様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　　元 年　　６ 月 ２０ 日現在

機関番号：３２６５２
研究種目：若手研究(B)
研究期間：2016～2018
課題番号：１６Ｋ１６０６２

研究課題名（和文）高性能・省電力な計算のための短尺浮動小数点表現の検討

研究課題名（英文）Reduced-precision formats for high-performance and energy-efficient computations

研究代表者
椋木　大地（Mukunoki, Daichi）

東京女子大学・理学研究科・研究員

研究者番号：９０７４２２８９

交付決定額（研究期間全体）：（直接経費）　　1,700,000円

研究成果の概要（和文）：本研究では数値計算において広く用いられている32/64ビットのIEEE浮動小数点フォーマットに対して，ビット長が短い短尺フォーマットを導入することにより，計算の高速化と省電力化が可能であるかを検討した．ソフトウェアによる軽量な実装方法を検討するとともに，主にGPUをターゲットとして，数値計算に用いられる基本的な線形計算カーネルで性能がデータアクセス律速となるものにおいて，計算速度と電力性能の両面での有効性を示した．

研究成果の学術的意義や社会的意義
浮動小数点表現を工夫することによる計算の高性能化は，さまざまな計算処理に応用できる汎用的なアプローチである．本研究の開始時と比べると，現在では本研究と類似のアプローチを試みた研究がいくつか見られるようになり，本研究の着想や議論には意義があったと考えられる．また，本研究期間内には達成できなかった混合精度計算手法への適用や，近年発展が著しいFPGAへの適用が今後期待される．

研究成果の概要（英文）：This study explored the possibility of reduced-precision formats which have shorter bit length against the IEEE 32/64 bit floating-point for enchance the performance of numerical computations in terms of both computation speed and energy efficiency. We proposed a light-weight implementation of reduced-precision formats on software and demonstrated the performance improvement, in terms of both speed and energy efficiency, on some data-intensive operations on basic linear algebra.

研究分野：高性能計算

キーワード：Reduced-precision　Mixed-precision　GPU

様　式　C－１９、F－１９－１、Z－１９、CK－１９（共通）

１．研究開始当初の背景

計算機ハードウェアの進化においてメモリや通信といったデータ移動に関連するコンポーネントの性能向上は，演算器の性能向上と比べるとペースが遅い．また演算器の動作周波数の頭打ちでシステムは高並列化に向かい，コア間・ノード間等のデータ移動の必要性が増している．そのためアプリケーション性能はデータ移動のコストに律速しやすい状況が生まれている．また，大規模計算機の開発において，その消費電力がシステムの性能を制約する要因となりつつあるが，データ移動には演算と比べて多くの電力を消費する．さらに近年では高性能計算のアプリケーションとして，データマイニングや機械学習といった多量のデータアクセスを要する処理が行われるようになった．このような背景から，データ移動やデータ表現の最適化に着目したプログラムの性能向上策が今後ますます重要となると考えられる．

一方で，科学技術計算プログラムの多くは，IEEE 754-2008 に基づく 32 ビット・64 ビット浮動小数点形式（FP32・FP64）によるデータ表現が用いられている．しかし，必要な精度での計算結果を得るために計算途中で必要となるデータ表現の精度は，この 2 つの浮動小数点型の精度とは必ずしも合致しないことが容易に想像できる．つまり FP32・FP64 型を用いたプログラムはある程度の不要な情報を含みながら計算を実行しており，データアクセスに無駄な時間や電力消費が発生している可能性があると言える．そこで，FP64 による計算において部分的に FP32 を利用する混合精度計算が研究されているほか，精度をそれほど必要としない深層学習計算において 16 ビット浮動小数点形式（FP16）の活用が検討されている．

２．研究の目的

本研究の目的は，FP64・FP32・FP16 の 3 段階の浮動小数点精度をさらに多段階化（言い換えれば FP64・FP32 に対してビット長を減らした短尺形式を導入）することで，数値計算プログラム中のデータ表現を最適化することにより，計算の高速化と消費電力の削減が可能であるかを検討することである．CPU・GPU 等の一般的な汎用プロセッサ環境を対象として，ソフトウェア的に実現可能な方法を検討する．そして主にデータインテンシブな数値計算カーネルにおいて，短尺形式を用いたプログラムの実装方式と性能・効果を検討する．

３．研究の方法

まず短尺形式の最も簡便な実現方法として，図 1 に示すような方法を検討した．この方法では 8/16/32 ビット長の整数型からなる構造体を用意し，そこに IEEE 形式（FP32・FP64）の符号部+指数部と仮数部のビット列を上位から入るだけ格納して，入りきらない仮数部は切り捨てる．これにより FP64 ベースで 6 段階，FP32 ベースで 2 段階，合計で IEEE 形式を含めて 11 段階の精度を表現することができる．これらの短尺形式は基本的に格納専用として，CPU・GPU 等の汎用プロセッサにおいては，ベースとなる IEEE 形式に応じて，FP32・FP64 演算器を使用して演算する．すなわち，DRAM メモリ上の短尺形式データはレジスタ上で FP32・FP64 型に変換され，FP32・FP64 演算器で計算し，再びレジスタ上で短尺形式に変換してメモリに書き戻される．IEEE 形式と短尺形式の変換は単純な論理演算命令によるコードで実装できるため，近年の演算器集積度が高いプロセッサでは性能ボトルネックとならないことが期待できる．なお，この方法で配列を構成する場合は，メモリアラインメントの観点から Structure of Array（SoA）形式でメモリ上に配置することとする．

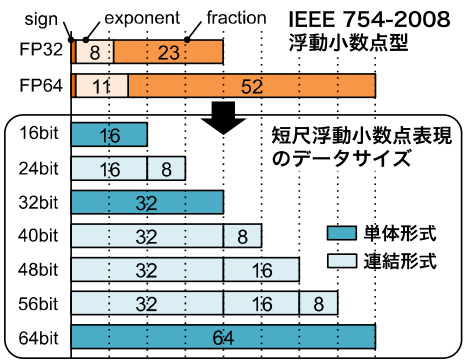

図1. 短尺浮動小数点表現

本研究ではまず，この方法に基づいていくつかの基本的な数値計算カーネルを実装し，その性能を検討した．また研究計画の段階では，分散並列プログラムへの適用や，4 倍・8 倍精度演算への適用，その他の表現・実装方式の検討も研究内容に盛り込んだが，後述のように実際に実施できた研究は一部にとどまった．

４．研究成果

(1) 基本線形代数演算への適用

本研究期間の開始に先立ち，3 に述べた手法を CPU・GPU 上の AXPY（ベクトルのスカラ倍と加算）および GEMM（行列積）に適用し，実行時間および電力性能を評価した（参考文献[1]）．AXPY では概ね期待通りの速度向上が確認できた一方で，CPU + OpenMP 環境において性能が不安定となるケースがあり解決の目処が立たなかったことや，CPU を対象とした類似研究が発表された（参考文献[2]）ことから，以降は GPU（CUDA）を対象に研究を進めた．その後，GEMV（行列ベクトル積），CSRMV（CSR 形式による疎行列ベクトル積）を評価に加え，IEEE 形式から短尺形式への変換時の簡易的な最近接偶数丸めの実装（ただしこの方法は二重丸めの問題がある）を行い，コンシューマ向け，HPC 向け，エンベデッド向けの 3 種類の GPU デバイスにおいて，計算速度および消費電力の評価を行った[雑誌論文②]．図 2 に NVIDIA Tesla K20 における GEMV・CSRMV の結果を示す．GEMV では短尺形式の使用による実行時間の削減と，それによる電力性能

の改善が確認できた一方で，CSRMV はメモリ律速であるものの，行列データへの間接参照に必要となるインデックス行列へのアクセスコストにより，短尺形式の効果は限定的であった．また特に複数ワードで構成される短尺形式では期待した性能が得られないケースが見られた．この原因には形式変換コストのほか，メモリアクセスに要する命令数の増加（すなわち 1 命令で移動できるデータ量の減少）や，キャッシュ効果の減少などが考えられる．なお，この評価に用いたコードは研究代表者が過去に開発を行った自動チューニング機構付き BLAS カーネル群 MUBLAS（参考文献[3]）に基づいている．短尺形式のサポートにおいては演算精度・データ型の異なるカーネル群を大量に実装・最適化する必要があるため，本研究では性能最適化への自動チューニングの活用をテーマに盛り込んだが，この事例においてはその有効性が発揮された．

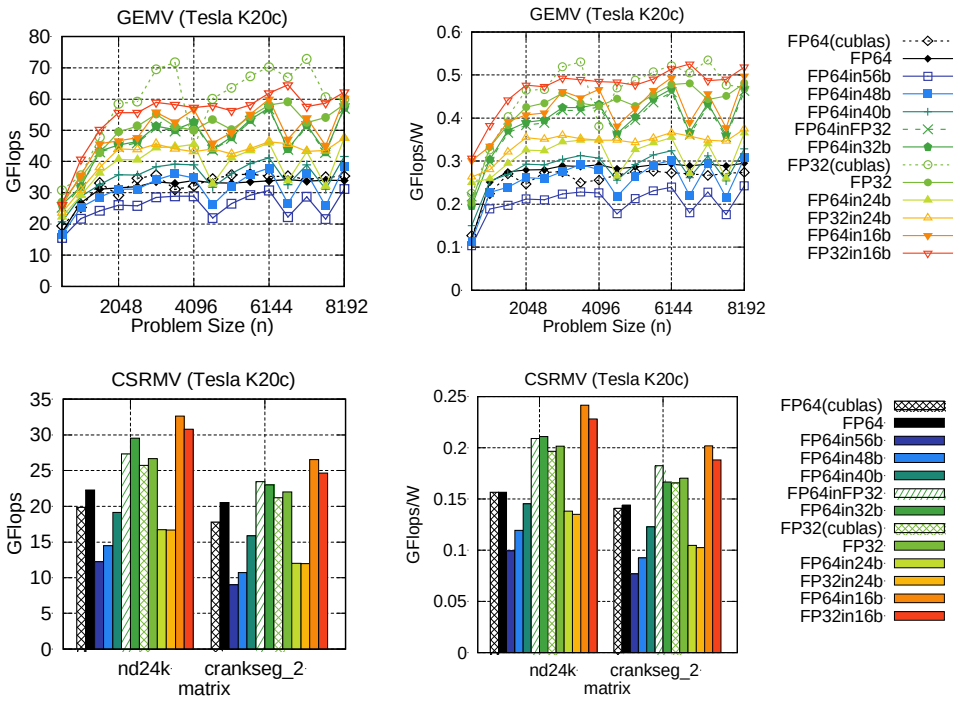

図 2. GPU（NVIDIA Tesla K20c）における GEMV・CSRMV の性能評価．この結果において，例えば「FP64in48b」は，FP64 を 48 ビットフォーマットに格納し，演算は FP64 で実行されている．CSRMV の疎行列は SuiteSparse Matrix Collection（https://sparse.tamu.edu）より取得．

(2) その他計算への適用

高精度行列積計算手法の一つである尾崎スキームにおいて，計算に必要な中間データの格納に短尺形式の適用を検討した[雑誌論文①]．この事例では当初 3 に示した方法を検討したが，データの指数部レンジが限定されていたため，最終的には指数部シフトにより固定小数点的に FP64 データを 32 ビット整数型に格納するという手法をとった．この方式のデータを FP64 で演算する GEMV カーネルを実装した結果，変換オーバーヘッドは十分に隠蔽され期待通りの性能が得られた．この他，非圧縮流体解析コードのベンチマークプログラムである姫野ベンチマークなどへ 3 に示した方法の適用検討も行ったが，短尺形式の効果を得るには元のコードが十分にメモリインテンシブでメモリバンド幅律速である必要があり，やや複雑なコードでは期待した効果が得られなかった．特に複数ワードからなる短尺形式は図 2 の CSRMV の結果と同様に期待した性能が得られず，メモリアクセスの最適化や型変換のコスト削減が課題となった．

当初の研究計画と比べると本研究の成果はやや乏しいと言わざるを得ないが，提案手法の有効性を示すことが技術的に予想以上に困難であったということに加え，本研究開始以降に本研究と類似のアプローチによる研究事例がいくつか見られるようになり，研究の方向性を修正したことが理由として挙げられる．例えば SIMD 型 CPU において短尺形式の実装・評価を行った研究（参考文献[2]）は本研究とほぼ同一の内容を先行しており，さらに応用例としても 21 ビット表現を含む複数の精度を活用した地震シミュレーションの混合精度計算（参考文献[4]）などが行われた．また深層学習向けに，本研究の FP32in16b と同様に仮数部の長さを除いて FP32 と互換性がある bfloat16（BF16）が提案され，ハードウェア実装も見られるようになった．このような関連研究の動向は，本研究の着想自体を肯定するものであるが，本研究の新規性を主張することが難しくなった．そこで研究期間中盤からは，コード中の必要精度を自動的に決定する方法の検討や，FPGA とのヘテロジニアス計算環境における短尺形式の実現方法の検討を開始した．前者は短尺形式や混合精度計算を実用化するために不可欠であるものの，まだ研究が成熟しているとは言い難い．また FPGA は任意の仮数部・指数部を持った浮動小数点演算をハー

ドウェア実装可能であるが，性能に関する議論やホストとのデータ交換手法などについて研究が不足している．これらについてはまだ具体的な成果は創出できていないが，これまでに関連する研究者と議論を行い，その結果として，本研究代表者が代表の科研費若手課題「超並列計算環境のための高精度かつ再現性のある行列計算ライブラリの開発」(#19K20286，2019-2021年度)の申請・採択に結びついた．本研究における成果と議論の一部は今後も新課題において発展させる．総じて，本研究期間内に発表できた具体的研究成果はやや乏しいと言わざるを得ないが，着想自体の妥当性は確認され，本研究においても一定の評価・応用例を示せた点，そして今後の研究に発展する議論が蓄積できたという点において，本研究の意義があったと結論付ける．

＜引用文献＞

[1] 椋木大地，今村俊幸，短尺浮動小数点形式の検討，情報処理学会研究報告：ハイパフォーマンスコンピューティング(HPC)，Vol．2015-HPC-152，No．4，pp．1-10，2015．
[2] A．Anderson，D．Gregg，Vectorization of Multibyte Floating Point Data Formats，Proc．2016 International Conference on Parallel Architectures and Compilation (PACT ’16)，pp.363-372，2016．
[3] D．Mukunoki et al.，Automatic Thread-Block Size Adjustment for Memory-Bound BLAS Kernels on GPUs，Proc．IEEE 10th International Symposium on Embedded Multicore/Many-core Systems-on-Chip (MCSoC-16)．pp．377-384，2016．
[4] T．Ichimura et al.，A fast scalable implicit solver for nonlinear time-evolution earthquake city problem on low-ordered unstructured finite elements with artificial intelligence and transprecision computing，Proc．International Conference for High Performance Computing，Networking，Storage，and Analysis (SC18)，pp.49:1-49:11，2018．

５．主な発表論文等

〔雑誌論文〕（計２件）

① 椋木大地，荻田武史，尾崎克久，Level-3 BLAS に基づく高精度行列積計算法による高精度かつ再現性のある BLAS ルーチンの実装とその最適化，情報処理学会研究報告：ハイパフォーマンスコンピューティング(HPC)，Vol．2018-HPC-166，No．9，pp．1-8，2018，http://id.nii.ac.jp/1001/00191328/（査読無）．
② Daichi Mukunoki，Toshiyuki Imamura，Reduced-Precision Floating-Point Formats on GPUs for High Performance and Energy Efficient Computation，Proc．IEEE International Conference on Cluster Computing (Cluster 2016)，pp．144-145，DOI：10.1109/CLUSTER.2016.77，2016（査読有）．

〔学会発表〕（計４件）

① Daichi Mukunoki，Takeshi Ogita，Katsuhisa Ozaki，OzBLAS：Accurate and Reproducible BLAS Based on Ozaki Scheme，GPU Technology Conference (GTC 2019)，San Jose McEnery Convention Center，San Jose，USA，March 17-21，2019．
② 椋木大地，今村俊幸，Reduced-/Extended-precision BLAS の実装方法の検討，第５回 大規模並列数値計算技術に関する研究集会，RIKEN AICS，神戸，2017 年 3 月 27 日．
③ Daichi Mukunoki，Toshiyuki Imamura，Daisuke Takahashi，Implementation Techniques for High Performance BLAS Kernels on Modern GPUs，SIAM Conference on Computational Science and Engineering (CSE17)，Hilton Atlanta，Atlanta，USA，February 28，2017．
④ 椋木大地，今村俊幸，高橋大介，Pascal アーキテクチャ GPU における線形計算カーネルの実装技術の検討，GPU Technology Conference Japan (GTC Japan 2016)， ヒルトン東京お台場，東京，2016 年 10 月 5 日．