T2R2 東京工業大学リサーチリポジトリ
Tokyo Tech Research Repository

論文 / 著書情報
Article / Book Information

題目(和文)	現代的なGPUにおける計算性能とメモリバンド幅の乖離の克服
Title(English)	Overcoming the Gap Between Compute and Memory Bandwidth in Modern GPUs
著者(和文)	ZhangLingqi
Author(English)	Lingqi Zhang
出典(和文)	学位:博士(理学), 学位授与機関:東京工業大学, 報告番号:甲第12510号, 授与年月日:2023年9月22日, 学位の種別:課程博士, 審査員:遠藤 敏夫,増原 英彦,脇田 建,坂本 龍一,横田 理央,松岡 聡,Attia Mohamed Wahib Mohamed
Citation(English)	Degree:Doctor (Science), Conferring organization: Tokyo Institute of Technology, Report number:甲第12510号, Conferred date:2023/9/22, Degree Type:Course doctor, Examiner:,,,,,,
学位種別(和文)	博士論文
Category(English)	Doctoral Thesis
種別(和文)	審査の要旨
Type(English)	Exam Summary

Powered by T2R2 (Tokyo Institute Research Repository)

（博士課程）

論文審査の要旨及び審査員

報告番号	甲第　　号		学位申請者氏名		ZHANG LINGQI	
論文審査 審 査 員		氏　名	職　名		氏　名	職　名
	主査	遠藤 敏夫	教授	審査員	横田 理央	教授
	審査員	増原 英彦	教授		松岡 聡	特定教授
		脇田 建	准教授		Attia, Mohamed Wahib Mohamed	Team Leader
		坂本 龍一	准教授			

論文審査の要旨（2000 字程度）

本論文は「Overcoming the Gap Between Compute and Memory Bandwidth in Modern GPUs（現代的な GPU における計算性能とメモリバンド幅の乖離の克服）」と題し、計算機の性能不均衡への対応のために計算カーネルの再構成と過剰な並列性の低減に基づく戦略を提案し、高性能計算分野におけるメモリバウンドな計算カーネルに対してその効果を実証している。本論文は以下のような英文 7 章で構成されている。

第 1 章「Introduction」では、本研究の主対象である計算機の計算性能とメモリバンド幅のバランスに関して、メニーコア化によるハードウェア並列度の向上およびメモリにおける三次元積層技術やキャッシュの深層化を含み、これまでの進展と現在の状況を CPU と GPU アーキテクチャについて述べている。そしてアーキテクチャのバランスの変化がメモリバウンドな計算カーネルの継続的な性能向上を困難にするという課題を述べ、その課題に対する本研究の提案と貢献を示している。

第 2 章「Background」では、本論文を理解するうえで必要な知識として、Flops per Byte 指標やリトルの法則を含む並列性に関する理論を紹介している。そして本論文で主対象とする GPU アーキテクチャについて、そのコア階層とメモリ階層、およびスレッド間の複数の同期種類を中心に述べ、また反復型のメモリバウンドな計算カーネルとしてステンシル計算と共役勾配法について述べている。

第 3 章「Microbenchmarking Nvidia GPUs: A Decade of Evolution Trends」では、複数世代の GPU アーキテクチャについて、計算カーネルに影響が大きい要素の詳細なマイクロベンチマーク測定結果を示している。スレッド間同期やカーネル起動の時間コストおよび、カーネルの並列度と速度性能の関係について詳細調査を行い、本論文の基本戦略である、性能不均衡の問題に対する適切な並列度の調整の重要性を示している。

第 4 章「PERKS: A Locality-Optimized Execution Model for Iterative Memory-Bound GPU Applications」では、反復型計算カーネルの局所性を高めることによって最適化を行う実行モデルである PERKS の提案と評価を行っている。まず PERKS の基本戦略として、適切な並列度の選択と GPU レベルでの同期機構を利用することにより、複数の反復を単一の GPU カーネルに統合することを述べている。この戦略上でデータを適切なメモリ階層に配置する複数の手法とそれらの性能への影響を分析している。以上の最適化手法により、カーネル起動のコスト削減やキャッシュの効率利用を可能としている。さらに PERKS に対応するための計算カーネルの再構成手法について、ステンシル計算と共役勾配法を中心に示し、性能評価によりそれらの速度性能の向上を示している。

第 5 章「EBISU: Epoch Blocking for Iterative Stencils, with Ultracompact Parallelism」では、反復型ステンシル計算を対象とした最適化手法である EBISU の提案と評価を行っている。EBISU は、PERKS の統合カーネルと適切な並列度選択の戦略に基づき、その上で複数時間ステップの計算を空間ブロックについて連続で行う時間ブロッキング手法を適用する手法である。それらを実現するためのリングバッファなどの実装および、並列度を考慮した性能モデルについて詳細に述べている。性能評価においては種々のステンシル形状を用いて複数の代表的なステンシルフレームワークとの比較を行い、速度性能の優位性を示している。

第 6 章「Discussion and Future Work」では、本研究の将来に向けた発展性について論じている。まず PERKS と EBISU の汎用化の方向性について述べ、それに向けたポインタマッピングの手法やレジスタのキャッシュ利用に必要なハードウェア機構などについて議論している。

第 7 章「Conclusion」では、本研究の総括を述べている。

以上のように、本研究は GPU アーキテクチャにおける反復型メモリバウンド計算カーネルの高速化のために、適切な並列度の選択とメモリ階層活用の効率化を行う戦略に基づく最適化技法を提案し、またその有用性を実証しており、理学上貢献するところ大である。よって本論文は博士（理学）の学位論文として十分価値があるものと認める。

注意：「論文審査の要旨及び審査員」は、東工大リサーチリポジトリ(T2R2)にてインターネット公表されますので、公表可能な範囲の内容で作成してください。