GTC Workshop Japan 2011 (2011 年 7 月 22 日)

NVIDIA C2050 による実用的かつ高速な倍々精度行列-行列積の実装

Implementation of a general purpose fast matrix-matrix product routine in double-double precision using NVIDIA C2050.

中田 真秀 [1)], 高雄 保嘉 [2)], 野田 茂穂 [1)], 姫野 龍太郎 [1)]

NAKATA, Maho, TAKAO, Yasuyoshi, NODA, Shigeho, and HIMENO, Ryutaro

1) 理化学研究所　情報基盤センター 2F （〒 351-0198 埼玉県和光市広沢 2-1）

2) JFE テクノリサーチ株式会社 ソリューション本部 (川崎)CAE センター (〒 210-0855 神奈川県川崎市川崎区南渡田町 1-1 京浜ビル 3F

Key Words: *CUDA, Rgemm, double-double, matrix matrix product*

1　概要

倍々精度の行列-行列積 “Rgemm” を NVIDIA C2050 上に実装した. パフォーマンスは GPU のみ, 括弧内は CPU-GPU 転送も含め, それぞれ 16.4(16.1)GFlops から 26.4(25.7)GFlops を達成し, 現在 C2050 を用いたものの中では最速である. Intel Xeon X3470 上の参照実装と比較して 160 倍から 260 倍高速になった. 実用を目標に, 転置を含めあらゆるサイズの行列にも対応した. 行列-行列積は, 多くの演算機で理論性能値に近い性能が出せるが, GPU を用いた利点は, GPU の高速性, FMA 命令のサポート, 演算密度が高く, PCIe の遅さをより隠しやすい, にある. 倍々精度 (ほぼ 4 倍精度) 演算は遅くはなく, GPU を用いると数年前の CPU の倍精度演算の速さ程度で実現できる [1].

2　はじめに

今後コンピュータにおける計算精度不足の問題が深刻になると考えられる. 例えば, ペタ, エクサスケールのシステムでは, 演算回数は一週間で 10^{20} から 10^{23} 回のオーダーで行うことになり, 倍精度演算では誤差の蓄積で誤った結果が出る可能性が考えられ, また, 規模が小さくても誤差に敏感な問題も多くあるからだ. 特に線形代数演算は様々な局面で使われるため, 高精度な汎用ライブラリの構築が重要となる.

それらを踏まえ, 著者の一人である中田は標準的な BLAS や LAPACK の高精度への拡張である, 高精度線形代数演算ライブラリ MPACK を開発してきた [2]. その中の倍々精度版は 8~24 回の倍精度演算のみでできるため手軽かつ高速なので, GPU 向けの最適化を行うことは重要かつ有用だと考えられる.

3　倍々精度と行列-行列積 Rgemm について

まず倍々精度は倍精度で 10 進約 16 桁の精度の数 2 つをそのまま配列として持つことで 10 進約 32 桁の精度が実現される. 倍々精度型の数 a は倍精度の数 a_{hi}, a_{lo} を使って

$$a = a_{hi} + a_{lo}$$

のように表す. 加算には 11 または 20 回, そして乗算は 8, 10 または 24 回の倍精度演算だけで行える [1].

行列-行列積 “Rgemm” は, A, B, C を行列とし, α, β を数とし, 次のような 4 つの行列-行列演算の中から一つを選んで行う.

$$C \leftarrow \alpha op(A) op(B) + \beta C$$

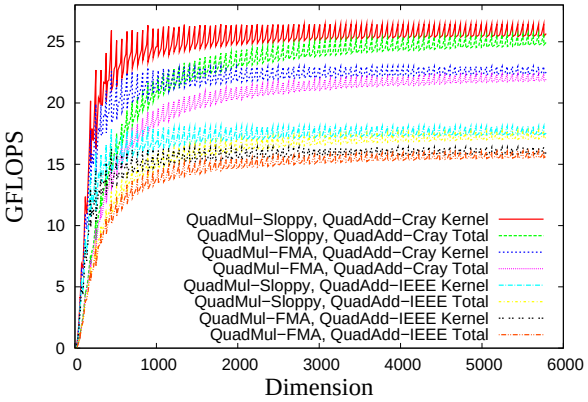

図 1: NVIDIA C2050 上での様々な演算精度を用いた倍々精度による行列-行列積 “Rgemm” のパフォーマンス. 最大で 26.4GFlops 得られた.

$op(\cdot)$ は行列を転置する, しない, を指定する.

4　Rgemm の最適化と性能評価

C2050 向けに “Rgemm” の最適化を行った. 椋木, 高橋の過去の研究と比して, C2050 向けの最適化を行い, 特にブロッキングを大きくしつつアクティブブロック数を 2 とし性能を引き出した. また Nath らの “Pointer redirecting” を実装し一般の行列サイズに対応した. いくつかの精度での倍々精度計算を行ったベンチマーク結果を図 1 示す. 結果については概要も参照のこと. ここには載せないが, 転置を行っても, テクスチャメモリから共有メモリにブロックを転送する手法を用いると, 性能劣化がほとんどみられなかった.

参考文献

[1] 中田真秀, 高雄保嘉, 野田茂穂, 姫野龍太郎:GPU による倍々精度行列-行列積の高速化, 計算工学講演会論文集, Vol.16, 2011. (印刷中, および引用文献参照)

[2] Nakata, M., “The MPACK (MBLAS/MLAPACK) a multiple precision arithmetic version of BLAS and LAPACK”, http://mplapack.sourceforge.net/, 2008-2010.