様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　元 年　 ６ 月 １０ 日現在

機関番号：１５４０１
研究種目：基盤研究(B)（一般）
研究期間：2014～2018
課題番号：２６２８０００２

研究課題名（和文）メモリマシンモデル上の並列計算理論の構築と次世代ＧＰＧＰＵアーキテクチャの提案

研究課題名（英文）Parallel Computation Theory for Memory Machine Models and Next Generation GPGPU Architecture

研究代表者
中野　浩嗣（Nakano, Koji）

広島大学・工学研究科・教授

研究者番号：３０２８１０７５

交付決定額（研究期間全体）：（直接経費）　　12,300,000円

研究成果の概要（和文）：GPUは画像を生成したり操作するために設計された集積回路である．最近のGPUは，従来CPUで処理していた汎用計算を行えるように設計されている．本研究の目的は，GPU処理のための適切な理論計算モデルを提案し，その理論モデルに基づいて効率良い並列アルゴリズムを開発し，性能評価を行うことである．本研究では，GPUのメモリアクセスに着目した理論モデルDMMモデル,UMMモデル，HMMモデルを提案した．このモデルに基づいて，GPU上の多くの効率良い並列アルゴリズムを開発した．特に，SKSSと呼ばれる動的計画法をGPUで処理する新しい手法を開発した．

研究成果の学術的意義や社会的意義
GPUの理論的モデルを提案することにより，理論研究者がGPU向けアルゴリズムを研究するためのベースを提供することができた．これまでは，並列アルゴリズムの理論研究者にとってGPU向け並列アルゴリズムの実装作業は困難であったが，これにより，GPU上での並列処理技法の研究が容易に行えるようになった．また，このモデルをベースに研究代表者らはGPUのいくつかの具体的なアルゴリズム手法，例えば，SKSS (Single Kernel Soft Synchronization)などを提案し，その有効性をGPUへの実装実験により実証することができた．

研究成果の概要（英文）：The GPU (Graphics Processing Unit) is a specialized circuit designed to accelerate computation for building and manipulating images. Latest GPUs are designed for general purpose computing and can perform computation in applications traditionally handled by the CPU. The main purpose of this research is to propose appropriate theoretical models for GPU computing, develop efficient parallel algorithms based on the theoretical models, and evaluate the performance. We have developed theoretical models, Discrete Memory Machine model, Unified Memory Machine model, and Hierarchical Memory Machine model which capture the essence of memory access to the shared memory and the global memory of the GPU. Based on these models, we have developed many efficient algorithms on the GPU. In particular, we have developed a new technique that we call SKSS (Single Kernel Soft Synchronization) technique. We have shown that this technique can be applied to accelerate dynamic programming algorithms on the GPU.

研究分野：情報工学

キーワード：Parallel Algorithms　GPGPU　Memory Machine Models

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

１．研究開始当初の背景

GPU(Graphics Processing Unit)は本来グラフィックス処理のための補助演算用の LSI であるが，これをグラフィクス以外の汎用計算に利用する技術 GPGPU が注目されており，さまざまな研究開発が行なわれている．現在の GPU のアーキテクチャは，複数の Processor Core と Shared Memory をもつ Streaming Multiprocessor が多数並んだものである（図１）．また，全 Processor Core がアクセスできる大容量の Global Memory を持っている．一方，1980 年から 2000 年にかけて並列計算の理論的研究が盛んに行なわれてきた．多くの並列アルゴリズム理論の研究は，共有メモリ型並列計算機の理論モデルである PRAM（Parallel Random Access Machine）を対象としてきた．例えば，1986 年に R. Cole は PRAM 上で n 個のデータを n 台のプロセッサを用いて O(log n)時間でソーティングを行なう最適並列マージソートアルゴリズムを示している．PRAM は１つのメモリ空間を持ち，全プロセッサが任意のアドレスに単位時間でアクセスできると仮定した並列計算機の理論モデルである．GPU は Streaming Multiprocessor 内の Processor Core が Shared Memory にアクセスすることができるので，各 Streaming Multiprocessor は１つの PRAM とみなすことができる．また，全 Processor Core が Global Memory にアクセスすることができるので，GPU 全体を１つの PRAM とみなすこともできる．しかし，PRAM 向けに最適化された並列アルゴリズムをそのまま GPU に実装しても，十分な性能を得ることができない．GPU のアーキテクチャや特性を考慮して，実装する必要がある．多くの汎用計算が，GPU に実装・評価されているが，その並列アルゴリズムの性能の理論的・解析的は行なわれていない．現在の GPU を用いた汎用計算の研究は，特定の GPU で実測した実行時間の比較ばかりである．実行時間は，GPU の型番，コンパイラの最適化，開発者のプログラミングスキルなどに大きく影響され，実行時間の実測値の比較では並列アルゴリズム優劣は明確にならない．そのため並列アルゴリズムの客観的な評価が困難である．

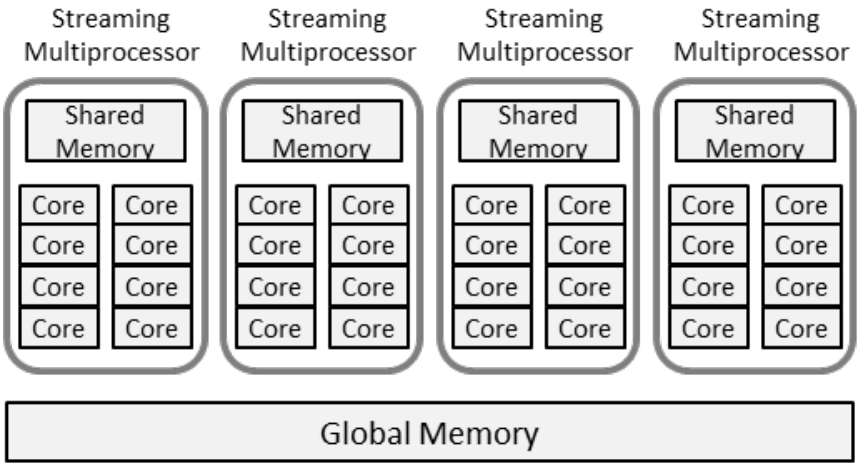

図１：GPU のアーキテクチャ

２．研究の目的

本研究の第一の目的は，GPU の本質をとらえた抽象理論計算モデルを提案することにある．図１の GPU アーキテクチャの特性を考慮し，GPU 上の並列アルゴリズムの性能の理論的解析を行えるようにする．その際，あまり重要な部分は切り捨て，性能に大きく影響する本質的な部分のみ抽出し，なるべく単純な理論計算モデルを提案する．また，そのモデルの上でさまざまな計算処理を効率よく行うアルゴリズムの設計手法を示し，実装実験によりモデルとアルゴリズムの妥当性を検証する．

３．研究の方法

GPU の本質をとらえた抽象理論モデルを構築し，基本的な並列アルゴリズム，例えば，Prefix-sums，SAT (Summed Area Table)，データ置換などの並列アルゴリズムを示す．この並列アルゴリズムは提案した抽象理論モデル上で動作するものであり，理論的な性能解析を行う．特に，計算時間の下界を示すことにより，並列アルゴリズムの最適性の検証も行う．さらに，より複雑な問題，例えば，画像のボロノイ図や，動的計画法などの処理について，実用的に高速な手法を開発し，GPU に実装して，性能評価を行う．

４．研究成果

GPU（図１）は複数の Streaming multiprocessor から構成される．各 Streaming multiprocessor は，計算を行うスレッドを実行する複数の Processor core と１つの Shared memory をもつ．この Shared memory は，小容量低レイテンシであり，各 Processor core が自由にアクセスすることができる．また，GPU は大容量高レイテンシの global memory をもち，すべての Processor core からアクセスすることができる．Streaming multiprocessor 内の

Shared memory は３２のメモリバンクにわかれている．１つのバンクの異なるアドレスに複数のメモリアクセスが同時に行われると bank conflict が発生し，メモリアクセスの処理が逐次的に行われ，スループットが低下する．よって，bank conflict がなるべく発生しないように，Shared memory のアクセスを行う必要がある．また，Global memory は，外部メモリであり，連続したアドレスへの同時アクセス（coalesced access）は効率よく行うことができる．一方，離散的なアドレスへのアクセス（stride access）はメモリアクセス処理が複数に分断され，スループットが低下する．よって，Global memory へのアクセスは stride access となるようにする必要がある．

これらのメモリアクセスの特徴を考慮し， GPU のための３つのメモリマシンモデル，Discrete Memory Machine(DMM)，Unified Memory Machine(UMM)，Hierarchical Memory Machine(HMM)を提案した（図２）．DMM は，GPU の１つの Streaming Multiprocessor 内の複数の Processor Core と Shared Memory を用いた並列計算のための 理論モデルである．UMM は，GPU 内の全 Processor Core が global memory を用いて並列計算するための理論モデルである．HMM は，複数の DMM から構成され，また全 Processor Core が UMM として動作する階層構造の並列計算モデルであり，現在の GPU のアーキテクチャに対応する．これらの並列計算モデルは，GPU の shared memory と global Memory へのアクセスでそれぞれ発生する Bank Conflict と coalesced access，およびメモリアクセスレイテンシの特徴を取り入れたものである．

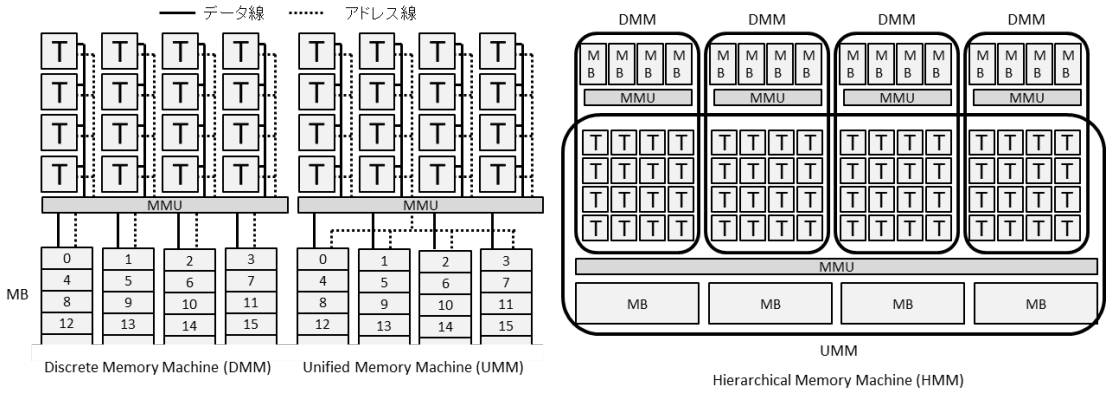

図２：メモリマシンモデル

このメモリマシンモデル上で，Prefix-sums，SAT (Summed Area Table)，データ置換，画像のボロノイ図，誤差拡散法，動的計画法，データ圧縮などのアルゴリズムを実装し，理論モデル上での性能評価を行った. 特に動的計画法や SAT 計算を効率よく行うための汎用手法である SKSS（Single Kernel Soft Synchronization）法を考案した．SKSS 法は，GPU 上で１つのカーネルだけを実行し，複数の CUDA Block（スレッドの集まり）がソフト的に同期しながら計算を行っていく処理である．通常の GPU の CUDA プログラムは，複数のカーネルを順に逐次的に起動する．これは CUDA Block 間の通信が直接行えないためで，デッドロックを発生させないようバリア同期のために，カーネルを終了させてすべての CUDA Block を一旦終了し，計算結果をグローバルメモリに書き出し，また，カーネル起動と CUDA Block の再スタートを行うという処理を繰り返す．そのため，グローバルメモリへの書き出し，読み出し，及び，カーネル起動・終了のオーバーヘッドが大きく，パフォーマンスを大きく低下させる．我々が考案した SKSS（Single Kernel Soft Synchronization）法では，これらのオーバーヘッドを最小限にすることができる．CUDA Block 間のソフト的な通信を実現するために，グローバルメモリ上のグローバルカウンタを用いて，起動 CUDA Block に動的にタスクを割り当てる．タスクが順次割り当てられるので，デッドロックが発生することがなく，CUDA Block の起動数，つまり並列度を最大化することができる．また，グローバルメモリへの書き出し，読み出しが少ないので，メモリ帯域を最大化することができる．

この SKSS（Single Kernel Soft Synchronization）法はさまざまな計算処理に応用が可能である．具体的には，Prefix-sums，SAT (Summed Area Table)，ボロノイ図，誤差拡散法，便パッキング問題などの処理が高速に行えることを理論的に証明し，GPU 実装により実証した．

５．主な発表論文等

〔雑誌論文〕（計１７件）（以下，すべて査読あり）

① Kohei Yamashita, Yasuaki Ito, Koji Nakano, Bulk execution of the dynamic programming for the optimal polygon triangulation problem on the GPU, Concurrency and Computation: Practice and Experience, 採録決定.

② Hiroki Tokura, Toru Fujita, Koji Nakano, Yasuaki Ito, and Jacir L. Bordim, Almost optimal column-wise prefix-sum computation on the GPU, The Journal of Supercomputing,

Vol. 74, No. 4, pp. 1510-1521, 2018.

③ Toru Fujita, Koji Nakano, Yasuaki Ito, Daisuke Takafuji, An Efficient GPU Implementation of CKY Parsing Using the Bitwise Parallel Bulk Computation Technique. IEICE Transactions on Information and systems, Vol. E100-D, No. 12, pp. 2857-2865, 2017

④ Daisuke Takafuji, Koji Nakano, Yasuaki Ito, and Jacir Bordim, C2CU: a CUDA C program generator for bulk execution of a sequential algorithm, Concurrency and Computation: Practice and Experience, Vol. 29, No. 17, 2016

⑤ Toru Fujita, Koji Nakano, and Yasuaki Ito, Fast Simulation of Conway's Game of Life using Bitwise Parallel Bulk Computation on a GPU, International Journal of Foundations of Computer Science, Volume 27, No. 8, 981-1003, 2016.

⑥ Hiroaki Kouge, Takumi Honda, Toru Fujita, Yasuaki Ito, Koji Nakano, and Jacir L. Bordim, Accelerating digital halftoning using the local exhaustive search on the GPU, Concurrency and Computation: Practice and Experience, Vol.29, No. 2, e3781, 2016.

⑦ Shunji Funasaka, Koji Nakano, and Yasuaki Ito, Fully Parallelized LZW decompression for CUDA-enabled GPUs, IEICE Transactions on Information and Systems, Vol. E99-D, No. 12, pp. 2986-2994, 2016.

⑧ Lucas Saad Nogueira Numes, Jacir Luiz Bordim, Koji Nakano, and Yasuaki Ito, A Memory-access-efficient Implementation for Computing the Approximate String Matching Algorithm on GPUs, IEICE Transactions on Information and Systems, Vol. E99-D, No. 12, pp. 2995-3003, 2016.

⑨ Takumi Honda, Yasuaki Ito, and Koji Nakano, GPU-accelerated Bulk Execution of Multiple-length Multiplication with Warp-synchronous Programming Technique, IEICE Transactions on Information and Systems, Vol. E99-D, No. 12, pp. 3004-3012, 2016.

⑩ Yuji Takeuchi, Koji Nakano, Daisuke Takafuji, Yasuaki Ito, A character art generator using the local exhaustive search, with GPU acceleration. International Journal of Parallel, Emergent and Distributed Systems, Vol. 31 No. 1, Pages 47-63, 2016.

⑪ Duhu MAN, Koji NAKANO, Yasuaki ITO, An Optimal Implementation of the Approximate String Matching on the Hierarchical Memory Machine, with Performance Evaluation on the GPU, IEICE TRANSACTIONS on Information and Systems, Vol.E97-D, No.12, pp.3063-3071, 2014.

⑫ Akihiko KASAGI, Koji NAKANO, Yasuaki ITO, Offline Permutation on the CUDA-enabled GPU, IEICE TRANSACTIONS on Information and Systems, Vol.E97-D, No.12, pp.3052-3062, 2014.

⑬ Akihiro Uchida, Yasuaki Ito, Koji Nakano, Accelerating ant colony optimisation for the travelling salesman problem on the GPU, International Journal of Parallel, Emergent and Distributed Systems, Volume 29, Issue 4, pp. 401-420, 2014.

⑭ Koji Nakano, Asynchronous Memory Machine Models with Barrier Synchronization, IEICE TRANSACTIONS on Information and Systems, Volume E97-D, No.3, pp.431-441, 2014.

⑮ Koji Nakano, Optimal implementations of the approximate string matching and the approximate discrete signal matching on the memory machine models, International Journal of Parallel, Emergent and Distributed Systems, Volume 29, Issue 2, pp. 104-118, 2014

⑯ Koji Nakano, Simple memory machine models for GPUs, International Journal of Parallel, Emergent and Distributed Systems, Vol. 29, No. 1 pp. 17-37, 2014.

⑰ Koji Nakano, Optimal implementations of the approximate string matching and the approximate discrete signal matching on the memory machine models, International Journal of Parallel, Emergent and Distributed Systems, Volume 29, Issue 2, pp. 104-118, 2014.

〔学会発表〕（計１８件）

① Lucas Saad N. Nunes, Jacir Luiz Bordim, Yasuaki Ito, Koji Nakano: A Prefix-Sum-Based Rabin-Karp Implementation for Multiple Pattern Matching on GPGPU, Proc. International Symposium on Computing and Networking 2018: 66-75

② Yutaro Emoto, Shunji Funasaka, Hiroki Tokura, Takumi Honda, Koji Nakano and Yasuaki Ito, An Optimal Parallel Algorithm for Computing the Summed Area Table on the GPU, Proc. of International Parallel and Distributed Processing Symposium Workshops, pp. 763-772, 2018.

③ Shunji Funasaka, Koji Nakano, Yasuaki Ito, Single Kernel Soft Synchronization Technique for Task Arrays on CUDA-enabled GPUs, with Applications, Proc. International Symposium on Computing and Networking, pp.11-20, 2017.

④ Hiroki Tokura, Yuki Kuroda, Yasuaki Ito, and Koji Nakano, A Square Pointillism Image Generation, and Its GPU Acceleration, Proc. International Symposium on Computing and

Networking, pp. 38-47, 2017

⑤ Takumi Honda, Shinnosuke Yamamoto, Hiroaki Honda, Koji Nakano, Yasuaki Ito: Simple and Fast Parallel Algorithms for the Voronoi Map and the Euclidean Distance Map, with GPU Implementations. Proc. of International Parallel Processing Sympoisum 2017: 362-371

⑥ Takahiro Nishimura, Jacir Luiz Bordim, Yasuaki Ito, Koji Nakano: Accelerating the Smith-Waterman Algorithm Using Bitwise Parallel Bulk Computation Technique on GPU. Proc. of International Parallel and Distributed Processing Symposium Workshops 2017: 932-941

⑦ Yi Yang, Yasuaki Ito, Koji Nakano: Photomosaic Generation by Rearranging Subimages, with GPU Acceleration, Proc. of International Parallel and Distributed Processing Symposium Workshops 2017: 942-951

⑧ Kohei Yamashita, Yasuaki Ito, Koji Nakano: A GPU Implementation of Bulk Execution of the Dynamic Programming for the Optimal Polygon Triangulation. Proc. of International Conference on Parallel Processing and Applied Mathematics (1) 2017: 314-323.

⑨ Ryouhei Murooka, Yasuaki Ito, Koji Nakano: Accelerating Ant Colony Optimization for the Vertex Coloring Problem on the GPU. Proc. of International Symposium on Computing and Networking, 2016: 469-475

⑩ Hiroki Tokura, Takumi Honda, Yasuaki Ito, Koji Nakano, Mitsuya Nishino, Yushiro Hirota, Masami Saeki: GPU-Accelerated Bulk Computation of the Eigenvalue Problem for Many Small Real Non-symmetric Matrices. Proc. of International Symposium on Computing and Networking 2016: 490-496

⑪ Shunji Funasaka, Koji Nakano, Yasuaki Ito: Light Loss-Less Data Compression, with GPU Implementation. Proc. of International Conference on Algorithms and Architectures for Parallel Processing 2016: 281-294.

⑫ Akihiko Kasagi, Koji Nakano, Yasuaki Ito: Parallelization Techniques for Error Diffusion with GPU Implementations. Proc. of International Symposium on Computing and Networking 2015: 30-39

⑬ Takumi Honda, Yasuaki Ito, Koji Nakano: A Warp-Synchronous Implementation for Multiple-Length Multiplication on the GPU. International Symposium on Computing and Networking 2015: 96-102

⑭ Koji Nakano, Yasuaki Ito: Optimality of Fundamental Parallel Algorithms on the Hierarchical Memory Machine, with GPU Implementation. Proc. of International Conference on Parallel, Distributed, and Network-Based Processing 2015: 626-634

⑮ Koji Nakano: A Time Optimal Parallel Algorithm for the Dynamic Programming on the Hierarchical Memory Machine. Proc. of International Symposium on Computing and Networking 2014: 86-95

⑯ Satoshi Okamoto, Yasuaki Ito, Koji Nakano, Jacir Luiz Bordim: Thorough Evaluation of GPU Shared Memory Load and Store Instructions. Proc. of International Symposium on Computing and Networking 2014: 614-616

⑰ Koji Nakano, Susumu Matsumae, Yasuaki Ito: Random Address Permute-Shift Technique for the Shared Memory on GPUs. Proc. of International Parallel Processing Symposium Workshops 2014: 429-438

⑱ Kazuya Tani, Daisuke Takafuji, Koji Nakano, Yasuaki Ito: Bulk Execution of Oblivious Algorithms on the Unified Memory Machine, with GPU Implementation. Proc. of International Parallel and Distributed Processing Symposium Workshops 2014: 586-595

〔図書〕（計１件）

① Koji Nakano: Theoretical Parallel Computing Models for GPU Computing. Open Problems in Mathematics and Computational Science 2014: 341-359

〔産業財産権〕

○出願状況（計０件）

① 特開 2017-091460：計算ノードネットワークシステム，中野 浩嗣，藤田 聡，鯉渕 道紘，藤原 一毅，2017

○取得状況（計０件）

〔その他〕

ホームページ等

http://www.cs.hiroshima-u.ac.jp/

６．研究組織

(1)研究分担者

研究分担者氏名：伊藤　靖朗

ローマ字氏名：(Yasuaki Ito)

所属研究機関名：広島大学

部局名：大学院工学研究科

職名：准教授

研究者番号（8 桁）：40397964

研究分担者氏名：高藤　大介

ローマ字氏名：(Daisuke Takafuji)

所属研究機関名：広島大学

部局名：大学院工学研究科

職名：助教

研究者番号（8 桁）：00314732

(2)研究協力者

なし