

GPUにおけるメモリアクセスの性能と消費電力に関する一考察

酒井 貴多[†] 山口 実靖[‡][†]工学院大学大学院 工学研究科 電気・電子工学専攻[‡]工学院大学 工学部 情報通信工学科

1 はじめに

描画演算処理装置である GPU(Graphics Processing Unit)は CPU よりも高いピーク性能を持ち、ハードウェアの特性を生かして処理を行えば非常に高い性能を発揮することができる。そのため GPU を汎用目的に使用する GPGPU(General-Purpose computing on GPU)技術に関する研究が盛んに行われている。GPGPU がこのような大きな注目を集める理由の 1 個に、その消費電力があげられる。GPU の消費電力は CPU と比較しても高いが、それを上回る程度で性能が高く、結果的に GPU は単位消費電力あたりの性能が高くなる。

本稿では GPGPU の消費電力に着目し、各種メモリアクセス手法、ブロック数やスレッド数における処理性能と消費電力の関係について考察する。

2 GPGPU(General Purpose computing on GPU)

GPGPU プログラミング環境として、NVIDIA 社が提供している CUDA がある。本稿では、CUDA デバイスと CUDA 環境を用いて考察を行う。

2.1 コアレスアクセス

GPU 上のプロセッサ群が同時並列にメモリアクセスを行うときに、各スレッドのアクセスアドレスが連続であればそれらメモリアクセスは大きな 1 個のメモリアクセスにまとめられ 1 回で処理される。これはコアレスアクセスと呼ばれるメモリアクセス手法であり、コアレスでないアクセス(アクセスアドレスが不連続であり、個々のアクセスが別々に処理される)と比較して短い時間でメモリアクセスを終えることが可能となる。

2.2 Shared Memory

GPU は、全プロセッサで共有する VRAM の他

に、同一 SM(Streaming Multiprocessor)内の 8 個の SP(Streaming Processor)で共有する "Shared Memory"という小容量のメモリを保持している。Shared Memory は VRAM と比較して各 SP の近くに配置されているため、VRAM よりも短時間でアクセスが可能である。よって、VRAM から読み込んだデータを Shared Memory にコピーし次回以降は Shared Memory からデータを読み込むことにより、毎回 VRAM からデータを読み込む手法と比べ高速なデータアクセスが可能となる。

3 消費電力計測

GPU ボードへの電力供給は図 1 の様に 2 種類の方法で行われている。1 つが PCI-Express スロットを通じて MB(マザーボード)よりされる電力供給であり、もう 1 つが ATX 電源から拡張ボード用 6pin 電源コネクタを通じて行われる電力供給である。GPU の消費電力を測定するには、これら 2 つの電力供給を計測する必要がある。我々は図 2,図 3 の様な実験環境を構築し供給電力量を測定した。MB の PCI-Express スロットから供給される電力の量は、GPU ボードと PCI-Express スロット間にライザーカードを挟みライザーカードに流れる電流をクランプメーターで計測することにより測定可能である。なお、PCI-Express スロットと GPU ボード間で流れる電流には 12V 線と 3.3V 線があり、別々に計測を行う必要がある[1]。拡張ボード用 6pin 電源コネクタから供給される電力量は、電力を ATX 電源から供給させるのではなく外部電源から供給させ、ワットチェッカーを用いることにより計測できる。

測定は、NVIDIA GeForce 8800GT(SP 112 基),Fedora 10(Linux 2.6.27-5),CUDA2.3 を用いて行った。

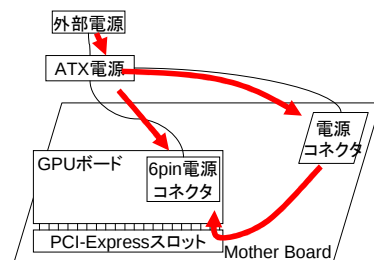


図1 GPUボードの電力供給

A Study on Memory Access Performance and Power Consumption on GPGPU Computing

Tota SAKAI[†], Saneyasu YAMAGUCHI[‡]

[†]Graduate school of Electrical and Electronics Engineering, Kogakuin University

[‡]Department of Information and Communications Engineering, Kogakuin University

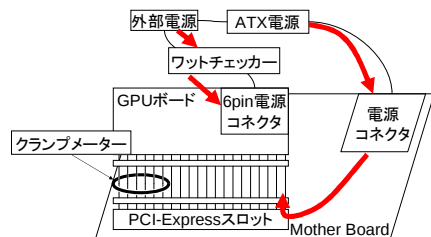


図2 GPUへの供給電力の測定環境(模式図)



図3 GPUへの供給電力の測定環境(写真)

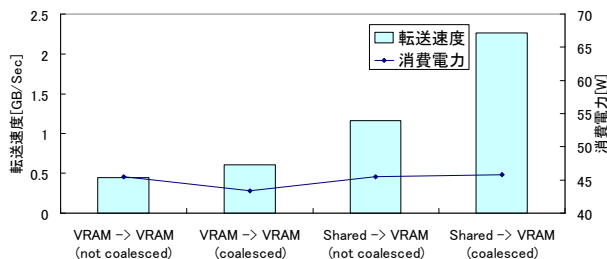


図4 性能と消費電力(メモリコピー)

4 実験

4.1 メモリコピーの性能と消費電力

VRAM から VRAM へのデータ転送処理および Shared Memory から VRAM へのデータ転送処理の性能と消費電力について調査を行った。本実験では、VRAM または Shared Memory から整数データ(4 バイト)を読み込み、それを VRAM に書き込む処理を繰り返すことにより性能と消費電力を測定した。メモリアクセスは 1Warp(32 スレッド)により並列に行った。測定結果を図 4 に示す。

図 4 より、メモリアクセスのコアレス化、Shared Memory の活用により、同程度の消費電力のままメモリアクセス性能のみを向上させることが可能であることが確認された。

次に、Shared Memory から Shared Memory へデータ転送処理にて得られた性能と消費電力について調査を行なった。Shared Memory アクセスは、1Warp で並列に行った。横軸の使用バンク数は並列に使用したバンクの数である。使用バンク数が少ないときは 32 個のスレッドのアクセスが少数のバンクに集中しておりバンク衝突が多く発生している状況である。測定結果を図 5 に示す。

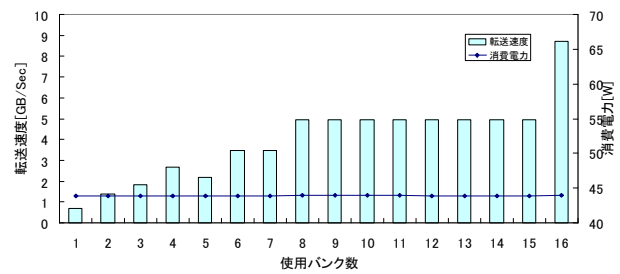


図5 性能と消費電力(Shared Memory)

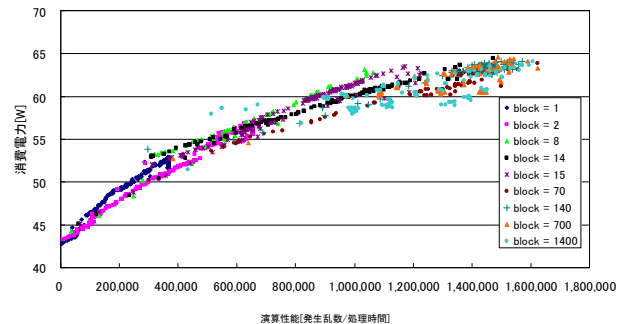


図6 演算処理における性能と消費電力

図 5 より、使用バンク数の増加により、同程度の消費電力のまま、性能のみを向上させることが可能であることが確認された。

4.2 プロセッサ演算の性能と消費電力

使用スレッド数と使用ブロック数を変化させて、プロセッサ演算時の性能と消費電力の関係を調査した。行った処理は Monte Carlo シミュレーションである。測定結果を図 6 に示す。図 6 より、SM の数(14)まではブロック数を増加させるにともない性能と消費電力が増加し、それ以降は消費電力を同程度に保ったまま性能のみを向上させられることが確認された。

5 まとめ

本稿では、GPGPU コンピューティングにおける消費電力と性能の関係について考察した。測定より、メモリアクセスのコアレス化、Shared Memory の使用、ブロック数を増加により消費電力を同程度に保ったまま性能のみ向上させられることが確認された。今後は、複数 GPU 環境での考察を行う予定である。

謝辞

本研究は科研費 (22700039) の助成を受けたものである。

参考文献

- [1] 長坂 仁, 丸山 直也, 額田 彰, 遠藤 敏夫, 松岡 聡: GPU における性能と消費電力の相関性の解析, 情報処理学会研究報告 2009-HPC-121(SWoPP 2009) (2009).