jh230053

高性能かつ高信頼な数値計算手法とその応用

荻田 武史（早稲田大学）

概要

今日の大規模かつ複雑化した情報基盤システムにおける数値計算では，演算速度・演算精度の最適化，メモリ・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・エネルギー効率の最適化に向けた検討が必須である．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ーや階層型行列（H 行列）演算等の高速計算に関する研究を推進し，同時にそれらの計算の信頼性及び電力効率を重視しながら，さらに悪条件な実問題に適用可能な実用的な手法の研究開発を実施するものである．

本研究では，疎行列ソルバー，H 行列演算，計算機システムと消費電力測定，精度保証と自動チューニング手法の研究項目を設定し，それぞれの研究成果を横断的に活用する方式で研究を推進する．

今年度は最終年度であり，各研究項目において，概ね最終的な目標を達成できた．

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

(2) 課題分野

大規模計算科学課題分野

(3) 共同研究分野(HPCI 資源利用課題のみ)

超大規模数値計算系応用分野

(4) 参加研究者の役割分担

I. 疎行列ソルバー

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

II. H 行列演算

主担当：横田，担当：岩下・伊田・石井・大川・大友・中村・Sa・Budikafa・Ma・Spendlhofer

III. 計算機システムと消費電力測定

主担当：坂本，担当：塙・星野・河合・近藤・大島・成瀬・堀越

IV. 精度保証と自動チューニング手法

主担当：荻田・片桐，担当：中島・河合・横田・伊田・近藤・尾崎・田中・今村・椋木・寺尾・Marques・Popovici・内野・福原・満田・羽生

2. 研究の目的と意義

本研究の目的は，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ーや H 行列演算等の高速計算に関する研究を推進し，同時にそれらの計算の信頼性及び電力効率を重視しながら，さらに悪条件な実問題に適用可能な実用的な手法の研究開発を実施することである．本研究の遂行は，来たるべきポストムーア時代，さらにそこで重要な役割を果たすことが予想される確率的コンピューティングの発展に貢献するものと期待される．

3. 当拠点の公募型共同研究として実施した意義

JHPCN は多様な計算機環境を備え，東大の Wisteria/BDEC-01 （Wisteria），Oakbridge-CX (OBCX)，名古屋大の不老等，幅広い多様な大規模システムを有し，本研究の目指す高性能・高信頼な数値計算手法の研究には最適である．Wisteria, OBCX では「ノード固定」における設定カスタマイズにより，個別ノードの消費電力測定が可能である．JHPCN は様々な分野の専門家を擁し，本研究の

ような学際的研究を推進する体制を容易に構築でき，北大，東大，東工大，名大，京大，九大各センターから様々な分野の研究者が参加した.JHPCN 各センターはオープンソースソフトウェア活用に積極的であり，本研究の成果を公開，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によって，利用者拡大及びソフトウェアのさらなる改良が可能となる.

4. 前年度までに得られた研究成果の概要

これまでに得られた成果の内，代表的なものについて以下に示す.

(1) 並列多重格子法前処理付き反復法における演算精度の効果及び精度保証についての論文を発表した．さらに，東大情報基盤センターに導入された A64FX 搭載の Wisteria (Odyssey)について，特に FP16 を使用した混合精度演算の評価を実施した．FP16 を前処理に適用することにより，ある程度までの悪条件問題でも安定に解を得られることを確認した.

(2) SuiteSparse Matrix Collection で提供されている，様々な条件を持つテスト行列に対して，FP64 のみに基づく従来法と提案法（FP64 と FP32 を用いた混合精度 GMRES(m)法）の比較実験を実施した．数値実験結果より，提案法は，従来法と同程度の収束性を示し，計算時間の面で優位となることが期待できることが確認された.

(3) Data Analytics アプローチにより過去の計算結果を前処理内で活用することで，Adaptive CG の更なる性能改善を実現した．地震シミュレーションにおいて開発手法を A64FX CPU に適した実装とともに用いることで，従来手法比 10.1 倍の高速化を達成した．また，本アプローチを地震発生の分析に使う粘弾性地殻変動解析に適用することで Adaptive CG ソルバーの性能改善に取り組んだ.

(4) AT 言語である ppOpen-AT の新 AT 機能として混合精度演算を活用し，電力最適化も行える新方式の提案を行った．NICAM を利用した性能評価の結果，全て FP64 の実行時間に対して，部分的に FP16 にすることで 1.12 倍の速度向上，かつ，1.06 倍のエネルギー量削減を得られた．また，演算精度については，部分的に FP16 にしても相対誤差が 3×10^{-4} に維持できることがわかった.

5. 今年度の研究成果の詳細

I. 疎行列ソルバー

(1) 並列疎行列解法について，所望の精度において計算時間，消費エネルギを最小化する最適演算精度を，アプリケーション・係数行列の性質，問題サイズ，ハードウェア環境等に基づき自動チューニング技術（AT）によって動的に制御するための検討を継続して実施した．昨年度は，最適な演算精度を選択する手法を開発するため，より多様な問題に，これまでに開発した精度保証手法を適用し，不均質性の度合いが低精度演算による前処理付き反復法の収束に強く影響することが明らかとなり，最適な演算精度を選択するためには，係数行列の固有値分布を知ることが不可欠であることがより明確となった．本年度はこれらの知見に基づき，岩下武史教授（北大）等の開発した，疎行列固有値ソルバを使用した検討を実施した．当該ソルバは問題サイズの制限があるため，問題の性質を的確に反映するための小規模モデル群作成を進めた.

(2) 反復改良法に基づいた混合精度反復法ソルバに関する研究を実施した．反復改良法の内部ソルバとして，低精度演算を利用した BiCGSTAB ソルバを用いた場合について，数値実験による性能評価を行い，その有効性を収束性の観点から検証した[P1]．内部ソルバとして低精度演算に基づく固定回の GMRES 反復を用いた場合と比較して，特定の問題で優位性を示すことを確認した.

(3) 長期間の地殻変動を評価する粘弾性地殻変動解析に対して，過去の時間ステップの求解結果を用いて反復法ソルバーの初期解を推定し解析を高速化する Adaptive CG を適用し，計算性能を評価した．ここでは，昨年度富岳向けに開発した計算手法をベースに，GPU に適したアルゴリズムを開発し実装することで，NVIDIA A100 GPU 上で高い性能が得られることを確認した[C3]．ここでは過

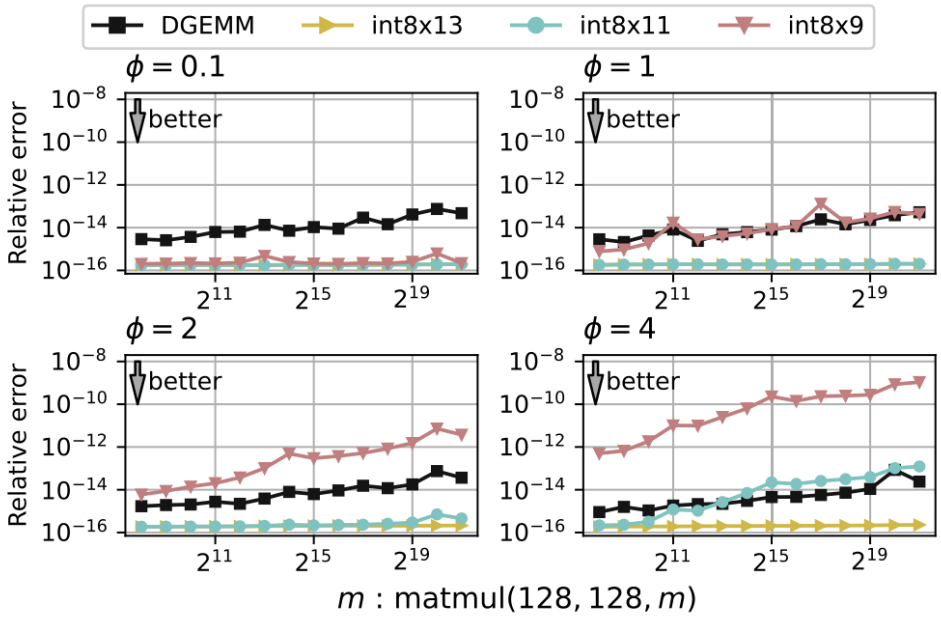

図 1： Int8 テンソルコアを用いたときの行列積の相対誤差

去のタイムステップ結果に基づくデータ駆動型予測手法を用いて初期解を高精度に求めることで，マルチグリッドソルバーの反復回数を減らし，計算コストの削減を実現している．過去のタイムステップの結果を圧縮し，複数の解析ケースを同時に解くことにより，GPU 上で高い実行性能・省メモリを実現し，GPU を使ったマルチグリッドソルバー比で 8.6 倍の高速化を達成した．

II. H 行列演算

低精度演算器を用いて高精度行列乗算を計算する尾崎スキームに着目し，IMMU を使用する利点と欠点を示した．整数(Int8)テンソルコアを用いた実験では，NVIDIA コンシューマーGPU 上の FP16 テンソルコアにおいて，cuBLAS や既存の尾崎スキームの実装よりも倍精度行列乗算を高速に計算できることを示した．さらに，FP64 の精度を維持しながら，量子回路シミュレーションを最大 4.85 倍高速化できることを実証した．この成果は International Journal of High Performance Computing Applications に採択された [P2].

図 1 に Int8 テンソルコアを用いて尾崎スキームにより倍精度演算を行った際の相対誤差を示す．尾崎スキームでは，複数の低精度型変数を用いて倍精度の演算を実現するが，図 1 の「int8x13」，「int8x11」，「int8x9」はそれぞれ 13 個，11 個，9 個の int8 を用いて double を表す手法を表す．図 1 を見ると，9 個では不十分であるが 11 個の int8 を用いることで倍精度を回復できることが分かる．また，図 1 に示される変数 ϕ は以下の式で定義される指数部のレンジを表す．

$$\mathbf{A}_{i,j}, \mathbf{B}_{i,j} = \mathrm{uniform}(-0.5, 0.5) \times e^{\phi \times \mathrm{normal}(0,\,1)}$$

図 2 に Int8 テンソルコアを用いたときの行列積の演算性能と電力消費を示す．GPU は A100，TITAN RTX，RTX A6000，RTX A6000 Ada の 4 種類を用いた．A100 GPU には FP64 のテンソルコアがあるため DGEMM の方が高い演算性能を示すが，その他の GPU では FP64 性能が低いため，提案手法の方がはるかに高い演算性能を示している．

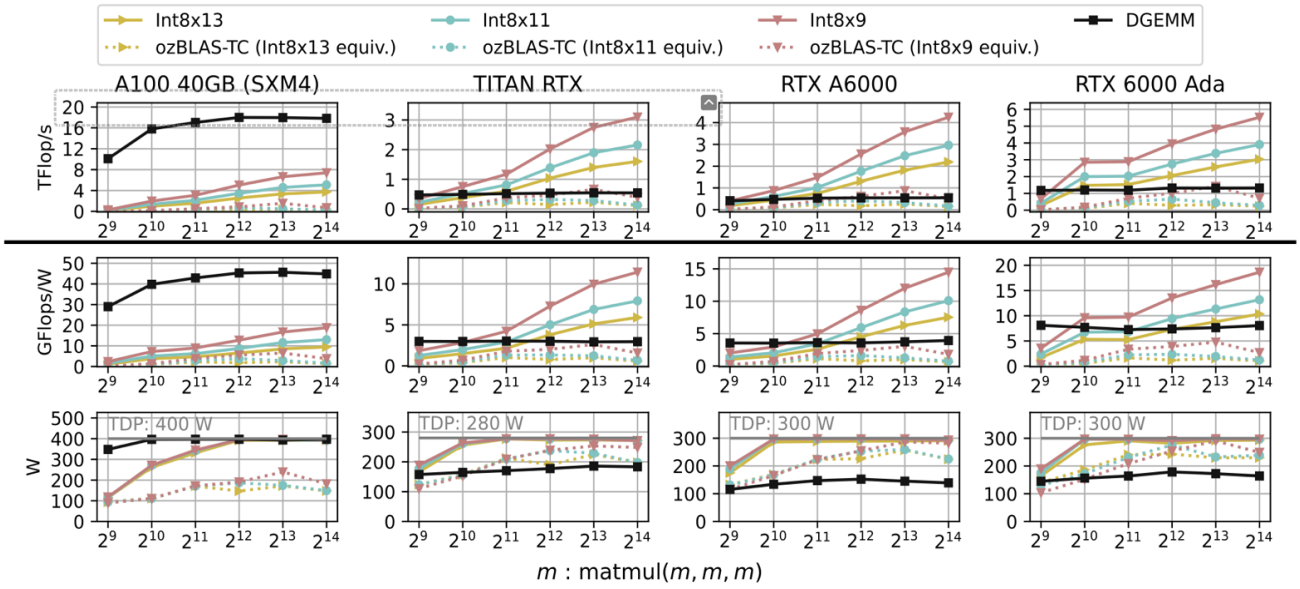

図 2:　Int8 テンソルコアを用いたときの行列積の演算性能と電力消費

III. 計算機システムと消費電力測定

様々な量子化手法が LLM 推論時のトークン生成速度や電力特性に与える影響について調査を進めている．これまでに LLM である Vicuna-v1.1 に対し，FastChat が提案する 8bit 量子化を行った際のトークン生成速度，消費電力，パワーキャップを行った際のエネルギー効率について調査を行った．新たに 1 種の LLM，4 種の量子化手法，ローエンド GPU を追加し，同様の調査を進めている．具体的には Vicuna-v1.5 の 7B,13B のモデルを追加し,bitsandbytes の 8bit/4bit，GPTQ の 8bit/4bit 量子化についての特性分析を行った．さらに，RTX4090 と RTX4060Ti の特性の差異を調査している．本調査内容は国際会議に投稿中である．

IV. 精度保証と自動チューニング手法

(1) 上記 I，II と連携して，疎行列ソルバー及び H 行列演算の実アプリケーションに精度保証法／精度推定法を組み込んでいる．

本研究では，より一般的な問題で精度保証手法の効果を確認するために，同精度保証手法を閾値付き不完全コレスキー分解(IC(t))を前処理とした CG 法に適用した．適用に際しては，これまでは精度保証法の適用が M 行列性を持つ係数行列に限定されていたが，H 行列性を持つ問題でも評価可能なように拡張した．実際に，Florida Sparse Matrix Collection から M または H 行列性を持つ問題を取得し，精度保証付き IC(t)CG 法を用いて精度保証を実施した結果を図 3 に示す．評価に使用した行列は 45 個であり，その全てで行列，ベクトルを単精度または倍精度で格納した場合の結果を示している．なお，図の ss は行列・ベクトルともに単精度で，sd は行列を単精度，ベクトルを倍精度で，dd は行列・ベクトルともに倍精度で格納した結果を示す．評価の結果，ベクトルを倍精度で格納すればある程度の精度が維持できている．対して，ベクトルを単精度化すると，演算精度の低下が確認できる．なお，いくつかの行列では，行列ベクトルの単精度化によって M,H 行列性を維持できなくなり，データが取れないパターンが発生している．さらに，行列番号の 38～40 では，行列のみの単精度化でも M ないし H 行列性を維持的なくなる現象を確認している．

なお，ここで紹介した精度保証の実装を補助するコード群および精度保証付き IC(t)CG 法は bitbucket にて公開している[L1,L2]．

(2) 前年度開発した ppOpen-AT を活用した最新 AT 手法を入れ込み，混合精度演算時時の演算精度と

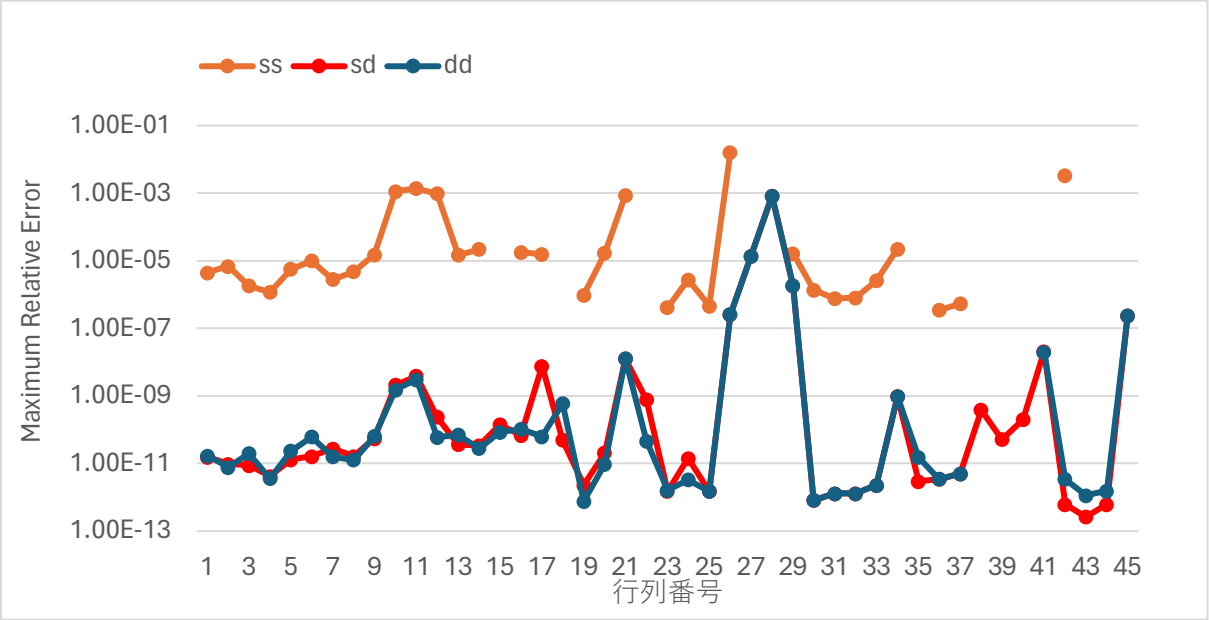

図 3: Florida Sparse Matrix Collection から取得した行列での精度保証付き IC(t)CG 法の適用結果（縦軸は最大相対誤差を，横軸は異なる行列を示す）

実行時間の最適化に資する新しい自動チューニング（AT）方式を開発した．具体的には，ppOpen-AT の従来機能には無い，指定された複数の対象領域に対する単精度化を行う混合精度演算について，どの領域への単精度演算適用をすると，精度と速度の観点から良いかを自動評価できる AT 機能を実現した．最適となる組合せの実装（ソースコード）を自動生成することが可能となる．本成果は，査読付国際会議論文に採録され，成果を国際会議で発表した[C2]．

以上に加えて，反復解法である PICCG 法の性能パラメタチューニングに適用できる AT 機構の開発を行った．本開発は昨年度成果の機能を拡張したものである．具体的には，PICCG 法の性能パラメタである，不完全コレスキー分解（IC）前処理の性能パラメタである，0 とみなす閾値とフィルイン深さの 2 つの性能パラメタを説明変数として，実行時間を予測する AI モデルを深層学習により生成した．この AI モデルが，反復解法アルゴリズム上の意味において妥当な回答をするかどうかの検証を，説明可能 AI（XAI）ツールにより分析する研究である．また本研究では，初めて倍精度計算と単精度計算の AI モデルを作成して，その AI モデルの妥当性検証を XAI で行った．本成果により，単精度と倍精度でどちらが高速化なのかを実行時に予測し，かつ，実行時に適する演算精度の実装に切替える AT 機構のための AI モデルが構築できることを示したものである．そのため，PICCG 法を用いるソフトウェアの混合精度演算の普及に寄与する本質的な研究成果といえる．なお本成果は，情報処理学会全国大会で発表し，全国大会学生奨励賞を受賞した[D6]．

6. 進捗状況の自己評価と今後の展望

低精度演算・データを積極的に活用する混合精度版アルゴリズムの研究開発について，混合精度演算を利用した反復改良法の枠組みにおいて，従来は内部ソルバとして不適と思われていた双ランチョス系の解法が有効性を持つ可能性を新たに示したことは当該研究分野において重要な進展と考えられる．一方で，線形反復法分野で利用が一般的な前処理を含めた混合精度ソルバについては，現時点で研究途上であり，今後の課題として位置付けたい．

精度保証と自動チューニング手法について，本年度に，PICCG 法における倍精度演算と単精度演

算の実行時間予測のための AI モデル構築のめどがついた．そのため，今後の進展としては，この倍精度演算と単精度演算の切替え機能の予測精度を,AT機構の観点から評価することがあげられる．

7. 研究業績

(1) 学術論文 （査読あり）

[P1] Y. Zhao, T. Fukaya, T. Iwashita: Numerical behavior of mixed precision iterative refinement using the BiCGSTAB method, J. Inf. Process., Vol. 31, pp. 860-874, 2023.

[P2] H. Ootomo, K. Ozaki, R. Yokota: DGEMM on Integer Matrix Multiplication Unit, International Journal of High Performance Computing Applications, accepted. (DOI: 10.1177/10943420241239588)

[P3] R. Yoda, M. Bolten(+), K. Nakajima, A. Fujii, Coarse-grid operator optimization in multigrid reduction in time for time-dependent Stokes and Oseen problems, Japan Journal of Industrial and Applied Mathematics, accepted, 2024. (DOI: 10.1007/s13160-024-00652-8)

[P4] A. Fujii, T. Tanaka, K. Nakajima: Light Weight Coarse Grid Aggregation for Smoothed Aggregation Algebraic Multigrid Solver, IEEE Access, accepted, 2024 DOI: 10.1109/ACCESS.2024.3386226

[P5] Y. Chen, K. Nakajima: A Cascadic Parareal Method for Parallel-in-Time Simulation of Compressible Supersonic Flow, IPJS Transaction on Advanced Computing Systems, in press, 2024

(2) 国際会議プロシーディングス （査読あり）

[C1] K. Nakajima: Communication-Computation Overlapping for Parallel Multigrid Methods, IEEE Proceedings of iWAPT 2024 in conjunction with IPDPS 2024 (in press), 2024

[C2] X. Ren, M. Kawai, T. Hoshino, T. Katagiri, T. Nagai: Auto-tuning mixed-precision computation by specifying multiple regions, CANDAR 2023, January 25, 2024. https://doi.org/10.1109/CANDAR60563.2023.00031

[C3] S. Murakami, K. Fujita, T. Ichimura, T. Hori, M. Hori, L. Maddegedara, N. Ueda: Development of 3D Viscoelastic Crustal Deformation Analysis Solver with Data-Driven Method on GPU, ICCS 2023. Lecture Notes in Computer Science, 14074. Springer, 2023.

(3) 国際会議発表（査読なし）

[I1] K. Nakajima: Communication-Computation Overlapping in Parallel Multigrid Methods, 21st Copper Mountain Multigrid Conference, Copper Mountain, Colorado, April 19, 2023.

[I2] Y. Zhao, T. Fukaya, T. Iwashita: Numerical Evaluation of Mixed Precision Iterative Refinement using Low Precision Krylov Methods: 10th International Congress on Industrial and Applied Mathematics (ICIAM 2023 Tokyo), poster, August 22, 2023.

(4) 国内会議発表（査読なし）

[D1] 中島研吾: 並列多重格子法における通信・計算オーバーラップの最適化, SWoPP 2023, 函館市, 2023 年 8 月 3 日.

[D2] 村上颯太, 藤田航平, 市村強, 堀高峰, 堀宗朗, M. Lalith, 上田修功: GPU におけるデータ駆動型手法を用いた粘弾性地殻変動解析手法の開発, 第 26 回応用力学シンポジウム, 2023 年 5 月 28 日.

[D3] 深谷猛, Z. Yingqi, 岩下武史: ILU(0)前処理付きGMRES(m)法に対する低精度計算の導入可能性の検証, 情報処理学会 研究報告ハイ

パフォーマンスコンピューティング, 2023-HPC-192, 36, 2023 年 12 月 6 日.

[D4] 深谷猛, Z. Yingqi, 岩下武史: 低精度計算を活用した混合精度型疎行列ソルバーの可能性, 第 15 回 自動チューニング技術の現状と応用に関するシンポジウム（ATTA2023）, 2023 年 12 月 22 日.

[D5] Y. Zhao, T. Fukaya, T. Iwashita: Performance Evaluation of Mixed Precision Iterative Refinement using Low Precision Krylov Methods, 日本応用数理学会 若手の会 第 9 回学生研究発表会, poster, 2024 年 3 月 7 日.

[D6] 中谷崇真, 河合直聡, 片桐孝洋, 星野哲也, 永井亨: ICTCG 法の実行時間予測モデルに対する説明可能な AI の適用, 情報処理学会第 86 回全国大会, 2024 年 3 月 15 日. （全国大会 学生奨励賞 受賞）

(5) 公開したライブラリなど

[L1] lib_verify: https://bitbucket.org/naosou/lib_verify/src/master/

[L2] ICTCG: https://bitbucket.org/naosou/ictcg/src/master/

(6) その他（特許, プレスリリース, 著書等）