jh250083

“All Core Computing” によるプログラム最適化に関する研究

大島 聡史[1], 野村 哲弘[2], 南里 豪志[1], 遠藤 敏夫[2], 深谷 猛[3], 伊田 明弘[4], 河合 直聡[5]

*1:九州大学　*2:東京科学大学　*3:北海道大学　*4:海洋研究開発機構　*5:東北大学

概要

- 背景：最新のGPUは大量の計算コアを有しており、使い切れず性能効率が悪いプログラムも珍しくない。GPUスパコンではホストCPUも高性能化しており、GPUだけ使うのはもったいない。
- 目的：GPUに搭載された大量の計算コアをフル活用する技術とCPUとGPUを適切に用いる技術をまとめて **All Core Computing** と称し、実現のための技術開発とそれを用いたプログラム最適化を行う。

要素技術とその活用

マルチプロセスGPU実行/複数GPUカーネル同時実行

- 最新GPUの計算性能を引き出すには高い並列度が必要。しかし並列度が高くないプログラムも多い。複数プロセス・複数GPUカーネルの同時実行により十分な並列度を確保することが有効。

MPSとMIG（Multi-Process ServiceとMulti-Instance GPU）

- プロセスごとに利用可能なGPU資源を制限できる技術。複数GPUカーネル実行時は計算資源の奪い合いにより演算効率が下がることがあるが、MPSやMIGの活用により性能改善できることもある。プロセスごとの最低性能の保証やスループット向上も期待できる。
- MPSの方が制限が緩く、MIGの方が制限が厳しい。組み合わせて利用することも可能。

Green Context

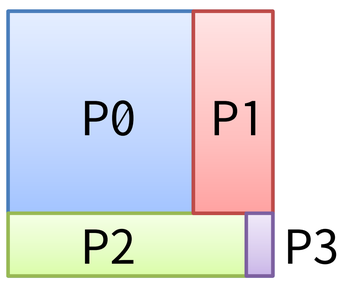
- GPU資源の制限はプロセスを起動する際に設定する必要がある。Green Contextをを使うと動的に制限値を変えることができる。（現状、CUDA Driver APIを使う必要がある。）

Grace Hopper

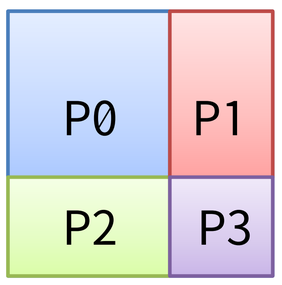
- CPUとGPUが1チップに封入。CPUとGPUが互いのメモリに高速にアクセス可能。従来のGPU計算環境よりもCPUとGPUによる共同作業が行いやすく、All Core Computingの観点でも期待大。

CPUとGPUのフル活用

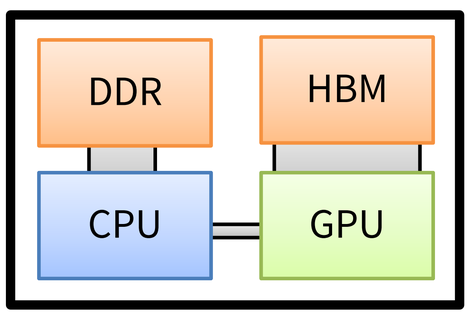
- なるべくGPUに計算を行わせCPUは介在しないことが従来のGPUプログラム最適化の基本戦略。高性能なホストCPUも適切に利用することでさらなる性能向上を目指す。プログラムの設計（記述方法など）とあわせて考える必要あり。



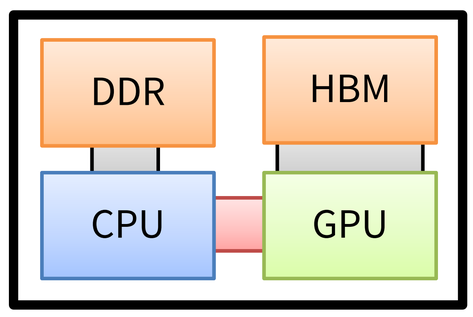
- 計算資源を奪い合うプロセス群



- プロセスごとに利用可能な計算資源を適切に制限することで性能の保証やスループットの向上が期待される



- 従来型GPU環境：CPU-GPU間のデータ転送が遅い。GPUの処理にCPUが介在すると性能低下しやすい。

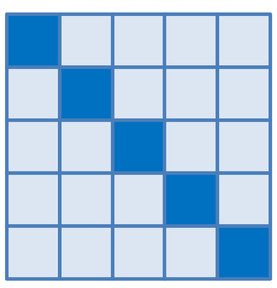


- Grace Hopper：CPU-GPU間のデータ転送が高速。互いのメモリに直接アクセス可能。GPUの処理にCPUが介在することでさらなる性能向上の可能性？

対象問題の例

ブロック低ランク（BLR）行列に対するQR分解

- ブロック化された行列に対する演算：個別の計算カーネルは小規模で並列度が低いが、全体としては十分な並列性がある
- バッチ計算が有効なケースもあるが、データ構造の再考が必要な場合は活用しにくい

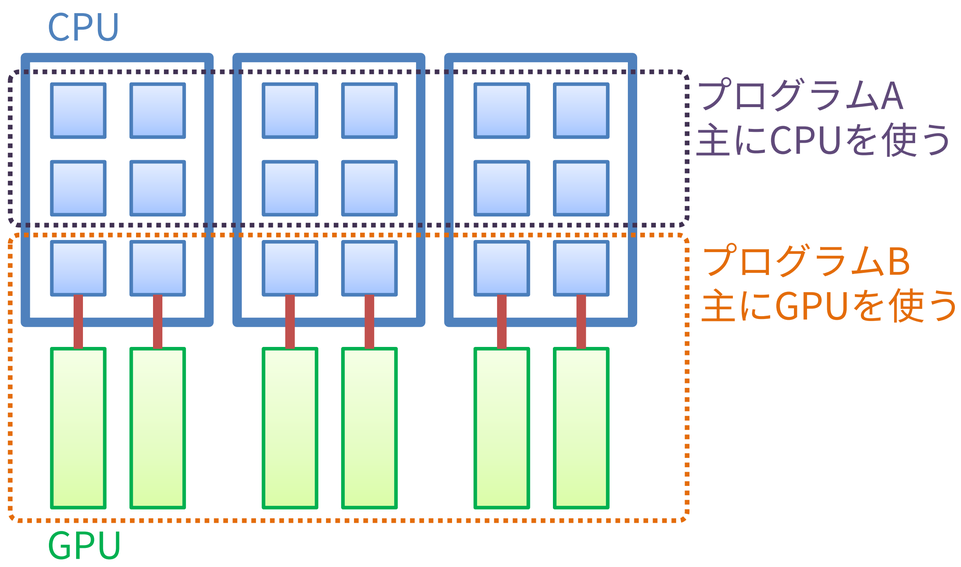


- BLR行列の例：濃いブロックは密行列、薄いブロックは低ランク近似行列
- 行列全体に対する並列計算は行わないが、列単位・ブロック単位・ブロック内の各計算に並列性があるため、マルチプロセスGPU実行/複数GPUカーネル同時実行とMPS/MIGが有用。
- GPUが得意としない計算も含まれており、All Core Computingの適用による高性能化の可能性があり実験を進めている。

連成計算

- 例えば主にCPUを使うプログラムと主にGPUを使うプログラムによる連成計算
  - シミュレーションと可視化 の連成
  - シミュレーションと機械学習 の連成

（ノード内・ノード間どちらもありえる）



- 連成フレームワークCoToCoAの活用をベースに検討中 https://github.com/tnanri/cotocoa

参考文献（関連する過去の発表）

- 大島聡史, 伊田明弘, 河合直聡, 横田理央, 山崎市太郎, "CUDA Fortran＋MIG＋UVMを用いたBLR行列QR分解の大規模高速化", 情報処理学会 研究報告(HPC-190) / SWoPP2023
- Satoshi Ohshima, Akihiro Ida, Rio Yokota, Ichitaro Yamazaki, "QR Factorization of Block Low-Rank Matrices on Multi-Instance GPU", PDCAT 2022, Springer LNCS 13798, 2023.04 発行, DOI:10.1007/978-3-031-29927-8_28