2014 年 1 月 6 日 情報システム基盤学合同輪講発表資料

GPUWattch : Enabling Energy Optimizations in GPGPUs

著者： Jingwen Leng et al.

出典： *Proc. of 40th International Sysmposium on Computer Architecture (ISCA'13), pp.487-498, 2013.*

発表者： 本多研究室　1353027 藤原 祐太

1　はじめに

近年，GPU の高い演算能力を汎用コンピューティングへ利用する GPGPU に注目が集まっている．これまでの GPU の用途は，主に PC ゲームなどのグラフィック処理であったため，より性能が重視された．しかし，スマートフォンなどの電源が限定されているモバイル端末への搭載を受け，性能だけでなく消費電力も重要な指標になってきている．これまで，GPU 向けの性能モデルはいくつか提案されているが，GPU のエネルギー効率を研究し最適化を行うための電力モデルは存在していない．

本論文では，GPU の電力モデルである GPUWattch を提案する．これは，CPU の電力モデルをベースとしながら実 GPU の電力測定値と比較しつつモデルを改良することでより正確な電力モデルを構築するものである．

2　従来の電力モデル

表 1 は，既存の電力モデルとその機能を示している．

表 1: 既存の電力モデル

Work	GPU	Configurable?	Cycle level?	Validated?
McPat	No	Yes	Yes	Yes
Hong and Kim	Yes	No	No	Yes
GPUWattch	Yes	Yes	Yes	Yes

これまで，GPGPU 向けの性能モデルや簡易的な電力モデルは存在していたが，ロバストな電力モデルは存在していなかった．ここで，ロバストな電力モデルとは，(1) 各コンポーネントのパラメータをユーザが任意に設定可能 (Configurable)，(2) サイクルレベルでシミュレーション可能 (Cycle-level)，(3) 実際のハードウェアと比較し，精度について妥当性の検証が行われている (Validated)，この 3 つの条件を満たしたものである．今回提案する電力モデルである GPUWattch はこれら 3 つの条件を満たしており，より自由度と精度の高いシミュレーションを行うことができる．

3　GPUWattch

3.1　概要

前節で述べた通り，GPUWattch は 3 つの機能を搭載している．(1) については，各コンポーネントの入力ポート数やビット幅，GPU のコア数などをユーザが任意に変更することで可能としている．

(2) について，サイクルレベルのパフォーマンスシミュレータである GPGPU-Sim[1] と統合を行うことで，サイクルレベルでの消費電力のシミュレートを可能にした．図 1 は，GPGPU-Sim と GPUWattch の連携を示したものである．

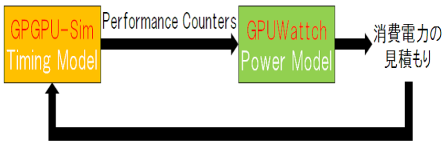

図 1: GPGPU-Sim と GPUWattch の連携

まず，GPGPU-Sim は，どのユニット (DRAM，キャッシュetc) を使用するのか，また，そのユニットは一定サイクルの間に何回使用されるのか (Performance Counter) の情報を GPUWattch へと転送する．そして GPUWattch はその情報を元に，各ユニットのアクセス当たりの消費エネルギーを見積もる．また，それらの情報を GPGPU-Sim へとフィードバックし上記の処理を繰り返すことで，プログラム終了するまでに消費する電力をシミュレートすることが可能となる．

(3) については，多くのベンチマークを使用して実際の GPU ハードウェアとの電力を比較し，精度について検証を行っている．

3.2　Initial Modeling

GPGPU 向けのロバストな電力モデルが存在していなかった理由の一つは，既存 GPU のアーキテクチャの詳細が公開されていないためであった．そのため，本研究では既存の CPU 電力モデルである McPat[2] などを基に GPU の構成を仮定して初期モデルを作成する．

図 2 は仮定した GPU の初期モデルの概略図を示している．GPU は大きく分けて 4 つのコンポーネントから

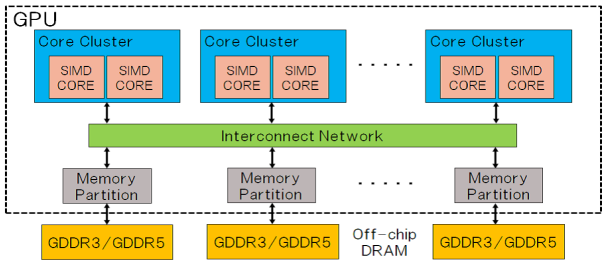

図 2: 初期モデルの概略図

構成されており，それぞれ大量の演算器を搭載している Core Cluster，コアとメモリをつなぐ Interconnect Network，L2 キャッシュから構成される Memory Partition，Off-chipDRAM の GDDR3/GDDR5 である．

初期モデル作成後，図 3 の (2) と (3) を繰り返すことで，電力モデルを完成させる．

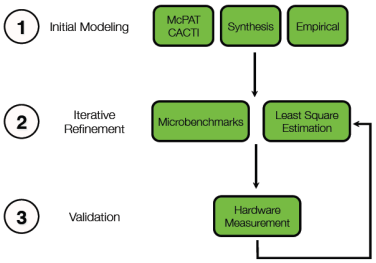

図 3: 電力モデル作成プロセス

3.3 Iterative Refinement

初期モデルを作成した後の次のステップとして，モデルの改良を行う．初期モデルは CPU の電力モデルを元に作成しているため，実際のハードウェアの違いによる誤差が生じる．例えば，GPU 特有のキャッシュであるコンスタントキャッシュやテクスチャキャッシュは CPU には搭載されていないため，初期モデルを作成する際はこれらを一般的なキャッシュとして定義する．しかし，実際は一般的なキャッシュの機能に加えて特殊な機能も搭載されているため，消費電力の挙動も完全には把握できない．このような初期モデルとは異なるハードウェアを実ハードウェアの電力測定と比較しつつモデル化する．具体的には，各コンポーネント毎に負荷を与える多数のマイクロベンチマークを利用し，実ハードウェアとモデルの電力を測定する．そらら電力値を最小 2 乗法を用いることで，モデルの係数を推定する．

3.4 Validation

仮のモデルを作成した後，実際のハードウェアにとモデルに対し実アプリケーションを使用して消費電力値を比べる．誤差が大きい場合，前節で説明したステップに戻り改良を繰り返すことでモデルを完成させる．

4 評価

本章では，モデルの精度を評価する．今回提案する電力モデルである GPUWattch は NVIDIA 社の GPU をベースとしている．評価対象としたハードウェアモデルは GTX 480 と Quadro FX5600 であり，どちらも異なったアーキテクチャである．使用するプログラムは，Rodinia，ISPASS，IPP ベンチマーク中のプログラムであり，合計 24 個のカーネルを使用した．図 4 は実ハードウェアと GPUWattch で測定した消費電力を示している．

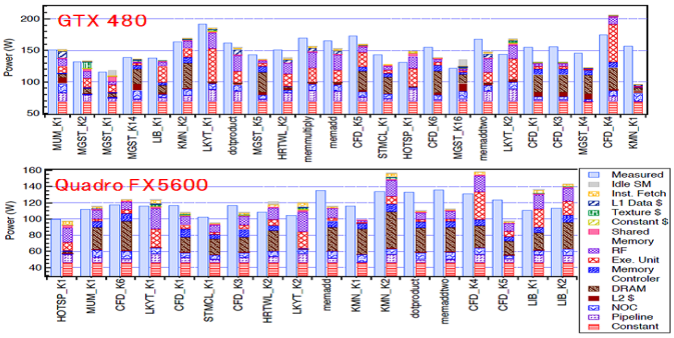

図 4: 実ハードウェアとの比較結果

GTX 480 との比較結果を見ていると，平均の誤差率は 9.9%以下である．ただし，KMN_K1 や CFD_K4 など，誤差率が 40%近くになってしまっているベンチマークもある．これらに考えられる共通の原因として，本研究で使用した性能モデルとハードウェアのメモリの階層部分の不一致によるものだと考えられる．

Quadro FX5600 に関しては，平均誤差率は 13.4%であった．また，GTX 480 と Quadro FX5600 のコンポーネント毎の消費電力の内訳を見てみると，演算器 (Exe.Unit) や L2 キャッシュ(L2$) による消費電力は GTX 480 の方が多い．これより，GPU の構成を違いによる電力を評価できることがわかる．

5 おわりに

この論文では，GPU の省電力化について研究を行うため，GPGPU 向けの電力モデルである GPUWattch を提案し，評価を行った．ハードウェアとシミュレータでの電力測定値を比較した結果，GTX 480 では誤差率が 9.9%以下に，Quadro FX5600 では 13.4%以下に抑えることができた．また，GPGPU-Sim と統合することでサイクルレベルでのシミュレーションも可能となった．

参考文献

[1] A. Bakhoda et al. Analyzing CUDA workloads using a detailed GPU simulator. In ISPASS, 2009.

[2] S. Li et al. McPAT: an integrated power, area, and timing modeling framework for multicore and manycore architectures. In MICRO, 2009.