

Air Computing Pool のタスク実行時間予測に 寄与するGPU 特性抽出の検討

源内 裕貴, 川口 峻平, 大下 裕一, 下西 英之
大阪大学

背景

GPUの需要増加により、クラウドサービスが広く普及

- AWS, Microsoft Azure, Google Cloud

研究機関同士でサーバーをシェアする

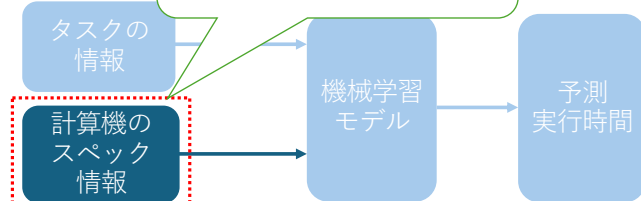
- ・低コスト
- ・GPU使用率の向上

"Air Computing Pool"の提案

- 共有するサーバーの種類が豊富
- ・未知のサーバーに対する実行時間の予測が必要

目的

GPUの特性抽出を行い、
未知のGPUに対しても
妥当な精度の予測を行う



課題

GPU性能を処理要素に分割せずに計測

→ 実行時間のボトルネックが不明

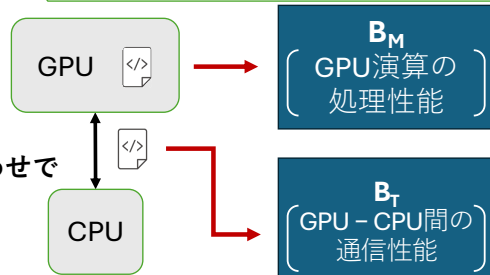
- ・GPU - CPU間の通信
- ・GPU演算の処理時間

アプリケーションによっては、
予測精度が下がる

提案手法

GPUを処理要素ごとに
テストベンチで計測

ベンチマークの組み合わせで
GPU性能をモデル化



テストベンチの事前評価

YOLOv8による動画の物体検出実行時間を線形予測モデルで
予測し、5種類の行列演算 (B_M) の比較評価を行った

線形モデルの実行時間予測式

$$T = B_T \times W_1 + B_M \times W_2 + W_3$$

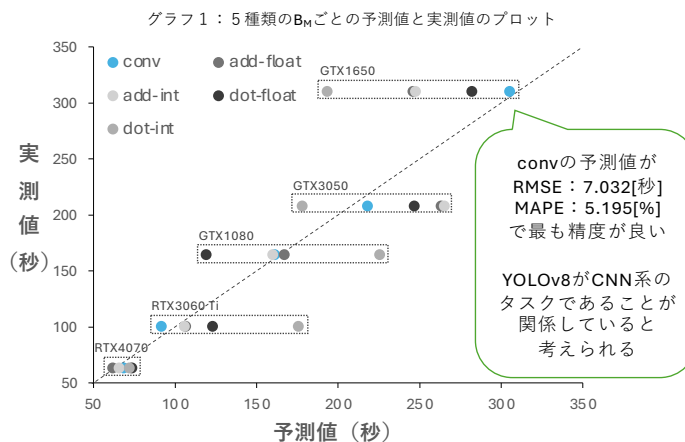
- ・ T : 予測実行時間 (秒)
- ・ B_T : 2つのfloat型 $10^4 \times 10^4$ 行列の転送時間 (秒)
- ・ B_M : 行列演算の実行時間 (秒)

GPUの種類による
特徴を吸収

タスクの種類による
特徴を吸収

B_M の行列演算の候補

- ・ conv : フィルタ 10×10 行列、対象 $10^4 \times 10^4$ 行列の畳み込み
- ・ dot-float(dot-int) : float型 $10^4 \times 10^4$ 行列の行列積 (int型)
- ・ add-float(add-int) : float型 $10^4 \times 10^4$ 行列の行列和 (int型)



未知のGPUに対する予測

B_M に conv を用いた際の線形予測モデルを
Leave-one-out cross validation法で評価した

- ・ 4種類のGPUで訓練、残りの1種類でテスト
- ・ 全てのGPUでテストを行うように繰り返す

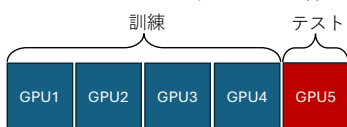
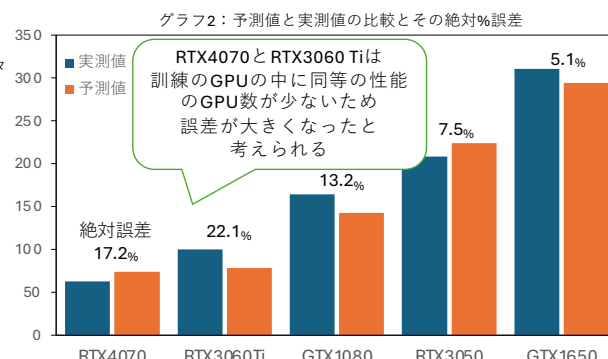


表1 : B_M が conv の場合の入力パラメータ

GPU	B_T	B_M
RTX4070	0.2	0.091
RTX3060 Ti	0.197	0.174
GTX1080	0.198	0.154
RTX3050	0.197	0.334
GTX1650	0.197	0.444



今後の研究方針

- ・ 機械学習モデルの入力に B_M , B_T を用いた際のGPUに対する汎用性の評価
- ・ CNNとは異なる特性を持つタスク (e.g. LLMの学習) に対してGPUの行列演算性能の測定に用いる処理の選定
- ・ GPU - CPU間の通信性能を個別で測ることのCPUに対する汎用性の評価