CPU/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

代表者：今村俊幸
所属：理化学研究所計算科学研究センター

１．研究目的

従来、HPC システムにおける数値計算には、IEEE 754 の単精度（FP32）および倍精度（FP64）の浮動小数点形式が標準的に用いられてきた。しかしながら、FP64 以上の高精度が求められる特定の応用においては、MPFR や QD などの高精度演算ライブラリが開発され、計算精度の向上が図られてきた。IEEE 754-2008 では四倍精度（FP128）および半精度（FP16）が新たに導入され、FP128 は IBM POWER9 / Power 10 にてハードウェア実装され、FP16 は深層学習分野において GPU 等で活用されるに至った。さらに、BF16 や TF32 といった独自の低精度浮動小数点形式も登場し、ハードウェアおよびコンパイラによる新たな数値表現の支援が進展している。このような技術の発展を背景に、数値計算分野においては混合精度計算法の研究が進み、複数の精度を活用する数値計算ライブラリの開発が活発化している。線形代数演算ライブラリにおいては、高精度演算ライブラリを組み込んだ MPLAPACK/MPBLAS、混合精度演算に対応する XBLAS、C++テンプレートによる混合精度対応 BLAS++などが開発されているが、並列化や性能最適化の余地を残している。特に FP64 を超える演算精度については、x86 や GPU 上ではソフトウェア実装を必要とするため、BLAS レベルのみならず算術演算の最適化が今後の課題となる。本プロジェクトでは、低精度から高精度までのさまざまな混合精度 BLAS ルーチンを最新の CPU/GPU 上に実装し、その性能評価を行う。また、高精度演算の実装技術として、浮動小数点演算ベースの手法（DD 演算）、整数演算を利用する手法（binary128 エミュレーション）、内積・行列積に特化した尾崎スキームなどを検討し、混合精度演算のさらなる発展に寄与することを目指す。

２．学際共同利用プログラムが果たした役割と意義

本研究プロジェクトは混合精度演算技術を基盤とし、その中心をなす低精度演算器の応用技術の発展、さらには低精度演算を活用した高精度演算の実現を目的とする。複数のデータ形式を並列に用いる複合型演算方式について、従来の正規化を要する完全方式に加え、正規化を省いた Pairwise 方式及び疑似方式の実装を完遂した。また、尾崎スキームに基づくエラーフリー変換を利用した高精度演算から低精度フォーマッ

トへのマッピングについて、本学際共同利用プロジェクトの計算資源（H100GPU）を用いて性能検証の一環として一部の実験を実施した。当該研究は、AI 演算能力の向上を目指す CPU 及び GPU の進化と相乗効果をなし、行列計算の分野における技術革新を促進するものと期待される

３．今後の展望

本研究プロジェクトは科研費基盤（B）「任意混合精度演算を基盤とする次世代浮動小数点数環境の構築」（２０２５年度から３か年）に採択された。審査過程において、本研究課題でとりあげる、複合型演算方式と尾崎スキームに対する国際競争力が高く評価されており、研究資源を集中してより汎用で強力な混合精度演算のソフトウェアならびにハードウェア技術の継続的な実践をすすめる。特に、尾崎スキームについては米国の GPU や TPU ベンダーを中心とした共同研究の可能性が十分にある。また、今後のプロセッサ設計にも強い影響力を与える技術革新ともいえるため、プロジェクトの大規模化も含めた再コーディネートも丁寧にすすめていく。

４．成果発表

(1) 学術論文：発表論文はなし

(2) 学会発表：６件

- 「様々な浮動小数点演算形式の評価プラットフォームとしての FPGA と GPU の比較」 情報処理学会 第 198 回 HPC 研究会, 2025, 3 月 19 日, 札幌, 中里直人, 河野郁也, 中田真秀
- “Portable Batched EigenSolver for multiple GPU environments,” 16th Joint Laboratory for Extreme-Scale Computing Workshop(JLESC16), 2024, 4 月 16 日, Kobe, Toshiyuki Imamura
- 「非正規化合成浮動小数点数を用いた計算方法の実装方法について」 日本応用数理学会 2024 年度年会, 2024, 9 月 16 日, 京都, 今村 俊幸, 尾崎 克久
- 「mX_Real: 非正規化合成浮動小数点パッケージについて」情報処理学会 第 197 回 HPC 研究会, 2024, 12 月 17 日, 沖縄, 今村 俊幸, 尾崎 克久
- 「INT8 Matrix Engine を用いた尾崎スキームに対する高速化」第 21 回日本応用数理学会研究部会連合発表会, 2025, 3/5~7, 岡山, 内野 佑基, 尾崎 克久, 今村 俊幸
- “mX_Real, yet another multi-component floating point package,” 2025 Conference on Advanced Topics and Auto Tuning in High-Performance Scientific Computing (ATAT2025), 2025, 3/21~22, Tainan, Toshiyuki Imamura (*invited)

(3) その他：なし

使用計算機	使用計算機に○	配分リソース※		
		当初配分	移行*	追加配分
Cygnus	○	280	0	0
Pegasus	○	1575	0	0
Wisteria/BDEC-01	N/A	0		0
	※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。 ＊バジェット移行を行った場合、「+2000」「-1000」のように記入			