

GPU における消費電力と性能に関する 一考察

酒井貴多[†] 山口実靖^{††}

GPU の処理性能が向上し、GPGPU に注目が集まっている。GPU は処理性能が高いだけでなく、消費電力あたりの性能が高いという特徴を持つ。本稿では GPU の各種メモリにアクセスしたときの性能と消費電力の関係や、各種アクセス手法によりアクセスしたときの性能と消費電力の関係についての調査結果を述べ、低消費電力化手法についての考察を行う。VRAM アクセスにコアレスアクセスと非コアレスアクセスを用いた場合の性能と消費電力を比較したところ、コアレスアクセスを用いる手法が性能、消費電力の両面において優れていることが確認された。また、Shared Memory を用いる効果を確認したところ、Shared Memory を用いることにより性能と消費電力の両方が改善されることが確認された。そして、スレッド数、ブロック数と性能、消費電力の関係を調査したところ、ブロック数を非常に大きくすることにより同一性能で比較したときの消費電力を低減できることが確認された。

A Study on GPU's Performance and Power Consumption

Tota Sakai[†] and Saneyasu Yamaguchi^{††}

GPU has much higher performance than CPU. GPGPU is expected to be a promising method to achieve high performance. In this paper, we discuss the relation between GPU's performance and its power consumption with several memory accessing method. Our experimental results demonstrated that coalesced accesses and utilizing the shared memory improves both the performance and power consumption.

1. はじめに

描画演算処理装置である GPU (Graphic Processing Unit) は CPU よりも高いピーク性能を持っており、ハードウェアの特性を考慮して処理を行えば非常に高い性能を得ることができる。そのため、GPU を描画演算処理以外の汎用計算に使用する GPGPU (General-purpose computing on graphics processing units) 技術に関する研究が盛んに行われている。これらの研究は大きな成果を上げており、例えば 2010 年 11 月におけるスーパーコンピュータランキング TOP500¹⁾ にて上位 4 システムのうち、日本の TSUBAME 2.0 などの 3 種のシステムが NVIDIA GPU を使用するシステムとなっている。GPGPU がこのような大きな注目を集めている理由の 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 は電力効率において優れている。省エネルギーの重要性はますます高まっており、今後も GPGPU の重要性は高まっていき、GPU の消費電力に関する考察や低消費電力化も重要になってくると考えられる。

本稿では GPU の消費電力に着目し、各種メモリアccess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について調査し、考察する。

2. GPGPU

GPGPU は GPU を用いて描画演算処理以外の汎用計算を行う技術である。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。CUDA を用いることにより、描画演算処理に対する専門的な知識が無くても GPU を用いた汎用計算が可能となる。以下、CUDA GPU に限定して解説を行う。

CUDA GPU は、図 1 の様な構成をしている。すなわち、GPU 内に複数の SM (Streaming Multiprocessor) が存在し、各 SM 内に 8 個の SP (Streaming Processor) が存在している。これらの SP が処理装置の単位となる。メモリとしては主に VRAM と Shared Memory が存在し、VRAM は全ての SM からアクセスが可能であり、Shared Memory は SM 内に存在し各 SM 内からのみアクセス可能である。VRAM は Shared Memory より容量が多い(数百 MB～数 GB 程度)が、Shared Memory と比較しアクセスに要する時間が長い。Shared Memory は SP からの高速アクセスが可能であるが、容量が少なく(16KB)、キャッシュとして使用されることが多い。つまり、複数回アクセスされるデータを VRAM から Shared Memory にコピーし、Shared Memory に対してアクセスを繰り返すことによりメモリアccess時間の短縮を実現する。

[†] 工学院大学大学院 工学研究科 電気・電子工学専攻
Electrical Engineering and Electronics, Kogakuin University Graduate School
^{††} 工学院工学部情報通信工学科
Department of information and Communications Engineering, Kogakuin University

高速 VRAM アクセス手法として、コアレスアクセスがある。GPU 上のプロセッサ群が同時並列にメモリに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のメモリアccessは大きな1個のメモリアccessにまとめられ1回で処理される。このように結合されて処理されるアクセスがコアレスアクセスである。コアレスアクセスを用いると、コアレスでない手法(アクセスアドレスが不連続であり、個々のアクセスが個別に処理される)と比較して非常に短い時間でメモリアccessを終えることが可能となる。通常のアクセスとコアレスアクセスの例を図 2 に示す。非コアレスアクセスの例では、各スレッドがアクセスする領域が連続的に確保されている。例えばスレッド A がアクセスするのは左端の 2 ブロックである。スレッド群が同時にアクセスするブロック群は非連続となっており、これらは結合されず個別に処理される。これに対してコアレスアクセスの例では、各スレッドがアクセスする領域は不連続となっているが、スレッド群が同時にアクセスするブロック群は図の様に連続領域となっている。この場合、これらのアクセス群は1個の大きなアクセスに結合され1回で処理される。

また、Shared Memory アクセス性能の低下要因にバンク衝突があり、性能向上手法としてバンク衝突の回避がある。GPU の Shared Memory は 16 個のバンクにより構成されており、各バンクは独立に動作可能である。よって、最大 16 個のバンクを並列に使用して 1 スレッドアクセス時(1 バンク使用時)の 16 倍の性能を得ることが可能となる。逆に多数のスレッドが並列に Shared Memory へのアクセスを行ったとしても、複数のスレッドが同一バンクに対してアクセス要求を発行したときはそれらの要求は該当バンクにより順次処理され、同時並列的には処理されない。よって、アクセスバンクが衝突すると Shared Memory アクセス性能は低下してしまう。各スレッドからは異なるバンクのデータへのアクセス要求が発行される様にプログラムを作成することが好ましい。

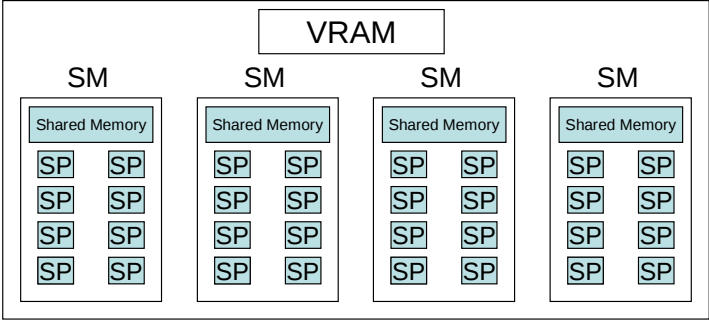


図 1 CUDA GPU の構造

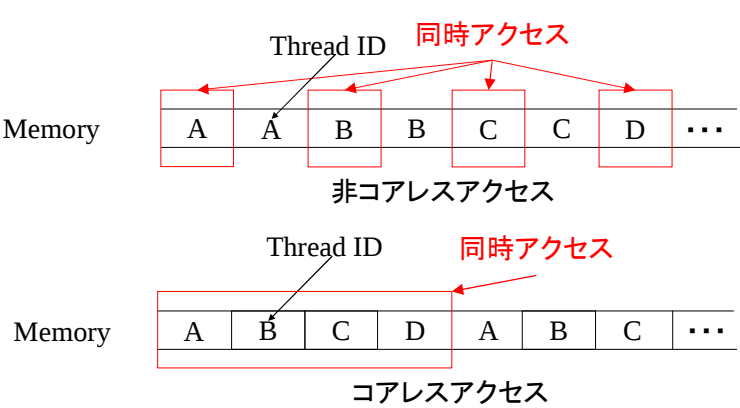


図 2 VRAM コアレスアクセス

3. 測定環境

本章にて、本研究で行った GPU 消費電力測定の測定環境について説明する。GPGPU は、GPU を用いて描画演算処理以外の汎用計算を行う技術である。

表 1 使用 GPU の仕様

"GeForce 8800 GT"	
CUDA Driver Version	2.3
CUDA Runtime Version	2.3
CUDA Capability Major revision number	1
CUDA Capability Minor revision number	1
Total amount of global memory	536150016 bytes
Number of multiprocessors	14
Number of cores	112
Total amount of constant memory	65536 bytes
Total amount of shared memory per block	16384 bytes
Total number of registers available per block	: 8192
Warp size	32
Maximum number of threads per block	512
Maximum sizes of each dimension of a block	512 x 512 x 64
Maximum sizes of each dimension of a grid	65535 x 65535 x 1
Maximum memory pitch	262144 bytes
Texture alignment	256 bytes
Clock rate	1.51 GHz

GPU ボードへの電力供給は図 3 の様に 2 種類の方法で行われる。1 つが PCI-Express スロットを通じて MB(マザーボード)よりな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 つの供給電力を測定する必要がある。

我々は図 4、図 5 の様な実験環境を構築し供給電力を測定した。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した。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。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て計測を行う。

測定は、NVIDIA GeForce 8800GT を用いて行った。GeForce 8800 の仕様、性能は表 1 の通りである。

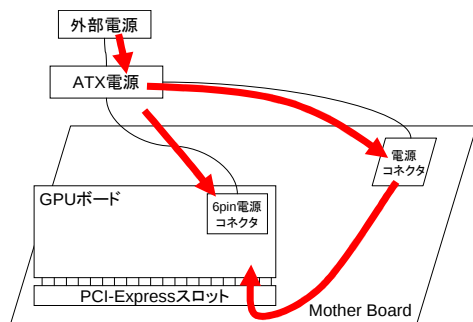


図 3 GPU ボードの電力供給

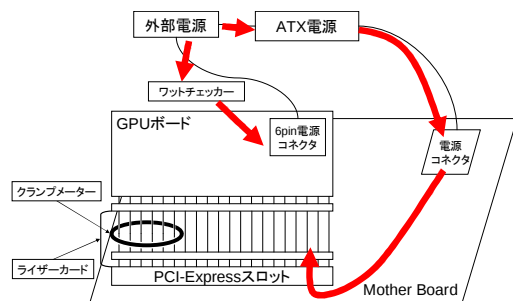


図 4 GPU ボードへの供給電力の測定環境 (模式図)



図 5 GPU ボードへの供給電力の測定環境 (写真)

4. 実験

4.1 メモリアクセスの性能と消費電力

GPU には VRAM と Shared Memory と異なる種類のメモリが搭載されている。また、VRAM アクセス手法には通常のアクセスとコアレスアクセスがあり、Shared Memory アクセス手法にはバンクコンフリクトが発生する手法としない手法がある。本章では、各種メモリへのアクセス時の性能と消費電力、各種手法でのメモリアクセス時の性能と消費電力について述べる。

最初に、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を示す。本測定では、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量を測定した。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 6 に示す。VRAM アクセス方法をコアレスアクセスにすることにより、消費電力を上昇させることなくメモリアクセス性能のみを大幅に向上させることが可能であることが確認された。また、読み込み元を VRAM から Shared Memory に変更することにより、消費電力を変えずにメモリアクセス性能を向上させることが可能であることも確認された。ただし、本実験では Shared Memory 使用時も書き込みは VRAM に対して行っているため、本測定結果は Shared Memory のみを用いた場合の性能ではない。

これらの性能を単位消費電力あたりの性能に換算したものを図 7 に示す。同図からも、性能と消費電力の両側面から考えたときもコアレスアクセスと Shared Memory の使用が有効な手法であることが確認された。

本実験では 1Warp により並列にメモリアクセスを行った。多数のスレッド、多数の Warp にて並列アクセスを行った場合は、メモリ待ち時間中に他の Warp の処理を行いメモリアクセス遅延時間を隠蔽できる。参考のために、512 スレッドで並列にメモリアクセスを行った実験の結果を付録に示す。

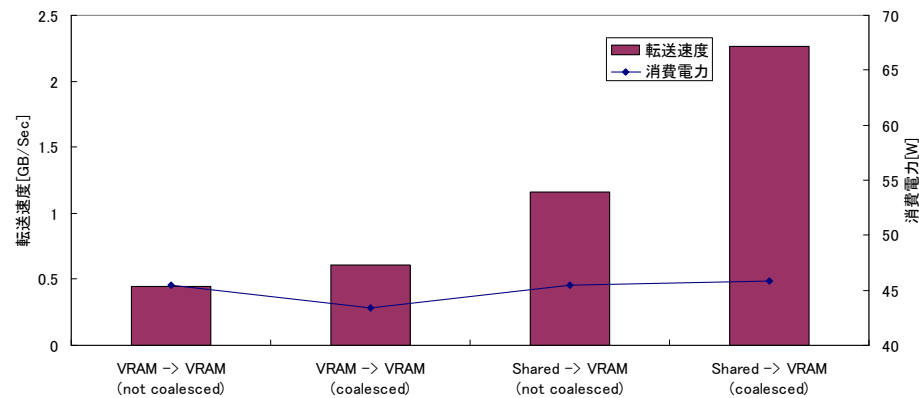


図 6 メモリコピーの転送速度と消費電力

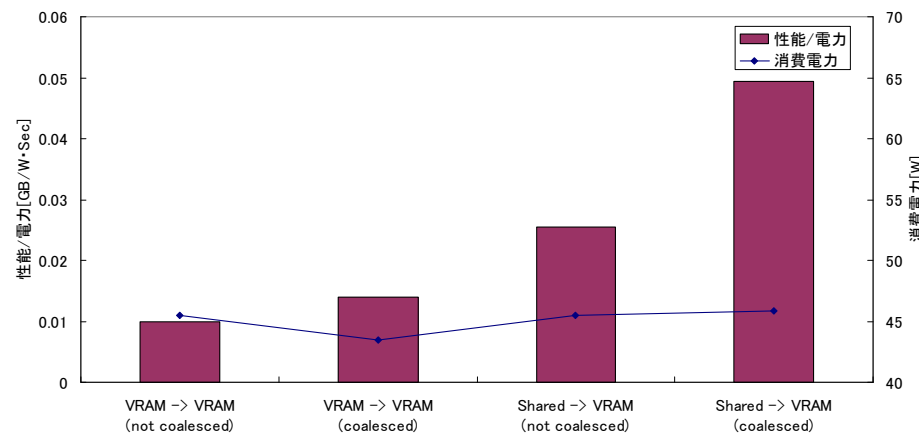


図 7 メモリコピーにおける単位消費電力あたりの転送速度

次に、Shared Memory から Shared Memory へのデータ転送処理にて得られた性能と消費電力の関係を図 8 に示す。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。使用バンク数が多いときは多くのバンクが並列に動作し、バンク衝突も少ない状況である。同図の結果より、バンク衝突を回避させることにより消費電力を増加させずに性能を大きく向上させることが可能であることが確認された。

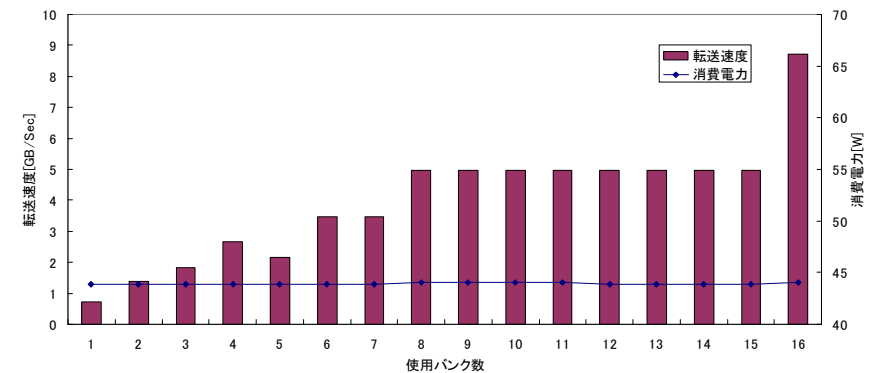


図 8 Shared Memory 間データ転送時の性能と消費電力

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。1 辺の長さが 1 の正方形の中からランダムに座標を選択し、その点が扇形の内部に入る確率を求め円周率を求めた。乱数は発生済みの状態でいい、発生処理は Monte Carlo シミュレーションに含めていない。

ブロック数は 1,2,8,14,140,700,1400 と変更させ、ブロックあたりのスレッド数は 1 から 512 まで変化させた。ブロック数、スレッド数、性能、消費電力の関係を図 9 に示す。図より、性能と消費電力には強い相関があり、性能が増えると消費電力は増加する傾向が非常に強いことが確認された。本実験で使用した GPU は 14 個の SM を持っているが、ブロック数を 1 から SM 数まで増加させていくと性能が向上するとともに消費電力も向上していき、SM 数以上に増加させていくと性能がなだらかに向上するとともに消費電力がなだらかに減少していく結果となった。すなわち同一性能で比

較すると、多くのブロックを用いた方が消費電力が低くなる傾向があり、性能を低下させずに消費電力のみを低下させるにはブロック数を増加させることが好ましいことが分かった。

また、前節の結果と比較することにより SP による演算処理の方がメモリアクセス処理よりも多くの電力を消費することが分かった。

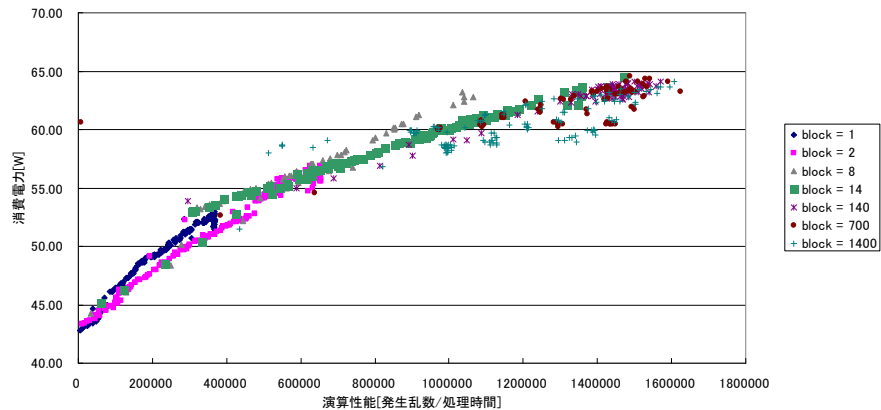


図 9 演算処理における性能と消費電力

次に、Matrix Multiplication 処理における性能と消費電力の関係を調査した。測定は、Shared Memory 使用時と VRAM のみ使用時に分けて行い、コアレスアクセス(各スレッドの同時アクセスメモリが連続)と各スレッドの同時アクセスメモリが 4 バイト間隔と 7 バイト間隔に変更して行った。測定結果を図 10(a)(b)に示す。図内の"SKIP4", "SKIP7"はアクセスメモリ間隔が 4 バイト、7 バイトであることを意味している。同図より Shared Memory を使用することにより (Shared Memory を使用しない場合と比較し) 性能が大幅に向上するとともに、消費電力も上昇していることが分かる。前節の実験にて、使用メモリを VRAM から Shared Memory へ変更することにより消費電力の向上を伴わずに性能のみを向上させることが可能であるとの結果が得られていたが、本節の実験では Shared Memory の使用により消費電力が向上する結果となった。これは Shared Memory の使用により処理速度が向上し単位時間あたりの SP の処理量が増加したためであると考えられる。また、各スレッドの同時アクセス領域が連続であるコアレスアクセスが性能と消費電力の両面において優れており、アクセスアドレス間の距離が 0 バイト(コアレス)、4 バイト、7 バイトと増加するにつれて消費電力も増加する結果となった。図 11(a)を単位消費電力あたりの性能に変換したものを図 11(b)に示す。

これらの図より、Shared Memory の使用、コアレスアクセスが性能と消費電力の両側面から見たときにも優れていることが確認された。

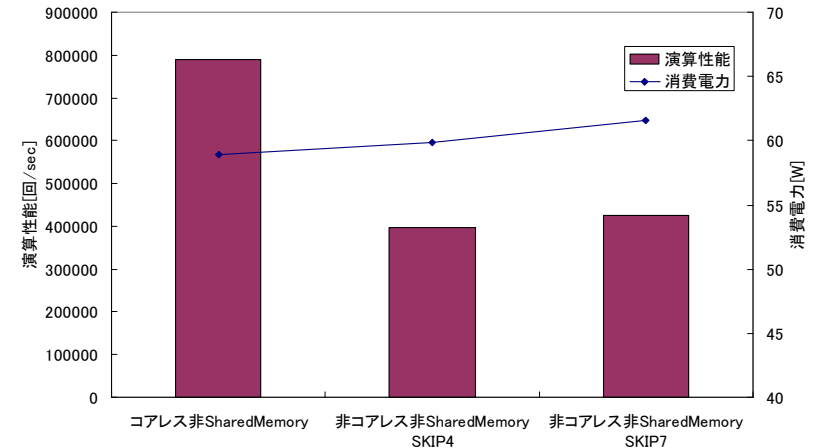


図 10 Matrix Multi における使用メモリ、メモリアクセス手法ごとの性能と消費電力(a)

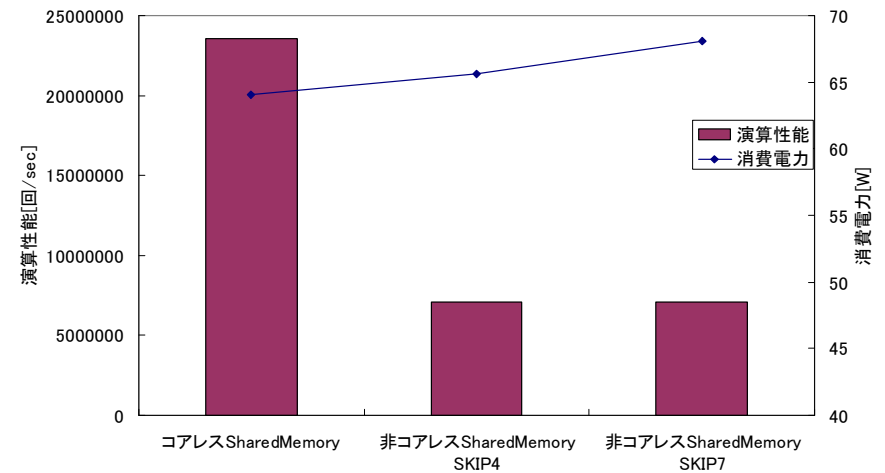


図 10 Matrix Multi における使用メモリ、メモリアクセス手法ごとの性能と消費電力(b)

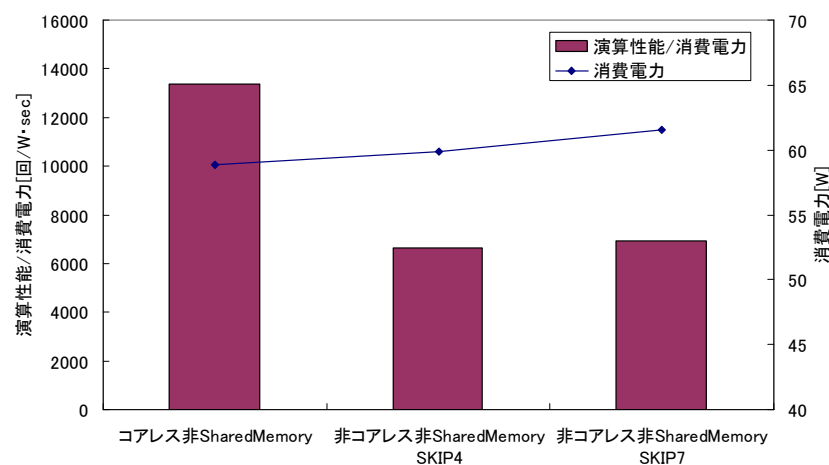


図 11 Matrix Multi 単位消費電力あたりの処理の性能(a)

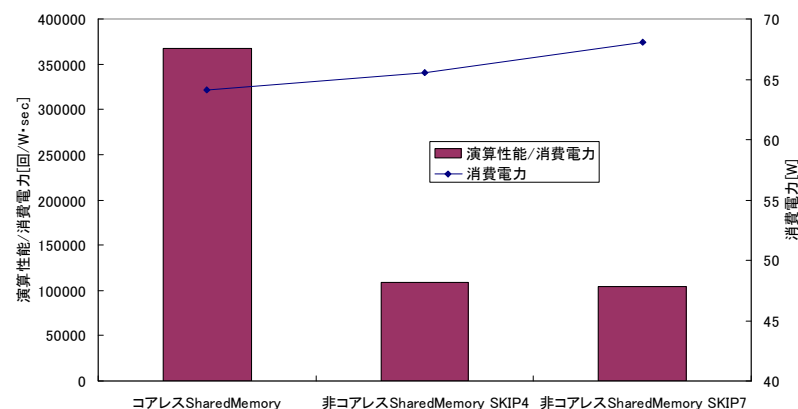


図 11 Matrix Multi 単位消費電力あたりの処理の性能(b)

5. 関連研究

GPU の電力消費に関する研究としては、松岡らによる消費電力量予測の研究がある²⁾。当該研究は GPU の消費電力に関する先駆的な研究であり、GPU の消費電力の予測方法や、GPU の消費電力の計測方法などを示している。我々の研究における電力消費の測定も、当該研究により示された方法に習っている。当該研究では、GPU プログラムのプロファイラの出力結果と消費電力の関係を調査し、それに基づく消費電力の予測方法を示している。

CPU に代わり GPU を用いることによる省電力に着目した研究として、中村らによるブル自動並列化の研究³⁾や、宇田川らによる GPU を用いた分子動力学法の高速度と省電力化の研究⁴⁾がある。当該研究では、GPU を用いて高速に分子動力学法の計算を行う方法を示し、GPU を用いることにより単位計算量あたりの消費電力を大幅に削減できることを示している。また、GPU を用いるスーパーコンピュータと、用いないスーパーコンピュータの比較を行い、GPGPU には性能面のみならず消費電力の面にも大きな優位性があることを述べている。

GPU による高速計算手法の確立としては、鈴木らによる高精度演算の並列化⁵⁾や、宗川らによる配列アライメント処理の高速化の研究⁶⁾がある。

6. おわりに

本稿では GPU の消費電力に着目し、GPU の各種メモリへ、各種アクセス手法アクセスを行ったときの性能と消費電力の調査を行った。調査の結果、コアレスアクセスや Shared Memory n 使用が性能と消費電力の両面において優れていることが確認された。また、演算処理時の性能と消費電力を調査し、演算処理時の消費電力がメモリアクセス時の消費電力よりも大きいこと、多数のブロックを作成することが省電力化につながることを示した。

今後は、本測定、考察結果を踏まえての省電力最適化手法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) <http://www.top500.org/>
- 2) 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).
- 3) 中村晃一, 林崎弘成, 稲葉真理, 平木敬: SIMD 型計算機向けループ自動並列化手法, 情報処理

学会研究報告, Vol.2010-HPC-126, No.10, pp.1-8(2010)

4) 宇田川拓郎, 関嶋政和: GPU を用いた分子動力学法の高速度化と省電力化, 情報処理学会研究報告, Vol.2010-HPC-127, No.5, pp.1-5(2010)

5) 鈴木智博, 高精度総和計算と高精度内積計算の GPU のための並列アルゴリズム: 情報処理学会論文誌, Vol.3, No.2, pp.48-56(2010)

6) 宗川裕馬, 伊野 文彦, 萩原 兼一: 統合開発環境 CUDA を用いた GPU での配列アライメントの高速度化手法, 情報処理学会研究報告, Vol.2008, No.19(2008-HPC-114), pp.13-18(2008)

付録

付録 A.1 メモリコピーの転送速度と消費電力(512 スレッド)

第 4.1 節 図 6 と同一の実験を 512 スレッドにて行い性能と消費電力を測定した。測定結果を図 2 に示す。図 6 同様に, コアレスアクセスを行うことにより消費電力を上げずに性能を向上させることが可能であることが確認された。

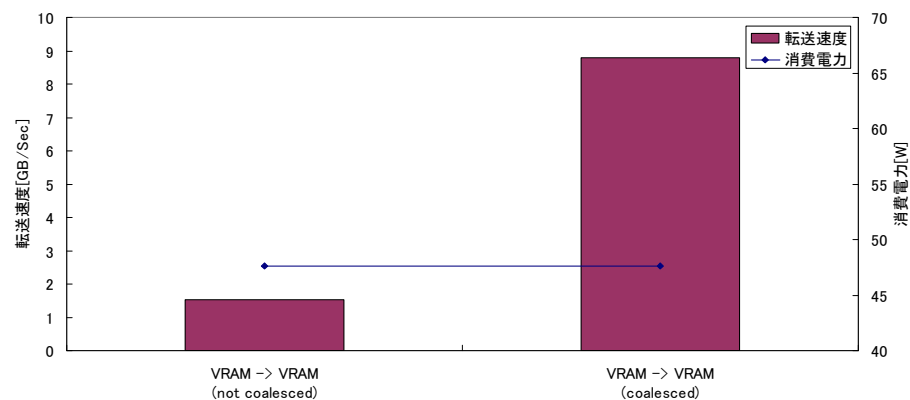


図 12 メモリコピーの転送速度と消費電力 (512 スレッド)