jh230010

QR 分解に関する高性能計算技術の研究

深谷 猛（北海道大学・情報基盤センター）

概要

線形計算は科学技術計算の基盤技術の一つであり、行列分解アルゴリズムはその代表例である。本研究課題では、主要な行列分解の一つである QR 分解とそれに関連する高性能計算技術について、参加研究者が協力して、数値計算アルゴリズムや実装方法等の研究開発を進める。2023 年度は、縦長行列の列ピボット付き QR 分解に対する高性能なコレスキー QR 型アルゴリズムを新たに開発するとともに、非縦長行列の QR 分解に対するコレスキー QR 型アルゴリズムの応用方法に関する研究を実施した。また、分散並列環境におけるタイル型アルゴリズムの実装および性能最適化に関する研究や、Block Low Rank 行列の QR 分解アルゴリズムの GPU 環境での実装方法に関する研究に取り組んだ。これらの研究を通して得られた成果を踏まえて、2024 年度も継続課題として、引き続き、QR 分解と関連高性能計算技術の研究開発を進める予定である。

1 共同研究に関する情報

1.1 共同研究を実施した拠点名

- 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
- 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

1.2 課題分野

- 大規模計算科学課題分野

1.3 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

1.4 参加研究者の役割分担

- 深谷 猛（北海道大学）：課題代表、全体統括、コレスキー QR アルゴリズム関連の研究開発
- 鈴木 智博（山梨大学）：課題副代表、統括補佐、タイルアルゴリズム関連の研究開発
- 大島 聡史（九州大学）：BLR 行列の QR 分解関連の研究開発、GPU 実装に関する助言
- 伊田 明弘（海洋研究開発機構）：BLR 行列の QR 分解関連の研究開発
- 岩下 武史（北海道大学・京都大学）：並列処理・並列実装に関する助言、階層行列（H 行列）関連の助言
- 門倉 陣之介（北海道大学・大学院生）：コレスキー QR アルゴリズム関連の研究開発

2 研究の目的と意義

与えられた行列を都合のよい行列の積に分解する計算（行列分解）は、数値線形代数分野の基盤技術の一つであり、様々なアプリケーションにおいて利用されるとともに、数値線形代数

アルゴリズムを構成する部品としての役割も持つ。そのため、行列分解計算の高性能化（高速化）は重要な課題であり、ハードウェアの特徴を考慮した上で、アルゴリズムから実装方法までの多岐にわたって、新規手法の開発や既存手法の改良を行うことが求められている。

本研究課題では、主要な行列分解計算の一つである、行列の QR 分解に着目する。QR 分解は、与えられた行列を直交行列と上三角行列の積に分解する計算であり、最小二乗問題の求解が代表的な応用例である。また、行列の固有値計算や特異値計算とも密接な関わりを持っている。加えて、ベクトルの直交化（直交基底の生成）が QR 分解と等価であるため、数値安定性を向上させる目的等で（ブロック版の）クリロフ部分空間法などでも頻出する。したがって、QR 分解の性能を向上させることで、様々な科学技術計算の効率化に貢献することができる。

QR 分解に関しては、異なる特徴を持った様々なアルゴリズムが存在する。一方、計算が行われる環境もマルチコア CPU、GPU、分散並列システムなど多種多様である。更に、計算対象の行列も縦長行列から正方行列まで多様な形状があり、加えて、最近では Block Low Rank（BLR）行列の QR 分解のような新しい問題設定も登場している。そのため、対象とする状況に応じて、既存のアルゴリズムを改良したり、新しいアイディアに基づいたアルゴリズムを開発したりすることが求められる。

以上のような背景を踏まえた上で、本研究課題では、QR 分解とその関連技術に関して、異なる知識や研究実績を有する研究者を集め、相互に協力しながら QR 分解とその関連技術の高度化に関する研究開発に取り組む。具体的には、各参加研究者と関わりの深い、HPC に適したアルゴリズム（コレスキー QR 型アルゴリズム）、超並列環境に適した実装技術（タイル型行列分解）、GPU 環境、新しい応用（BLR 行列）をベースにした研究開発を、JHPCN で利用可能な多様な計算機を活用して実施することで、QR 分解とその関連技術の高性能化に資する新しい知見や技術を創出することを目指す。最初から特定の限られた問題設定のみを考えるのではなく、「QR 分解」を共通のキーワードとして、各参加者のアイディアを柔軟に組み合わせることで、様々な状況における QR 分解の高性能化の可能性を追及することが、本研究課題の大局的な目的である。

3 当拠点公募型研究として実施した意義

JHPCN では特徴の異なる多種多様な計算機環境を使用可能であり、QR 分解に代表される基本的な線形計算技術の研究開発に最適である。課題代表者が過去に発表した論文（T. Fukaya, PDCAT 2022）のように、統一的な視点で、各計算機上でのアルゴリズムの性能を評価・分析することで、各アルゴリズムの特徴（長所・短所など）が明確になり、アルゴリズム開発者とアルゴリズム利用者の双方に有益な知見を得ることができる。また、各センターのシステム運用担当者等と協力しながら、本研究課題を通して開発したアルゴリズムをライブラリやベンチマークとして整備・提供することで、各センターのスパコンで実行されているアプリケーションの高性能化や新規システムの性能評価に貢献することも期待される。

4 前年度までに得られた研究成果の概要

新規課題のため、該当せず。

5　今年度の研究成果の詳細

各実施項目に対する 2023 年度の主な研究成果は以下の通りである。

(1) コレスキー QR に基づく列ピボット付き QR 分解の開発とマルチコア CPU 環境での性能評価

縦長行列に対して、Rank Revealing QR 分解の一種で行列の低ランク近似等の応用を持つ、列ピボット付き QR 分解（QRCP: QR factorization with Column Pivoting）を計算するコレスキー QR 型アルゴリズムの開発を行う。縦長行列の通常（列ピボットなし）の QR 分解に関して、コレスキー QR 型アルゴリズムが現在の計算機環境に適していることが知られており、その HPC 向きの特徴を維持した上で、縦長行列の QRCP を計算する新しいコレスキー QR 型アルゴリズムを開発することが目標である。

2023 年度の成果として、目標としていたアルゴリズムを開発することができた。QRCP に対するコレスキー QR 型アルゴリズムは、数学的には自明であるが、数値計算の場合、丸め誤差の影響により、ピボットの選択を誤るという問題が生じる。これに対して、いくつかの予備実験を通して、ピボット選択の正誤に関する知見を収集し、経験的ではあるが、十分な信頼性が期待でき、同時に、従来のコレスキー QR 型アルゴリズムが有する HPC に適した特徴を維持した形でアルゴリズムを構成することに成功した。

開発したアルゴリズムのマルチコア CPU 環境での性能評価として、本研究課題に配分された JHPCN 構成機関のいくつかのスパコンシステムの 1 ノードを用いた実験を実施した。実験の結果、（Intel MKL や富士通製の）LAPACK で提供されている既存アルゴリズム（列ピボット付きハウスホルダー QR）に対して、状況次第では 30 倍近く高速であることが確認でき、新しく開発したアルゴリズムの有効性を示すことができた（図 1）。また、東京大学情報基盤センターの大規模 HPC チャレンジの制度を利用して、大規模分散並列環境（OBCX および BDEC-Odyssey）における性能評価も行った。我々のアルゴリズムは、既存アルゴリズムに対して、多くの条件で高速となり、特に、Intel CPU で構成されたシステム（OBCX）では、最大で 25 倍強の高速化が確認できた（図 2）。今回開発したアルゴリズムは、従来のコレスキー QR 型アルゴリズムが持つ「実装の容易さ」の特徴も維持しているため、今後、多くのアプリケーションプログラムで活用されることが期待される。

以上の成果について、基本となるアイディアと初期段階の性能評価結果は ISC2023 のポスター [3] で発表し、主たる結果は IPDPS2024 の論文 [2] で発表予定である。また、成果の概要は、SIAM PP24 で口頭発表 [5] し、東京大学情報基盤センターの大規模 HPC チャレンジに関連した内容については、同センターが発行するスーパーコンピューティングニュースで報告 [11, 12] している。

(2) 非縦長行列の QR 分解に対するコレスキー QR 型アルゴリズムの適用と分散並列環境における性能評価

項目 (1) の部分で述べたように、縦長行列の QR 分解に対して、コレスキー QR 型アルゴリズムは非常に有効である。しかし、計算対象の列数が行数に対して十分に少なくない（非縦長行列）の場合、コレスキー QR 型アルゴリズムの性能に限界があることが分かっている。この状況に対して、本研究課題の開始前に、我々は Block Gram-Schmid のアルゴリズムとコレスキー QR 型アルゴリズムを組み合わせる手法

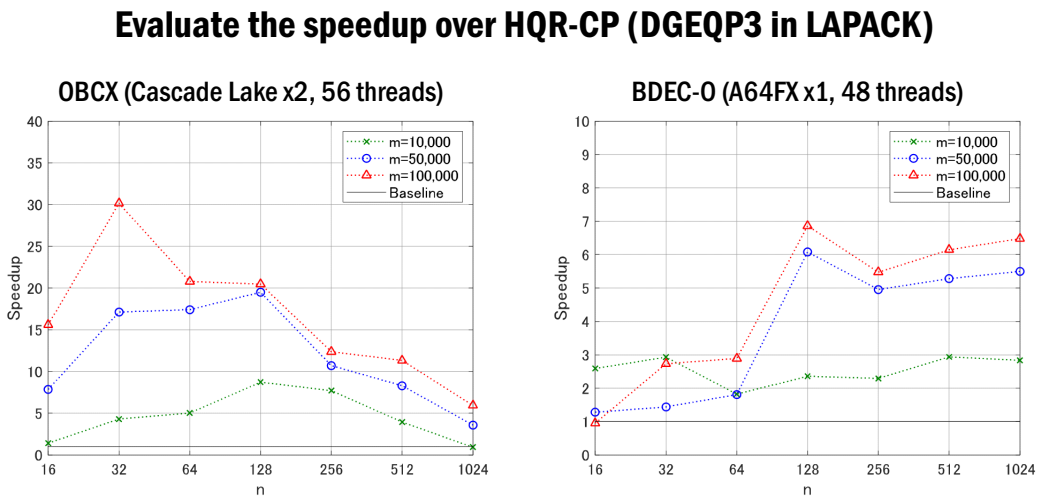

図 1　縦長行列の QRCP に対するコレスキー QR 型アルゴリズムの性能評価結果の例：マルチコア CPU 環境（[5] の発表スライドより引用）

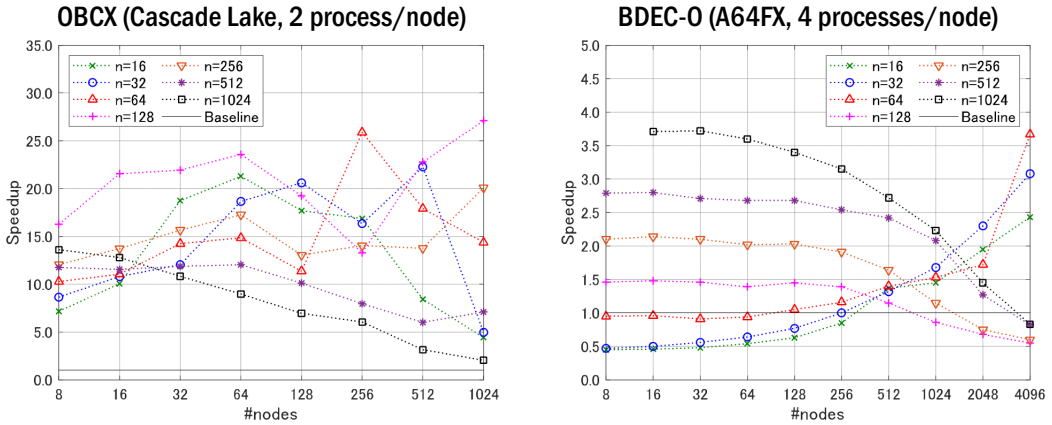

図 2　縦長行列の QRCP に対するコレスキー QR 型アルゴリズムの性能評価結果の例：分散並列環境（[5] の発表スライドより引用）

を提案し、マルチコア CPU 環境における有効性を確認している。この成果を踏まえて、本研究課題では、MPI を用いた分散並列実装を行い、実際のシステム上でその有効性を検証することを目標とする。

2023 年度の成果として、再直交化付き Block Classical Gram-Schmidt（BCGS2）アルゴリズムと反復型コレスキー QR アルゴリズム（Ite-

CholQR）を組み合わせた手法を、列方向の 1 次元ブロックデータ分散を用いて、MPI による分散並列実装した。その上で、本研究課題に配分された北海道大学の Grand Chariot と東京大学の BDEC-Odyssey を用いて、実行可能であったノード数までの性能評価を実施した。性能評価の結果、分散並列環境においても、我々のアプローチが有効であり、非縦長行列の QR 分解に対して、既存のコレスキー QR 型アルゴリズムを直接適用するよりも、BCGS2 と併用する形で適用する方が高速となることを確認できた（図 3)。また、BCGS2 におけるブロック幅と計算時間の関係についても詳しい実験を行い、適切なブロック幅の決定方法の考案に向けて有益となる知見を収集できた。

上述した、本項目に関する主たる成果は、2023 年 12 月開催の HPC 研究会で報告 [8] した。また、本成果は、修士課程の学生が主に担当したものであり、修士論文としてまとめるとともに、日本応用数理学会若手の会主催の学生研究発表会（2024 年 3 月）でポスター発表 [10] を行っている。

(3) 最新のマルチコア CPU 環境におけるタイル QR アルゴリズムの詳細な性能評価

行列分解に対するタイル型アルゴリズムは、従来アルゴリズム（とそれに対する fork-join 型の並列実装）と比べて、超並列環境に適したアルゴリズムおよび並列実装であり、QR 分解をはじめとする種々の行列分解計算に関して、その有効性が確認されている。本研究課題では、最新のマルチコア CPU を含む多様な計算機環境において、タイル型アルゴリズム（タイル QR）の性能を改めて検証する。具体的には、行列の形状（行と列の比率など）やアルゴリズム内のパラメータ（タイルサイズなど）と性能の関係を調査する。これにより、タイル型アルゴリズムの有効性や課題を明確にし、その利用や将来的な分散並列化において有益となる知見を収集することを目指す。

2023 年度の成果として、これまでに鈴木が中心となって開発を進めてきたタイル QR アルゴリズムをベースに、アルゴリズムと実装方法のレビューを行うとともに、本研究課題に配分されたいくつかの計算機環境における動作確認を行った。また、これと並行して、BLR 行列のタイル型コレスキー分解アルゴリズムの分散並列実装を行い、実機上での性能評価を実施した。開発したプログラムに関して、プロセス数，タスク数，タイルサイズなどのパラメータと性能の関係を調査・分析し、それぞれのチューニングを行った（図 4)。コレスキー分解と QR 分解は、タイル型アルゴリズムの内部の計算カーネルは異なるが、タイル型アルゴリズムの実装の観点では共通点が多く、今後のタイル QR アルゴリズムの分散並列実装の開発やチューニングに向けて有益となる知見を収集することができた。

上記の BLR 行列に対する分散タイルコレスキー分解に関する研究成果は、東京で開催された応用数理分野最大の国際会議（ICIAM）で概要を報告 [4] するとともに、HPC Asia2024 の会議論文 [1] として採択された。

(4) GPU 環境における Block Low Rank 行列の QR 分解アルゴリズムの性能評価

近年、BLR 行列をはじめとする行列の低ランク近似を活用した手法とそれに関する行列計算の研究が活発である。本研究課題の一部のメンバーが参加している JHPCN の別課題において、既存ライブラリのルーチンを利用する形で、BLR 行列の QR 分解の GPU 実装が進められており、本研究課題では、QR 分解に関するより専門的な立場から、この高度化を目指す。

2023 年度の成果として、大島を中心に開発

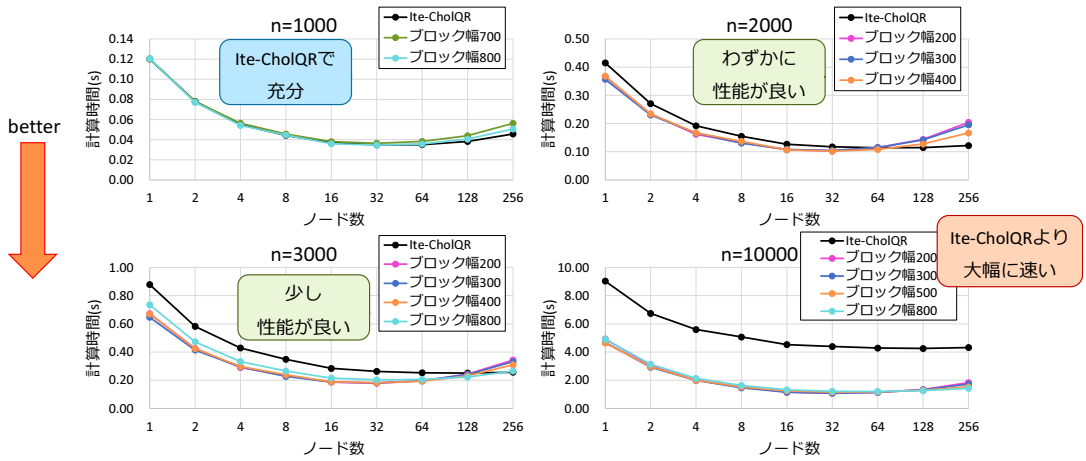

図 3　非縦長行列の QR 分解に対する BCGS2 とコレスキー QR 型アルゴリズムを組み合わせた手法の分散並列実装（一次元ブロックデータ分散採用）の性能評価結果（東大 BDEC）の例（[8] の発表スライドより引用）

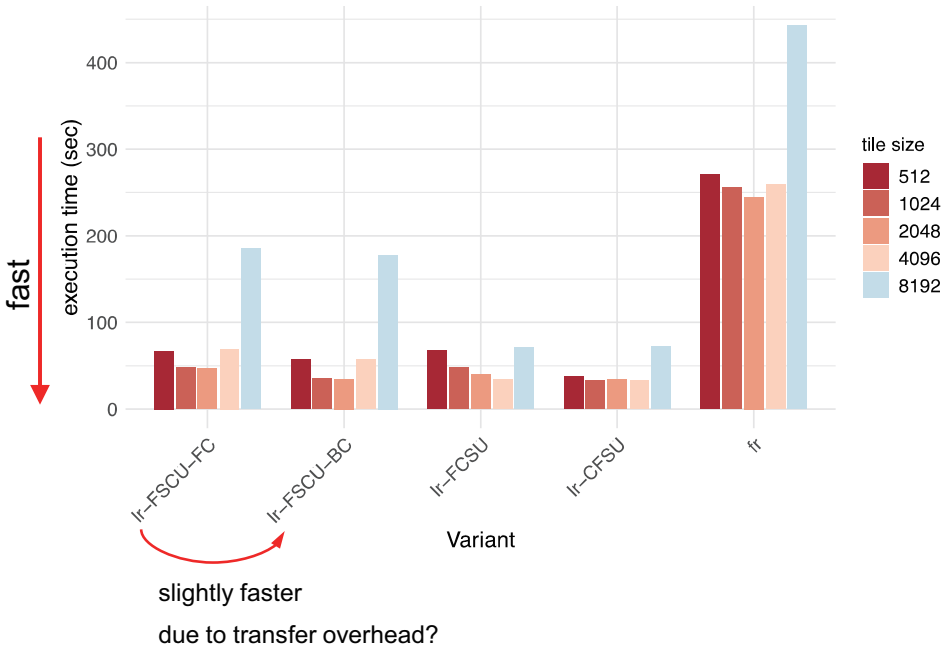

図 4　BLR 行列に対するタイルコレスキー分解の分散並列プログラムにおけるタイルサイズと実行時間の関係の検証結果例（[1] の発表スライドより引用）

が進められている BLR 行列向けの QR 分解の GPU 実装について、本研究課題のメンバーで、現在のアルゴリズムや実装方法の検討を行うとともに、実機上での性能評価に取り組み、今後の GPU 実装の改良に向けた課題を整理した。また、BLR 行列の QR 分解の計算を GPU 上で

行う場合、素朴な形で GPU 化を行うと、最新の GPU の性能を十分に活用することが難しいことが分かっている。この課題を解決するために、MPS や MIG といった複数プロセスによる GPU の共有利用の方法を適用し、GPU 利用を効率化による計算の高速化を行った（図 5）。本アプローチについて、実機上での性能評価の結果、十分な有効性が期待できることが確認できた。

上述の成果の一部は、2023 年 8 月開催の HPC 研究会で報告 [7] している。また、自動チューニングと HPC に関する国内外の会議においても本成果を発表 [9, 6] している。

6　今年度の進捗状況と今後の展望

本研究課題の申請時に設定した具体的な実施項目については、前節の記載の通り、十分な成果が得られたと考える。項目間で申請時に設定した目標の達成率に多少の差は生じているが、一方で、申請時に想定していなかった成果も得られている。項目 (1) と (3) については、研究成果を査読付き国際会議論文として発表できており、項目 (2) と (4) についても、HPC 研究会の研究報告として発表済みで、査読付き論文を十分に目指せる成果が既に得られている。項目 (1) に関しては、JHPCN 以外の制度を活用して、当初の計画を大きく上回る成果を得ることができている点は特筆できる。以上の状況を踏まえると、2023 年度の進捗状況は、全体として十分な結果であると考える。

本研究計画の申請時、初年度の目標として、「課題参加者同士が協力して研究開発を進めるための土台を整備すること」を掲げており、これは十分に達成できたと考える。継続課題として、2024 年度も無事に採択されたので、二年目は研究者間の協力をより強化して、研究開発に取り組む。特に、計画している各実施項目に関して、項目間で知見や技術を交換しながら、研究開発を進めることで、QR 分解に関して異なる知見や技術を持つ研究者が参加している本研究課題の強みを生かせるようにしたい。

配分された計算資源の利用に関して、一部の資源を十分に活用できなかった点は一年目の反省点である。本研究課題は、申請段階で計画したシミュレーションを実行する種の課題と比べて、アルゴリズムから実装（プログラム）までの開発が伴う課題であるため、申請時の計画通りに配分された計算資源を使用するのが難しいこともある。しかしながら、研究開発の進捗状況と配分資源の利用状況を定期的に確認し、必要に応じて計画の見直しを行うことで、配分資源をできるだけ有効利用できるように努めたい。

7　研究業績一覧（発表予定も含む）

学術論文 （査読あり）

なし

国際会議プロシーディングス（査読あり）

[1] Han Jiao, Jilin Zhang(+), Tomohiro Suzuki, Task-based low-rank hybrid parallel Cholesky factorization for distributed memory environment, International Conference on High Performance Computing in Asia-Pacific Region (HPCAsia '24), pp. 107–116, 2024.

[2] Takeshi Fukaya, Yuji Nakatsukasa(+), Yusaku Yamamoto, A Cholesky QR type algorithm for computing tall-skinny QR factorization with column pivoting, 2024 IEEE International Parallel and Distributed Processing Symposium (IPDPS 2024), 13-page, 2024. (to be published)

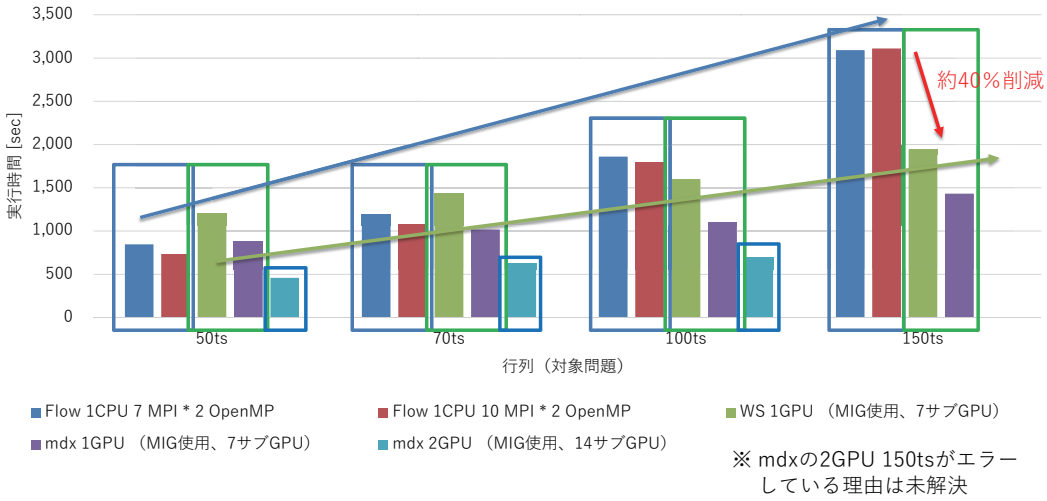

図 5　BLR 行列の QR 分解に関する MIG を用いた GPU 実装の有効性の評価結果例（[7] の発表スライドより引用）

国際会議発表（査読なし）

[3] Takeshi Fukaya, Yuji Nakatsukasa(+), Yusaku Yamamoto, Tall-skinny QR factorization with column pivoting by a Cholesky QR type algorithm, ISC High Performance 2023, Hamburg, Germany, May, 2023. (reviewed, poster)

[4] Tomohiro Suzuki, Task-based hybrid parallel matrix factorization for distributed memory environment, 10th International Congress on Industrial and Applied Mathematics (ICIAM '23 Tokyo), Minisymposia on Progress and Challenges in Extreme Scale Computing and Big Data, Tokyo, Japan, August, 2023.

[5] Takeshi Fukaya, Yuji Nakatsukasa(+), Yusaku Yamamoto, QRCP of a Tall-skinny Matrix by a Cholesky QR Type Algorithm, SIAM Conference on Parallel Processing for Scientific Computing (PP24), Baltimore, USA, March, 2024.

[6] Satoshi Ohshima, Considering multi process calculations on current GPU, ATAT in HPSC 2024, Hsinchu, Taiwan, Marchi, 2024.

国内会議発表（査読なし）

[7] 大島 聡史, 伊田 明弘, 河合 直聡, 横田 理央, 山崎 市太郎 (+), CUDA Fortran ＋ MIG ＋ UVM を用いた BLR 行列 QR 分解の大規模高速化, 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（SWoPP2023）, Vol. 2023-HPC-190, No. 14, pp. 1–8, 函館市, 2023 年 8 月.

[8] 門倉 陣之介, 深谷 猛, 佐竹 祐樹, 岩下 武史, 分散並列環境における CholeskyQR と BCGS2 を用いた非縦長行列の QR 分解, 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, Vol. 2023-HPC-192, No. 20, pp. 1–15, 那覇市, 2023 年 12 月.

[9] 大島 聡史, 伊田 明弘, 横田 理央, 山崎 市

太郎 (+), 一万計算コア超時代の GPU に向けたプログラム最適化と自動チューニングを考える, 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）, 東京都, 2023 年 12 月.

[10] 門 倉　陣 之 介, 深谷 猛, 佐竹 祐樹, 岩下 武史, CholeskyQR と BCGS2 を用いた非縦長行列の QR 分解. 日本応用数理学会 若手の会 第 9 回学生研究発表会, 長岡市, 2024 年 3 月.（ポスター）

公開したライブラリ等

なし

その他（特許，プレス発表，著書等）

[11] 深谷 猛, 大規模分散並列環境におけるコレスキー QR 型アルゴリズムによる縦長行列の列ピボット付き QR 分解の性能評価, 東京大学情報基盤センター スーパーコンピューティングニュース, Vol. 25, No. 4, pp. 20–28, 2023.

[12] 深谷 猛, 大規模分散並列環境におけるコレスキー QR 型アルゴリズムによる縦長行列の列ピボット付き QR 分解の性能評価（続）, 東京大学情報基盤センター スーパーコンピューティングニュース, Vol. 26, No. 2, pp. 52–58, 2024.