2014 年 4 月 2 日 情報システム基盤学専攻中間発表資料

使用コア数と動作周波数の動的制御による GPU の省電力化に関する研究

所属： 高性能コンピューティング学講座
発表者： 1353027 藤原 祐太
主任指導教員： 本多 弘樹

1 概要

近年，半導体の微細化技術の恩恵を受け，GPU の性能向上が著しい．従来，GPU はグラフィック処理を行うデバイスとして普及してきたが，その高い演算能力を HPC 分野などの汎用的な計算処理に応用することが一般的になってきている．

GPU は CPU に比べ多数の演算器を搭載しており，並列性の高い処理を行うことができる反面，演算回路の大規模化により消費電力の増加が問題となってきている．高性能計算環境においても，消費電力の増大は発熱量の増加による冷却コスト増加や信頼性の低下に繋がってしまう．そのため，GPU の省費電力化は重要な課題である．

本研究では，実行するアプリケーションに応じて使用コア数を調整させ，さらに周波数を動的に制御することで GPU の省電力化を目指す．

2 研究の背景

2.1 GPU のアーキテクチャモデル

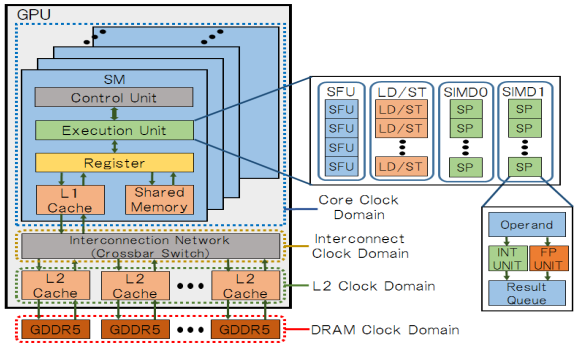

図 1: GPU のハードウェア構成

図 1 は一般的な GPU のハードウェア構成を示している．GPU には Core Clock Domain，Interconnect Clock Domain，L2 Clock Domain，DRAM Clock Domain の複数のクロックドメインが存在する．各ドメインはそれぞれ異なる周波数で動作する．また，本研究では Streaming Multiprocessor（SM） と呼ばれるプロセッサコアをコアと称して扱う．

2.2 GPU の電力消費傾向

図 2 は文献 [1] で報告されている GPU の各コンポーネントの平均消費電力の割合を示している．この図より，コアによる消費電力（図中黒枠）は全体の約 60%以上を占めている．このことから，コアの消費電力削減を考えることが，GPU の消費電力削減の効果が大きいと分かる．

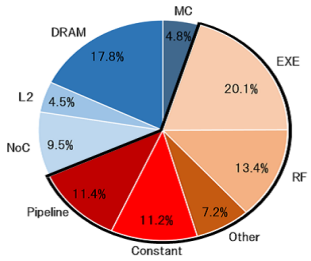

図 2: GPU の消費電力の内訳

また，実行するアプリケーションによっては，GPU に多数のコアが存在していてもそれら全てを使い切れないものも多い．従って，アプリケーションの特性に応じて使用コア数を調整することで消費エネルギーを削減できる可能性がある．

2.3 LSI における電力削減手法

一般的に，LSI の総消費電力は，動的消費電力であるスイッチング電力と，静的消費電力のリーク電力の総和で表すことができる．スイッチング電力削減技術として，LSI に要求される負荷に応じて電源電圧と動作周波数を変更する Dynamic Voltage and Frequency Scaling（DVFS）が広く使われている．DVFS は動作周波数低下に比例して，また電源電圧低下の 2 乗に比例して電力を削減できるため，有効な省電力手法である．

また，リーク電力削減技術として，使用されていない回路への電源供給を遮断する Power Gating（PG） が広く使われている．

3 本研究の目的

本研究では，アプリケーションに応じて使用するコア数と動作周波数を動的に制御する手法の考案を目指す．その第一ステップとして，いくつかのベンチマーク群からアプリケーションを選択し，それらのアプリケーションをシミュレータ上で実行し，どのような制御を行えばよいのか検討する．具体的には，GPU のコア部分による消費電力をターゲットとし，アイドル状態のコアに対しては PG を行い，メモリアクセスなどコア部以外の処理が主にな

る際には DVFS を用いる．また，本研究で検討する PG および DVFS は，コア部のみを対象とし，その他のドメインは周波数・電圧共に一定として評価を行う．

4　進捗状況

これまで，GPU の省電力化手法を検討するための予備評価として，GPU の性能と電力を評価できるシミュレータである GPGPU-Sim[2] を用いて，コア部の周波数と電圧および使用コア数を制御した場合の性能と消費エネルギーを測定し，アプリケーションの特性と合わせてその傾向について調査を行った．評価に用いた GPU は NVIDIA GeForce GTX480 に従った．ベンチマークについては，ISPASS2009 ベンチマークより，CUDA で記述された 7 つのアプリケーションプログラムを用いて評価を行った．

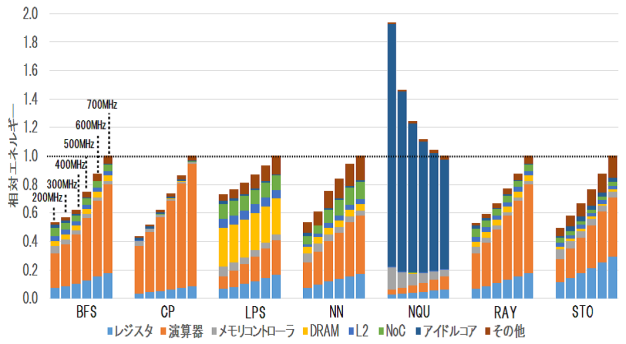

図 3: アプリケーション毎の消費エネルギーの内訳

図 3 は，各アプリケーションを動作周波数 700MHz で実行した際の消費エネルギーを基準とした相対消費エネルギー，およびその内訳を示している．

NQU を除く全てのアプリケーションにおいて，周波数を下げることにより消費エネルギーが削減されている．一方で，NQU はアイドル状態のコアの数が多く，一部のコアのみが動作するという特徴がある．そのため，周波数・電圧制御対象ドメインのスイッチング電力が小さく，周波数低下で実行時間が長くなると合計の消費エネルギーが増加してしまっている．

使用コア数を変化させた際の消費エネルギーに特に異なる傾向が見られたアプリケーションの評価結果として，NN と NQU の相対実行時間および相対消費エネルギーをそれぞれ図 4，図 5 に示す．各周波数において，右端が使用コア数 15 個で左に行くにつれて 1 個ずつコア数を減らしている．

NN は特に並列性の高いアプリケーションであり，コア数の削減に比例して実行時間が長くなるため，コア数を減らすと消費エネルギーが大きく増加してしまう．一方，NQU はコア数を減らしても実行時間に変化がなく，また，消費電力の大半がアイドルコアによるものであるため，コア数削減によりアイドル電力を大幅に削減できた結果，全体の消費電力削減に貢献している．

これらの結果から，アプリケーション毎の特性を掴み，使用するコア数や周波数を制御することによって消費電力を削減できる可能性を確認することができた．

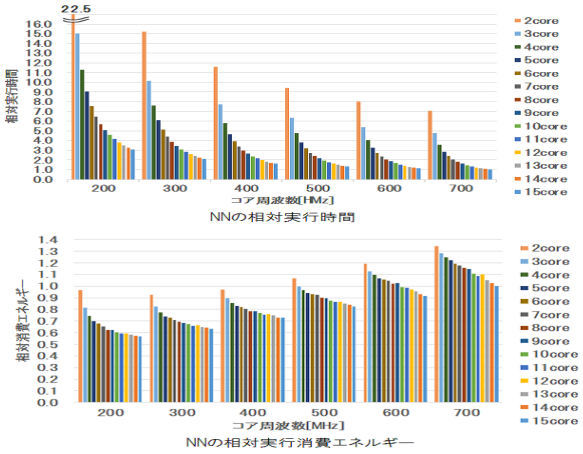

図 4: NN の評価結果

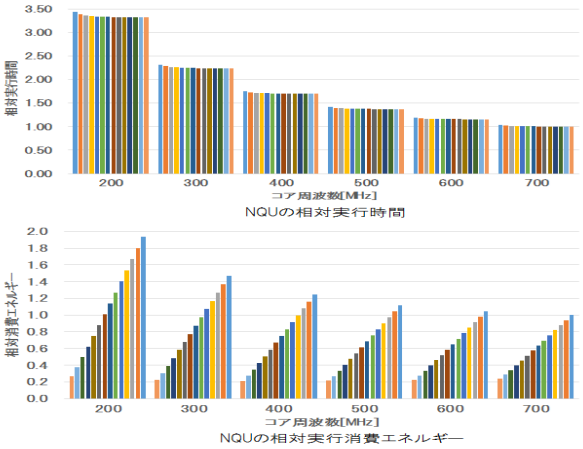

図 5: NQU の評価結果

5　まとめと今後の方針

予備評価により，アプリケーション毎に DVFS および PG の有効性を確認することができた．

今後，予備評価によって得られたアプリケーション毎の特性を踏まえ，シミュレータ上に使用コア数動的制御機構および DVFS 機構を実装し，具体的な省電力化手法を考える．特に，アイドルコア数やメモリアクセス頻度などに重点をおいて研究を進めていく．

上記の予備評価については，2013 年 12 月 16 日に行われた第 199 回計算機アーキテクチャ・第 142 回ハイパフォーマンスコンピューティング合同研究発表会（情報処理学会）で発表を行った．

参考文献

[1] J. Leng et al. GPUWattch: Enabling Energy Optimizations in GPGPUs. Proc. ISCA'13, pp.23-27, 2013.

[2] A. Bakhoda et al. Analyzing CUDA workloads using a detailed GPU simulator. Proc. ISPASS-2009, pp.163-174 , 2009.