2E16

グラフィックボードによる密度汎関数計算の加速

（名大院・情報科学）　　○安田耕二

【汎用計算機の現状】

電子状態計算は分子科学で不可欠な地位を占めており、最近では生体分子さえ計算対象になっている。この様な大規模計算には、多数の商用 PC をネットワーク結合したクラスター計算機が使われる。この汎用 CPU は指数関数的に高速化（2 年で 2 倍）しているが、それはプロセスルールの微細化や、多段パイプラインアーキテクチャにより、より多くの素子をより高速動作させた結果である。これに対して主記憶(DRAM)は大幅に遅いため、高速な CPU にデータを供給できない状況である。そのため CPU 内部に高速メモリー（キャッシュ）を置き、頻繁に使うデータを保存する[1]。これは多種類のプログラムを高速実行するために、汎用プロセッサには必須であるが、チップ面積のかなり（9 割）を占めている。演算を実際に行っているのは、実は CPU 上の僅かなトランジスタだけである。

【GPU とは】

科学技術計算では、プログラムの特定箇所が計算時間の殆どを占め、また演算量に比べて通信量が少ない例が良く見られる。この場合キャッシュは不要なため、演算素子のみを高密度に集積すれば、汎用 CPU より高速計算が可能である。安価で大量生産されている、その様なプロセッサの例として、GPU（graphics prosessing unit）がある。GPU とは全ての PC が備えている描画素子で、文字や図形データをディスプレーに表示する役目を持つ。コンピューターは表示すべき任意の曲面を、多数の頂点をつなぐ平面で近似している。多数の 3 次元頂点座標を、ディスプレー上の 2 次元座標に投影するため、GPU には高速な積和計算回路が、多数組み込まれている。GPU の演算回路を多目的に使う試み（general-purpose GPU）は以前から報告されていたが、極めて低精度の数しかサポートしなかったため、科学技術計算には使えなかった。

数年前 Microsoft により GPU の新しい規格が提案され、単精度浮動小数演算をサポートする事になった。NVIDIA はこの規格の GPU を開発、昨秋から販売している[2]。この GeForce 8800 GTX グラフィックボードの価格は約 8 万円で、Windows や Linux の PC のバスに挿して使う。ピーク演算性能は 300～500 GFLOPS に達し、最速の汎用 CPU の 5～10 倍である。本研究の目的は、この GPU を用いて密度汎関数（DFT）計算を高速化する事である[3]。

【アーキテクチャ，アルゴリズム】

この GPU は、16 個の独立な、単一命令流－複数データ流（SIMD）型 multiprocessor から成る（図1）。各 multiprocessor（点線）は単精度浮動小数演算器を備えた 8 個の processor と、16KB の共通メモリー、読み込み専用キャッシュから成る。これら 16×8=128 個の prosessor は 1.35 GHz で動作し、異なるデータに対して同じ命令を実行する。グラフィックボードは 768MB の外部 DRAM メモリーを持つ。共通メモリーへの読み書きは速いが、外部 DRAM メモリーへの読み書きには 500 clock cycle 以上かかり遅い。この遅延を隠すため、多数（1 processor あたり 24 個以上）の独立な計算（thread）を実行する事が推奨されている。並列計算機用に拡張された C 言語でプログラムを記述すると、付属のコンパイラーは GPU 用のデバイスコードと、GPU を制御する API 関数を生成する。これをホストプログラムと link し実行する。

この GPU を使う際の問題点は、(1)単精度浮動小数のみサポート (2)僅かな作業領域（共通メモリー）(3)細粒度で等質に数千 thread に並列化が必要な事である。汎用 CPU 用のソースコードを、この条件を満たすよう自動変換する事は現在不可能なため、アルゴリズムを再設計する必要が有る。500～1000 基底程度の密度汎関数計算では、近接 Coulomb ポテンシャル、交換相関ポテンシャルの計算時間が大半を占めるので、これらのアルゴリズムを再設計した。Gaussian03 で、6-31G 基底を用いてエネルギー計算を行う事を目標にした。

近接 Coulomb 項は Hermite Gauss 基底と Rys 求積法を用いた direct J 法[4]を用いた。この方法は McMarchie-Davidson, Obara-Saika 法より高速で、メモリー使用量も少なく、primitive Gaussian を使っても高速に計算できるという特徴がある。また値の大きな 2 電子積分のみホスト上で倍精度計算して、単精度問題を回避した。GPU での計算手順は次の通り。各 thread は自分担当の shell pair P と、全 thread に一斉放送された共通の shell pair Q に対して、2 電子積分の Schwartz 上限値を計算する。この値が $\lambda_{cut}=10^{-10}$ 以上 λ_{GPU} 未満なら 2 電子積分を並列に計算する。その後一斉放送された密度行列値を掛け、共通メモリー上の Fock に加算する。全ての Q shell pair を送ったら、積算した Fock 行列をホストに回収する。他方ホストは Schwartz 上限値が λ_{GPU} 以

上の 2 電子積分を計算し、Fockに加算する。

求積にはRys多項式の零点と重みが必要である[5]。これらはshell 4 つ組で値が変わるβ= αR^2の関数である。我々は2電子積分上限値の不等式を使い、積分の絶対誤差を 10^{-7}にする零点と重みの精度を、βの各区間で決めた。必要な精度は、各零点や重みでかなり異なった。そしてこの精度で零点と重みを近似する、SIMD型processorに適した補間表を作った。

交換相関項は (1)3 次元グリッド上で電子密度、密度勾配を求め (2)交換相関汎関数値をグリッド上で求め、(3)それにグリッド重みと基底関数の値を掛けて求和して得られる。計算時間の殆どを占める (1)と(3)のみ GPU で行い、(2)はホストで行った。基底関数やグリッドは沢山あるので、細粒度並列化は容易である。また交換相関項は比較的小さいため、単精度で計算する事にした。GPU 上の作業領域が不足したため、グリッド点での基底関数値を一部再計算して使った。更にグリッド重みの再計算をやめ、グリッドと重なりを持つ基底の選択も工夫した。

【結果】

計算精度と時間を調べるため、Taxol と Valinomycin を、PW91 functional と 3-21G や 6-31G basis を用いて計算した。ここでは Valinomycin/6-31G 基底を、ホスト計算機 (Pentium 4 2.8GHz, 2GB memory) ＋GeForce 8800 GTX グラフィックボードで計算した結果を示す。

まず全ての近接Coulombポテンシャルを GPUで計算した場合、収束したSCFエネルギー誤差は-1.1×10^{-3} a.u. (350 K)、近接Coulombポテンシャルの計算時間は 1/6 になった。他方 $\lambda_{GPU}=10^{-3}$ とすると、凡そ 10%の積分がホスト上で倍精度計算され、全エネルギー誤差は-6.7×10^{-7} a.u. (0.2 K)、計算時間は 1/3 になった。SCF 収束速度はλ_{GPU} に殆ど依存しなかった。この結果から、SCF初期反復では全計算をGPUで行い、最後の数反復のみ $\lambda_{GPU}=10^{-3}$ とすれば、精度を失う事無くGPUでDFT計算が加速できる。GPUの理論ピーク性能はホスト計算機の 60 倍以上だが、2電子積分計算はGPUで 10 倍程度しか加速できていない。GPUでは現在 2 電子積分の対称性を使っていない事、最適化が不十分な事が理由に考えられる。

次に交換相関ポテンシャルをGPUで計算すると、収束したSCFエネルギー誤差は 3.6×10^{-5} a.u. (11 K)となり、単精度で計算しても実際上問題無い事が分かった。交換相関ポテンシャルの計算時間は全体で 1/40、加速対象部分のみ比較すると 1/120 となった。今回加速しなかったグリッド重みの計算と、グリッド上での汎関数値計算が実行時間の大半に残ったが、これらもGPUで加速可能である。以上の結果からGPUでDFT計算を大幅に加速できる事が分かった。

[1] ヘネシー, パターソン「コンピュータ・アーキテクチャ」(日経 BP 社)

[2] http://www.nvidia.com

[3] K. Yasuda, J. Comp. Chem, in press

[4] C. A. White, H. Head-Gordon, J. Chem. Phys. 104, 2620 (1996).

[5] M. Dupuis, J. Rys, H. F. King, J. Chem. Phys 65, 111 (1976).

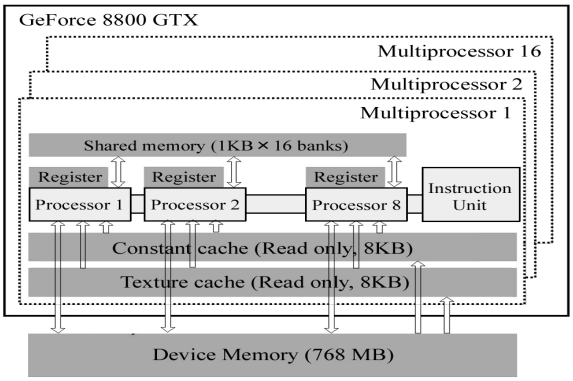

図 1：GeForce 8800 GTX チップは 16 個の multiprocessor（点線）からなる。各 MP は 8 個の processor、16KB の共通メモリ、読み出し専用キャッシュからなる。チップ外に 768 MB のメモリーを持つ。全 processor は同じコードを実行する。

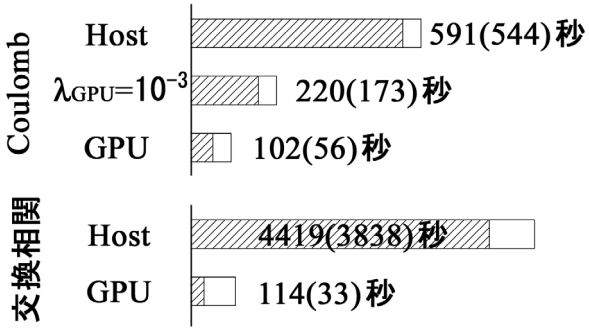

図 2：Valinomycin/PW91/6-31G の計算時間。斜線部と () 内は加速対象のみの時間。ホスト計算機は Pentium 4 (2.8 GHz), メモリー 2GB。GPU は Gaussian03 の DFT 計算を大幅に加速する事が分かる。