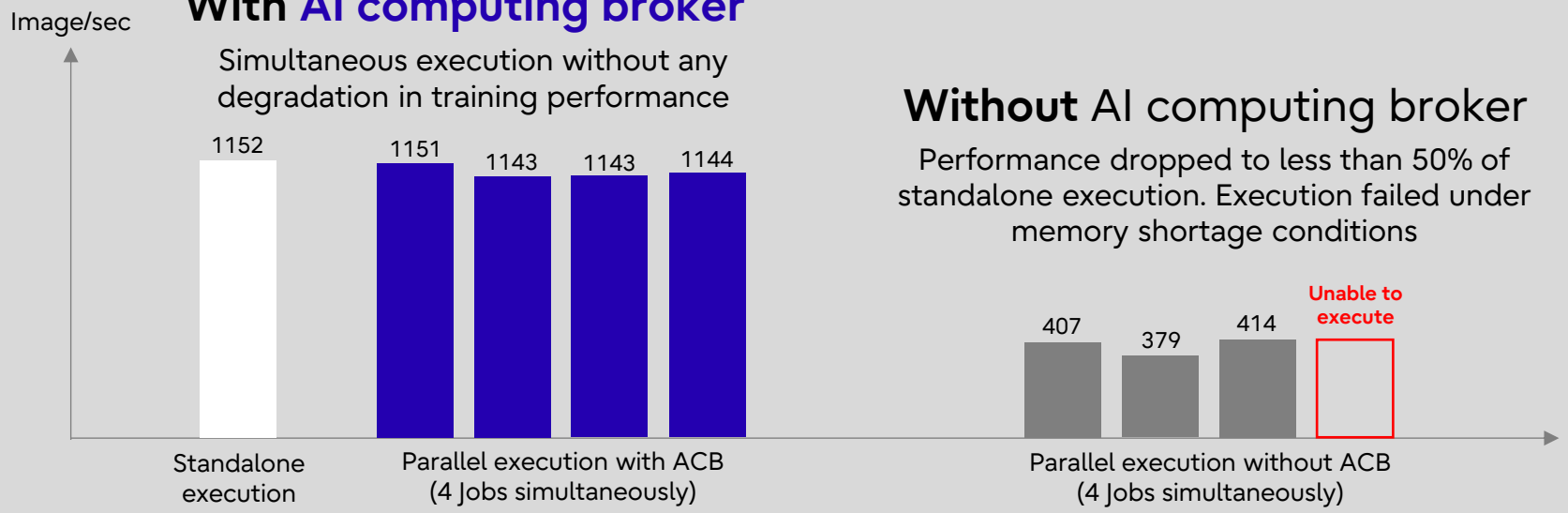
TRADOM Inc.: https://www.tradom.jp/company

Use Case: Xtreme-D

Xtreme-D

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

Verification test in Raplase laaS



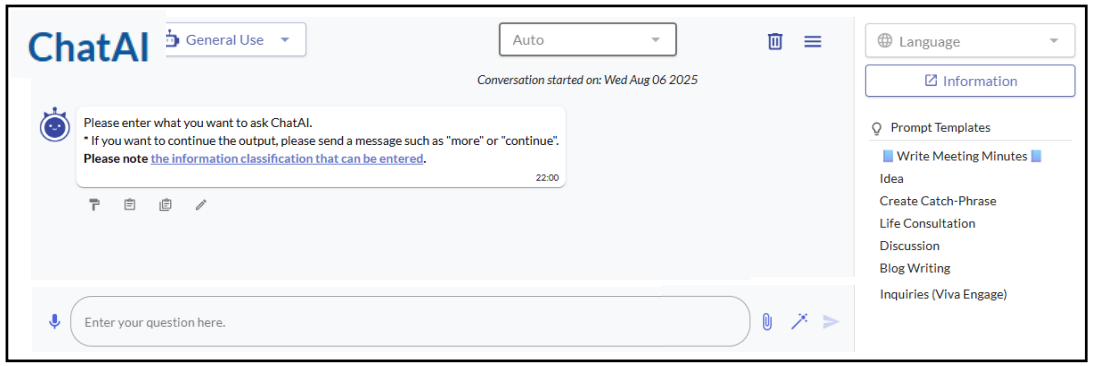
- Validation Approach: Demonstrated through Xtreme-D's proprietary PyTorch benchmark tests. A computational load was applied to the GPU and the measured CUDA utilization, memory consumption, FP16/FP32 performance, and training indicators
- Measurement Environment: CPU: AMD EPYC 7743 x 2, Memory: 1 TB,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

ACB in Action II: Memory Oversubscription for Multi-LLM hosting decreasing TOC for On-Prem Deployment



Model Selection:

1. Command R: Global Taxes
2. Takane: M&A
3. Takane: M&A Contracting
4. Command R: IT Outsourcing
5. ...

Chat AI interface enables selection of domain-specific models (e.g., Global Taxes, M&A contracts)

Efficient Inference Serving with **vLLM** Integration

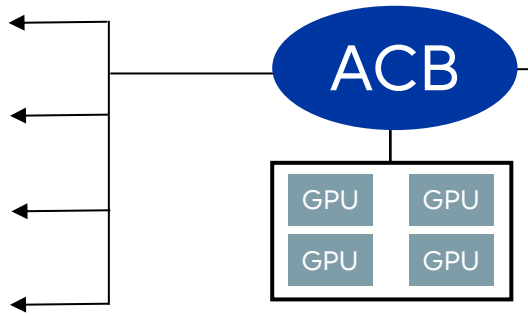
Example:

LLM1: Command R Global Taxes + **RAG 1:** Tax report

LLM 2: Takane M&A

LLM 3: Takane M&A Contracting + **RAG 2:** M&A

LLM 4: Command R: IT Outsourcing



Only Single Server!

LLM 5: Takane IP

LLM 6: Takane-Employment

LLM 7: Takane-Compliance

LLM 8: Takane-Privacy-GDPR

ACB Delivers SLA Parity, Minimal Swap Overhead, and up to 4X Consolidation for Major TCO Impact

FUJITSU

Methodology

Problem

Serving benchmarks assume one "hot" model per GPU → unrealistic for production with many quiet models and spiky traffic.

Adopted the **InferenceMAX benchmark** to validate SLA parity and swap overhead in multi-model scenarios.

Solution

vLLM + **AI Computing Broker (ACB)** enables dynamic model state swapping → consolidates idle GPU time across tenants.

Implications

Traffic Pattern Matters

- Best case: Spiky traffic with non-overlapping requests
- Near-perfect consolidation factor

Cost & Infrastructure Impact

- Up to 4x consolidation → ~75% TCO reduction
- 1 GPU / model → 1 GPU / all models (no SLA regression)
- Smaller hardware footprint

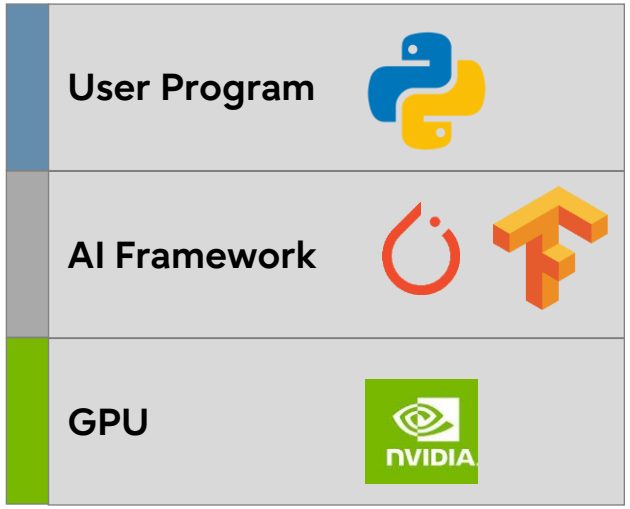
Performance & Risk Envelope

- Worst case: frequent swaps
- Negligible impact on throughput and SLA

Fast Deployment with Minimal Disruption: No Code Changes Required in User Programs

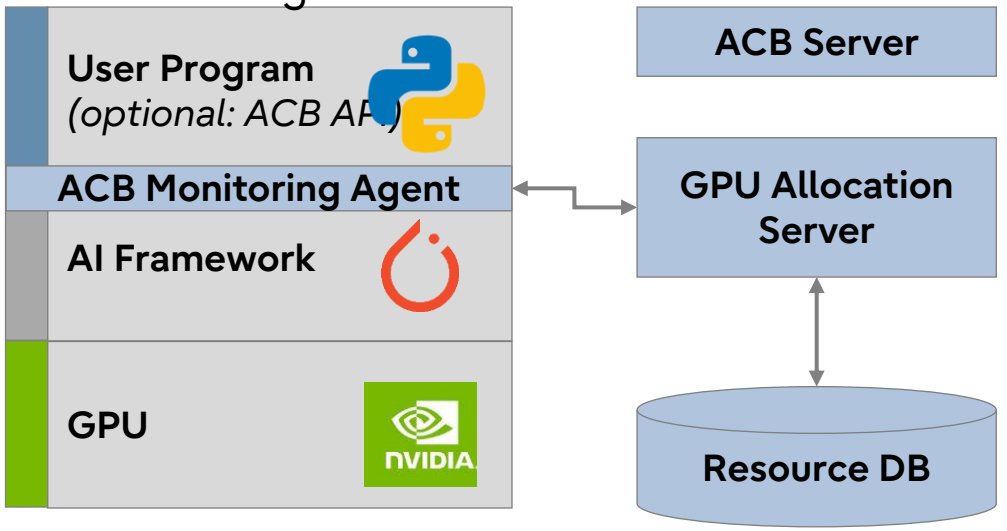
Traditional AI Execution

The program calls GPU directly through AI framework.



AI Execution Using ACB

ACB monitors function calls to the AI framework and tracks GPU usage for dynamic resource assignment.



TensorFlow support limited
AMD ROCm support on the roadmap

Integration with Minimal Disruption

- **Framework Compatibility:**
  Works with existing AI applications (PyTorch, TensorFlow) without code modification
- **Rapid Installation:**
  Quick deployment in on-premises, cloud, or hybrid environments
- **Workflow Integration**:
  Integrates seamlessly with existing AI workflows
- **ACB PyTorch Auto Client:**
  No code changes required in the user program and automatically monitors kernel calls

What Makes ACB Stand Out: Unique Task-Level Optimization Across CPU & GPU

Feature	ACB	Run: ai	Exostellar ai	Slurm
Runtime-aware GPU Allocation	🟢	🔴	🔴	🔴
Memory partitioning	🟢	🟢	🟢	🟡
Memory oversubscription	🟢	🟡	🟢	🔴
Cluster-level orchestration	🔴	🟢	🟡	🟢
Plug-and-play integration	🟢	🔴	🔴	🔴
Focus	GPU Eff.	GPU Mgmt.	GPU Mgmt.	Job Sched.

Let's Partner to Unlock Full GPU Potential

FUJITSU

1. **ACB Works Best When:**
   - CPU/GPU tasks alternate frequently
   - You batch or queue GPU jobs
   - GPU memory usage spikes intermittently
   - You host multiple LLMs with uneven demand

2. **Explore the Solution:** Visit the Fujitsu AI Compute Broker website for in-depth information and technical white paper.

3. **Sign up for a Free Trial:** Evaluate ACB yourself in our 30 days free trial.

4. **Questions?** Please contact us via our inquiry form or email.

Website: AI computing broker - Fujitsu Research Portal
Email: fra_acb_support@fujitsu.com

Appendix and Backup Slides

Engineered for Broad Compatibility Across Enterprise AI Stacks

FUJITSU

Category	Details
GPU Support	Consumer (RTX 30/40 Series) · Mid-Tier (A10, L40S) · High-End (A100, H100) (NVIDIA only)
Driver	NVIDIA driver 515+ · CUDA 11.4+
OS Compatibility	Ubuntu 20.04+ · CentOS 8+
Python	3.10+
Frameworks	PyTorch · TensorFlow (limited)
Deployment Modes	Bare Metal · Docker · Slurm · Kubernetes (planned)
Security Model	Encrypted license key tied to GPU type
Integration Points	vLLM· NVIDIA MPS · MIG

Complement Technologies

FUJITSU

Technology	Layer	Definition	How It Complements ACB
NVIDIA MIG	Driver	Partitions a single physical GPU into isolated GPU instances with dedicated compute and memory resources.	Allows ACB to dynamically assign pre-partitioned resources for workload isolation.
NVIDIA MPS	Driver	Enables multiple CUDA processes to share a GPU concurrently by avoiding context switching.	ACB can monitor and schedule these processes more effectively to improve throughput.
Run:ai Time-Slicing	Driver	Alternates GPU access between jobs over time without hardware partitioning.	ACB optimizes which jobs are active to reduce idle slices and boost efficiency.
vLLM	Application	An optimized inference engine for large language models using PagedAttention for fast throughput.	ACB allocates GPU resources to vLLM jobs based on runtime demand, improving responsiveness.
Triton Server	Application	A multi-framework model server that handles concurrent inference jobs efficiently.	ACB ensures Triton has access to active GPU resources, avoiding contention or idling (currently not supported by ACB).
Alluxio	Data Layer	A data orchestration system that caches and unifies data access across multiple storage backends.	By speeding up I/O, Alluxio removes data bottlenecks, allowing ACB-managed GPUs to stay busy.
Slurm	Middleware	A traditional HPC job scheduler that allocates compute resources in clusters.	ACB can work alongside Slurm to handle finer-grained GPU allocation within scheduled jobs on same node(s).
Kubernetes (with GPU Operator)	Middleware	Orchestrates containers and can manage GPU provisioning with plugins.	ACB enhances K8s by optimizing GPU utilization dynamically within the assigned pods.

Complement Technologies

FUJITSU

- **Fujitsu ACB (middleware) + NVIDIA MIG (driver level)**: While MIG partitions a GPU into multiple instances for isolated workloads, ACB can dynamically allocate these instances to different tasks based on real-time demand, maximizing overall GPU utilization.

- **Fujitsu ACB + Triton Inference Server/vLLM (application layer)**: ACB can manage and allocate GPU resources dynamically to Triton/vLLM, ensuring that inference workloads receive the necessary resources when needed, leading to efficient model serving.

- **Fujitsu ACB + NVIDIA MIG + vLLM**: Combining all three allows for a robust system where GPUs are partitioned for isolation (MIG), resources are dynamically allocated based on demand (ACB), and models are served efficiently (vLLM).