jh250019

次世代計算機の潜在能力を引き出すための科学技術アプリケーションの刷新

横田理央（東京科学大学）

これからの半導体は深層学習分野の巨大な需要に応えるべく Tensor Core のような低精度行列演算器にますます特化することが予想される。欧米では以前からスパコンに GPU が導入されているため主要アプリケーションの GPU 化は既に行われているが、Tensor Core 化まで行われているものはまだ少ない。国産の計算科学アプリケーションの Tensor Core 化に今から着手すれば、そこから世界を牽引する研究が多く派生することが期待される。本研究の目的は、これから主流になる深層学習向けのプロセッサに対し、代表的な６つの計算科学アプリケーションの高速化を世界に先駆けて行い、その技術を国内に広く波及させることで、次世代計算基盤における科学技術成果創出の最大化を図ることである。

1. 共同研究に関する情報

⑴ 共同利用・共同研究を実施している拠点名

北海道大学 情報基盤センター
東京科学大学 情報基盤センター
京都大学 学術情報メディアセンター

(2) 課題分野

大規模計算科学課題分野

(3) 参加研究者一覧と役割分担

横田 理央(全体統括・分子アプリ)
Mohamed Wahib（医療アプリ）
西澤 誠也（気象アプリ）
岩下 武史（電磁気アプリ）
伊田 明弘（密行列ライブラリ）
金森 逸作（量子アプリ）
芝 隼人（材料アプリ）
深谷 猛（疎行列ライブラリ）
尾崎 克久（低精度からの精度回復）
Shiyao Xie（密行列ライブラリ）
Jiamian Huang（N 体ライブラリ）

2. 研究の目的と意義

これからの半導体は深層学習分野の巨大な需要に応えるべく Tensor Core のような低精度行列演算器にますます特化することが予想される。H100 から B100 にかけて Tensor Coreのみの性能が向上しており､FP64 やFP32 の性能はむしろ低下している。つまり、Tensor Core が利用できないアプリケーションコードは今後半導体の進歩による性能向上が期待できないことを表している。本研究の目的は、これから主流になる深層学習向けのプロセッサに対し、代表的な６つの計算科学アプリケーションの高速化を世界に先駆けて行い、その技術を国内に広く波及させることで、次世代計算基盤における科学技術成果創出の最大化を図ることである。2025 年度は、気象・電磁気・分子・量子・材料・医療のそれぞれのアプリケーションに対して、まずは Tensor Core を用いない GPU 化を行う。また、できるものに関しては Tensor Core を活用できるよう密行列積の形にアルゴリズムや離散化手法を改良し CPU 実装を行う。

3. 当拠点の公募型共同研究として実施した意義

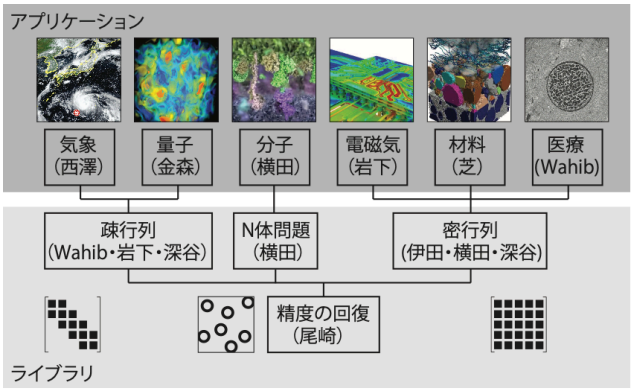


図 1 研究開発の体制

本課題は東京科学大学、京都大学、北海道大学、芝浦工業大学、兵庫県立大学、理化学研究所、海洋研究開発機構の間での共同研究であり、気象・電磁気・分子・量子・材料・医療のアプリケーション開発者と密行列・疎行列・N 体問題のライブラリ開発者、および GPU スパコンの運用担当者を含む体制となっている。これは公募型共同研究として実施するのに相応しい研究課題および研究体制であるといえる。特に、Tensor Core を有する GPU スパコンを運用しているセンターとの共同研究を行うことは本課題を進める上で必須となる。

4. 前年度までに得られた研究成果の概要

該当なし

5. 今年度の研究成果の詳細

2025 年度は気象・電磁気・分子・量子・材料・医療のそれぞれのアプリケーションに対して、まずは Tensor Core を用いない GPU 化を行った。また、Tensor Core を活用できるよう密行列積の形にアルゴリズムや離散化手法を改良し CPU 実装を行った。ここでは、それぞれのアプリケーションごとに研究成果の詳細を述べる。また、疎行列・密行列・N 体問題・精度回復のライブラリに関しても研究成果の詳細を述べる。

5.1 気象アプリケーション（西澤）

2025 年度は、気象モデルにおける Tensor Core 活用の可能性を明らかにするため、代表的アプリケーションとして SCALE-RM を対象に、計算構造の解析および GPU 上での性能特性の調査を行った。まず、既存 GPU 実装の性能評価およびボトルネック解析を実施し、力学コアにおけるフラックス計算や鉛直方向解法、ならびに物理過程における放射・雲微物理計算が主要な計算負荷を占めることを確認した。

これらの計算を行列・テンソル的観点から再整理し、Tensor Core 適用の観点で演算パターンを分類した結果、有限体積法におけるフラックス計算は小規模な密行列演算のバッチ処理として再構成可能であり、また HE-VI 法（Horizontally Explicit Vertically Implicit 法）における多数の三重対角行列解法もバッチ処理により高スループット化の余地があることを示した。さらに、同部分について、低精度演算と反復改良（iterative refinement）を組み合わせた手法の検証も行ったが、三重対角系は直接法で解けるため反復回数自体が少なく、低精度化による単発の高速化よりも反復に伴うオーバヘッドが支配的となり、全体性能の向上は得られなかった。さらに、不連続ガラーキン法（Discontinuous Galerkin 法、 DG 法）のような高次離散化手法では要素内演算が本質的に密行列積として表現されるため、Tensor Core との親和性が高いことを明らかにした。

一方で、セルごとに係数行列が異なることによるデータ配置の複雑さや、低精度演算に伴う数値誤差の蓄積、メモリ帯域律速といった実装上および数値的課題も整理した。以上により、気象モデルにおいてもアルゴリズム再設計により Tensor Core を有効活用できる可能性を示した。今後は、DG 法の GPU 実装および Tensor Core を用いたバッチ密行列演算の導入を進め、実アプリケーションにおける性能向上と精度影響の定量評価を行う予定である。

5.2 量子アプリケーション（金森）

量子アプリケーションでは、素粒子であるクォークの振る舞いを量子力学に基づいて取り扱う格子 QCD アプリに取り組んでいる。格子 QCD アプリでは、計算時間の大部分がディラック方程式と呼ばれる偏微分方程式の解法に費やされ、メモリ及び通信帯域律速になっている。カーネル部の構造は離散化の詳細によるが、格子点上に 12～100 程度の複素数が乗る 9 点ステンシルになっている。そのため、テンソルコアによる高速化は難しい。一方で、単精度演算との混合精度による高速化は広く用いられている。そこでメモリ帯域の効率的な利用を目的として、GPU がサポートする半精度演算を用いた場合の反復解法の安定性について検討した。

ベースラインには BiCGStab 法を用いた倍精度ソルバーと FP32 での BiCGStab 法に FP64 の残差反復を組み合わせた混合精度ソルバーを選び、FP32 のかわりに FP16 を用いた混合精度ソルバーについて調査した。カーネルには、最も単純なカーネルの一つであるウィルソン型のものを用いた。単純に FP32 でのソルバーを FP16 でのソルバーに置き換えるだけでは、系が大きくなると反復数が大幅に増えてしまい高速化はできなかった。FP16 では、反復法の作業ベクトル（解ベクトル・残差ベクトルを含む）の多くの成分が、途中でアンダーフローによって 0 になっていた。残差ベクトルもアンダーフローによって正しい値より小さくなるため、本来は収束条件を満たすべきではない解で BiCGStab 法が打ち切られていた。そこで、適切なリスケーリングをソルバーの中に組み込むことで、アンダーフローを回避する手法を提案した。FP16 が利用可能な CPU（A64FX）でのベンチマーク結果は良好であり、GPU での実装を準備中である。

5.3 分子アプリケーション（横田）

分子動力学（MD）における長距離クーロン相互作用の計算に対し、従来広く用いられてきた PME 法に代わり、FMM（Fast Multipole Method）の実用化と GPU 最適化を進めた。特に、GROMACS への統合を通じて、既存の近接相互作用計算カーネルと FMM による遠距離相互作用を組み合わせたハイブリッド構成を採用し、精度と性能の両立を図った。

この際に、FMM の計算ボトルネックである M2L（Multipole-to-Local）演算を密行列積（GEMM）として再定式化することで、Tensor Core を活用可能な形へ変換した。これにより、従来の FFT ベース手法では活用できなかった低精度高スループット演算資源を有効に利用できるようになった。また、空間分割構造を GPU 上で構築し、通信を局所化する Local Essential Tree（LET）により、大規模並列環境での通信ボトルネックを大幅に削減した。さらに、MD 計算における時間発展の連続性に着目し、ツリー構造の更新頻度を pair-list 更新と同期させることで、FMM の定数コストを削減した。加えて、CUDA ストリームを用いた非同期実行により、近接相互作用計算と遠距離相互作用計算の重畳実行を実現し、GPU 利用効率を向上させた。性能評価では、数千万原子規模の大規模系において PME を上回るスケーラビリティを示し、最大で約 2。5 倍の高速化を達成した。**GROMACS の PME は極限までチューニングされており、それに対する優位性を示すことができたことは意義深い。**

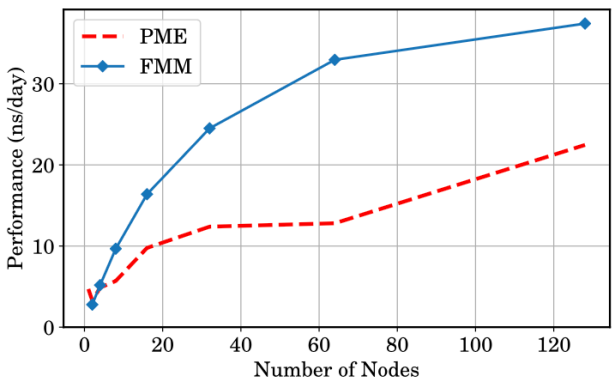


図 2 PME と FMM の GROMACS におけるスループット (ns/day)

5.4 電磁気アプリケーション（岩下）

一般に計算電磁気学でよく用いられる解法として、有限要素法、FDTD 法、境界要素法が挙げられる。本研究課題では、このうち境界要素法による電磁場解析に焦点をあてる。境界要素法は無限遠を含む解析対象を高精度に扱うことができる点などに優位性がある。

境界要素解析では、一般に密行列を係数とする連立一次方程式を解く必要がある場合が多い。このとき、本方程式を高速に解くための技術として、高速多重極展開法や行列の低ランク近似(H行列、H^2行列、BLR 行列など）があるが、今回は汎用性の点、また将来の H 行列への応用を見据えた予備的な検討の観点から、密行列をそのまま取り扱う方式について検討を行った。

具体的に、本研究では、複数の導体に誘起される表面電荷を、ラプラスの方程式を基礎方程式とする境界要素解析で導出することとした。ここで、一般の電磁気シミュレーションでは、単一のモデルを一回だけ解くようなケースはまれで、たくさんのモデルを解くことが多い現況を鑑み、これらの複数モデルの同時求解のケースにおける高速化について検討を行った。本研究では、複数のモデルに対してこれらを逐次的に求解する従来の手法に対して、一度に複数のモデルを扱うことにより、求解に使用される反復ソルバの密行列ベクトル積カーネルを行列行列積カーネルへと変更し、NVIDIA GPU の Tensor コアの効果的な利用を行う高速化を試みた。本試みについて、約 1 万自由度の境界要素解析を NVIDIA A100 GPU 上で行い、その効果について検証した。その結果、まず、複数の右辺ベクトルをまとめて取り扱うことによる高速化の効果が 3 倍程度あり、さらに、行列行列積の実行において、Tensor コアを使うことによる高速化の効果が 12 倍程度あった。また、Tensor コアの活用にはデータのメモリレイアウトが重要であることが明らかとなった。最後に、行列行列積カーネルに cuBLAS ライブラリを利用する実装としたところ、元の逐次的な解析手法による場合と比べて、約 25 倍の高速化が達成された。

5.5 材料アプリケーション（芝）

材料科学ではしばしば多彩な構造・組成から有用な性能･特性を持つ物質を絞り込む。そのため、物質群のデータを機械学習・深層学習で解析するアプローチそのものに意義がある。本項目の研究では、科学技術計算の密行列化（Tensor Core 化）を深層学習の利用より進めていく立場から、各種の材料科学のニューラルネットワークモデルに対する評価を行った。

2025 年度は、ガラスと呼ばれる乱雑な構造を持った物質の動力学的特性を予測する GNN モデル“BOTAN”に対して、積極的な TF32 の量と、自動混合精度計算を用いた評価を行った。BOTAN は、node、edge、それぞれに対応した embeddings を２層 MLP で処理した出力を、相互にメッセージ交換する単純な GNN である。64x64 の MLP では TF32 を積極的に利用した加速が２割程度にとどまっていたところ、MLP サイズにすることで、５割以上の TF32 による加速が実現されることが見出され、広範なガラスをカバーできるパラメター数領域における Tensor Core の有用性が示された。

他に、機械学習力場に対する GNN のベンチマーキングを実施した。機械学習力場とは、量子力学に基づく密度汎関数計算の結果を学習したモデルにより、局所的な構造のフィンガープリントから高精度に推論され分子動力学計算に利用可能となる。現在の代表的な機械学習力場である DeepMD と Allegro（+NequiP）に対する評価を実施するために、DeepMD のグループが以前に実施したベンチマーキングのデータセットから、両者が学習できるようにフォーマットと深層学習アプリの接続を行った。それぞれの当初実装のモデルと、それから幾何学的同変ニューラルネットワークなどのモデルなどに対し、GPU 上の演算負荷のプロファイルについて評価解析を進めることができた。

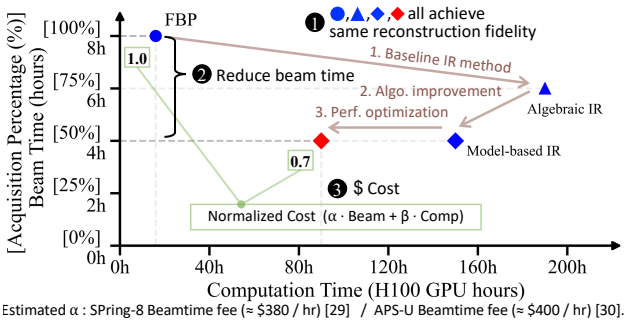
5.6 医療アプリケーション（Wahib）

2025 年度は高エネルギーシンクロトロンイメージングにおけるスパースビュー反復再構成のためのスケーラブルなフレームワークを開発した。その中心的な焦点は、実用的なスキャン・演算のトレードオフを実現するためのテンソルコアの効率的な活用にある。これは、従来の反復再構成が計算コストが高すぎて現実的でない、利用率の高い施設における主要なボトルネックに対処するものである。

物理ベースのモデリング､ADMM ベースの再構成、および Tensor Core 主導のシステム最適化にまたがる包括的な再設計を導入した。特に、デュアルモード実行設計やメモリ効率の高いデータレイアウトを含む、Tensor Core に最適化された疎行列カーネル（SpMM）を開発し、不規則な疎性パターンにもかかわらず高い利用率を実現した。これらの最適化により、ワークフローは、最新の GPU アーキテクチャ上で、通信制約型から計算制約型の実行へと移行した。

その結果得られたシステムは、計算処理が実測を効果的に代替し得ることを実証している。50%のスパースビューサンプリングにおいて、高い再構成精度を維持しつつ、大規模な運用において実測＋計算の総コストを 0。71 倍に低減した。

本研究は、Tensor Core による高速化がシンクロトロンイメージングを変革し、ビームタイムの短縮、施設のバックログ解消、そして大規模かつ高解像度のイメージングワークフローの支援を可能にする。



Estimated α : SPring-8 Beamtime fee (≈ $380 / hr) [29] / APS-U Beamtime fee (≈ $400 / hr) [30].
Estimated β : AWS EC2 p5.48xlarge (≈ $4 / GPU hr) [31] / TSUBAME 4.0 (≈ $7 / GPU hr) [32].

図３ Spring-8 における実測の代替効率

5.7 疎行列ライブラリ（深谷）

大規模疎行列を係数とする連立一次方程式に対するクリロフ部分空間法をはじめとする疎行列向けの数値解法は、部分空間を用いた解の近似に基づく構造をしており、数学的にはある種の最小二乗問題に帰着されることが多い。この背景を踏まえて、本年度は最小二乗問題の求解における GPU の活用を念頭におき、近年、応用数理分野で活発に研究されている乱択アルゴリズム（Randomized Algorithm）の GPU 実装に関する研究を進めた。

乱択アルゴリズムでは、乱数を利用したスケッチと呼ばれる操作により、元の問題（行列）を低次元の問題に変換し、確率的に実用上十分な精度の解を効率的に得ることを目指す。ここで､「確率的」とは、想定される全ての入力に対して必ずしも計算結果の精度が保証されるのではなく、多くの入力に対しては十分な精度が期待できる、という意味である。

最小二乗問題に対する乱択アルゴリズムの計算の主要部は行列のスケッチ部分の処理であり、ガウス分布から得られる乱数を要素とする単純な方法（ガウシアンスケッチ）や、離散コサイン変換（フーリエ変換）を活用した SRCT（Subsampled Randomized Cosine Transformation）などが知られている。これらの操作は、一般的な数値線形代数アルゴリズムで従来必要とされてきたカーネルと特徴が異なり、乱数自体の生成を含めて、適切な GPU 実装が自明ではない。そこで、本年度の研究として、CUDA による離散コサイン変換の実装と cuBLAS、cuSOLVER、cuRAND などの利用可能な数学ライブラリ関数を併用する形で最小二乗問題に対する乱択アルゴリズムの GPU 実装を行い、性能評価を実施した。

性能評価の結果､上記の方針で実装した SRCT を用いた乱択アルゴリズムは、従来の（決定論的）アルゴリズムと比べて、同程度の計算精度の結果を得ることができるパラメータ（スケッチサイズ）を設定した際、計算時間の観点で十分に有効であることが確認できた。この結果を踏まえて､次は、疎行列向けの反復解法等への応用を目指す。

5.8 密行列ライブラリ（伊田）

境界要素法（BEM）の離散化で生じる密行列は、O（N^2）のメモリと計算コストを要し、大規模シミュレーションにおけるスケールを制限する課題がある。この制約を緩和し、メモリ要件を O（$N^{1.5}$）へと削減する手段として、遠距離相互作用を圧縮するブロック低ランク（BLR）行列などによる近似が有効である。本研究では、この BEM における「BLR 行列と縦長密行列の積」に対し、NVIDIA Tensor Core を活用するための最適化戦略を検討した。

Tensor Core の利用（Batched GEMM）には行列次元の統一が必須だが、BLR 行列はブロックサイズやランク数が不均一であるという課題があった。これを解決するため、均一なブロックサイズを生成する「FDE 法」と、行列のゼロパディング（不要な余白）を最小化する「SRA 法」の 2 つの最適化手法を提案・実装した。

NVIDIA A100 GPU を用いた評価の結果、提案手法（FDE+SRA）はメモリ使用量を O（$N^{1.5}$）であり、既存の BLR 実装の最大 3 倍の演算速度を達成した。また、図??に示す通り、縦長密行列の列数（右辺ベクトルの数 M）を変化させた評価において、$M \geq 64$ の領域で CUDA コアのピーク性能を突破し、Tensor Core による演算加速が明確に確認された。問題サイズと列数を最大化した条件では 14,529 GFLOPS（Tensor Core ピーク性能の約 75%）に到達している。

本研究の成果は、国際会議 SCA/HPCAsia 2026 にて査読付き論文として発表された。

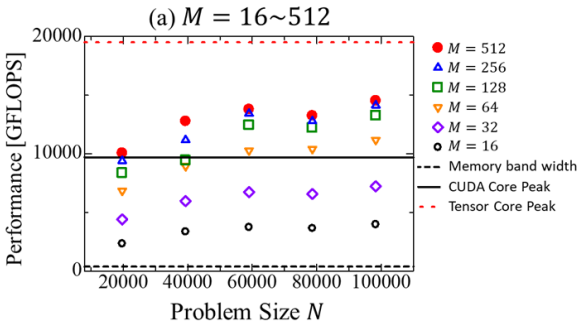


図４　右辺ベクトルの数Mおよび問題サイズNを変化させた際の達成 FLOPS

5.9 N 体問題ライブラリ（横田）

分子動力学を含む広範な科学計算に共通する基盤技術として、N 体問題ライブラリ ExaFMM の高度化を進めた。特に、FMM を中心としたアルゴリズムの GPU 最適化および大規模並列化を実現し、次世代計算機環境に適したライブラリ設計を行った。

従来の N 体計算では、計算量削減のための FMM が理論的には O(N)の優れたスケーリングを持つ一方で、実装上の定数コストや通信の複雑さが障壁となり、実アプリケーションでの利用は限定的であった。本研究では、これらの課題に対し、空間分割構造の GPU 常駐化、通信の局所化（LET）、および行列演算への変換といった複数の技術を統合することで、実用的性能を達成した。

特に重要なのは、遠距離相互作用の中核である M2L 演算をバッチ型行列積として再構成し、Tensor Core による低精度演算を活用した点である。これにより、AI 計算向けに進化したハードウェアアーキテクチャとの親和性が飛躍的に向上した。また、MPI 通信においても、近接領域は point-to-point 通信、遠距離は collective 通信を組み合わせるハイブリッド戦略を導入し、スケーラビリティを改善した。

さらに、粒子分布の時間的変化が小さいことを利用し、ツリー構造や相互作用リストの再構築頻度を削減することで、計算効率の向上を図った。これにより、理論計算量だけでなく実効性能の観点でも優れたスケーリングを実現している。

本ライブラリは、分子動力学のみならず、重力計算、電磁場解析、粒子法など多様な N 体問題に適用可能であり、今後の GPU 中心・低精度演算志向の計算基盤における中核的役割を担うことが期待される。

5.10 精度回復ライブラリ（尾崎）

近年の計算機アーキテクチャでは、FP16、BF16、INT8 などの低精度演算に対する性能向上が著しい一方で、FP64 演算性能の向上は相対的に鈍化している。このような状況のもと、低精度演算の高いスループットを活用しつつ、最終的には FP64 相当の精度を持つ計算結果を得るためのエミュレーション技術の重要性が高まっている。とくに、数値線形代数においては高い精度が要求される場面が多く、低精度ハードウェアを高精度計算に有効活用する枠組みの構築が重要な課題となっている。本研究では、この課題に対するアプローチとして、低精度演算を活用する 尾崎スキーム II を提案した。尾崎スキーム II は、中国式剰余定理（CRT: Chinese Remainder Theorem）に基づいて行列積を計算する方法であり、行列要素を複数の法に関する剰余表現へ変換したうえで、低精度行列積を複数回実行し、その結果から最終的な計算値を復元する。この構成により、従来の 尾崎スキーム I と比較して、必要となる行列積の回数を削減でき、エミュレーション全体の計算コストを抑えることが可能になった。

この点は、低精度演算器と FP64 演算器の性能差が極めて大きい GPU において特に重要である。たとえば H200 では、INT8 Tensor Core による行列積性能が FP64 行列積性能の約 32 倍に達する。しかし、尾崎スキーム I では、FP64 相当の精度を得るために多数の低精度行列積を必要とするため、この性能差を十分に活かしても、Native な DGEMM を上回る速度で FP64 エミュレーションを実現することは容易ではなかった。これに対し、尾崎スキーム II では、必要な行列積回数の削減に成功したことで、低精度演算器の高いスループットをより直接的に活用できるようになった。その結果、H200 においても、Native な DGEMM より高速に、FP64 相当の精度を持つ行列積を実現できた。

6. 進捗状況の自己評価と今後の展望

2025 年度の目標は、気象・電磁気・分子・量子・材料・医療の 6 つの科学技術アプリケーションに対して、まず Tensor Core を用いない GPU 化を進め、将来の Tensor Core 化の効果を定量的に評価するためのベースラインを構築することであった。また、Tensor Core を活用するために、各アプリケーションの主要計算を密行列積またはバッチ密行列積の形に再構成するためのアルゴリズム・離散化手法の検討および CPU 実装を進めることも目標であった。

本年度の進捗を総合的に評価すると、当初計画はおおむね達成され、一部のアプリケーションおよびライブラリでは当初計画を前倒しして Tensor Core を用いた性能評価まで進めることができた。特に、分子アプリケーション、電磁気アプリケーション、医療アプリケーション、密行列ライブラリ、N 体問題ライブラリでは、単なる GPU 化やアルゴリズム変換の検討にとどまらず、実際に Tensor Core を活用可能な行列積型の計算構造を導入し、性能向上を定量的に示すことができた点は大きな成果である。

分子アプリケーションでは、GROMACS における長距離クーロン相互作用計算について、従来の PME 法に代わる FMM の実用化を進め、FMM の主要カーネルである M2L 演算を密行列積として再定式化した。さらに、GPU 上での空間分割、Local Essential Tree による通信局所化、CUDA ストリームによる近距離・遠距離相互作用計算の重畳などを実装し、数千万原子規模の大規模系において PME を上回るスケーラビリティと最大約 2.5 倍の高速化を示した。これは、2025 年度の目標である GPU 化および密行列積化を大きく超え、Tensor Core 時代の分子動力学計算の新しい方向性を示す成果である。

電磁気アプリケーションでは、境界要素法における密行列ベクトル積を、複数右辺を同時に扱う行列行列積へ変換し、NVIDIA A100 GPU 上で評価を行った。その結果、複数右辺化による高速化、Tensor Core 利用による高速化、さらに cuBLAS を用いた実装によって、逐次的な解析手法に比べ約

25 倍の高速化を達成した。これは、当初計画で想定していた「複数右辺を用いた密行列積化」が実アプリケーションにおいて有効であることを明確に示す成果であり、今後の H 行列・BLR 行列への展開にもつながる重要な進捗である。

医療アプリケーションでは、高エネルギーシンクロトロンイメージングにおけるスパースビュー反復再構成を対象に、ADMM ベースの再構成手法と Tensor Core に最適化された疎行列カーネルを組み合わせたスケーラブルなフレームワークを開発した。不規則な疎性パターンを含むにもかかわらず、データレイアウトやデュアルモード実行設計の工夫により GPU 上で高い利用率を実現し、50%スパースビューサンプリングにおいて高い再構成精度を維持しつつ、実測と計算を合わせた総コストを 0.71 倍に低減できることを示した。これは、単なる計算高速化にとどまらず、実験施設におけるビームタイム短縮や測定効率向上に直結する成果である。

材料アプリケーションでは、ガラスの動力学的特性を予測する GNN モデル BOTAN を対象として、TF32 および自動混合精度計算の有効性を評価した。小規模な MLP では加速効果が限定的であった一方、より大きな MLP 構成では TF32 により 5 割以上の高速化が得られ、材料科学におけるニューラルネットワークモデルにおいても Tensor Core の効果がモデルサイズや演算構造に強く依存することを明らかにした。また、DeepMD や Allegro/NequIP などの機械学習力場に対しても、データセット接続および GPU 上の演算負荷プロファイル解析を進めることができた。これにより、材料アプリケーションにおける Tensor Core 活用の対象と条件を整理できた。

気象アプリケーションでは、SCALE-RM を対象として既存 GPU 実装の性能評価とボトルネック解析を実施し、力学コアにおけるフラックス計算、鉛直方向解法、物理過程における放射・雲微物理計算が主要な計算負荷であることを確認した。また、有限体積法のフラックス計算や HE-VI 法の三重対角行列解法について、バッチ処理や低精度演算の適用可能性を検討した。さらに、不連続ガラーキン法のような高次離散化手法では要素内演算が密行列積として表現され、Tensor Core との親和性が高いことを明らかにした。一方で、DG 法の実アプリケーションへの GPU 実装や Tensor Core を用いた定量評価は今後の課題として残っており、本年度は主としてボトルネック解析とアルゴリズム設計段階の成果であったと評価できる。

量子アプリケーションでは、格子 QCD におけるディラック方程式ソルバーを対象に、FP16 を用いた混合精度反復解法の安定性を検討した。単純に FP32 を FP16 へ置き換えるだけでは、系が大きくなると反復数が増加し、さらに作業ベクトルや残差ベクトルのアンダーフローによって収束判定が不正確になることを明らかにした。これに対し、適切なリスケーリングをソルバー内部に組み込むことでアンダーフローを回避する手法を提案し、FP16 が利用可能な CPU 上で良好な結果を得た。GPU 実装は準備段階であり、2025 年度目標に対しては一部未達成の要素が残るが、低精度化に伴う数値安定性の課題を具体的に特定し、その解決策を得た点は、今後の GPU 化・Tensor Core 化に向けた重要な進展である。

ライブラリ開発においても、アプリケーション横断的な成果が得られた。疎行列ライブラリでは、最小二乗問題に対する乱択アルゴリズムの GPU 実装を行い、SRCT を用いた手法が従来の決定論的手法に対して計算時間の面で有効であることを確認した。密行列ライブラリでは、BLR 行列と縦長密行列の積に対して、Tensor Core 利用に適した FDE 法および SRA 法を提案・実装し、既存 BLR 実装の最大 3 倍の演算速度と、Tensor Core ピーク性能の約 75%に相当する高い性能を達成した。N 体問題ライブラリでは、FMM の M2L 演算をバッチ型行列積として再構成し、GPU 常駐化、通信局所化、相互作用リスト更新頻度の削減を組み合わせることで、分子動力学を含む多様な N 体問題に展開可能な基盤を整備した。精度回復ライブラリでは、CRT に基づく尾崎スキーム II を提案し、低精度 Tensor Core 演算を用いながら FP64 相当の精度を得るための

行列積エミュレーションの計算コスト削減を実現した。

以上を踏まえると、2025 年度は、6 つのアプリケーションに対して Tensor Core を用いない GPU 化を進めるという当初目標に加え、多くの対象で Tensor Core 活用に向けたアルゴリズム変換と初期性能評価まで進めることができた。一方で、アプリケーション間で進捗には差があり、気象アプリケーションでは DG 法など高次離散化手法の実装、量子アプリケーションでは FP16 混合精度ソルバーの GPU 実装と実機評価が今後の重要課題として残っている。また、Tensor Core を有効活用するためには、単に演算を行列積の形に変換するだけでなく、データレイアウト、バッチサイズ、メモリ帯域、通信、低精度化による誤差蓄積を含めた総合的な最適化が必要であることも明らかになった。

今後は、2025 年度に構築した GPU 実装および行列積化の基盤をもとに、各アプリケーションにおいて Tensor Core を本格的に利用した実装と性能評価を進める。気象アプリケーションでは、DG 法の GPU 実装とバッチ密行列演算の導入を進め、実際の気象シミュレーションにおける性能向上と精度影響を定量的に評価する。量子アプリケーションでは、リスケーリングを組み込んだ FP16 混合精度ソルバーを GPU 上に実装し、格子サイズや物理パラメータを変えた場合の安定性と高速化効果を検証する。分子および N 体問題では、FMM の Tensor Core 実装をさらに高度化し、GROMACS との統合を通じてより広範な分子系・並列規模での評価を行う。電磁気アプリケーションでは、複数右辺化と Tensor Core 化の成果を H 行列、BLR 行列、H2 行列などの低ランク近似手法へ拡張し、より大規模な境界要素解析への適用を目指す。材料アプリケーションでは、GNN および機械学習力場における Tensor Core 活用条件を整理し、精度を維持しながら高速化できるモデル構造や混合精度戦略を明らかにする。医療アプリケーションでは、スパースビュー再構成フレームワークをより大規模な実測データへ適用し、施設運用上の測定時間削減効果を定量化する。

さらに、ライブラリ群については、個別アプリケーションに閉じない共通基盤として整備を進める。密行列・疎行列・N 体問題・精度回復の各ライブラリを連携させることで、小規模密行列積の大量実行、低ランク行列演算、混合精度反復解法、低精度からの精度回復を統合的に扱える環境を構築する。これにより、Tensor Core を単なる高速演算器として利用するのではなく、アルゴリズム、離散化、データ構造、数値精度を含めたアプリケーション全体の再設計を可能にする。

本研究課題の最終的な目標は、深層学習向けに進化する次世代 GPU の潜在能力を、科学技術計算において最大限に引き出すことである。2025 年度の成果により、代表的な科学技術アプリケーションの多くで、Tensor Core を活用するための道筋が具体化された。今後は、個別カーネルの高速化にとどまらず、アプリケーション全体での実効性能、精度、メモリ使用量、通信コストを総合的に評価し、GPU 中心・低精度演算志向の次世代計算基盤に適した計算科学アプリケーションの刷新を進めていく。