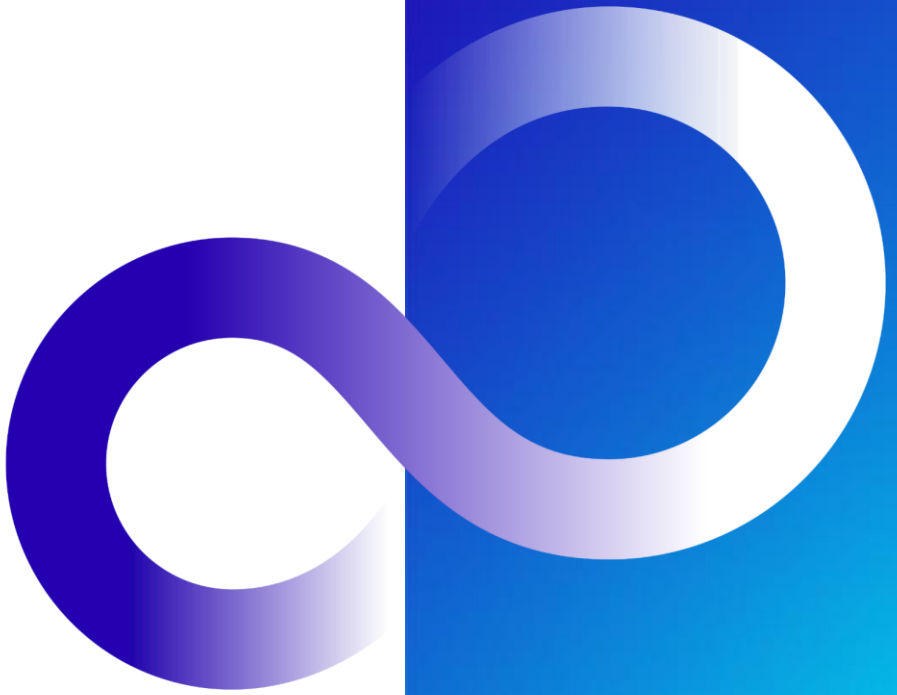
AI Computing Broker Walking Deck

Fujitsu Research of America



Fujitsu Public

エグゼクティブサマリ

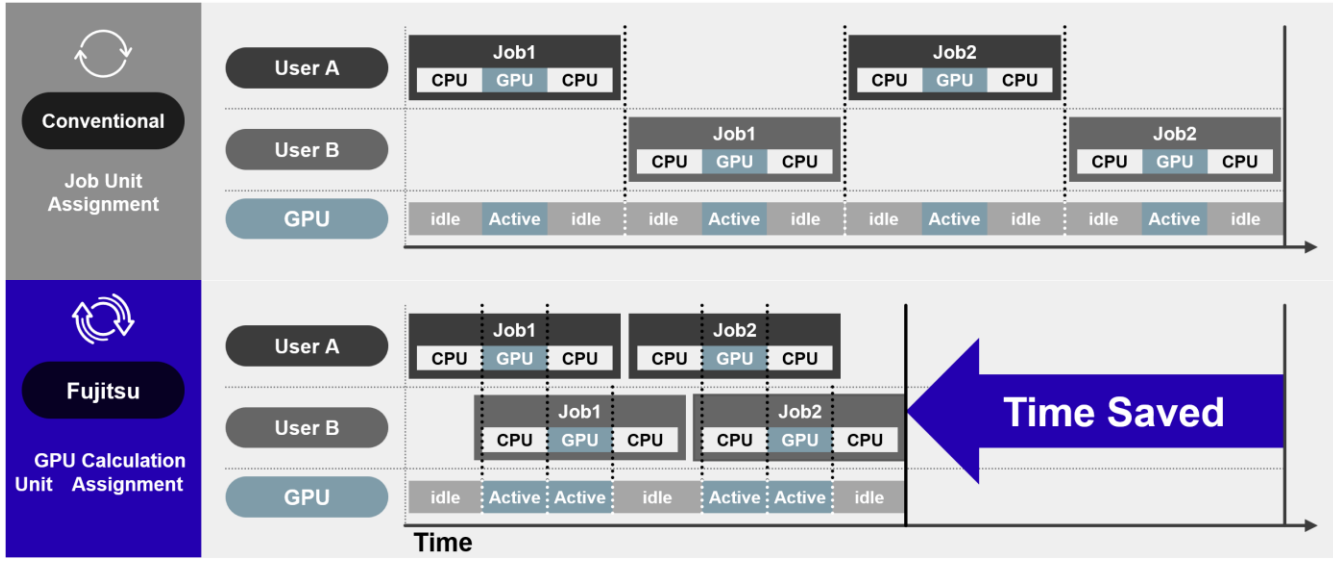
FUJITSU

Challenge

AIに不可欠なGPU、効率的に活用できていますか？

多くのAI処理において、GPUの待機時間が発生し、インフラ投資のROI低下、さらにAI開発スピードが遅くなっている可能性があります

GPU
FUJITSU
AI Computing Broker
(ACB)



Automated workload distribution

ワークロードのアクティビティに基づき、GPUを動的に割り当て

Plug-and-Play integration

アプリケーション層のコード変更不要

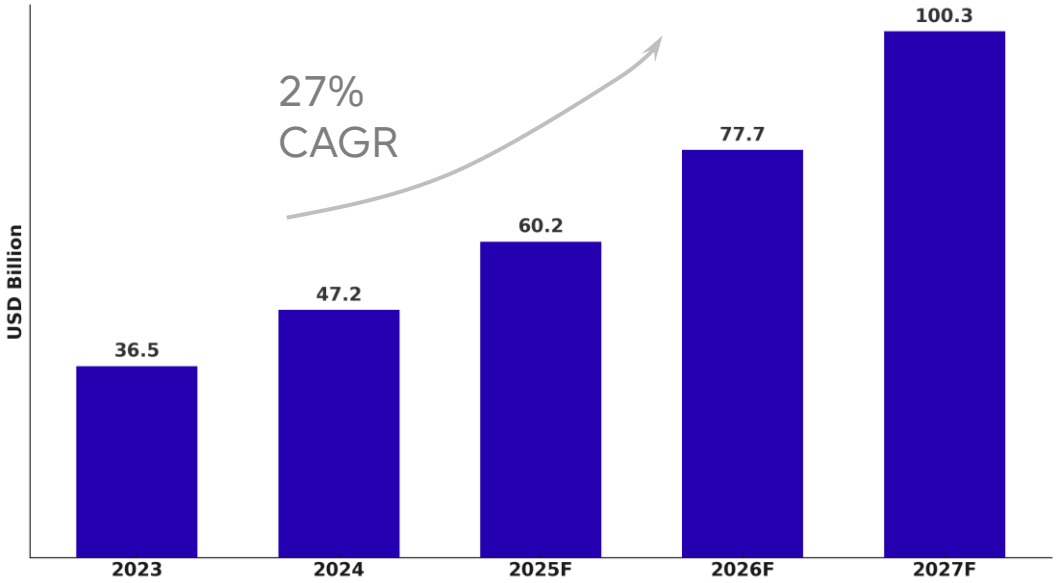
Efficiency gains

複雑なAIワークロードに対し最大45%のGPU利用効率向上

今すぐGPUインフラのROIを最大化し、AI開発を加速しましょう！

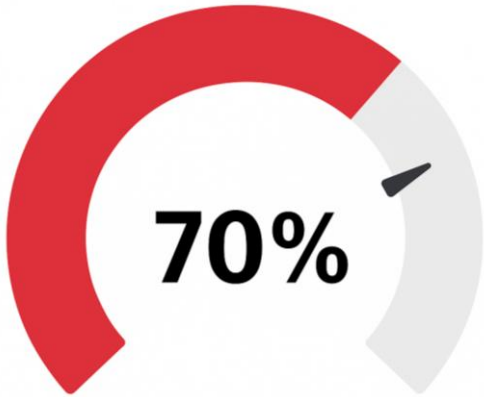
非効率なGPU利用が、企業AI ROIを阻害

AIインフラ市場は爆発的な成長を遂げようとしており、年平均成長率27%で拡大し、2027年には1,000億ドルに迫ると予測されています。



Key Insight:
爆発的なインフラ成長は、複雑なAIモデル、企業におけるAI導入の拡大、そしてGPUコストの上昇によって牽引されています。

稼働率の低いGPUが、インフラ投資のROI（投資対効果）を損なっています。

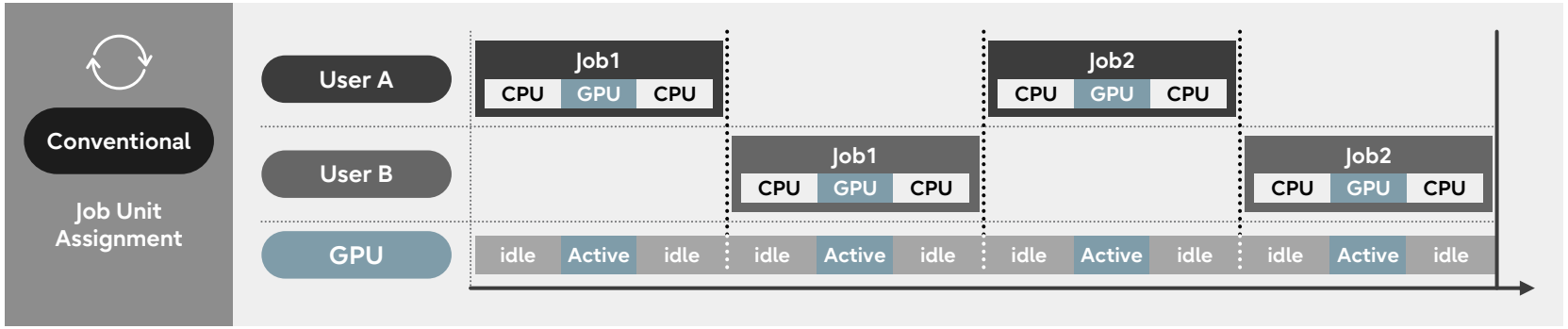


- ✕**高額なインフラコスト：**
  非稼働GPUへの無駄な支出
- ✕ **AI開発の遅延：**
  コンピューティングリソースの活用率の低さが、AIモデルの配置・展開を阻害
- ✕**最先端AI活用の制約：**
  既存のインフラでは、最先端のAIモデルを効果的に実行することが困難

企業の平均GPU使用率はピーク時でさえ70%を下回っています。
(The State of AI Infrastructure at Scale 2024 Report)

非効率なGPU利用を招くAIワークロードの複雑性

主な課題	原因
静的な割り当て	GPUはジョブ全体を通じて占有され続けるため、CPU処理が多い時間帯でGPUアイドル時間が発生する
異質なコンピューティングプロファイル	AIタスクは各段階でCPU/GPU比率が変動するため、リソース要求が均一でない 例：AlphaFold2、リスク予測
非効率なスケジューリング	単純なスケジューラでは共有GPUを十分に活用できない

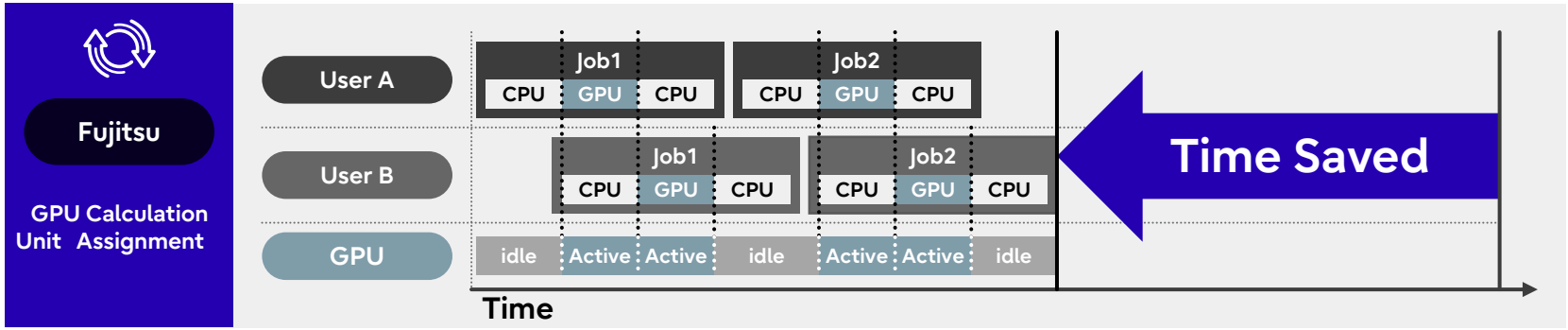


Source: https://en-documents.research.global.fujitsu.com/ai-computing-broker/documents/ACB_WhitePaper_en.pdf

ACBが複雑なAIワークロード全体でGPU利用効率を最大化

FUJITSU

主な利点	ACBの仕組み
GPU利用効率の向上	**ランタイム対応GPU割り当て:** AIフレームワークの稼働状況を監視し、必要なGPUを割当てる
インフラコストの削減	**フルメモリアクセス:** アクティブなプログラムは、GPUメモリ全体にアクセスできる
AI開発の加速	**高度なスケジューリングアルゴリズム:** 高度なスケジューリングアルゴリズム (バックフィル等) でジョブ配置を最適化し、総利用率を最大化する



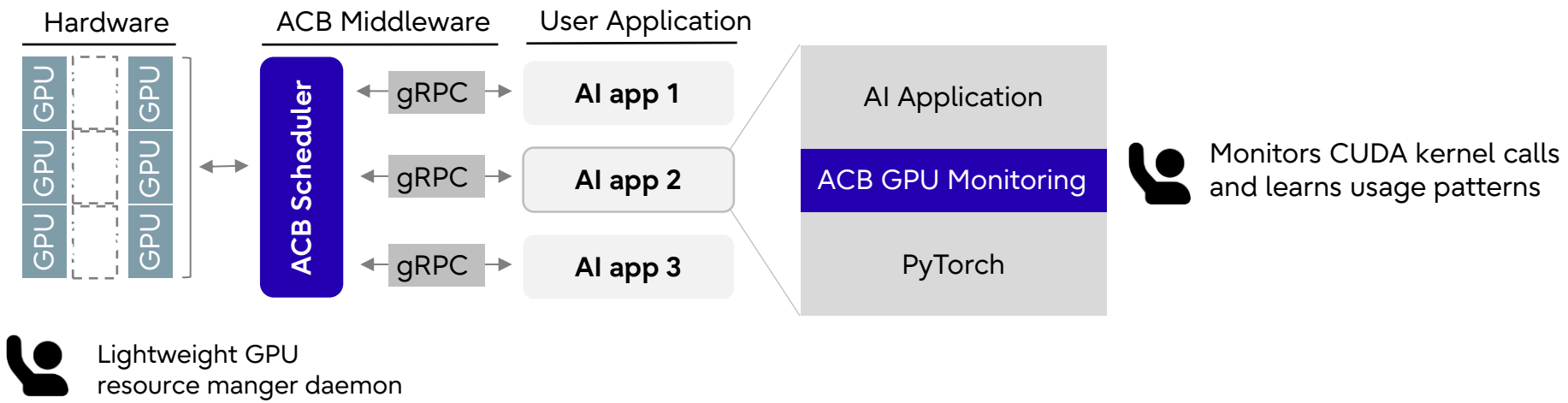
Source: https://documents.research.global.fujitsu.com/ai-computing-broker/documents/ACB_WhitePaper_ja.pdf

ACBによるGPU利用効率の最大化

ACBの主な機能

- ランタイム対応 GPU スケジューリングミドルウェア
- ユーザープログラムでのコード変更不要
- Docker サポート
- マルチサーバー

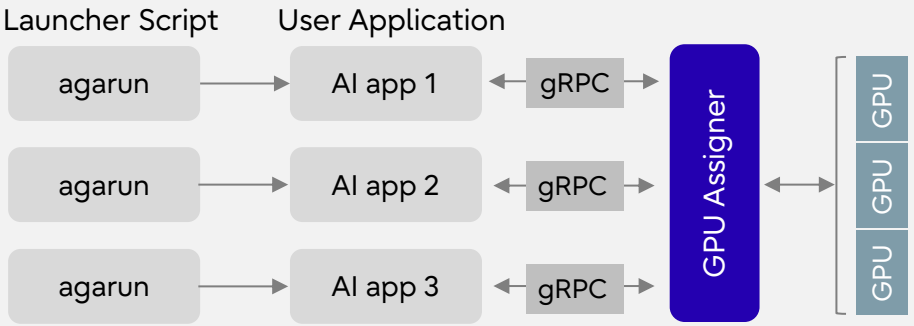
ACB はランタイム対応GPUスケジューリングミドルウェア



ACBアーキテクチャ概要

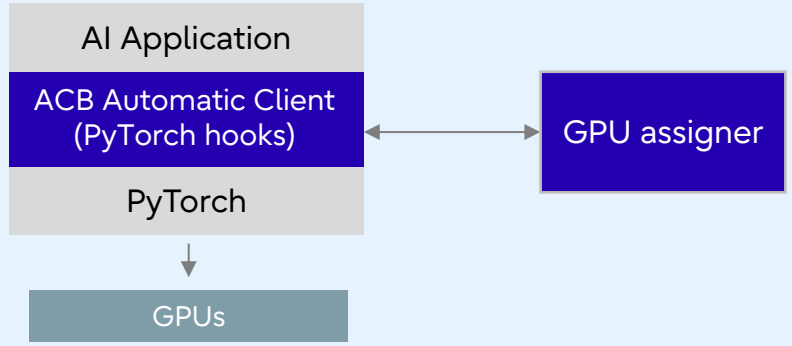
ACBは多様なAIアプリケーション間でシームレスかつ柔軟なGPU共有と効率的な利用を実現

Server-side (lightweight app)



GPU assigner デーモンがGPUプールを管理し、gRPCプロトコル経由でアプリケーションにリソースを割り当てる

Client-side



ACB clientは、Pytorchの呼び出しをフックし、GPU使用パターンを監視・分析する

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

実証されたACBがもたらす価値

実証された顧客インパクト

- AI学習ワークロードで**2倍のスループット向上**
- LLM推論で**5倍のメモリオーバーサブスクリプション**
- **複数のAIアーキテクチャ、データセット、業界**にまたがって提供

AI Models:

Transformers, YOLO, GAN, LLM, hybrid workflows

Industry Verticals:

GPU cloud provider, Finance, AI Tech

GPU Type:

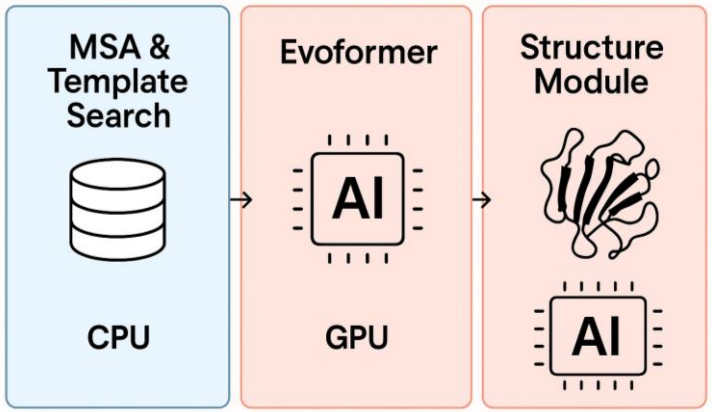
H100, A100, L40S, RTX

Pain point:

Model training throughput, GPU requirement, GPU scheduling

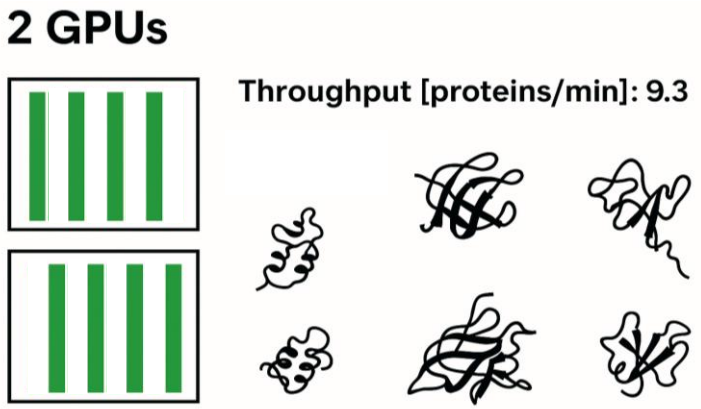
ACBの実例I: AlphaFold2でGPU利用効率を45%向上

AlphaFold2：
AIで構造生物学における大きな課題を解決



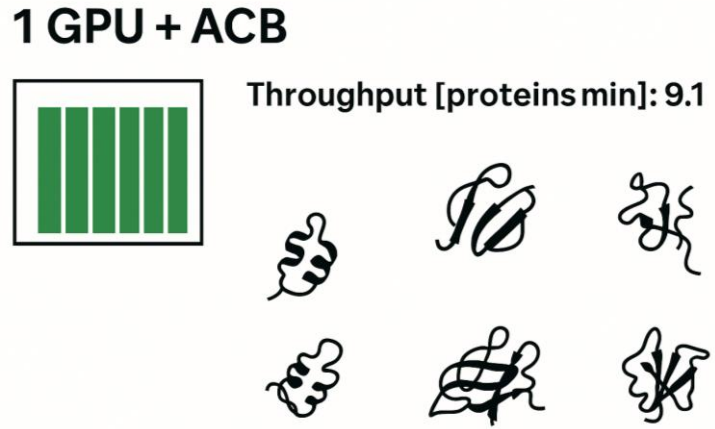
- **構造生物学への革命：**アミノ酸配列からタンパク質の構造予測
- **創薬への影響：**タンパク質工学における大規模なコンビナトリアルライブラリーのスクリーニング
- **複雑なアーキテクチャ：**多様なCPU/GPU要求を持つ多段階プロセス

AlphaFold2推論におけるGPU使用率は、一貫して高くない



- **ベースライン性能：**2つのGPUで9タンパク質/分を生成
- **テンプレート検索：**GPUのアイドル時間が大幅に発生
- **静的割り当て：**GPUがジョブの全期間にわたって予約される

AI Compute Brokerは、単一のGPUで同等のスループットを実現

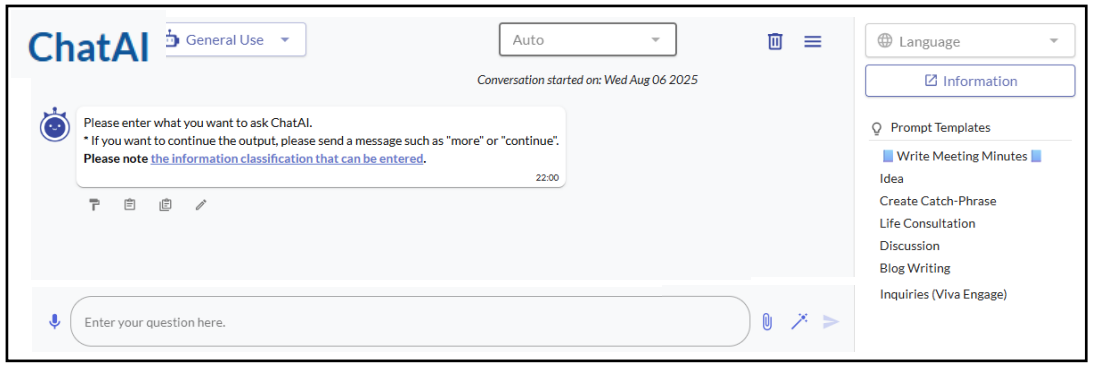


- ACBは、アイドル状態のGPU時間を回収し、2つ目のワーカーに割り当て
- GPU使用率を45%向上*

*T4 GPUでテスト済み

ライブデモはこちらをご覧ください

ACBの実例II：マルチLLMホスティング向けオーバーサブスクリプションによるオンプレミス展開のTOC削減



Model Selection:

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

Chat AI インターフェースで、特定分野向けモデルの選択を可能にする (例: Global Taxes, M&A contracts)

vLLM との連携で効果的な推論サービング

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

ACB

GPU GPU
GPU GPU

Only Single Server!

LLM 5: Takane IP

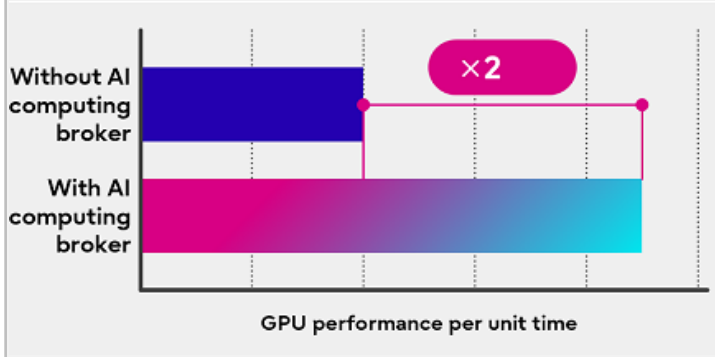
LLM 6: Takane-Employment

LLM 7: Takane-Compliance

LLM 8: Takane-Privacy-GDPR

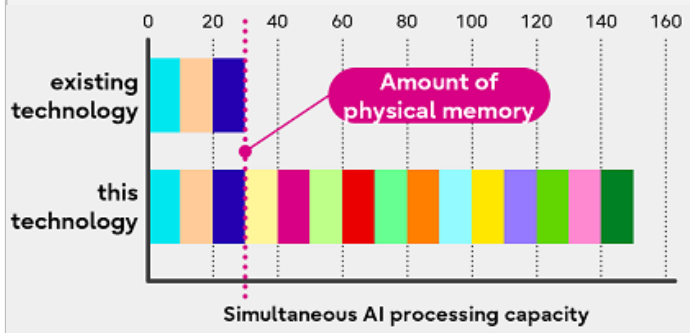
金融・クラウド・IT分野でACBの価値を提供

Japanese FinTech



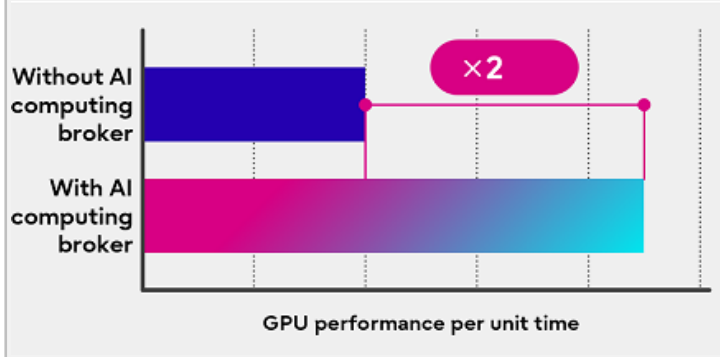
- **ユースケース：**外国為替リスク予測のためのAIモデルプロトタイピング
- **主な成果：**スループットが2倍に向上

Japanese cloud computing



- **ユースケース：**単一のGPUノード上で複数のAIモデルをホスト
- **主な成果：**物理GPUメモリ容量の5倍を管理

Japanese AI Tech



- **ユースケース：**IaaSビジネス向けの物体認識モデルのトレーニング
- **主な成果：**スループットが2倍に向上

お客様事例: TRADOM

GPUリソースを最適化し、より多くのAI学習を実行可能に

AI Computing Broker の効果

+25%

コード変更なしで
AIモデル実行時の
GPU利用率を向上

x 2

モデル実行時の
スループット
向上

"ACBは、AIモデルを生成するうえで必要なGPU計算資源を手軽に効率化でき、AI学習処理を多重化することでより精度の高いモデルを短期間で開発できる可能性をもっている"

-トレーダム株式会社執行役員
チーフデータサイエンスオフィサー
栢本淳一

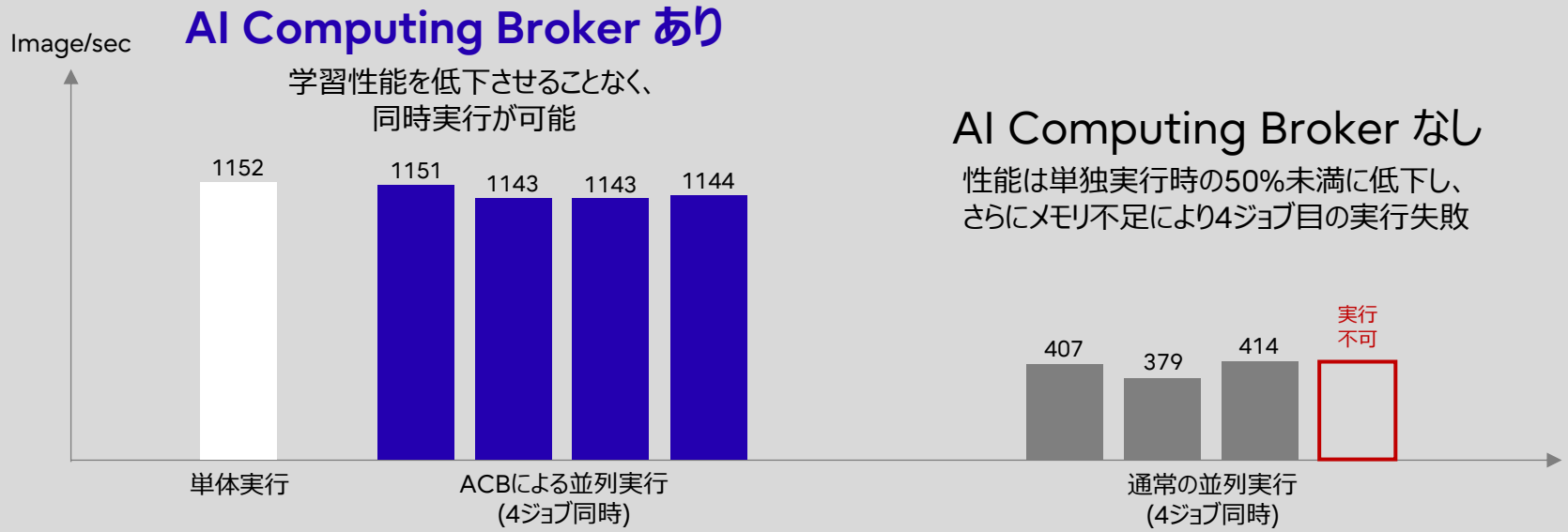
About TRADOM

AIを活用し、為替リスク管理のお悩みをトータルで解決するクラウドサービスを提供

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

お客様事例: Xtreme-D

GPUを利用するAIワークロードの効率とスループットが向上。計算コスト削減に貢献



- [検証アプローチ] Xtreme-D社独自のPyTorchベンチマークテストで実証。GPUに計算負荷をかけ、CUDA利用率・メモリ量・FP16/FP32性能および学習速度を測定
- [測定環境] CPU: AMD EPYC 7743 x 2・メモリ: 1TB、GPU: NVIDIA V100 (16GB) x 、オペレーティングシステム: Ubuntu 24.04

About Xtreme-D

マルチクラウド対応の高速AIクラウドサービス Raplase (Ra+) を提供

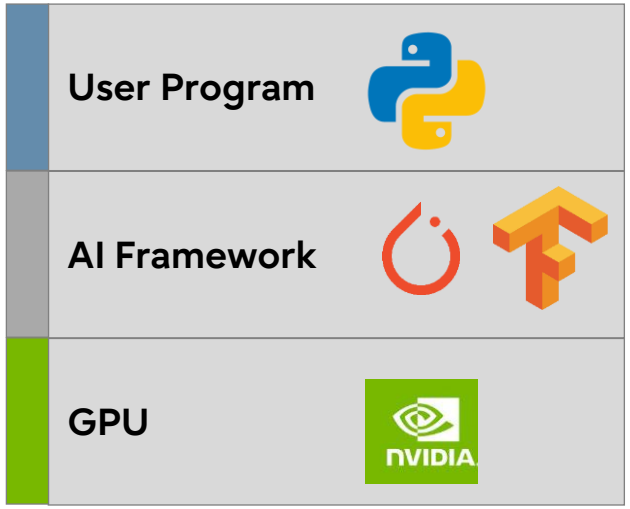
Xtreme-D Inc.: https://xtreme-d.net/

最小限の設定変更で迅速な導入を実現

FUJITSU

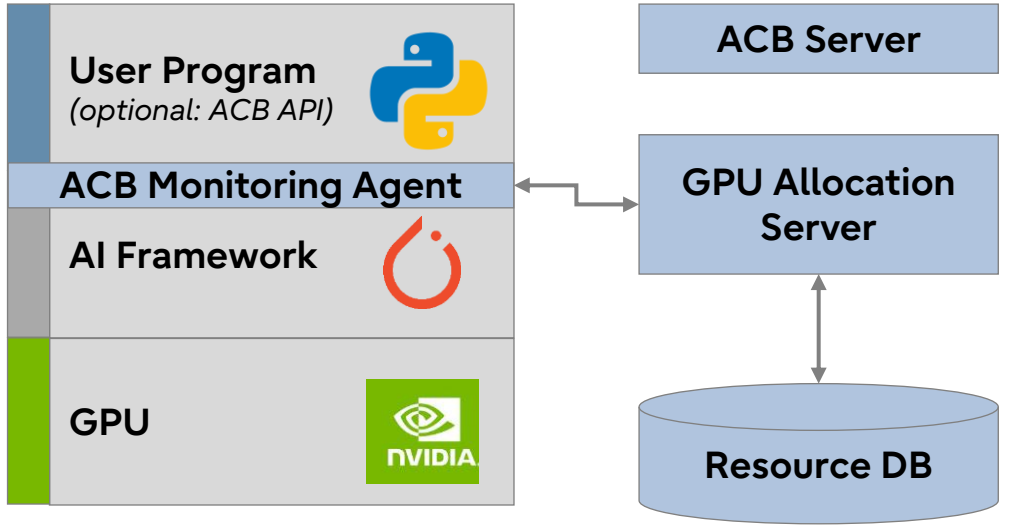
従来のAI実行

プログラムはAIフレームワークを通じてGPUを直接呼び出し。



ACBを利用したAI実行

ACBはAIフレームワークへの関数呼び出しを監視し、動的なリソース割り当てのためにGPU使用状況を追跡。



TensorFlow support limited
AMD ROCm support on the roadmap

最小限の変更での統合

- **フレームワーク互換性:**
  既存のAIアプリケーション（PyTorch、TensorFlow）とコード修正なしで連携
- **迅速なインストール:**
  オンプレミス、クラウド、またはハイブリッド環境への迅速な配置・展開
- **ワークフロー統合:**
  既存のAIワークフローとのシームレスな統合
- **ACB PyTorch Auto Client:**
  ユーザープログラムのコード変更は不要で、カーネル呼び出しを自動的に監視

ACBの強み：CPUとGPUをまたぐタスクレベル最適化

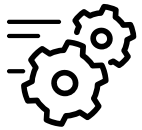
Feature	ACB	Run: ai	Exostellar ai	Slurm
Runtime-aware GPU Allocation	🟢	🔴	🔴	🔴
Memory partitioning	🟢	🟢	🟢	🟡
Memory oversubscription	🟢	🟡	🟢	🔴
Cluster-level orchestration	🔴	🟢	🟡	🟢
Plug-and-play integration	🟢	🔴	🔴	🔴
Focus	GPU Eff.	GPU Mgmt.	GPU Mgmt.	Job Sched.

GPUの可能性を最大限に引き出すパートナーシップ

FUJITSU

1. 主なポイント:
   - ランタイム対応GPU割り当て
   - フルメモリアクセス
   - 高度なスケジューリング

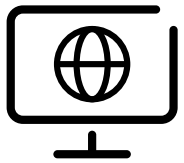
2. 詳細は、Fujitsu AI Compute Broker ウェブサイトやテクニカルホワイトペーパーをご覧ください



3. 30日間無料トライアルも受付中

4. ご質問は? お問合せフォーム または メール で受付いたします

AI Computing Broker Walking Deck



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

Appendix and Backup Slides

エンタープライズAIスタック全体での幅広い互換性を実現

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

補完技術①

技術	レイヤー	定義	どのようにACBを補完するか
NVIDIA MIG	Driver	単一の物理GPUを、専用の計算リソースとメモリリソースを持つ、分離されたGPUインスタンスに分割します。	これにより、ACBはワークロードの分離のために、事前に分割されたリソースを動的に割り当てることができます。
NVIDIA MPS	Driver	コンテキストスイッチを回避することで、複数のCUDAプロセスがGPUを同時に共有できるようにします。	ACBは、これらのプロセスをより効果的に監視およびスケジュールして、スループットを向上させることができます。
Run:ai Time-Slicing	Driver	ハードウェアパーティショニングなしで、ジョブ間でGPUアクセスを時間的に切り替えます。	ACBは、どのジョブをアクティブにするかを最適化して、アイドル時間を削減し、効率を高めます。
vLLM	Application	PagedAttentionを用いて高速なスループットを実現する、大規模言語モデル向けの最適化された推論エンジン。	ACBは、実行時の需要に基づいてvLLMジョブにGPUリソースを割り当て、応答性を向上させます。
Triton Server	Application	複数のフレームワークに対応し、同時実行される推論ジョブを効率的に処理するモデルサーバー。	ACBは、TritonがアクティブなGPUリソースにアクセスできるようにし、競合やアイドル状態を回避します（現在ACBではサポートされていません）。
Alluxio	Data Layer	複数のストレージバックエンドにわたるデータアクセスをキャッシュし、統合するデータオーケストレーションシステム。	I/Oを高速化することで、Alluxioはデータのボトルネックを取り除き、ACBによって管理されるGPUが常に稼働状態を維持できるようにします。
Slurm	Middleware	従来のHPCジョブスケジューラは、クラスタ内の計算リソースを割り当てます。	ACBは、Slurmと連携して、同じノード上でスケジュールされたジョブ内で、よりきめ細かいGPU割り当てを処理できます。
Kubernetes (with GPU Operator)	Middleware	コンテナをオーケストレーションし、プラグインを使用してGPUプロビジョニングを管理できます。	ACBは、割り当てられたPod内でGPU使用率を動的に最適化することにより、K8sを強化します。

補完技術②

- **富士通ACB（ミドルウェア）+ NVIDIA MIG（ドライバレベル）：**MIGはGPUを複数のインスタンスに分割してワークロードを分離しますが、ACBはリアルタイムの需要に基づいてこれらのインスタンスを異なるタスクに動的に割り当て、GPU全体の利用率を最大化します。

- **富士通ACB + Triton Inference Server/vLLM（アプリケーション層）：**ACBはGPUリソースをTriton/vLLMに動的に管理および割り当て、推論ワークロードが必要なときに必要なリソースを受け取れるようにし、効率的なモデルサービングを実現します。

- **富士通ACB + NVIDIA MIG + vLLM：**これら3つを組み合わせることで、GPUが分離のためにパーティション分割され（MIG）、リソースが需要に基づいて動的に割り当てられ（ACB）、モデルが効率的に提供される（vLLM）堅牢なシステムが実現します。