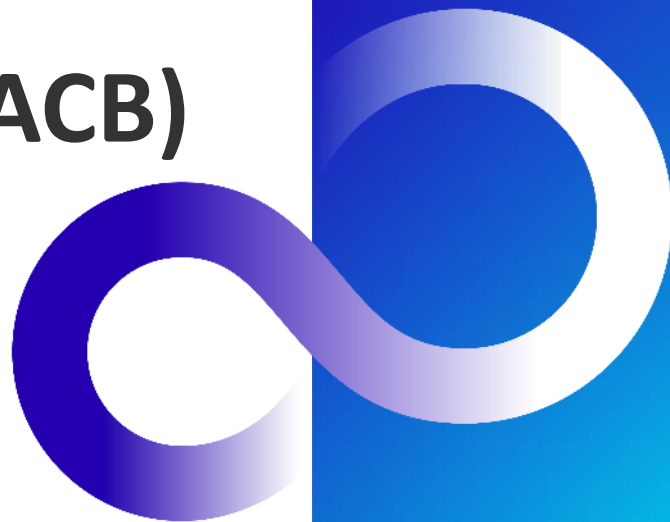


GPU不足を緩和する技術

AI Computing Broker (ACB)

富士通株式会社
コンピューティング研究所
CWBプロジェクト

吉田 英司



Computing Workload Broker

FUJITSU

- 量子やHPCといった特徴のある計算機の利用には専門家が必用
- 先端の計算資源を最適に組み合わせ、様々な分野の開発者が簡単に利用できるように



写真提供: 理化学研究所

AI Computing Broker (CPU/GPUハイブリッド計算)

ワークロードの特性を自動的に分析し、
高価なGPUを最適に配分、AIを高速実行



急速なGPU需要の拡大

生成AIや深層学習などの急速な普及を背景としたGPU需要の高まり

課題

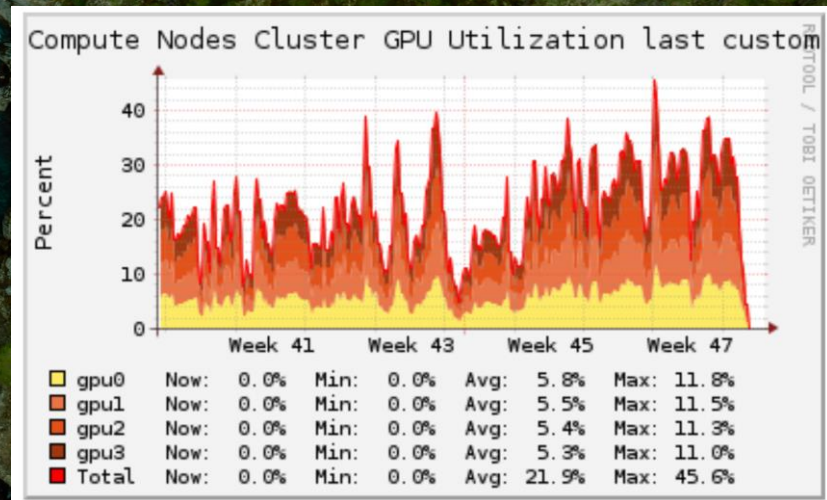
GPUの
確保

GPUの調達・
利用コスト増

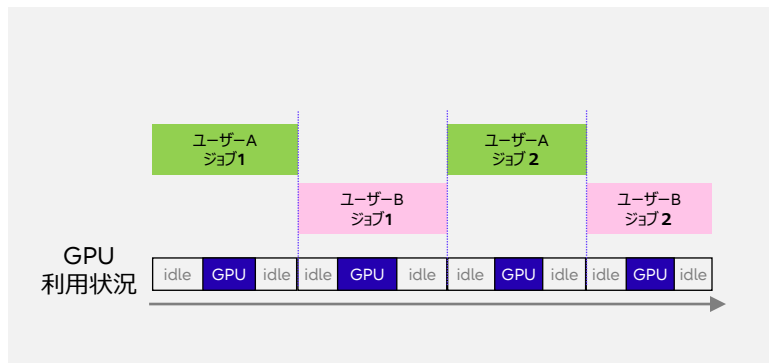
高効率なGPU利用への期待

多くの環境でGPU使用率は高くないことが多い
種々の効率化方策をとっているTSUBAMEでも30%以下

TSUBAME webから
<https://www.t3.gsic.titech.ac.jp/monitoring>



GPUの稼働率向上



GPUをジョブ単位で占有するため、他のジョブはGPUがidle状態でも占有解除されるまでGPUが利用できない

利用できるGPUメモリ量の最大化

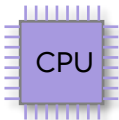


ハード的な機能でGPUを共有する場合、各ジョブが利用できるメモリ容量は小さくなる

世界的なGPU需要に対応する高効率なGPU利用技術の実現

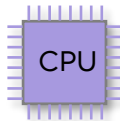
AI job

Loading data

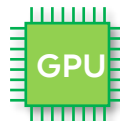


Pre-process

(Frame split, color correction, noise removal, etc.)

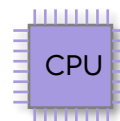
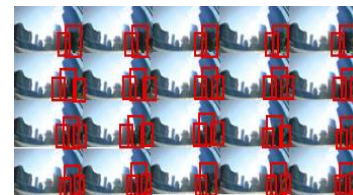


Inference



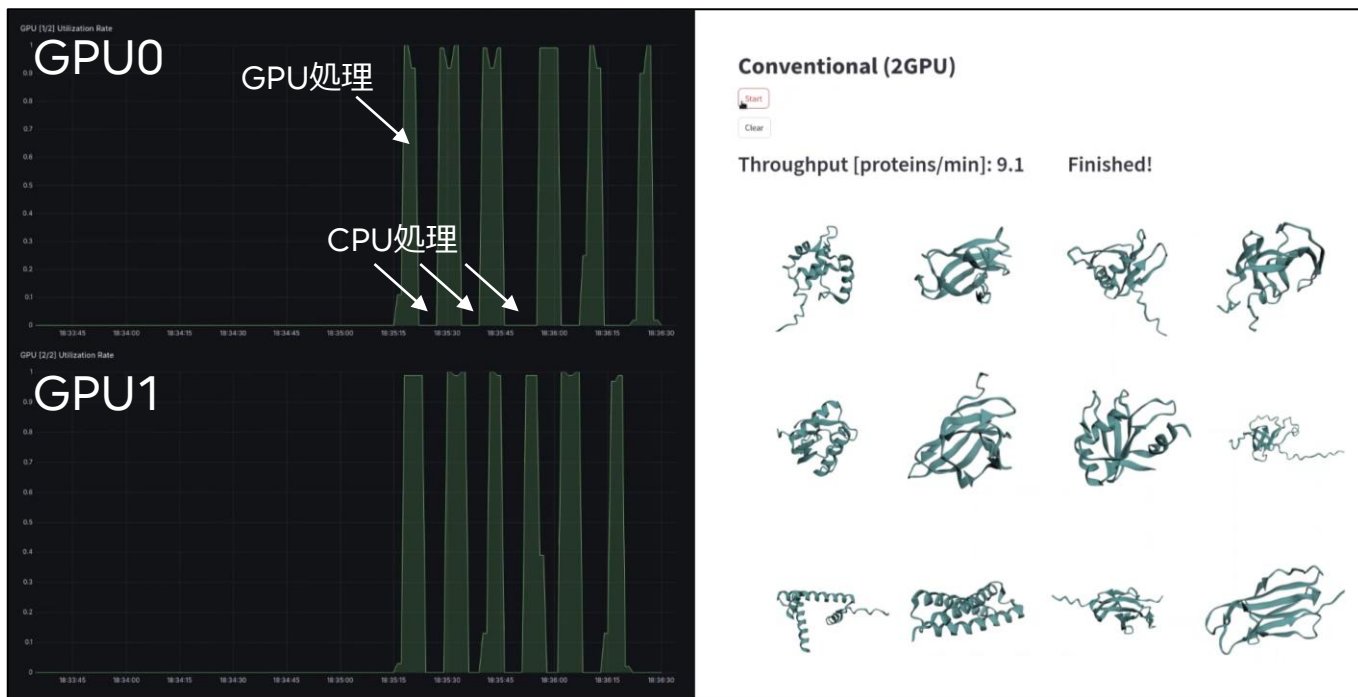
Post-process

(Analysis of movement and meaning, etc.)



参考: 低いGPU利用率の一因

- AlphaFold2の例 : GPU使用率30%程度
 - DB検索等の前処理(CPU処理)のため



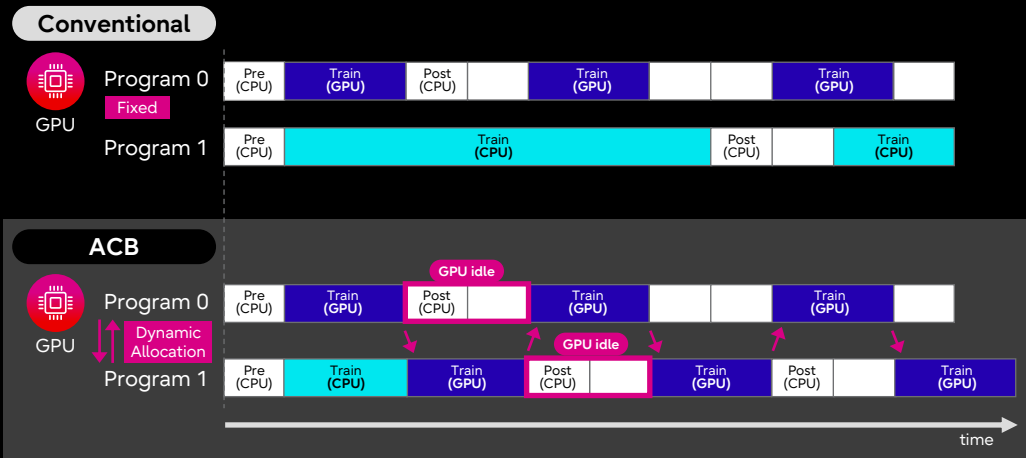
Overview of AI Computing Broker

Features of AI Computing Broker

- GPU scheduling specialized for deep learning
- Control of GPU allocation at the processing unit level within the program
- Improves GPU utilization rate, reducing overall processing time
- More processes can be executed with the same GPU resources
- Unlike sharing of GPUs through virtualization technology, the program can fully utilize the GPU-equipped memory

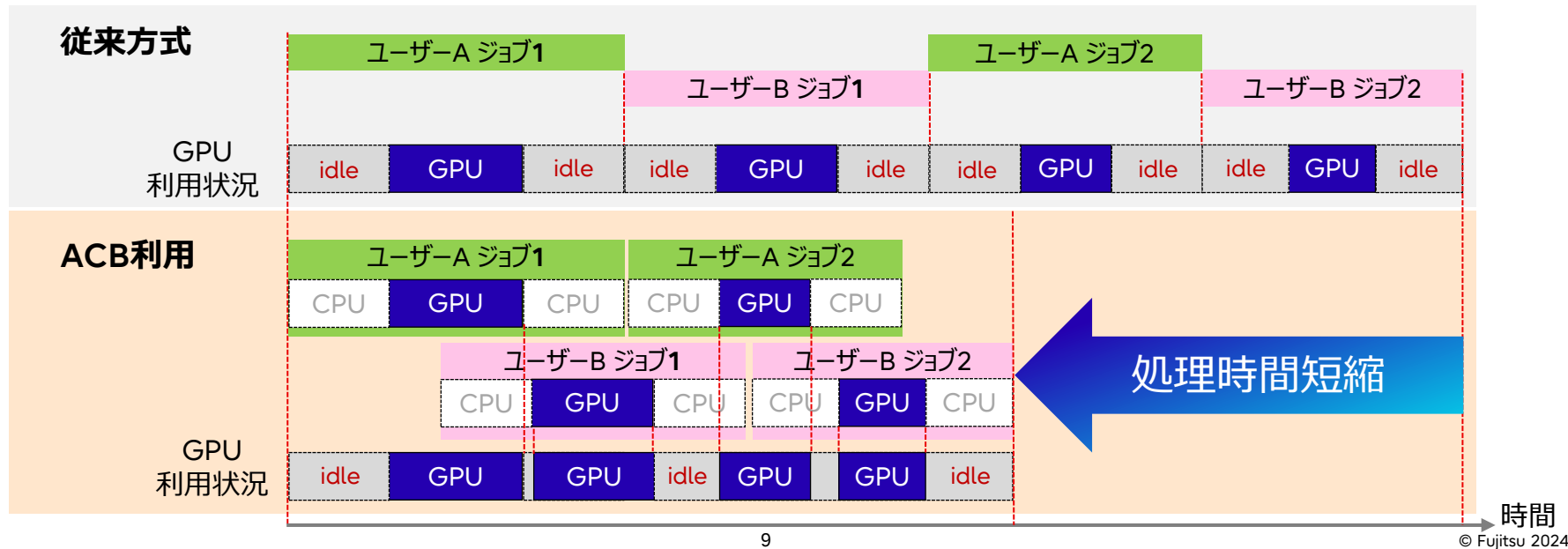
Technology of AI Computing Broker

Analyze AI processing content, dynamically allocate GPU to processes that require GPU



高効率なGPU利用で複数ジョブの同時実行時間を短縮

- 計算の中身を分析し、実際にGPUを利用する処理を検出
- GPUを利用する処理に着目し、GPUの動的割り当てを制御



ジョブ数に依存せずGPUの最大メモリ容量を利用可能

- 仮想化技術によるGPU共有と異なり、アプリ動作と連動してソフトでGPUを共有
- そのため、ジョブ数に依らずGPUの最大メモリを利用可能

従来方式：ハードレベルでのGPU共有

GPUメモリはジョブ数で分割
各ジョブは**小さいモデル**しか計算できない

GPUカード

GPUメモリ (H100:80GB)

ジョブA

ジョブB

ジョブC

GPU演算器

ジョブA/B/Cで共有

ACB利用：アプリ動作を見ながらGPUを共有

GPUメモリは各ジョブが占有
ジョブ数に関係なく**最大のモデル**を計算できる

GPUカード

GPUメモリ

ジョブA

GPU演算器

ジョブA

メモリ入替

GPUカード

GPUメモリ

ジョブB

GPU演算器

ジョブB

メモリ入替

GPUカード

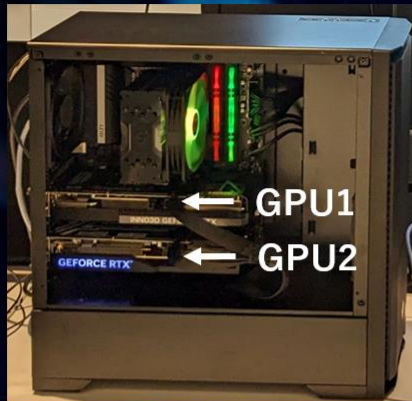
GPUメモリ

ジョブC

GPU演算器

ジョブC

time →



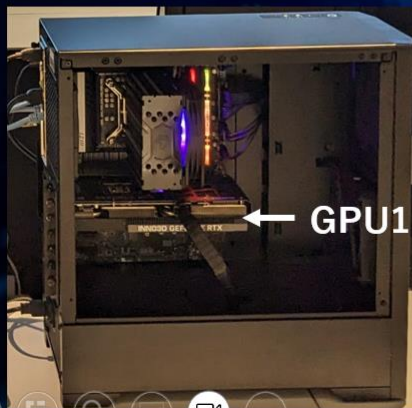
streamlit-m - Streamlit

localhost:8502

Conventional 2GPU AlphaFold2 Demo Dashboards

2GPU それぞれのプログラムでGPUを利用

20.33 (fps) 20.35 (fps)



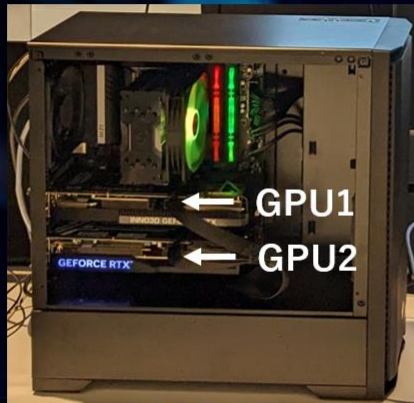
streamlit-m - Streamlit

localhost:8501

Conventional 1GPU AI Computing Broke... AlphaFold2 Demo Dashboards

1GPU GPUを利用 GPUが使えずCPU処理のため遅い

21.11 (fps) 7.52 (fps)



2GPU

20.95 (fps)

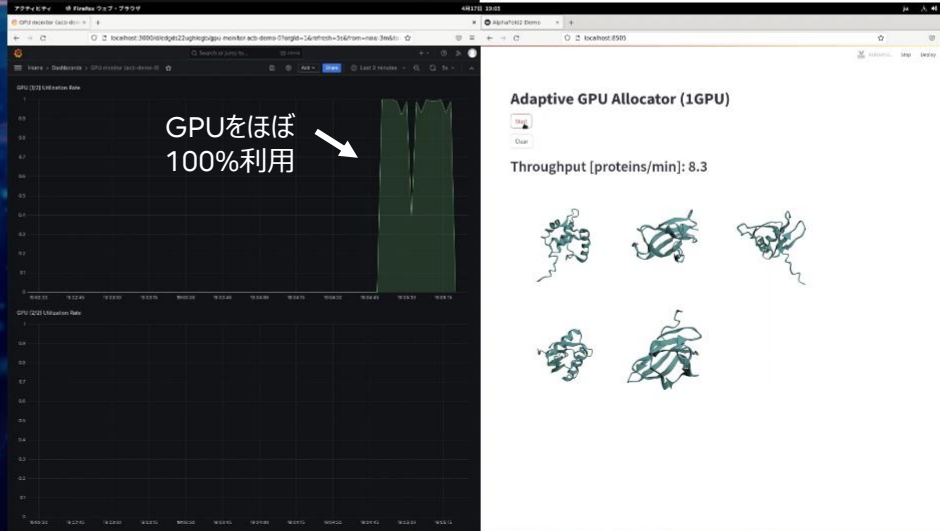
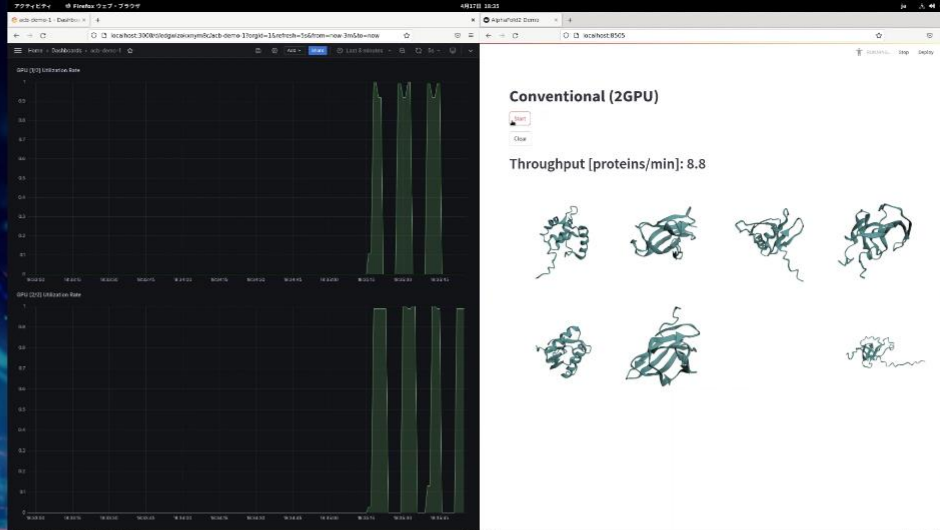
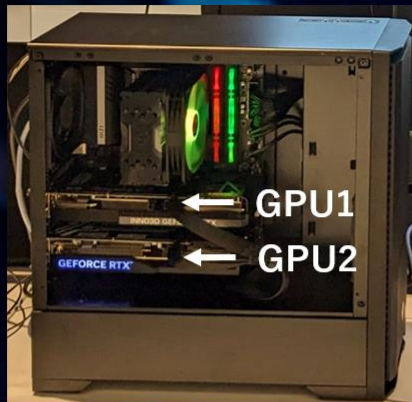
20.91 (fps)

1GPU + AI Computing Broker

20.33 (fps) ↙

20.00 (fps) ↙

1枚のGPUを効率的に利用



- **AI Computing Broker** : GPUを最大限に有効活用
 - AIジョブ間でGPUを効果的に切り替え、GPU処理の待ち時間を低減する技術
 - GPUの利用率が向上し、全体的な処理スループットが向上
- **導入のメリット** : GPU導入・利用コスト削減、処理スループット向上
 - 同じ時間内に同じ量のジョブを実行するのに、より少ないGPUで実行可能
 - 同じ数のGPUを使用して同じ時間内に、より多くの処理が実行可能

Thank you

