T2R2 東京科学大学 リサーチリポジトリ
Science Tokyo Research Repository

論文 / 著書情報
Article / Book Information

題目(和文)	GPU上の仮想シストリックアレイ
Title(English)	A Virtual Systolic Array on GPUs
著者(和文)	陳 鵬
Author(English)	Peng Chen
出典(和文)	学位:博士(学術), 学位授与機関:東京工業大学, 報告番号:甲第11530号, 授与年月日:2020年3月26日, 学位の種別:課程博士, 審査員:松岡 聡,遠藤 敏夫,額田 彰,脇田 建,横田 理央
Citation(English)	Degree:Doctor (Academic), Conferring organization: Tokyo Institute of Technology, Report number:甲第11530号, Conferred date:2020/3/26, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Category(English)	Doctoral Thesis
種別(和文)	審査の要旨
Type(English)	Exam Summary

Powered by T2R2 (Science Tokyo 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	陳　鵬	
論文審査 審 査 員		氏　名	職　名	氏　名	職　名
	主査	松岡　聡	特任教授	横田　理央	准教授
	審査員	遠藤　敏夫	教授		
		額田　彰	准教授		
		脇田　建	准教授		

論文審査の要旨（2000 字程度）

本論文は「A Virtual Systolic Array on GPUs（GPU 上の仮想シストリックアレイ)」と題し、高性能と高いエネルギー効率を達成できる演算器であるシストリックアレイをCUDAアーキテクチャ上にソフトウェアにより実現し、メモリバウンドな問題の演算を最適化することを目的としている。ソフトウェアにより GPU 上に仮想シストリックアレイを構築し、それを生かして各種メモリバウンドな問題を高速化する為の手法を提案する論文であり、英文で全９章から構成されている。

第１章「Introduction」では、本論文の導入として、本研究で対象とする、GPU 上でメモリバウンドな問題を解く際の現状の課題、それに対する仮想シストリックアレイによる解決策の提案と、それを実現するための技術的課題を挙げている。最後に本論文の構成を述べている。

第２章「Background」では、本論文の内容を理解するために必要な研究背景について述べている。具体的には、先ず GPU の歴史と NVIDIA 社 GPU の CUDA アーキテクチャを紹介している。続いて、今まで深く研究されているシストリックアレイの構成と代表的なシストリックアレイのアプリケーションについて説明している。その後 GPU のレジスターをキャッシュとして使う動機を説明している。最後に、課題としてよく取り上げられるメモリバウンドな問題の例と､GPU 上で仮想的なシストリックアレイを実現してこの問題を解決する可能性について解説している。

第３章「Software Systolic Array Execution Model」では、CUDA アーキテクチャの特性を利用し、 Warp とレジスターにより仮想シストリックアレイの構築を提案している。具体的には、レジスターはシストリックアレイの処理要素の機能を模倣し､CUDA の Warp 内で shuffle 命令を利用してシストリックアレイのデータ通信メカニズムを実現している。

第４章「Two-dimensional Convolution in SSAM」では、画像処理と深層学習の代表例として、二次元コンボリューションのアルゴリズムを仮想シストリックアレイを用いて実現し、その性能評価を行なっている。具体的には、二次元コンボリューションアルゴリズムを GPU により高速化する動機を紹介し､CUDA の Warp 内でレジスターキャッシュを構成する方法とレジスターを用いた部分和の並列計算とそのデータの移動を提案している。その上で、性能モデルを構築し、マイクロベンチマークにより取得した GPU のパフォーマンスカウンタ値を用いて、提案手法の理論的な有効性を証明している。評価では、最新の Tesla GPU により、複数の高速化されたライブラリ（ArrayFire、NPP 等）の比較を行い、提案手法の有効性を実証している。

第５章「Stencils in SSAM」では、科学計算の代表例としてステンシル計算を仮想シストリックアレイを用いて実現し、その性能評価を行なっている。具体的には、代表的な二次元と三次元ステンシル計算のベンチマーク（2d5point、3d7point、poisson 等）をブロッキングと仮想シストリックアレイを用いて実現している。ステンシル計算の最適化によく使われる時間ブロッキングと空間ブロッキング手法も仮想シストリックアレイで実現できることを示している。評価では、最新の Tesla GPU により、時間と空間ブロッキング手法を用いて高速化された複数ライブラリ（Halide、PPCG、StencilGen 等）との比較を行い、提案手法の有効性を実証している。

第６章「Summed Area Tables Primitive in SSAM」では、機械学習と画像処理の代表例として、Summed Area Tables（SAT）計算を仮想シストリックアレイを用いて実現し、その性能評価を行なっている。

具体的には、先ず SAT 計算アルゴリズムと SAT の活用について説明する。そして、既存の GPU 上での SAT 計算の高速手法を紹介し、それらの仮想シストリックアレイを用いた実現方法を提案している。さらには、仮想シストリックアレイによる既存手法の実現には、計算の依存関係の課題があるため、それを解決する、SAT 計算アルゴリズムである BRLT を提案している。性能モデルを用いることにより、BRLT アルゴリズムの既存手法に対しての並列計算性能の向上および算術演算の減少を確認している。評価では、最新の Tesla GPU により、複数の高速されたライブラリ(NPP, OpenCV)との比較を行い、提案した仮想シストリックアレイ手法、および BRLT アルゴリズムの有効性を実証している。

第７章「Related Work」では、本論文に関する重要な先行研究について述べている。具体的には、レジスターキャッシュ手法と部分和手法、それらの様々なアプリケーションでの活用と実績、GPU 性能モデルを含んでいる。

第８章「Discussion & Conclusion」では、GPU 上での仮想シストリックアレイを用いて最適化されたアルゴリズムの性能特性が、GPU の世代（Tesla P100,V100）により異なる原因を議論している。仮想シストリックアレイを用いたアルゴリズムの最適化手法を踏まえて、GPU アプリケーションの自動ソースコード生成と最適化ガイドラインをまとめている。

第９章「Future Work」では、本研究をより発展させる今後の方向性を示している。

以上のように、本研究では GPU 上で仮想シストリックアレイを活かした、メモリバウンドな問題の高速化を実現する為の具体的な手法について提案し、また実験によってその有効性を確認しており、学術的に貢献するところが非常に大きいことが評価される。よって、本論文は博士（学術）の学位論文として十分価値があるものと認められる。

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。