Vol.2009-HPC-121 No.27
2009/8/5

GPUにおける性能と消費電力の相関性の解析

長坂 仁[†1,†2] 丸山直也[†1,†2] 額田彰[†1,†2]
遠藤敏夫[†1,†2] 松岡聡[†1,†2,†3]

GPU の演算性能の飛躍的な発達により、画像処理だけでなく汎用計算にも用られるようになるにつれて GPU の消費電力削減の重要性が高まっている。 我々は、GPU の省電力化に向けた第一歩として消費電力と性能の相関を調べ、性能値より電力を予測するモデルを提案する。GPU 上で実行されるアプリケーションの特性に応じた省電力化を測るため、モデル化にはプログラム実行から得られるパフォーマンスカウンタ値を用い、それらを説明変数とした線形回帰分析により電力を予測する。評価の結果、回帰分析により 92.8%の精度で消費電力を予測できた。また、正の相関が強いものとして、命令スループット、メモリアクセス、レジスタ数、負の相関が強いものとして分岐実行数が特定された。

Correlative Analysis of Performance Counters and Power Consumption on GPUs

Hitoshi Nagasaka,[†1,†2] Naoya Maruyama,[†1,†2]
Akira Nukada,[†1,†2] Toshio Endo[†1,†2]
and Satoshi Matsuoka [†1,†2,†3]

GPUs are being employed in large-scale supercomputing environments, where their power consumption is a first-class design constraint. To reduce their power consumption, we propose a prediction model that leverages application behavior observable through performance counters. It predicts the power consumption of a given GPU kernel by a liner regression that uses the performance counter values when the kernel is executed, such as instruction throughput, register usage, memory accesses, and number of branches. Our experimental studies show that the model achieves up to 92.8% accuracy. We also found that, among others, instruction throughput and memory accesses are the most positively correlated with power, while number of executed branches is the most negatively correlated one.

1. はじめに

近年、GPU の高い演算能力をグラフィックス処理以外の汎用処理に応用する技術 (GPGPU) が注目されている。科学技術計算だけでなく医療や金融シミュレーション等にも用いらている[1)]。また、本学のスーパーコンピュータ TSUBAME にも昨年度末に搭載されるなど GPU は HPC 分野で利用が広がりつつある[2)]。そのような利用の拡大につれて、GPU の低消費電力化が非常に重要な問題になってきている。例えば今回の実験で用いた最新の GPU である NVIDIA GeForce GTX285 や Tesla S1070 中の 1 枚の GPU の消費電力は最大で 200W 程度にもなり、 一般的な CPU の消費電力を優に超える。

GPU の低消費電力化のためにはその電力消費の特徴を知ることが重要である。例えば、消費電力パターンが特定できた場合、それにより GPU の消費電力を予測し、複数の実装手法中より電力効率の良い実装を選択する等の電力最適化の可能性がある。しかし、GPGPU に関するこれまでの研究は主にその性能に関するものがほとんであり、消費電力に着目された研究は我々の知る限り行われていない。

我々は、GPU の消費電力の特徴を解析し、その予測手法を提案する。GPU の消費電力は実行プログラムによって 100W 程度と大幅に異なりうるため（2 節参照）、消費電力を予測するためには実行プログラムの特徴を捉える必要がある。そのために GPU 上でのプログラム実行時のパフォーマンスカウンタの値を取得しそのプログラムの特徴として用いる。取得したカウンタ値を説明変数とした線形回帰分析により消費電力を予測する。実際に 53 種類のカーネルを用い、予測モデルの制度を評価し、カウンタと電力の相関を調査した。予測の精度としては、平均誤差率が 7.2%という結果を得た。それぞれのカウンタの相関として、命令スループットが多いプログラム程電力が大きくなることが分かった。

2. GPU 消費電力の予備評価

予備実験として GPU のカーネル関数の特徴の違いからどれほど消費電力に差が生じるかを調査した。その結果を図 1 に示す。それぞれの電力は最大電力を表している。

FMA プログラムは、長崎大学濱田氏によるものであり、fma 演算の繰り返しにより GPU のピーク性能に近い性能を達成できるプログラムである。同プログラムはメモリ負荷はほぼ

†1 東京工業大学
†2 科学技術振興機構
†3 国立情報学研究所

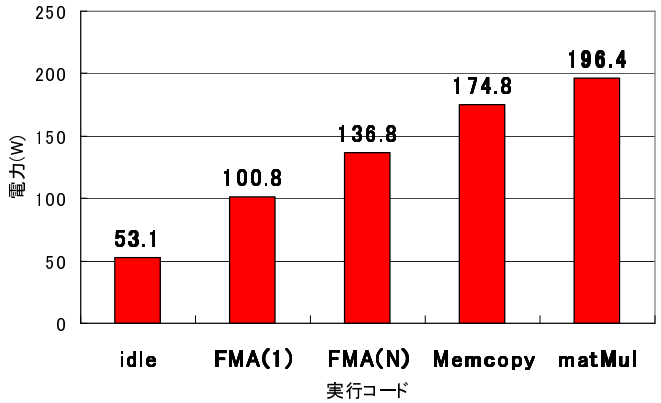

図 1 実行コードによる消費電力の違い

ゼロであるため、計算による消費電力の特徴を観測することができる。括弧の中はスレッド数であり、図中の N というのは消費電力が最大となるスレッド数を表している。今回の実験では N=256 である。Memcopy はデバイス内でのメモリ転送を行うプログラムであり、計算は行わない。アプリケーションとして 2880x2880 の行列積を求めるプログラム (matMul) についても予備評価を行った。、主として計算を行うが、メモリ転送も両方行う例として取り上げた。ここでの idle 時とは何もプログラムを実行していない状態の電力である。idle 時と比べ、約 2 倍のものから約 4 倍までと実行コードの特徴によりその消費電力には大きな開きがあることが分かる。そこで、我々はパフォーマンスカウンタの値をその特徴とし、それらの値から消費電力の相関が見られないかを調査した。また、2 つの実行内容や実行時間が同じ時にどちらのプログラムの方が電力の効率が良いか考えるために、どのカウンタの値が消費電力と相関が強いかを調査した。

3. 相関性解析手法

我々は GPU 上でカーネルの消費電力とそのカーネル実行のパフォーマンスカウンタとの相関を解析し、消費電力予測モデルを構築する。パフォーマンスカウンタは 1 回のカーネル実行全体にわたる総計のみ取得可能なため、消費電力は実行時の平均値とする。また、カーネル毎の実行時間のばらつきに対応するため、パフォーマンスカウンタの値を実行時間で割り、単位時間当たりの値を用いる。以下、それぞれの詳細を述べる。

3.1 消費電力の取得

GPU での消費電力を取得には以下がある。

電源からの供給電力: 実験に用いた GPU は GeForce GTX 285 であり 6 ピン電源コネクタ 2 本が接続されている。供給電圧は BIOS から調べた値を使用した。

PCIExpress からの供給電力: GPU とマザーボード間にをはさみ、電力供給線における電力を測定する。供給電圧は規格で定められている 12V と 3.3V をそれぞれ使用した。電圧は時間とともに変化しうるが、我々の調査により数%の変動であることから今回の実験ではその差は無視した。

3.2 パフォーマンスカウンタ値の取得

CUDA ではパフォーマンスカウンタの値は CUDA Profiler を用いることで容易に取得可能である[3)]。今回実験に用いたパフォーマンスカウンタの種類を表 1 に示す。ただし、occupancy はカウンタではないがプロファイラより同様に取得でき、かつ CUDA においてよく知られた代表的な指標なため、以下の回帰分析において説明変数として用いる。

表 1 実験に用いたパフォーマンスカウンタの種類

occupancy	スケジュール可能最大スレッド数に対するアクティブスレッド数の割合
gridSize	グリッドサイズ
blockSize	ブロックサイズ
dynSmemPerBlock	動的に割り当てられた共有メモリバイト数
staSmemPerBlock	静的に割り当てられた共有メモリバイト数
registerPerThread	スレッド毎に使用されるレジスタの数
gld_coherent	コヒーレントなグローバルメモリへのロード
gst_coherent	コヒーレントなグローバルメモリへのストア
branch	分岐の数
divergent_branch	divergent branch の数
instructions	実行された命令数
warp_serialize	共有メモリまたはコンスタントメモリアクセスにおいて 競合するアドレスの為、シリアル化されたスレッドワープ

パフォーマンスカウンタ値はカーネル関数 (メモリーコピー関数含む) 実行毎に取得できるが、1 度に値を取得可能なカウンタは 4 種類以下に限定されている為、複数回実行して取得する必要がある。また、パフォーマンスカウンタは単一の SM のみから読まれる。その為 SM 間でスレッドブロック割り当てにばらつきがある場合、読み込まれたカウンタ値が全体の SM の代表としては必ずしも適しているとは言えない。今回は一時的な解決策とし

てブロック数が SM 数未満であった実行は解析対象から除外した。

3.3 相関性の解析

平均消費電力を目的関数、単位時間あたりのカウンタ値を説明変数として線形回帰分析にかけ消費電力を予測する。すなわち、消費電力を P、カウンタの種類を n、パフォーマンスカウンタ値を p として

$$P = c_0 + \sum_{i=1}^{n} c_i * p_i \quad (1)$$

と表せる、最も P を高い精度で予測できる c_i を求める。また、その精度を解析するために leave-one-out 手法を用いる。具体的にはまずサンプル **i** を排除し残りのサンプルで回帰分析を行い、その結果からサンプル **i** の消費電力の予測精度を調べ、この操作を全サンプルに対して行う。この時、カウンタ値は実行時間に依存するもの (命令数等) は単位時間あたりとし、さらに種類によりサイズ等が異なる為標準化 (平均 0, 分散 1) した後に回帰分析にかける。また、どのカウンタ値が消費電力との相関が強いかを調査する為、回帰係数の比較を行う。

4. 準備・実験

4.1 実 験 環 境

表 2 GeForce GTX 285 の詳細

Total amount of global memory	1Gbyte
Number of multiprocessors	30
Number of cores	240
Total amount of constant memory	64Kbyte
Total amount of shared memory per block	16Kbyte
Total number of registers available per block	16384
Warp size	32
Maximum number of threads per block	512
Maximum sizes of each dimension of a block	512x512x64
Maximum sizes of each dimension of a grid	65535x65535x1
Maximum memory pitch	256Kbyte
Texture aligment	256byte
Clock rate	1.48GHz

今回用いた GPU は NVIDIA 社製 GeForceGTX285 でありアーキテクチャの詳細は以下の通りである。また、使用したマシンの OS は OpenSUSE11.0(kernel:2.6.25.20-0.4-pae) CPU は AMD Phenom(tm) 9500 Quad-Core Processor(2.2GHz) である。CUDA ドライバ 2.2、NVIDIA ドライバ 185.18.08 を用いた。

図 2 に実験環境の全体図を示す。

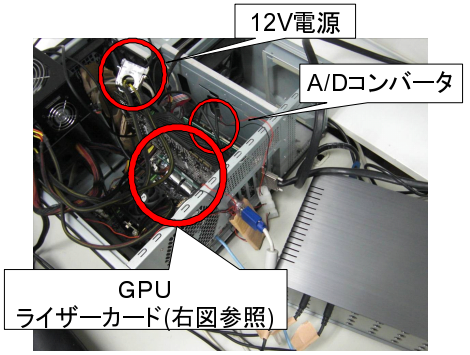

図 2 マシン全体図

図 3 ライザーカード

GPU における消費電力を測定するには ATX 電源の 12V 線から供給される電力、PCIe から供給される電力の 2 箇所を測定する必要がある。12V 線での測定は図 2 に示すように電流センサを装着するだけで可能となる。一方、PCIe から供給される電力は 3 に示すようにライザーカードをはさみ、さらにその中から 12V、3.3V の電力を供給している配線を測定する必要がある。

電流計には株式会社シナジェテック製 ST-30600 を用いる。これは、計測の際に配線に加工を必要としないクランプセンサを用いている。また、電流計と GPU コードのカーネル関数のタイムスタンプの差異を最小限に抑えるために同一のマシンに接続している。サンプリング間隔は 1ms とした。

4.2 計　　測

実験に使用したコードは CUDA SDK 付属のサンプルコードである。カーネル関数呼び出しの前後でタイムスタンプを取得し、後に電流計測の時間と照らしあわせて電力を算出する。これらの元のコードではカーネル関数の実行時間が非常に短いものが多いため、計測の誤差を小さくする為カーネル内処理を繰り返し実行するように変更し、すべてカーネルの個々の実行時間が 1 秒間となるようにした。

5. 評　　価

5.1 消費電力予測

図 4 に leave-one-out 手法を用いて得た消費電力の予測値と実測値の比較をしたものである。平均の誤差率は 7.2%という結果となった。サンプル 40 は予測値と非常に大きく離れている。比は 23.4 である。この大きな誤差原因として考えられるのは、特異なカウンタ値が存在することである。実際に、このサンプル 40 の warp_serialize は他のサンプルがほとんど 0 なのに対し、時間当たり 200 もの値を示している。回帰分析による予測した結果を調査したところ、やはり他のカウンタ値と比べ、非常に大きな値を示していた。こうした誤差はサンプルをより増やすことによって特異なサンプルを減らすことで削減可能であり、今後の課題である。

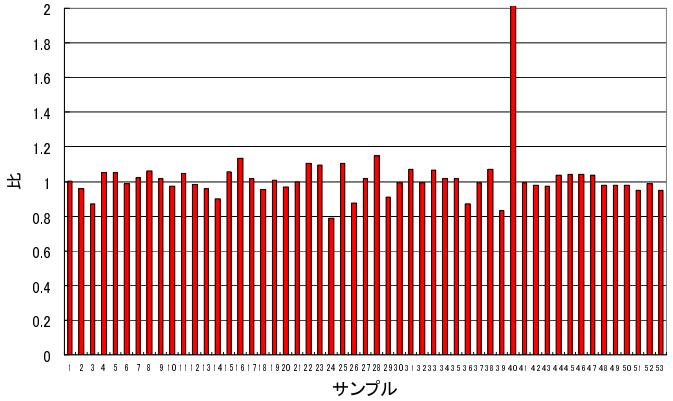

図 4　予測値と実測値の比

5.2 カウンタ値との相関性

図 5 にそれぞれのカウンタ値の回帰係数を示す。instructions というカウンタ値が最も相関が強い結果となった。これは命令スループットが高ければ消費電力も上がるということを示す。逆に branch が多い、つまり分岐の数が多い場合消費電力が抑えられる事が分かった。GPU は投機実行などの複雑な機構を持たないため、分岐命令があるとその結果が判明

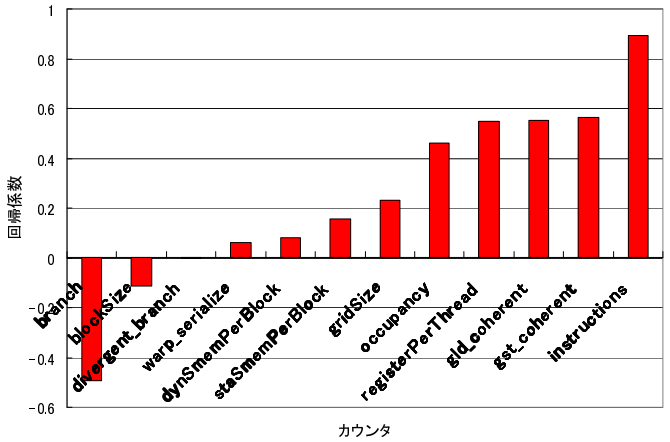

図 5　各カウンタ値の回帰係数

するまで後続の命令を実行することができない。他のスレッドも実行可能な状態にない場合にはストールし、その間消費電力が抑えられると考えられる。

6. 関 連 研 究

我々の知る限り GPGPU における GPU の電力の評価は報告されていない。以下は本研究と同様にパフォーマンスカウンタを用いた性能・電力モデリングに関する取り組みについて述べる。

浅井らによる学習と DVFS を用いた消費電力削減手法では予めパフォーマンスカウンタを用いて得られる値と性能の相関関係を回帰分析による学習させておく[4]。目的のプログラムを走らせる際にインターバル毎にパフォーマンスカウンタの値を受け取り、先に学習させた結果から次のインターバルの実行において、予め設定した性能を下回らない最低の周波数で動作 (DVFS) させ、消費電力を抑えている。理論値ではあるが、 性能は設定値を達成しつつ最大 24.9%の消費電力削減となることが確認された。

Maury らはパフォーマンスカウンタ値の線形回帰分析による実行時間予測に基づき、マルチコアプロセッサにおける性能的・電力的最適並列度の推定手法を提案している[5]。一般にコア間でキャッシュが共有されたプロセッサではキャッシュコンフリクトの影響により常

にコア数と同数のスレッドが最適性能となるとは限らない。提案手法では OpenMP の並列リージョンについて予備実行によるパフォーマンスカウンタの値から実行時間を推定し、最適実行時間を達成する並列度を動的に選択する。これは、実験により常にコア数と同数のスレッドを用いた場合に比べて 17%の性能向上を達成し、26%のエネルギー削減が示されている。

7. おわりに

7.1 まとめ

本研究では、実行プログラム毎に GPU における消費電力が大きく異なることに注目しその関係について調べた。、消費電力とカーネル関数の特徴を示すパフォーマンスカウンタの値の間の相関性を解析し、かつどのカウンタ値と相関性が強いかを調査した。その結果誤差は平均して 7%程度に推定ができ、命令スループットの高いプログラムでは消費電力が大きいことが分かった。逆に分岐の数が多いプログラムでは消費電力が抑えられる傾向があるという結果を得た。

7.2 今後の課題

今回の実験では相関性の解析には線形回帰分析を用いたが、他の解析手法、例えばニューラルネットワークによる予測モデル等を検討し、精度の向上を目指す。また今回は 53 個のカーネル実行をサンプルとして用い、90%超の精度を達成することができたが、一方で特異なカウンタ値をもったカーネルの予測については大幅な予測エラーが観測された。今後は単に多くのサンプルを取得するだけでなく、様々な特徴を持ったカーネルについて偏りなく計測対象とすることでこのような予測エラーの軽減を目指す。

また、今回は省電力化に向けた第一歩としてカーネル実行時の平均電力のみを考慮し、エネルギー、実行時間等は考慮していない。今回の結果からはスループットが最も正に相関が強かったが、かといって命令スループットを下げる方針が省電力化に適しているとは必ずしも言えず、アプリケーションの実行完了時間も考慮に入れた、ED 積等の指標による評価も必要である。今後の課題としては、Baghsorkhi らによるカーネルの性能モデルを用いた GPU 向け実行時間予測手法などを応用することで[6]、GPU の電力性能の最適化に取り組む。

謝辞 本論文を執筆するにあたり、ライザーカードを用いた電力測定において東京大学須田礼仁先生、Da-Qi Ren 様には多大な協力をして頂きました。誠にありがとうございます。GPU 電力消費評価に用いた FMA プログラムは長崎大学濱田先生に頂きました。感謝いたします。本研究の一部は科学技術振興機構戦略的創造研究推進事業『ULP-HPC: 次世代テクノロジのモデル化・最適化による超低消費電力ハイパフォーマンスコンピューティング』，及び Microsoft Technical Computing Initiative “ HPC-GPGPU: Large-Scale Commodity Accelerated Clusters and its Application to Advanced Structural Proteomics ” によるものである。

参考文献

1) Samuel S. Stone, Justin P. Haldar, Stephanie C. Tsao, Wen-Mei W. Hwu, Zhi-Pei Liang, and Bradley P. Sutton. Accelerating advanced mri reconstructions on gpus. In *CF '08: Proceedings of the 2008 conference on Computing frontiers*, pp. 261–272, 2008.
2) 遠藤敏夫. 東京工業大学 tsubame におけるアクセラレータ活用事例. 情報処理, Vol.50, No.2, pp. 100–106, 2009.
3) NVIDIA. Cuda profier, 2009.
4) 浅井雅司, 池田佳路, 佐々木宏, 近藤正章, 中村宏. 統計処理に基づくコンパイラ協調型 DVFS 手法. 情報処理学会論文誌, No.8, pp. 43–48, 2006.
5) Matthew C. Maury, F. Blagojevic, C. D. Antonopoulos, and D. S. Nikolopoulos. Prediction-based power-performance adaptation of multithreaded scientific codes. *Parallel and Distributed Systems, IEEE Transactions on*, Vol.19, No.10, pp. 1396–1410, 2008.
6) Sara Baghsorkhi and Wen mei Hwu. Analytical performance prediction for evaluation and tuning of GPGPU applications. In *Workshop on Exploiting Parallelism using GPUs and other Hardware-Assisted Methods (EPHAM'09), In conjunction with The International Symposium on Code Generation and Optimization (CGO) 2009*, 2009.