

CUDA での高速化についての tips

2014 年 11 月 5 日

広島大学大学院理学研究科・石川健一

NVIDIA による GPU 利用環境 CUDA を使った高速化にあたって知っておくとよい点をまとめる。

行った計算：格子 QCD 計算

使用言語：CUDA

参考になるサイトや文献：

1. レジスタの効率的利用

- 配列はレジスタではなくメモリに割り当てられる。配列変数には `register` 修飾子をつけられない。レジスタを使うにはスカラー変数に `register` 修飾子を付けて使う。カーネルループ内で使いまわすデータは一度 `register` 修飾子付スカラー変数に読みだして使いまわすようなベタ書きコードにすると速くなる傾向がある。
- カーネル内部で使えるレジスタ数が不足するとレジスタ溢れが起こり効率が低下する。これを防ぐために、C 言語ではスコープをうまく使うとスカラー変数の名前が見通しがよくなり上記データ有効利用の手動最適化がしやすい。
- CUDA(GPU)には豊富なレジスタがあるが、スレッドあたりのレジスタ数は、ブロック内のスレッド数で変化する。1 ブロック内で多くのスレッドを生成する場合 1 スレッドあたりで使用できるレジスタ数は少なくなる。このため上記のレジスタ変数によるデータ有効利用（使いまわし）をレジスタ溢れなくうまく活用するには、スレッド長さやカーネル内部の複雑さをコーディングの際に調整する必要がある。また使えるレジスタ数は GPU カードの種類や世代によって異なるので最適化対象の GPU のハードについてある程度知っておく必要がある。

2. キャッシュの有効利用

- あるアドレスから読み出すデータが読み出しだけでそのアドレスのデータを変更しない、かつ、そのアドレスのデータを再び使う可能性がある場合、`tex1Dfetch` などの関数を通じてデータを読み出すと読み出し専用キャッシュである `Texture` キャッシュにデータが残り計算速度の向上に有効である（2-3 割速くなる場合有り）。

以上