jh240021

高性能・高信頼な数値計算手法と応用の新展開

片桐孝洋（名古屋大学）

ポストエクサスケール時代に向けて、計算機システムはますます複雑化し多様化が進んでいる。特に数値計算では、演算速度や演算精度の最適化、メモリやネットワークの階層の深化に対応した通信最適化、そして電力やエネルギ効率の最適化が避けられない。エネルギ節約の観点から、スーパーコンピューティングにおいても、計算効率がますます重要になっている。特に、近年需要が高まる生成 AI などの振興計算需要への考慮も喫緊の課題である。ゆえに低精度・変動精度演算の積極的な活用、そのための精度保証手法の確立が急務である。本研究は、3 年間継続した前課題の成果を最大限に活用し、科学技術シミュレーションに現れる大規模行列に対して有効な疎行列ソルバや階層型行列（H 行列）演算等の高速計算に関する研究成果を引き継ぎ、新展開を目指す。計算の信頼性及び電力効率を重視しながら、適用可能な実用的手法の研究開発も継続する。

1. 共同研究に関する情報

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
東京大学 情報基盤センター
東京工業大学 学術国際情報センター
名古屋大学 情報基盤センター
京都大学 学術情報メディアセンター
九州大学 情報基盤研究開発センター

(1) 課題分野

大規模計算科学課題分野

(2) 参加研究者一覧と役割分担

I. 疎行列ソルバー

主担当：岩下・中島・藤田，担当：市村・深谷・鈴木・星野・河合・八代・荒川・大島・Wellein・Basermann

II. H 行列演算

主担当:横田,担当:岩下・伊田・Huang・Spendlhofer

III. 計算機システムと消費電力測定

主担当：坂本・塙，担当：近藤

IV. 精度保証と自動チューニング手法

主担当：片桐・荻田，担当：中島・星野・河合・横田・伊田・近藤・尾崎・今村・Marques・内野・任・森下・中谷

2. 研究の目的と意義

ポストエクサスケール時代に向けて、計算機システムはますます複雑化し多様化が進んでいる。特に数値計算では、演算速度や演算精度の最適化、メモリやネットワークの階層の深化に対応した通信最適化、そして電力やエネルギ効率の最適化が避けられない。エネルギ節約の観点から、スーパーコンピューティングにおいても、計算効率がますます重要になっている。特に、近年需要が高まる生成 AI などの振興計算需要への考慮も喫緊の課題である。ゆえに低精度・変動精度演算の積極的な活用、そのための精度保証手法の確立が急務である。

本研究は、3 年間継続した前課題の成果を最大限に活用し、科学技術シミュレーションに現れる大規模行列に対して有効な疎行列ソルバや階層型行列（H 行列）演算等の高速計算に関する研究成果を引き継ぎ、新展開を目指す。計算

の信頼性及び電力効率を重視しながら、適用可能な実用的手法の研究開発も継続する。2024 年導入の JHPCN 最新計算機環境もターゲットに加え、以下の課題を実施する：

I. 疎行列ソルバ、H 行列演算等の代表的な数値計算アルゴリズム、各アプリケーション（地震学、大気海洋科学、構造力学、流体力学、電磁気学等）について、メモリアクセス最適化及び分散メモリ通信最適化に着目し、安定で高性能な手法を各システムに実装し、消費電力の効果を検証しつつ、変動精度演算を実用的レベルまで引き上げることを目指す。

II. 疎行列ソルバ、H 行列や FMM の演算に加えて、基本線形計算カーネル群（BLAS、テンソル演算、疎行列-ベクトル積）を対象として、実用的な精度保証法／精度推定法のアプリケーション適用拡大を目指す。I の各アルゴリズム、アプリケーションについて所望の結果精度達成という条件下で、計算時間や消費電力を最小化する最適演算精度を自動チューニング（Auto-tuning, AT）技術により、動的に制御する手法の確立を目指す。

計算機としては、Wisteria/BDEC-01（東大）、スパコン「不老」（名大）等に搭載された A64FX を中心に、NVIDIA V100／A100／H100、Intel Xeon／Sapphire Rapids、AMD 等の最新 GPU と CPU 及び、その後継機種を対象にして、コード最適化の知識集積と公開による拠点ユーザへの還元を強力に推進する。また、得られた成果は東大情報基盤センターを中心に開発を進めている「(計算＋データ＋学習）融合を実現する革新的ソフトウェア基盤 h3-Open-BDEC（科研費基盤(S) 19H05662、研究代表者：中島研吾（東大）、http://nkl.cc.u-tokyo.ac.jp/h3-Open-BDEC/)」の最終成果も活用し、HPCI の多様な計算機環境へ展開させ、新たな技術展開の境地を切り開く。

3. 当拠点の公募型共同研究として実施した意義

本課題で目指す高性能・高信頼な数値計算手法の研究には、JHPCN 拠点の多様な計算機環境の活用の必要性に加えて、東大「Wisteria/BDEC-01（Odyssey)」、名大「不老」TypeI・TypeII サブシステム、東京科学大「TUBAME4.0」、京大「Camphor3」など、多様な最先端大規模システムの活用が必要である。さらに、JHPCN 拠点は多様な学際分野の専門家を擁していることから、本課題の学際的研究を推進する体制を容易に構築できる。加えて、北大、東大、東京科学大、名大、京大、九大、理研 R-CCS の各センターから様々な分野の研究者が参加している。JHPCN 各センターはオープンソースソフトウェアの活用や開発に意欲的に取り組んでおり、本研究の成果を公開し、各センターのスパコンにデプロイし、講習会等の普及活動を協力して行うことが可能となる。このことで、利用者拡大及びソフトウェアのさらなる改良が期待される。

本研究は、最先端のスパコン向けに開発された高性能数値計算アルゴリズムに対して、変動精度演算を適用し、精度保証/精度推定及び自動チューニング(AT)手法を開発する試みとしては、国内のみならず国際的にも唯一のものである。本研究では昨年成果で得られた知見を基に、疎行列/H 行列を係数行列とする実問題に適用可能な実用的数値解法、精度保証法/精度推定法、自動チューニング手法の研究開発を実施するとともに、機械学習等も含めたより広範囲なアプリケーションに適用する。これにより、科学技術シミュレーションにおける有効性を検証できる。

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

<u>I. 疎行列ソルバー</u>

(1)代表的な前処理付き反復法である ICCG 法について，前処理後の係数行列の Remainder matrix の行列ノルム，固有値分布を算出し，反復回数との関係性を調査した．行列データベースから取得した 5 種の行列を用いた数値実験において，前処理行列を単精度化しても，固有値分布や反復回数に大きな影響がない結果が得られた[1].

(2) 高速な低精度演算器のさらなる活用に向け，整数演算を用いた陽解法シミュレーションの高速化手法を開発した．ここでは，Tensor Core による INT8 計算を多段で用いることで陽解法構造格子有限要素法における疎行列ベクトル積を倍精度相当の精度で計算できるアルゴリズムを開発した．提案手法において INT8 を 8 段で用いることで FP64 と同等の疎行列ベクトル積計算の精度が得られることを確認し，A100 GPU における INT8 Tensor Core を用いた高速計算により，通常の FP64 コアを使った場合と比較して，疎行列ベクトル積部で 4.5 倍の高速化，また，FP64 計算と比べてシミュレーション全体で 3.4 倍の高速化を実現した．これは計算全体を FP32 で計算した場合よりも高速であり，高速な INT8 計算コアを用いることが可能な高精度計算手法を用いることで，通常の FP32 計算よりも高速かつ高精度で波動計算が実現できることを示した[Ichimura et al. ICCS (International Conference on Computational Science), 2025].

(3) ICCG 法の代表的な並列化手法であるブロックヤコビ IC 前処理について，その性質に関する研究を行った．同前処理はスレッド数（並列実行数）の増加に対して，反復回数が増加する性質を持っているが，その詳細は明らかでなかった．そこで，前処理後の係数行列の固有値分布を計算し，調査した結果，スレッド数の増加に対して，小さい固有値，大きい固有値の双方がより小さく，あるいはより大きく，条件数や固有値分布の状態が悪化する方向に変化することが分かった．また，重要な知見として，これらの変化が少数の固有値に留まらず，多くの固有値に見られるもので，例えば，減次法によって，それを補償することが難しいことが明らかになった．

II. H 行列演算

(1)混合精度演算を用いた H 行列の Cholesky 分解を反復改良法の前処理に用いる手法を開発した．これまでに Tensor Core による低精度演算の誤差を反復改良法を用いて回復する手法は提案されているが，H 行列による階層的低ランク近似による誤差を回復する手法は検討されていない．本研究では LU 分解を用いた前処理に H 行列を用いた場合と GMRES による反復法による前処理に H 行列を用いた場合の収束性への影響を検証した．図 1 に低ランク近似精度と収束性の関係を示す．

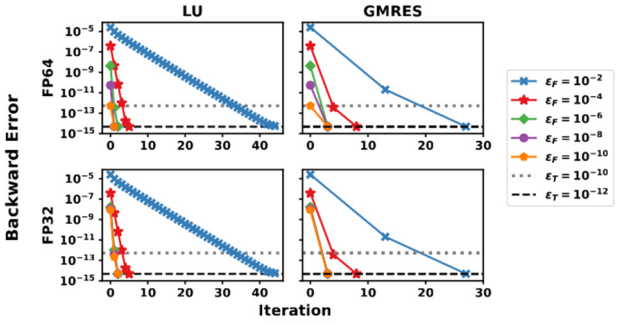

図 1 LU 分解と GMRES による前処理に H 行列を用いた場合の低ランク近似精度と収束性の関係

低ランク近似の精度が $\varepsilon = 10^{-2}$ のときは収束性が大幅に低下するが低ランク近似の精度が上がるにつれて収束性も向上することがわかる．また，$\varepsilon = 10^{-10}$ 以上の近似精度であれば収束性に大きな違いはないことがわかる．さらに，近似精度が $\varepsilon = 10^{-2}$ のときは LU 分解に比べて GMRES の方が収束性が良いが，近似精度が上がるにつれて LU 分解の方が収束性が良くなることもわかる．この結果は ISC に採択された[Thomas Spendlhofer, Qianxiang Ma, Yasuhiro Matsumoto, Rio Yokota, On the Interplay Between Precision, Rank, Admissibility, and Iterative Refinement for Hierarchical Low-Rank Matrix Solvers, ISC High Performance, June 2025].

(2)Fast Multipole Method(FMM)は分子動力学や重力体問題などの N 体問題の計算量を $O(N^2)$から O(N)に低減する高速化手法である．GROMACS などの分子動力学ソフトウェアでは Particle Mesh Ewald(PME)を用いてクーロン力の計算を行うが，FFT に依存する PME では並列化効率に限界がある．そこで PME を FMM で置き換えることで並列化効率を向上させることができる．また，FMM の多重極展開を局所展開に変換する部分はセルの間の相対的な位置関係が同じ場合をまとめて計算することで行列積の計算になり Tensor Core 用いることができる．本研究では，このような Tensor Core を用いることができる FMM の実装を開発中した．これは GROMACS の開発者らと共同で進めており，まだ論文にはなっていないものの完成すれば大きな波及効果が期待できる．

III. 計算機システムと消費電力測定

LLM の細粒度な電力特性の理解を進めるため，中規模 LLM (gpt-neo-2.7B) に対して，レイヤごとの計算量，計算時間，消費電力を計測するための作業を進めた．これまでに作成した，gpt-neo-2.7B を各レイヤを一つずつ実行するためのプログラムを拡張し，電力特性を分析するための計測関数などを追加した．さらに，各レイヤ内で行われる主要な 4 つの処理（Self-Attention，Residual＋$Norm_1$，MLP，Residual＋$Norm_2$）について，より詳細な実行時間，計算量の調査を行った．研究室内の CPU と GPU を用いてこれら消費電力計測を含む初期評価を開始した．

また，一般的に利用される LLM フレームワークである Hugging Face の Transformers を用いた推論と精度を比較すると，作成した推論プログラムは著しく精度が低いため，精度を向上するための改良を進めた．

IV. 精度保証と自動チューニング手法

(1)疎行列を係数とする線形方程式に適用可能な精度保証法の開発を検討した．具体的には，最近発表された行列分解に基づく精度保証法の有効性について調査し，既存手法との比較・検討を開始した．また，線形方程式及び固有値問題に対して，反復改良法によって近似解の精度そのものを向上させる方式と精度保証法の融合について検討し，発表を行った[3][4]．

(2)閾値付不完全コレスキー分解前処理付 CG 法において，閾値の選択の AT を，疎行列画像と深層学習による AT 方式を採用した AI モデルに対する，説明可能 AI(XAI)研究を行った．成果に査読付国際論文が受理され発表を行った[1]．本成果は，疎行列反復解法の AT に AI を用いる場合の XAI の適用事例として世界初の成果である．XAI の適用により，AI を用いた AT 方式が妥当な解答を得ることを客観的に示すことができ，実利用可能な AT 方式を構築することができた．

加えて本年度はモデル構築に使用する実測データの問題設定を，係数行列のサイズを 32×32×32，最大 fill-in 値を 0,1,2，閾値を 1.0e-05～1.0(100 種)，λ を 1～1.0e04 (100 種)とした．この実測データから行列の画像情報，最大 fill-in 値，閾値を入力として実行時間予測を行うモデルを倍精度，単精度，混合精度の演算精度ごとに，予測の誤差が小さい高性能モデルと大きい低性能モデルの 2 種を構築した．このようにして得られた計 6 種のモデルについて，係数行列のサイズと最大 fill-in 値は実測データと等しく，閾値を 1.0e-07～2.0(50 種)，λ を 1.0e-04～5.0e04 (50 種)とした場合を入力として与えることで実測データとの比較検証を行った．

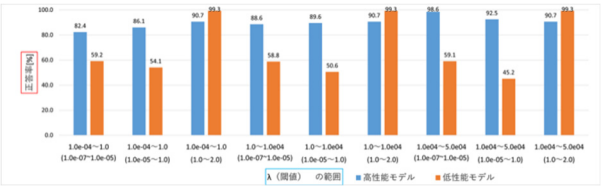

図 2 高精度および低精度のモデルに対する演算精度選択の AT に対する正答率

図 2 から，演算精度の選択の AT において，学習時に与えたデータ範囲内外を問わず，高い正答率を得られることがわかった．

以上の研究を通して，要求資源を十分に活用して研究が遂行された.

6. 進捗状況の自己評価と今後の展望

I. 疎行列ソルバー

本年度の進捗は 100%である。

現在開発したプログラムに基づき，実応用問題のテスト行列やより大規模な問題において，前処理後の係数行列の固有値分布について調査する.

II. H 行列演算

本年度の進捗は 80%である.

(1)の混合精度 H 行列を前処理に用いた反復改良法の成果は ICS に採択された. (2)では GROMACS の PME 法を FMM で置き換え，Tensor Core を用いることができる FMM の実装を開発した. まだ，Tensor Core 上での性能最適化の余地があるため論文化はされていないが，GROMACS の開発者からは大きな期待が寄せられている.

III. 計算機システムと消費電力測定

本年度の進捗は 100%である.

LLM の細粒度な電力特性を理解するため，中規模な LLM に対し Hugging Face の Transformers 等のフレームワークを用いず，各レイヤを 1 つずつ計算する LLM 推論プログラムを作成し，各レイヤ内の主要な処理の実行時間と消費電力を計測する基本的な電力特性分析環境を構築した. これらを用いて，LLM の CPU 環境と GPU 環境における実行時間と消費電力の計測を開始した. 今後は計測結果をもとに，詳細な分析，様々な省電力化手法の適用の検討を進める.

IV. 精度保証と自動チューニング手法

本年度の進捗は 100%である。

疎行列を係数とする線形方程式に対する精度保証法について, SuiteSparse Matrix Collection にある比較的小規模な問題に対する数値実験は実施されているが，大規模な問題に対しての有効性を確認する必要がある. また，反復解法に基づく精度保証法についても引き続き検討する.

本年, 研究を通して XAI における AI モデルの精度の問題が明らかになった. 今後，生成される AI モデルの入力データに対する不確定性を考慮した AI モデル生成，および，その AI モデルを活用した AT 方式の開発を進める必要がある. この課題は AT の側面では萌芽性が高い. また，単精度演算，混合精度演算を切替える AT 方式は近年の計算機アーキテクチャでは重要であり，その観点から研究を進める予定である.