# PU/GPU/FPGA における Reduced-/Extended-Precision BLAS ルーチンの実装と評価

## Implementation and Evaluation of Reduced-/Extended-Precision BLAS Routines on CPUs/GPUs/FPGAs

今村俊幸
理化学研究所計算科学研究センター

１．研究目的

IEEE 754 で規定される各種浮動小数点数に加え，ベンダ独自の低精度表現が実装された例もよく知られている(Bfloat16, TF32 等)．新しい浮動小数点表現のサポートとあわせて，複数精度を混用した混合精度計算法についてもその重要性が高まっている．本研究プロジェクトでは，数値計算ライブラリにおいてもこれら拡張・混合精度を可能とするソフトウェア部品の調査，それらを設計実装する上で必要となる技術要件について議論し実装を通じて検証を進める．

特に，低精度から高精度までのさまざまな精度およびそれらの混合精度（データ精度と算術演算精度が異なるような場合）をサポートする BLAS ルーチン（行列積等）を，最新のメニーコアプロセッサ（CPU および GPU）上に実装し，その性能を評価する．高精度演算として，(1) 浮動小数点演算ベースの手法（double-double 演算（Dekker 1971）等），(2) 整数演算を用いた手法（GCC/ICC の binary128 エミュレーション等），(3) 内積・行列積に特化した手法（尾崎スキーム（Ozaki et al. 2012）等）を実装する．いくつかの実装はすでにプロジェクトメンバーらによる過去の研究で開発されているものを使用する．

２．研究成果の内容

本年は主に 4 つの成果としてまとめられる．

1) DD や整数などによる拡張演算をサポートする BLAS を生成するための基盤研究並びに SYMV(対称行列ベクトル積)に適用した高性能 GPU カーネルの作成.
2) 表現データと演算データ型の異なるタイプを実現するソルバ構築手法の提案
さらに同アイディアを実現した低精度データ向けメモリアクセサ版疎行列積ルーチンの評価等にも活用
3) テンプレートを活用した真の混合精度演算 BLAS(tmBLAS)のプロトタイプを作成
4) IEEE754 とは異なる新しい数値フォーマット POSIT 演算の GPU 実装評価

３．学際共同利用プログラムが果たした役割と意義

MCRP は，私たちが必要とする最新鋭の CPU と GPU 環境を提供する極めて有用なプラットフォームである．従来技術と新技術の融合を可能とし実証するためには書くことができなかった．特に Pegsus が搭載する GPU, CPU は国内外でも本プログラム以外では利用は困難であった．また，本プロジェクト提案を通じて将来的な他分野との共同アプリケーションの機会を見出すことが可能となり，分野横断科学を創出する重要な機会(Venue)の提供を行ってもらったと考えている．これらの意義は，他のスーパーコンピュータ公募事業とは異なるものであり，高く評価できる．

４．今後の展望

本年度は設定した研究内容について，既存ソフトウェアの拡張ならびにプロトタイプの作成が実現されている．実用に供する部分に対する追加支援，特に更なる最適化が必要なものに対しては早急に対応する．更に，DD フォーマットから正規化処理を除いた PairArithmetic ならびにその 3 語ならびに 4 語版への拡張を進めたい．可能であれば浮動小数点だけでなく整数も組み合わせた真の混合演算による高精度演算機構の実現を CPU+GPU で実現するのみではなく，FPGA などの他プロジェクトと連携して進めていきたい．

５．成果発表

(1) 学術論文

- "Evaluation of POSIT Arithmetic with Accelerators", N. Nakasato, Y.Murakami, F. Kono, & M. Nakata: The International Conference on High Performance Computing in Asia-Pacific Region, 2024, Jan 25 - 27, Nagoya, Japan https://doi.org/10.1145/3635035.3635046
- “Sparse Matrix-Vector Multiplication with Reduced-Precision Memory Accessor”, Daichi Mukunoki, Masatoshi Kawai & Toshiyuki Imamura: 2023 IEEE 16th International Symposium on Embedded Multicore/Many-core Systems-on-Chip (MCSoC), DOI:10.1109/MCSoC60832.2023.00094

(2) 学会発表

- "tmBLAS：a Mixed Precision BLAS by C++ Template", A. Suzuki, D. Mukunoki, T. Imamura: Research Poster, ISC2023, Hamburg, Germany. May 2023
- "Evaluation of various arithmetic for linear algebra on GPU and FPGA", N. Nakasato: ICIAM 2023 TOKYO, Mini-symposium, [01060] Exploring

Arithmetic and Data Representation Beyond the Standard in HPC, 2023 August 20 - 25, Tokyo, Japan

- “Multiple- and Mixed-Precision BLAS with C++ Template”, Toshiyuki Imamura, Daichi Mukunoki, Atsushi Suzuki, 10th International Congress on Industrial and Applied Mathematics (ICIAM TOKYO), Waseda University, Tokyo, Japan, August 20-25, 2023

(3) その他

以下のオープンソースソフトウェアを公開中

- tmBLAS: GitHub - RIKEN-RCCS/tmblas https://github.com/RIKEN-RCCS/tmblas
- mX_real: GitHub - RIKEN-RCCS/mX_real: Yet another C++ compound multiprecision https://github.com/RIKEN-RCCS/mX_real
- RpFp: Large-scale Parallel Numerical Computing Technology Research Team, R-CCS (riken.jp) https://www.r-ccs.riken.jp/labs/lpnctrt/projects/rpfp/

| 使用計算機 | 使用計算機に<br>○ | 配分リソース※ | |
|---|---|---|---|
| | | 当初配分 | 追加配分 |
| Cygnus | | | 0 |
| Pegasus | ○ | 19000 | 0 |
| Wisteria/BDEC-01 | | | |
| ※配分リソースについてはノード時間積をご記入ください. | | | |