論文 / 著書情報
Article / Book Information

題目(和文)	GPUクラスタにおけるメモリ階層を考慮した時間依存性があるアルゴリズムのための最適化手法
Title(English)	Optimization Methods for Efficient Utilization of Memory Hierarchy for Algorithms with Temporal Dependences on GPU Clusters
著者(和文)	金光浩
Author(English)	GUANGHAO JIN
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9423号, 授与年月日:2014年3月26日, 学位の種別:課程博士, 審査員:松岡　聡,青木　尊之,遠藤　敏夫,増原　英彦,渡辺　治
Citation(English)	Degree:Doctor (Science), Conferring organization: Tokyo Institute of Technology, Report number:甲第9423号, Conferred date:2014/3/26, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Category(English)	Doctoral Thesis
種別(和文)	論文要旨
Type(English)	Summary

（博士課程）
Doctoral Program

論 文 要 旨

THESIS SUMMARY

専攻： Department of	数理・計算科学　専攻	申請学位（専攻分野）： Academic Degree Requested	博士 Doctor of Science　（理学）
学生氏名： Student's Name	金光浩	指導教員（主）： Academic Advisor(main)	教授　松岡　聡
		指導教員（副）： Academic Advisor(sub)	

要旨（和文 2000 字程度）
Thesis Summary（approx.2000 Japanese Characters ）

本論文は「Optimization Methods for Efficient Utilization of Memory Hierarchy for Algorithms with Temporal Dependences on GPU Clusters」と題し、GPU クラスタなどの深いメモリ階層を持つ計算機システムにおいて、高性能性と大規模性を両立する演算を可能とする技術を提案するもので、英語で全７章から構成されている。

第１章「Introduction」では、まず本論文の研究背景および高性能性と大規模性の両立を困難とする原因が述べられ、本研究の貢献について述べられている。さらに本論文の構成が述べられている。

第２章「Background」では、本研究の背景として、広範な科学分野のシミュレーションにおいて重要な演算カーネルの一つであるステンシル計算およびGPUのアーキテクチャとCUDAプログラミングモデル、およびその上でのステンシル計算の典型的な実行方法が紹介されている。また本論文の性能評価に使われる、GPUを採用したスーパーコンピュータであるTSUBAME2.0およびTSUBAME2.5について述べられている。さらに高性能性と大規模性の双方が重要となる実アプリケーションの存在にも関わらず、深いメモリ階層を持つ計算機システム上で、それらの両立が困難な理由が議論されている。

第３章「Related work」では、通信量削減手法を中心とした関連研究について述べられている。ステンシル計算においてはテンポラルブロッキング手法が知られており、種々のメモリ階層に適用されているが、その計算コスト、データ移動コスト等において改善の余地が大きいことが議論されている。

第４章「Optimization methods for stencil computation on single node」では、単一ノード環境を対象として、GPUメモリ容量を超えるような大規模ステンシル計算の最適化手法について提案されている。テンポラルブロッキングをホスト・デバイス間通信の削減に用いる手法を基本としつつ、そのオーバヘッドを削減する手法を提案している。提案には、計算結果の再利用のためのバッファ効率化・カーネル内テンポラルブロッキングとの統合・配列データバッファのサイズ削減手法を含む。TSUBAME2.0/ TSUBAME2.5上での評価実験を行うことにより、単純な手法に比べ20倍以上の高速化を実現しており、提案手法の有効性が実証されている。

第５章「Optimization methods for stencil computation on multiple nodes」では、第4章に述べた手法の複数ノード環境への拡張について論じている。テンポラルブロッキングの挙動を妨げず、かつノード間通信を効率的に保つデータ分割手法が提案されている。この分割手法の適用により、GPUメモリ容量と利用GPU数の積よりも大規模で効率的なステンシル計算を可能としており、TSUBAME2.0/TSUBAME2.5の最大256GPUを用いた評価実験により、大規模性と良好なスケーラビリティの両立がなされていることが実証されている。

第６章「Optimization methods for band SpMV on single node」では、第4章、第5章で提案したアプローチの、ステンシル計算以外への適用について議論されている。その試みの一つとして帯行列ベクトル積計算を取り上げ、GPUメモリ容量を超える問題サイズに対応可能なソフトウェアの実装と、TSUBAME2.5上のシングルGPUでの評価実験を通して、有効性が実証されている。

第７章「Conclusion and directions for the future」では本論文の成果を総括し、将来取り組むべき課題について述べられている。

以上のように、本論文は深いメモリ階層を持つ大規模計算機システムにおいて、高性能性と大規模性を良好に両立可能とする技術を提案し、その有効性を確認しており、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。

備考 ： 論文要旨は、和文 2000 字と英文 300 語を 1 部ずつ提出するか、もしくは英文 800 語を 1 部提出してください。
Note : Thesis Summary should be submitted in either a copy of 2000 Japanese Characters and 300 Words (English) or 1copy of 800 Words (English).

（博士課程）
Doctoral Program

論 文 要 旨

THESIS SUMMARY

専攻： Department of	数理・計算科学　専攻	申請学位（専攻分野）： Academic Degree Requested	博士 Doctor of Science　（理学）
学生氏名： Student's Name	金光浩	指導教員（主）： Academic Advisor(main)	教授　松岡　聡
		指導教員（副）： Academic Advisor(sub)	

要旨（英文 300 語程度）
Thesis Summary（approx.300 English Words ）

This dissertation work, named "Optimization Methods for Efficient Utilization of Memory Hierarchy for algorithms with Temporal dependences on GPU Clusters" describes communication-reducing optimization methods to allow computations on the domain that is bigger than the memory capacity of GPUs while maintaining high performance. It is organized in 7 chapters, outlined below.

Chapter 1: Introduction

In this chapter, we describe motivation of this work, formulate problem statement, list our contributions and give outline for the rest of the dissertation.

Chapter 2: Background

This chapter provides general information about application, GPU and CUDA program model. Then, we provide information about TSUBAME2.0 and TSUBAME2.5 supercomputers that are used to evaluate the optimization methods. We also explain the demand of computing bigger domain in real application and the memory limitation of GPU clusters by using common way. Then, we introduce our objective which is to efficiently use GPUs to compute bigger domain while maintain high performance.

Chapter 3: Related work

This chapter reviews related works that provides the possibility in reducing communication between host and device memory, improving the performance in kernel level.

Chapter 4: Optimization methods for stencil computation on single node

This chapter first analyzes the existing optimization methods and their limitations. Then, we introduce our optimization methods that can enable the computation on the bigger domain while maintaining higher performance than existing optimization methods in single node case.

Chapter 5: Optimization methods for stencil computation on multiple nodes

This chapter applies the optimization methods on multiple nodes. Furthermore, we proposed a new optimization method that efficiently uses the register of GPU to maintain higher performance.

Chapter 6: Optimization methods for band SpMV on single node

This chapter utilizes the optimization methods to band sparse matrix-vector case.

Chapter 7: Conclusion and directions for the future

Finally, we give some thoughts on the previous chapters and summarize contributions made by our work. We conclude the overall results of this thesis and discuss about the possible directions for the future works.

備考 ： 論文要旨は、和文 2000 字と英文 300 語を 1 部ずつ提出するか、もしくは英文 800 語を 1 部提出してください。
Note : Thesis Summary should be submitted in either a copy of 2000 Japanese Characters and 300 Words (English) or 1copy of 800 Words (English).