

GPUのための並列計算モデル

小池 敦^{1,a)} 定兼 邦彦^{1,b)}

1. 概要

本報告では GPU を使用した汎用的な並列計算のための計算モデルを提案する。一般的な GPU アーキテクチャは SIMD と MIMD の特徴を併せ持っている。すなわち、GPU デバイスは 1 つの大域メモリと複数のマルチプロセッサを備えており、各マルチプロセッサは独立に処理を実行する (MIMD)。一方、各マルチプロセッサは 1 つの共有メモリ (内部メモリ) と複数のコアを用いて計算を行っており、マルチプロセッサ内の全コアは常に同期して同一の命令を実行する (SIMD)。このアーキテクチャにより、GPU はプログラムを効率よく実行できる。

本研究の目的は、適切なモデルを提案することにより GPU 向けアルゴリズムの漸近解析をできるようにすることである。漸近解析により、デバイスの仕様や入力データの値に依らない汎用的なアルゴリズムの性能を知ることができる。また、実際にプログラムを実行せずに評価を行うことができるため、アルゴリズムを設計する際に有用である。逐次アルゴリズムの評価では、これらの理由から、漸近解析が多く使用されている。逐次アルゴリズムの漸近解析では、RAM(Random Access Machine) モデルと呼ばれる計算モデルを仮定して評価を行っている。RAM モデルはすべての逐次実行マシンに対する抽象化となっており、RAM モデルを用いることで、汎用的なアルゴリズムの評価を行うことができる。

一方、並列実行マシンには、RAM モデルのような共通の抽象化が存在しない。一般的に並列アルゴリズムの漸近解析には PRAM モデル [1] が使用されているが、PRAM モデルは 1 つのメモリとそのメモリに同時アクセス可能な複数のコアを持つモデルであり、GPU のような SIMD と MIMD の両方の特徴を併せ持つアーキテクチャとは異なっている。そのため、GPU 向けアルゴリズムを PRAM モデルを使用して評価しても、有益な結果は得られない。例えば、GPU 向けアルゴリズムにおいては、メモリアクセスにおいて、バンクコンフリクトやコアレスリングと呼ばれる事象を正しく考慮することが重要であるが、PRAM モデルではこれらを考慮した評価を行うことができない。

そこで、本報告では GPU アルゴリズムの漸近解析に適した計算モデルとして AGPU モデルを提案する。AGPU モデルのアーキテクチャを図 1 に示す。AGPU モデルは、デバ

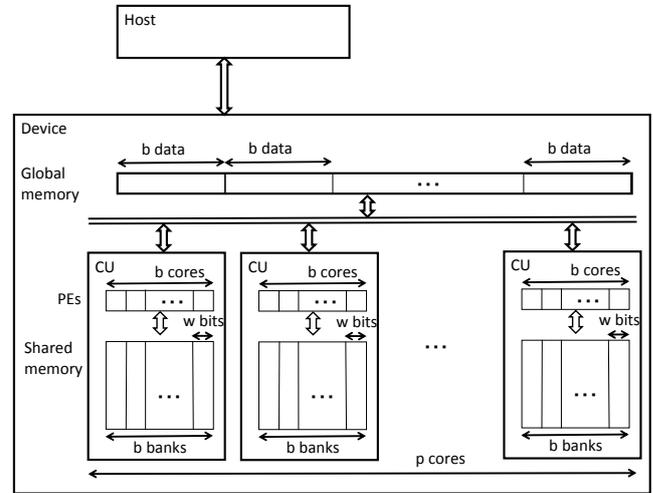


図 1 AGPU モデルのアーキテクチャ

イス内のコア総数 p 、マルチプロセッサ (図の CU; Compute Unit) 内のコア数 b 、メモリのワード長 w の 3 つのパラメータで規定される。マルチプロセッサ内の b 個のコアは同期実行しており、 b を用いることで、バンクコンフリクトやコアレスリング等の評価を行うことができる。

また、本モデルを使用したアルゴリズム解析の例として、リダクション計算を取り上げる。GPU を用いたりダクション計算においては、Tree-based と Cascading の 2 種類のアルゴリズム [2] が知られている。実際の GPU での計算では後者は前者よりも高速に動作する。我々は AGPU モデルによるアルゴリズムの解析を行い、本モデルにおいても後者のほうが時間計算量が小さくなっていることを示す。また、本報告では非可換演算を用いたりダクション計算に対する新しいアルゴリズムを提案し、本モデルおよび実際の計算時間が共に小さくなることを示す。

先行研究として、DMM and UMM モデル [3] は、GPU のメモリアクセスについて汎用的な評価を行うことができるモデルである。本報告はこれを発展させ、GPU に対するアルゴリズムの設計と評価を適切に行えるようにしたものである。

参考文献

- [1] Fortune, S. and Wyllie, J.: Parallelism in random access machines, *Proceedings of the tenth annual ACM symposium on Theory of computing*, STOC '78, New York, NY, USA, ACM, pp. 114–118 (1978).
- [2] Harris, M.: Optimizing Parallel Reduction in CUDA (2008).
- [3] Nakano, K.: Simple Memory Machine Models for GPUs, *IPDPS Workshops*, pp. 794–803 (2012).

¹ 国立情報学研究所 〒101-8430 東京都千代田区一ツ橋 2-1-2
^{a)} koike@nii.ac.jp
^{b)} sada@nii.ac.jp