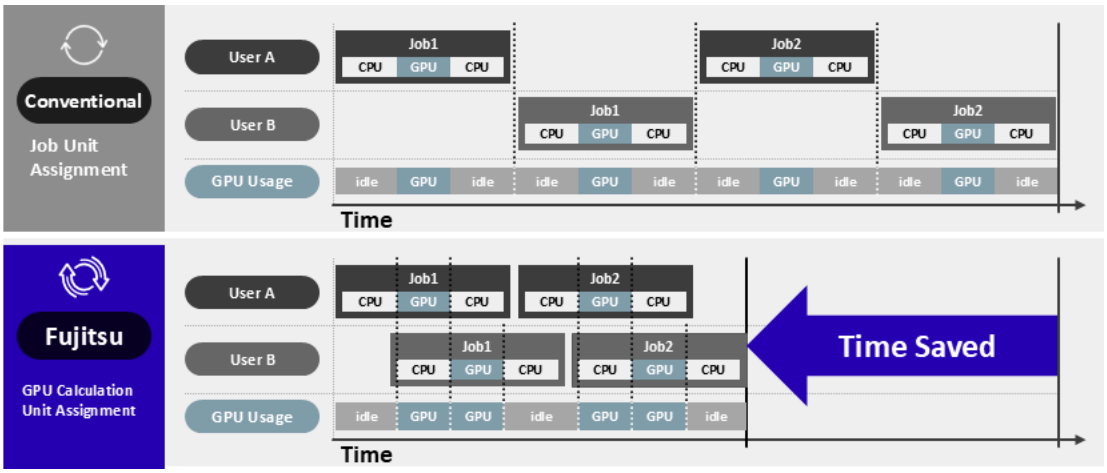
AI Computing Broker

ACB は GPU の稼働率を高め、コストを抑えながら、パフォーマンスを向上させます

ACB website

1. 富士通 AI Computing Broker (ACB) とは？

富士通 ACB は、GPU のリソース配分を最適化し、GPU メモリの物理的な上限を超えた AI ワークロードの要求に対しても適切に管理するインテリジェントなミドルウェアです。ACB は、既存のスケジューラーを補完することで、GPU 不足や利用率の低さといった課題を解決します。このソリューションを導入することで、追加のハードウェアを導入することなく、より大規模な AI ワークロードの実行が可能となります。また、既存の環境にシームレスに統合できるため、中断することなく GPU 利用効率を向上させ、迅速な結果を導き出すことができます。単独のソリューションとしても柔軟に稼働させることが可能です。



2. お客様の課題

- **GPU 資源の浪費とコスト増大:** GPU の利用率が低く、無駄なコストが発生している。
- **メモリ制限による AI 開発の制約:** GPU メモリの制限が、大規模モデルの利用や学習の妨げとなっている。
- **スケジューリングの競合による重要ジョブの遅延**: 非効率なスケジューリングがリソース競合を引き起こし、重要な AI ジョブの遅延と GPU 利用率の低下を招いている。
- **デプロイの遅延と設定の複雑化**: コード変更や設定の複雑さが、AI 開発の迅速性を損なっている。

3. 主な機能

- **Runtime-aware GPU allocation: AI フレームワークの稼働状況を把握し、必要な GPU を割り当てます。**
- **Full Memory Access: 稼働中のプログラムは、GPU メモリ全体にアクセスできます。**
- **Advanced Scheduling: バックフィルなどの技術を活用し、ジョブ配置の最適化と全体的な利用率を最大化します。**
- **Fast Deployment: ユーザープログラムのコード変更なしで統合可能。**

4. 事例

- 複数の AlphaFold2 ジョブ間で GPU リソースを動的に共有し、スループットを維持しながら GPU 使用効率を 45%向上。
- 単一の GPU サーバー上で複数の LLM を管理し、メモリ使用量を最適化し、遅延なく同時モデルサービングを可能に。

5. 更に詳しく知りたい方へ

- **まずは、ACB 技術紹介ページ(ACB website)をご確認いただき、その後、30 日間無料トライアルにご参加ください！**
  https://documents.research.global.fujitsu.com/ai-computing-broker/