jh250018

高性能・高信頼な数値計算手法と応用の新展開(2)

片桐孝洋（名古屋大学）

概要

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。本研究は、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演算等の高速計算、大規模言語モデル関連演算、精度保証、自動チューニング技術により新展開を目指す。

1. 共同研究に関する情報

(1) 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
東北大学サイバーサイエンスセンター
東京大学 情報基盤センター
東京科学大学 情報基盤センター
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

(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

I. 疎行列ソルバー

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

II. H 行列・テンソルトレインと LLM

主担当：横田、担当：岩下・伊田・星野・椋木・大島・河合・Huang・Xie

III. 精度保証

主担当：荻田、担当：片桐・尾崎・中島・椋木・寺尾・内野・Marques

IV. 自動チューニングと計算機システム

主担当：片桐、担当：大島・坂本・塙・近藤・任・中谷・林

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・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、演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。そのため、低精度・変動精度演算の積極的な活用と精度保証手法の確立が急務となっている。

この背景から本研究では、昨年を最大限に活用したうえで、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さらに発展させることで新展開を試みる。計算の信頼性及び電力効率を重視しながら、適用可能な混合精度演算などの実用的手法の研究を継続する。JHPCN 最新計算機環境を考慮し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、数値的に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、低精度・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、大規模言語モデル(LLM)などの演算に加えて、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、実用的な精度保証と混合精度演算のアプリケーション拡大を目指す。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。

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に意義がある。

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ことにある。

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点に意義がある。

4. 前年度までに得られた研究成果の概要

本年度は以下の昨年度成果をさらに発展させ、高性能・高信頼な数値計算手法の確立を行うことを目指した。

岩下らは、IC 分解前処理後の係数行列の固有値分布について、調査を行った。ブロックヤコビ IC 前処理のスレッド数(並列数、ブロック数)の影響を調べ、スレッド数を増やすと条件数が悪化することを確認した。また、固有値は小さい側も大きい側も(条件が悪い方向に)シフトしており、その数は多く、減次法や Subspace correction 法による対応は困難と判明した。また、IC 分解前処理部分を単精度化した場合の影響を調べた。結果として、テストした行列の範囲では影響は軽微であった。これまでにも提案があるが、IC 前処理の単精度化による高速化は有効の可能性が大きいことを明らかにした。

横田らは、(1)混合精度演算を用いた H^2 行列の Cholesky 分解を反復改良法の前処理に用いる手法の研究を行った。H^2 行列は密行列をブロック分割しそれぞれのブロックを低ランク近似する手法であり、各ブロックは既に近似になっているので、低精度演算を用いても精度低下は起きない。本研究により従来の HSS 行列と比べてランクあたりの精度が高い H^2 行列を用い

ることで、HSS 行列より高い精度を実現した。また、(2)N 体問題の高速解法である FMM に Tensor Core を適用する研究では、FMM のセル間の相互作用において並進不変性や回転対称性を考慮し、本来行列 - ベクトル積になる計算を行列 - 行列積に変換し、Tensor Core を用いる手法の研究を行なった。

中島らは、(1)疎行列演算における最適前処理・演算精度の選択を行い、対象問題の材料定数分布とメッシュ幅分布等から計算実施前に計算時間を最短とし、消費エネルギを最小とする最適前処理・最適演算精度を選択する手法の開発を行った。簡易形状の小規模問題を使用した事前学習結果から，実問題における大規模疎行列の固有値分布を予測する手法の開発を試みた。しかし、手元にある固有値ソルバが密行列向けで逐次処理であったため、数百自由度程度の行列のみ扱った。一方、jh240058 でパイプライン型前処理付き反復法について行い、低精度・混合精度演算を適用した場合に関する検討を実施した結果 Residual Replacement［Ghysels et al. 2014］[Yamazaki et al. 2022］によって悪条件問題解決への目処を付けた。

片桐開発の AT 言語である ppOpen-AT に、混合精度演算向けに拡張させた新 AT 方式を開発し、ジャーナル論文発表を行った[5]。八代ら開発の大気海洋分野の NICAM ベンチマークを FX64 と X86 アーキテクチャで動作させ、新 AT 方式の性能評価を行った。FP64 の実行時間に対して複数個所を部分的に FP16 にする混合精度計算の最適化の自動化を行った。性能評価の結果、FX64 と X86 アーキテクチャそれぞれ 1.31 倍と 1.12 倍の速度向上を達成した。

5. 今年度の研究成果の詳細

I. 疎行列ソルバー(主担当：岩下・藤田・中島)

(1) **悪条件問題対応**：年度前半は、主としてパイプライン型アルゴリズムに関する検討を実施した。単精度、混合精度に関する検討を実施し、「混合精度＋Residual Replacement（RR)」は「条件の良い問題であれば」倍精度演算とほぼ同じ内部反復回数で収束することが示された。「単精度＋Iterative Refinement（IR)」を適用した場合は。条件の良い問題は良好であるが（1 反復当たり 2 倍の効率)、悪条件問題では「混合精度＋RR」（同：1.33 倍）に劣る。「単精度＋IR＋RR」は条件の良い問題では「単精度＋IR」と比較して良好であるが、悪条件問題では却って悪化する。内積における通信を隠蔽する「パイプライン型 CG 法」について、大規模並列環境における、特定のハードウェア、コンパイラ、MPI ライブラリに依存しない実装手法として、マスタースレッドに内積の集団通信と Halo 通信を割り当て、純内点における疎行列ベクトル積とオーバーラップさせる手法を提案し、Wisteria/BDEC-01（Odyssey）を最大 2048 ノード使用した場合、CC-Overlapping を全く適用しない場合と比較して 27％、Halo 通信のみに CC-Overlapping を適用した場合と比較しても 6%程度の性能向上が得られた。

(2) **疎行列演算における最適前処理・演算精度の選択に関する基礎技術**：
電磁場有限要素解析において、ブロックヤコビ IC 前処理による並列化 ICCG 法を利用した場合について、固有値分布を調査し、国際会議において報告した。

(3) **陽解法有限要素法による波動伝播解析**：
精度混合計算手法を開発した。ここでは、主要計算部となる疎行列ベクトル積を、複数段の整数演算で計算できるよう分解し、計算段数を調整することで計算精度を FP64 相当から FP32 相当まで可変としたうえで、整数演算部を INT8 Tensor Core で加速計算した。波動解析において場所ごとに計算精度を調整することで、FP32 計算では精度劣化が

生じる問題において FP64 相当の計算精度を、FP64 CUDA core を用いた場合の 3.3 倍の速度で計算できることを示した(図 1 . Miyabi-G 1 ノードを使用)。精度可変計算による追加のノード間通信は不要であるため、本手法は Miyabi-G 256 ノードまで高い効率でスケールすることが分かった (図 2)。

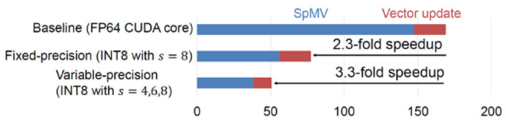


図 1 精度可変計算による
陽解法有限要素法の高速化
(Miyabi-G 1 ノードでの実行時間)

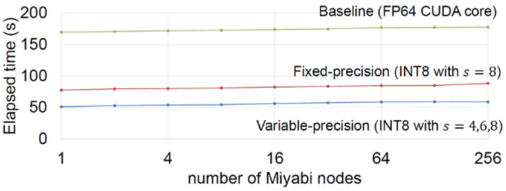


図 2 精度可変計算による
陽解法有限要素法の高速化
(Miyabi-G におけるウィークスケーリング)

(4) **混合精度演算を用いた線形反復ソルバ**：半精度演算利用による線形反復ソルバの性能改善について試みた。Nested Krylov 法の枠組みを利用し、倍精度、単精度、半精度演算を各階層で利用することで、従来の倍精度浮動小数点数に基づくソルバと同一の求解精度を保ちつつ、メモリバンド幅を節約した混合精度演算による線形反復ソルバ：F3R を開発した。本ソルバは異なる演算精度を用いた FGMRES 法とリチャードソン法を用いた反復ソルバである。本ソルバを CPU、GPU 上で性能評価し、その結果について、高性能計算に関するトップ会議である SC25 で発表した。

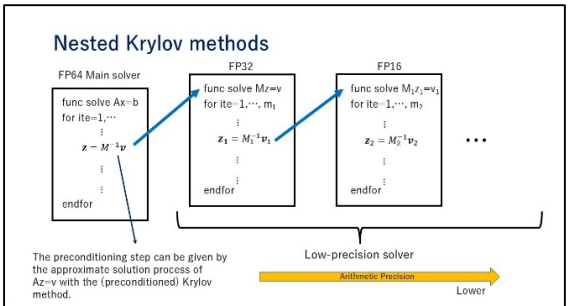


図 3 Nested Krylov 法と混合精度演算

II. H 行列・テンソルトレインと LLM (主担当：横田)

(1) 高速多重極展開法(FMM)は N 体問題の高速化手法であり、H^2 行列-ベクトル積と見做すことができる。FMM のおけるホットスポットは多重極展開を局所展開に変換する M2L の計算であるが、この部分は密行列積の形で書くことができ、Tensor Core を活用できる可能性がある。本課題では、FMM における M2L の計算を Tensor Core を用いて実装し、大幅な高速化を図る。

今年度前半には、GROMACS の開発者と連携し、GROMACS の PME を FMM で置き換え、FMM における M2L の計算を Tensor Core を用いて実装することに成功した。年度の後半では、GROMACS と FMM をより密に統合することで余計な配列の確保やコピーを削減し、CPU-GPU 間通信、ノード間通信も大幅に削減した。その結果、当初は PME と比べて 100 倍程度遅かった FMM を PME よりも 2 倍程度速くすることができた。GROMACS の高度に最適化された PME よりも高速な FMM の実装を開発することができたことは大きな成果であり、その内容をまとめた論文は現在 SC26 に投稿中である。

(2) LLM の学習・推論は入力系列長の二乗に比例する計算コストがかかる。近年、この密行列積の計算コストを低ランク近似、

Tensor Train (TT)、Block Tensor Train (BTT)、Monarch 行列などを用いて低減する手法が提案されているが、その近似が実際の学習に及ぼす影響を十分に調べた例は少ない。本課題では、これらの密行列積の高速近似手法を Tensor Core 上で実装し、混合精度演算・行列近似がもたらす学習精度への影響を調べる。

今年度前半には、LLM に Tensor Train (TT)、Block Tensor Train (BTT)、Monarch 行列などを実装し、通常の Transformer モデルに対して精度を維持しながらも高速な学習が実現できることを示した。年度の後半には、Mixture of Experts (MoE)アーキテクチャと BTT を組み合わせた実装を行い、通常の MoE に比べて僅かに高速な BTT-MoE を実現することができた。

III. 精度保証(主担当：荻田)

(1) 正則な係数行列Aに対する連立 1 次方程式$Ax = b$に対する精度保証付き数値計算、つまり近似解$\hat{x}$に対する誤差$|x - \hat{x}|$の上限を求める数値計算法の調査・研究を行った。特に係数行列が疎行列の場合の計算法について報告する。

疎行列系連立 1 次方程式の精度保証付き数値計算法では、LDLT 分解とシルベスターの慣性則を用いた方法が知られており、拡大方程式

$$\begin{pmatrix} O & A \\ A^T & O \end{pmatrix}\begin{pmatrix} 0 \\ x \end{pmatrix} = \begin{pmatrix} b \\ 0 \end{pmatrix}$$

を用いることで、行列サイズが倍になるデメリットがあるが、非対称行列にも適用が可能である。一方で、LDLT 分解はシルベスターの慣性測に対して数値的な不安定性が報告されており、大規模な問題への適用は困難であった。

本課題では、行列の平衡化を用いることで LDLT 分解の不安定性を解消できることを発見し、その調査・研究を遂行した。

Suite Sparse Matrix Collection から、20 個の非対称行列を用いて調査を行った結果、先行研究では精度保証付き数値計算に失敗した 5 件の行列について、平衡化を行う提案手法が精度保証付き数値計算に成功することを示した。また、両方が成功する 15 件のうち、12 件で平衡化を用いることで計算時間が低減されることも示した。これは、平衡化による軸交換の最適化が行われたため、LDLT 分解の高速化が行われ、それと比較して平衡化のコストが十分に小さかった結果だと思われる。

また、精度保証付き数値計算が成功した問題に対して、反復改良を用いることで、効率的に高精度な数値解を保証することに成功した。その結果、直接法を用いた近似計算の数倍から十数倍の計算時間で、高精度な数値解を精度保証することに成功した。

(2) GPU 環境に適した行列積の高性能エミュレーション技術（Ozaki-I・Ozaki-II）の開発とその誤差解析を行なった[V1,V2,V3]。誤差解析では、Ozaki-I・Ozaki-II の精度をコントロールするパラメータと問題サイズを用いて厳密な誤差限界を定式化した。導出された誤差限界は実際の誤差をタイトに包含することを確認した(図 4)。

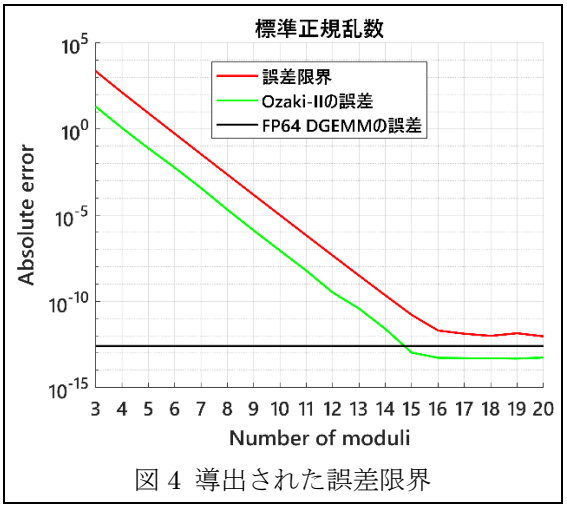


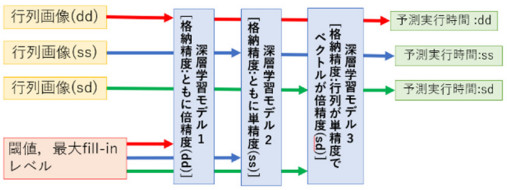
図 4 導出された誤差限界

IV. 自動チューニングと計算機システム(主担当：片桐)

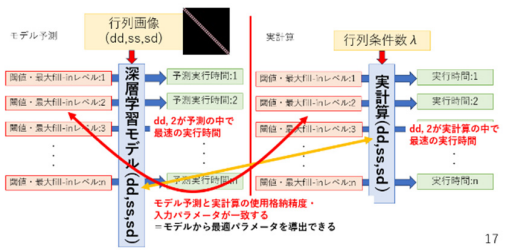
(1) **AT 方式開発**：

ICCG 法の疎行列反復解法について、倍精度、混合精度、単精度を AT により選択できる手法を提案した(図 5)。深層学習モデルを利用し、疎行列形状の画像を入力するだけで、最も高速となる演算精度による実行時間を予想することができる。それにより、最速の演算方式を自動選択する AT 機能を実現した。加えて、妥当な予想ができるかのモデルの品質評価を行った(図 5)。

本結果を、情報処理学会研究報告として発表するとともに、名古屋大学の修士論文を執筆した。



(a) 提案手法



(b) 品質の検証方法

図 5 提案手法と品質の検証方法

(2) **電力特性分析**：

LLM 推論では、入力フェーズにおいて入力トークン列を一括して処理する一方、出力フェーズでは生成済みトークンに対応する KV キャッシュを参照しながら 1 トークンずつ逐次的に生成を行う。そのため、入力フェーズと出力フェーズでは処理の並列性や計算特性が異なり、電力・エネルギ特性にも差異が生じることが予想される。そこで本実験では、モデルサイズの違いが入力フェーズおよび出力フェーズの電力特性に与える影響を調査した。実験には、これまで構築してきた電力特性計測基盤を用い、LLM モデル構成が類似しつつレイヤ数の異なる複数のモデル、具体的には GPT-Neo-125M（12 層）、GPT-Neo-1.3B（24 層）、GPT-J-6B（28 層）を対象とした。

実験の結果、本実験条件では、入力フェーズはトークン列を一括して並列処理できるためトークンあたりのエネルギが小さい一方、出力フェーズは 1 トークンずつ逐次的に生成するため、生成時のエネルギ消費は主に出力フェーズに支配される傾向が確認された。また、小規模モデルでは GPU の計算資源を十分に使い切れていない傾向が見られ、同一 GPU 上で複数リクエストをバッチ処理または並列実行することで、エネルギ効率を改善できる可能性が示唆された。

6. 進捗状況の自己評価と今後の展望

本研究は広範な分野と内容に富む。限られた時間における研究進捗と論文発表を考慮すると、**自己評価による達成度は 100%**であると判断する。

今後の展望を以下に述べる。

I. 疎行列ソルバー(主担当：岩下・藤田・中島)

(1) **悪条件問題対応**：低精度演算によって倍精度演算(FP64)をエミュレートした実装を適用し、有効性について検討することで、手法の適用可能性を拡げることができる。

(2) **混合精度演算を用いた線形反復ソルバ**：これまでに開発してきた整数演算に基づく線形反復ソルバや複数の演算精度による浮動小数点数演算を用いた線形反復ソルバの性能改善が今後の課題である。具体的には、特定の部分空間内の誤差の修正による反復法の収束性改善、自動的な演算精度の選択と切り替えの実現などが挙げられる。また、反復ソルバの基礎カーネルである疎行列ベクトル積の性能改

善による取り組みも行っていく予定である。

(3) **陽解法有限要素法による波動伝播解析**：現状の精度混合計算手法においては、シミュレーションをする前に決め打ちで場所ごとの計算精度を設定していたが、問題特性に応じて自動的に計算精度を調整できるようにすることで手法の適用性・ロバスト性を改善できる手法の提案とつながることが期待できる。

II. H 行列・テンソルトレインと LLM(主担当：横田)

（1） FMM の Tensor Core 実装を GROMACS 内に組み込み、単一の GPU においては Tensor Core を用いない場合に対して大幅な高速化が実現できることを示した。また、マルチ GPU 実装においては FMM の PME に対する並列化効率の優位性を十分に発揮することができ、初めて GROMACS の PME よりも time-to-solution で優位性を示す FMM の実装を開発することができた。これは、当初計画で予定していたよりも遥かに大きな成果であるといえる。

（2） LLM への構造化行列の適用では、TT、BTT、Monarch などの限定的な構造しか試していない。今年度の後半には、様々なブロック構造、活性化層・転置層・正規化層の位置、MoE などの並列構造も含めて網羅的な探索を行い、真に最適な構造を見つけることができた。これは、Transformer アーキテクチャにもまだ改善の余地があることを示唆する結果であり、意義深いといえる。

III. 精度保証(主担当：荻田)

(1) **疎行列向けの精度保証付き数値計算法**：疎行列系連立 1 次方程式に対する精度保証付き数値計算法は、区間演算のグランドチャレンジに数えられるほどの難問であるが、本成果は直接法の枠組みで、世界初の解決例である。

今後の展望として、並列計算機向けのアルゴリズム開発と実装を実施し、超大規模問題やアプリケーションの問題に対する適用を計画している。また、応用上で重要な、反復解法における精度保証付き数値計算の多くの問題が未解決であり、その解決に取り組む。

(2) **低精度演算を活用した高精度行列積のエミュレーション技術**：一般行列乗算エミュレーションを応用し、対称行列などの特定の構造をもつ行列の乗算やランク k/2k 更新などのエミュレーション手法を開発する。

IV. 自動チューニングと計算機システム （主担当：片桐)

(1) **AT 方式開発**：本年度提案した、高速な演算精度選択が事前にできる深層学習モデルによる AT 手法は、当該分野にとっての汎用性は大きい。今後、多様なアプリケーションへの適用が期待できる。

(2) **電力特性分析**：同一 GPU 上で複数リクエストをバッチ処理または並列実行した場合の電力特性について調査を行う。

以上