jh230005

ソフトウェア工学による自動チューニング技術の新展開

片桐孝洋（名古屋大学）

概要

高性能な数値計算ソフトウェアを開発する際、性能チューニングは大きな開発コスト（工数）が必要となるだけでなく専門知識が必要とされ、誰もが行えるわけではない。一方、高性能数値計算ソフトウェアの開発工数を削減する目的でのソフトウェア工学の研究は、我が国ではほとんど行われていない。そこで本提案では、自動チューニング（AT）、GPU 最適化、ソフトウェア工学を専門とするコンピュータサイエンス学者と、計算化学と量子アルゴリズム、固有値・非線形ソルバを専門とする応用数理（計算科学）者との協業による学際領域研究を推進し、新しい AT 研究の境地を開拓する。

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

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

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

- **代表者**:片桐孝洋（研究統括・AT 開発）
- **副代表者**:星野哲也（GPU 最適化、AT 開発）
- **共同研究者**:森崎 修司(ソフトウェア工学)、大島聡史(GPU 最適化、AT 開発)、中島研吾(連立一次方程式ソルバー)、Osni Marques(AT 開発、固有値問題)、Weichung Wang（AT 開発、固有値問題)、Feng-Nang Hwang（非線形ソルバ)、望月祐志（化学計算、量子計算)、杉崎研司（化学計算、量子計算）
- **共同研究者（大学院生）**:森下誠(AT 開発、量子計算)、福原諒河 （AT 開発、量子計算)、羽生達郎（AT 開発、AI 適用)、湯淺義尚（AT 開発、AI 適用)、IVAN LUTHFI IHWANI（非線形ソルバ)、任軒正博（AT 開発、AI 適用)、植野貴大(AT 開発、AI 適用)、水島慎吾(AT 開発、AI 適用)

2. 研究の目的と意義

【目的】高性能な数値計算ソフトウェアを開発する際、性能チューニングは大きな開発コスト（工数）が必要となるだけでなく専門知識が必要とされ、誰もが行えるわけではない。一方、高性能数値計算ソフトウェアの開発工数を削減する目的でのソフトウェア工学の研究は、我が国ではほとんど行われていない。そこで本提案では、自動チューニング（AT)、GPU 最適化、ソフトウェア工学を専門とするコンピュータサイエンス（CS）学者と、計算化学と量子アルゴリズム、固有値・非線形ソルバを専門とする応用数理（計算科学）者との協業による学際領域研究を推進し、新しい AT 研究の境地を開拓する。

【意義】

Ⅰ）ソフトウェア工学：数値計算ソフトウェアにおけるソースコードの更新、コンパイル、静的解析、ビルド、自動テストの実行といった一連の

更新手順を自動化し、効率化する継続的インテグレーションにおいて、自動テストの実行順序を工夫することにより、デバッグを含む開発効率を高める手法が提案されている。これまで、Fault-prone module prediction といったバグを含みやすいソフトウェアモジュールを予測する研究はあったが、予測の対象はソフトウェアモジュールでありテストケースではない。また、実行する前にどのようなテストケースによって問題が見つかりやすいかを予測することは対象ソフトウェア、開発プロセス、開発者に依存する可能性が高く、その予測手法は明らかではない。

本課題では、数値計算ライブラリ LAPACK を事例にして、テストケースにおいて本課題の解決を狙う。加えて、AT 技術の研究は、高性能化のためのパラメタチューニングやコード自動生成で工数削減に寄与するのが自明であり、生産性の観点からソフトウェア工学上の意義をなす。

本提案では、代表者の片桐が開発した AT 言語の ppOpen-AT フレームワーク、および IDEAS-ECP プロジェクト開発の AT ツールを活用する。

Ⅱ）実用アプリケーション評価：本課題では、計算化学における量子アルゴリズムへの適用を想定し、量子回路シミュレーション（特に GPU で稼働する NVIDIA cuQuantum）の高速化に寄与する性能パラメタチューニングを対象にする。

一方、AT の適用アルゴリズムを拡大するため、非線形ソルバのアルゴリズム上に現れる性能パラメタチューニングも取り扱う。

Ⅲ）チューニングノウハウの共有：本提案では、名大「不老」と東大「Wisteria/BDEC-01」の CPU と GPU を利用する。そのため、「富岳」の CPU である ARM A64FX と GPU の V100 と A100 の高性能実装技法やパラメタチューニングのノウハウが集約される。そのノウハウを論文出版等で公開することで、拠点や国家スパコンの利用技術の知見が共有化される。

Ⅳ）国際連携：参画する Marqus 博士、王教授と黄教授は応用数理分野の著名な研究者であるため、本課題の成果が国際的にも周知されることで国際連携を強力に進めることができる。

3. 当拠点の公募型共同研究として実施した意義

本提案は CS 学者と計算科学者との協業で学際性に富み、学際共同研究として実施する意義を有する。また CS 学者においても、高性能計算(HPC)とソフトウェア工学の専門家との協業は、米国ではエネルギー省の IDEAS-ECP プロジェクトなど大規模ファンドの支援があるが、我が国では皆無といえる。そのため学際的観点からもインパクトが大きい。本課題は、実用的な数値シミュレーションソフトウェアと大規模並列化された数値計算の性能評価を含んでいる。そのため、拠点の計算機資源の活用が必須となる課題である。

4. 前年度までに得られた研究成果の概要

本年度が初年度である。

5. 今年度の研究成果の詳細

【問題設定】本提案は 3 年計画であり、本年は１年目である。ソフトウェア工学、AT、アルゴリズム研究の各側面から行う。

① ソフトウェア工学：ソフトウェア（LAPACK）が提供する機能や能力を期待通り満たすかどうかを確かめるためにソフトウェアテストを実施する。ソフトウェアに与えられる入力値を幅広くカバーし、様々な入力に対して期待通りの出力が得られるか確認するため、様々なテストケースを実行する。バグ等の原因で期待通りの結果とならない場合、その部分を修正して再びテストケースを実行し、すべてのテストケースにおいて問題がないことを確かめる。このとき、バグが検出されやすいテストケースを選ぶことができれば、テストケースの実行時間を短縮することが期待される。そこで、本研究では HPC におけるテストケースを優先順位付けし、問題を検出しやすいテストケースがあるかどうかを調査する。その後、この本調査をもとに、汎用化した

テストケース最適化手法の提案を試みる。

② AT 研究：片桐開発の ppOpen-AT、および、IDEAS-ECP プロジェクト開発の GPtune 等を活用する。本年度は、メンバが今まで取り扱ってきたアプリケーション（NICAM（気象）、ABINIT-MP（FMO 法））に加え、量子回路シミュレータ cuQuantum、および、量子アニーリング WEB インタフェース Amplify の性能パラメタに既存 AT 技術を適用することで既存 AT 方式の効果検証を行い、かつ AT 方式の改良を検討する。

③ アルゴリズム・実装技法研究：黄開発の inexact Newton 法による非線形ソルバの並列アルゴリズムの改良を行うとともに性能評価を行う。また、アプリの NICAM、ABINIT-MP、量子回路シミュレータ cuQuantum、および AT に必要な AI の GPU 最適化を行うとともに、AT が必要となる性能パラメタを明らかにする。

【成果概要】

① ソフトウェア工学：科学技術計算において数値計算ソフトウェアのテスト実行時間の短縮は非常に重要である。ここでは、典型的な数値解析ライブラリである LAPACK を取り上げ、固有値計算におけるテストシーケンスの最適化を目的とした。

本研究のテストシーケンスとは、固有値計算の解析解がわかっている問題を与え、計算結果がその範囲内に入っているかをチェックするテスト問題の系列である。特定のソフトウェアを前提としないテストシーケンスの最適化の研究はあるが、数値計算ソフトウェアの評価はあまりない。本研究ではまず、LAPACK のライブラリに意図的にバグを発生させ、固有値ライブラリのテスト STCollection のシーケンスを入れ替える最適化によって、バグを検出するまでのテスト実行時間が短縮できるか検討した[23]。

本実験ではバグを模倣するため、BLAS ライブラリに含まれる dgemm ルーチンの *alpha* の値を 1.0 にした場合と、0.01 にして、STCollection のテストルーチンの結果を検証した。その実行結果を表 1 に示す。表 1 は、STCollection の Case17 のテストにおける演算精度を示している。ここで case17 とは、分割統治法(dgesv)で固有値問題を解く場合のテストである。

表 1　STCollection の Case17 のテストにおける演算精度

Alpha 値	実行時間	残差	直交性
1.0（正常）	6.32	5.00×10^{-3}	1.97×10^{-2}
0.01	5.09×10^{-1}	6.82×10^{11}	1.36×10^{12}

表 1 から、残差ベクトルのノルムと直交性の値が極端に大きくなっており、このテストによりバグが検出できることが確認できる。

次にテストシーケンス入替による最適化の効果を検証した。表 2 は、case17 実行時間（最適化実施例）と case17 までの実行時間（デフォルトの STCollection によるテスト実行シーケンスによりバグを検出した時間）を示している。

表 2　テストシーケンス最適化の効果

Case17 を最初に行う(a)	Case17 迄累計時間(b)	高速化率 (b/a)
73.13［秒］	385.61［秒］	5.27 倍

表 2 から、テストシーケンスの入替による最適化で、case17 を先頭としてテストを実行することにより 5.27 倍の高速化を達成できる可能性があることを明らかにした[23]。

今後、本事例をもとに、ソフトウェア工学の観点から汎用的な最適化方法の構築を試みる。

また 2023 年 8 月 30 日に、本 JHPCN 課題が主催の国際ワークショップ**数値計算とソフトウェア生産性に関するワークショップ(WNCSP2023)**を開催し、情報交換を行った[8]-[12]。

② AT 研究

【GPTune 適用】WNCSP2023 において米国 ECP プロジェクトの発表[9]、および、数値計算ライブラリ ScaLAPACK のパラメタチューニングに GPTune を適用した事例[10]を発表した。また、疑似量子アニーラのハイパーパラメタチューニングの成果として、[13][17][24]での発表がある。ここでは、GPTune を使った成果の概要を以下に説明する。

GPTune は、ベイズ推定により性能パラメタチューニングが行える AT ツールであり、その対象の詳細を知ることなく、チューニングが行えるブラックボックス型の AT ツールである。特に MPI 対応がされている唯一の AT ツールであり、スーパーコンピュータでの AT の適用に好適である。本研究では、いままで GPTune による最適化が行われていない、ScaLAPACK の LU 分解ルーチンに AT を実装したのが、成果である。

この AT では、LU 分解時のブロックサイズに加えて、プロセッサ・グリッドの最適化を行えることが大きな特徴である。性能評価結果の一部を、図 1 に示す。

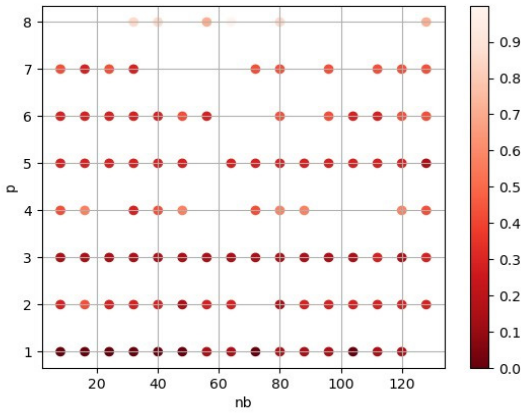


図 1 GPTune による 2 パラメタチューニング実行例（LU 分解ルーチン、行列サイズ 1000x1000）。ブロック幅 *nb*、プロセッサ・グリッド *p* x *q* の調整を行う。なおこの例では、プロセス数は 8 に限定している。

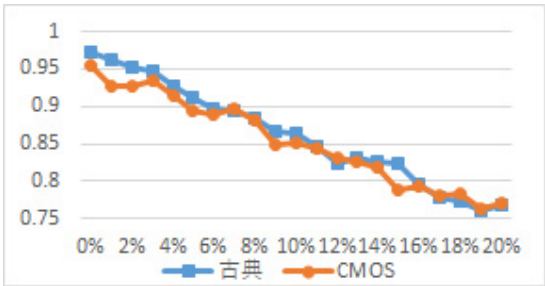
図 1 では、ベイズ推定により、データが少なく、高速となると予想される点を優先的に探索する。また、GPTune により発見された、最適となるパラメタを表 1 に示す。

表 1 最適なパラメタ値（LU decomposition.

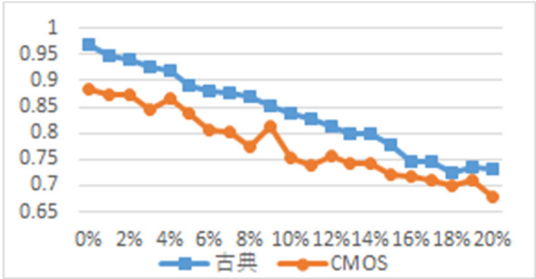
行列サイズ	(nb, p)
1000	(16, 1)
2000	(64, 1)
3000	(64, 1)

表 1 より、デフォルト値ではない、最適なパラメタ値の発見に成功している。

【量子コンピュータ関連技術と AT】 疑似量子アニーラへの AT 技術の展開の研究[21][2②]を行った。具体的には、CMOS アニーラへの、Suport Vector Machine(SVM)の疑似量子アニーラ(CMOS アニーラ）への適用評価を、世界で初めて行った。性能評価として、線形分離可能問題 1 種、線形分離不可能問題 2 種で評価を行った。結果を図 2 に示す。



(a) 線形分離可能の場合



(b) 線形分離不可能の場合

図 2 実験結果（X 軸は誤差，Y 軸は正答率[%]）

図 2 から、線形分離可能な問題は古典と CMOS ア

ニーラで精度はほぼ同等であったが、線形分離可能な問題では、古典のほうが精度が高いことを明らかにした[22]。

なお、今後の課題として、CMOS アニーラの実行時間が古典に対して 1000 倍ほど大きな時間がかかることが判明した。そのため、その原因を解析中である。

【AT フレームワーク】AT を行うにあたり、パラメタサーベイを多数行うことが多く、スーパーコンピュータシステムへの多数のジョブの投入が必要となる。この作業は煩雑であり、HPC ソフトウェアの開発生産性を低下させる。そこで、ジョブスケジューラと連携し、パラメタサーベイを簡便に行うプログラム環境である Xcrypt と連携し、AI のハイパーパラメタの最適化を簡便に行うプロトタイピング[4]に成功した。本課題は、GPTune などを組み込むことで、AT フレームワークとして活用できる。そのため、今後の展望に期待できる成果である。

具体的には、ResNet50 のハイパーパラメタであるバッチサイズのチューニングを、Xcrypt を用いて実装した。スーパーコンピュータ「不老」TypeII サブシステムの複数 GPU を活用し、並列実行を行うフレームワークのプロトタイピングを行った。

	Batch size	Loss	Execution time [s.]
1st	210	8.4305	71.3
2nd	223	8.4332	74.8
3rd	202	8.4385	68.1
4th	192	8.4415	80.5
5th	231	8.4458	89.6
default	256	8.5044	95

(a) デフォルト実行と、TOP1～5 のバッチサイズと実行時間[秒]

(b) 全体の傾向（Loss 基準）

図 3 ResNet50 のバッチサイズの自動チューニング（Xcrypt による並列実行）

図 3 の並列実行により、逐次でチューニングすると 1321[秒]かかる処理が 27.9[秒]と、43.7 倍高速化できることを明らかにした[4]。

【計算化学と量子コンピュータ】発表[19][20]でキーノート講演を含む口頭発表を行った。また、論文発表[1]では、スーパーコンピュータ「不老」TypeII サブシステムの GPU での cuQuantum による量子シミュレーションの活用により、18 量子ビットの量子回路シミュレーションで CPU での実行に対して、42.7 倍の高速化に成功している。

一方 ABINIT-MP においては、GPU 化の検討をミニアプリで実施した[2]。その結果をもとに、AT を適用する性能パラメタの抽出の検討を行った。

③アルゴリズム・実装技法研究：黄開発の inexact Newton 法による非線形ソルバの数値実験を東京大学のスーパーコンピュータを活用して進めた。本年度の実施内容は、以下の通りである。

有限格子のような複雑な構造における波の挙動を正確にシミュレートすることは、さまざまな科学および工学アプリケーションにとって非常に重要である。本研究では、有限格子などの複雑な形状における非線形ヘルムホルツ問題に対処するための有力なアプローチとして、マルチスケール有限要素法（MsFEM）を研究した。 MsFEM は、粗いメッシュを使用して回折格子の全体構造を捕捉し、同時に各要素内に特殊な基底関数を構築する方法である。これらの機能は、格子の周期的性質や局

所的な材料特性など、格子の複雑な形状を表現できる。この原理から、領域全体にわたって過度に洗練されたメッシュを必要とせずに、細かいスケールの特徴を組み込むことができるため、高い計算効率の維持が期待できる。本研究では、非線形ヘルムホルツ方程式に合わせてチューニングされた MsFEM フレームワークを提案し、有限格子の解析における有効性を評価した[14][16]。

Ⅳ）国際連携では、数値計算とソフトウェア生産性に関するワークショップ(WNCSP2023)を開催し、ソフトウェア工学と数値計算ライブラリ開発に関する米国 ECP プロジェクト関連の招待講演 2 件を行った。特に GPTune の利活用成果が多く創出されており、国際連携は良好と判断できる。

6. 進捗状況の自己評価と今後の展望

本課題は①～③と多岐にわたるが、多くの論文と口頭発表の成果を創出した。そのため、総合的に本年度の達成率は 100%と自己評価する。

以下に、今後の展望を記す。

① ソフトウェア工学：本年度のケーススタディを基に、LAPACK 関連のルーチンの効率的なデバックを行うための、テストケース選定の汎用的な方法と、そのソフトウェア工学上の方法論の検討を、ソフトウェア工学の共同研究者と行う。本年度、数値計算テストの事例で、テストシーケンスを入れかえることで、処理の高速化に効果がある事例を、初めて明らかにすることができた。この成果は、ソフトウェア工学の観点では自明ではなく、今後の進展が期待できる。そのため、この事例の結果を拡張することで、汎用的な手法の提案が可能となると期待される。

② AT 研究：いままで記述の通り、AT と量子関連技術（疑似量子アニーラ、量子回路シミュレーション）で、口頭発表などの多くの成果が創出された。なお、量子関連技術におけるハイパーパラメタチューニングの AT 適用、および AT の有効性を指摘したのは、本研究が初めてといえる。ハイパーパラメタチューニングは AT 分野の得意とする分野であり、AT による量子関連技術への展開は、まさに今、始まったばかりである。そのため将来性があり、成果創出が期待できる。

③ アルゴリズム・実装技法研究：本研究での inexact Newton 法による非線形ソルバの結果をまとめており、今後の進捗が期待できる。

7. 研究業績

(1) 学術論文 （査読あり）

[1] K. Sugisaki, V. S. Prasannaa, S. Ohshima, T. Katagiri, Y. Mochizuki, B. K. Sahoo, and B. P. Das, "Bayesian phase difference estimation algorithm for direct calculation of fine structure splitting: accelerated simulation of relativistic and quantum many-body effects", Electr. Struct., 5 (2023) 035006-1-10.

[2] 望月祐志, 中野達也, 坂倉耕太, 奥脇弘次, 土居英男, 加藤季広, 滝沢寛之, 成瀬彰, 大島聡史, 星野哲也, 片桐孝洋,"FMO プログラム ABINIT-MP の整備状況 2023", J. Comp. Chem. Jpn., 23 (2024) 4-8.

(2) 国際会議プロシーディングス （査読あり）

(3) 国際会議発表（査読なし）

[3] M. Morishita, O. Marques (+), Y. Liu, T. Katagiri, "Experimenting with GPTune for Optimizing Linear Algebra Computations", HPC Asia 2024, (2024/1/26). (ポスター発表, アブストラクト査読)

[4] T. Hanyu, M. Kawai, T. Katagiri, T. Hiraishi, T. Hoshino, T. Nagai, "Auto-tuning of Hyperparameters by Parallel Search Using Xcrypt", HPC Asia 2024, (2024/1/26). (ポスター発表, アブストラクト査読)

[5] R. Fukuhara, M. Morishita, T. Katagiri, M. Kawai, T. Hoshino, T. Nagai,

"Performance Evaluation of Support Vector Machines with Quantum-inspired Annealers", HPC Asia 2024, (2024/1/26). (ポスター発表，アブストラクト査読)

[6] T. Katagiri, Adaptation of XAI to Numerical Libraries: A Case Study for Automatic Performance Tuning, ICIAM2023 (10th International Congress on Industrial and Applied Mathematics), Waseda University, Tokyo, Japan, (2023/8/23).

[7] M. Morishita, T. Katagiri, S. Ohshima, T. Hoshino, T. Nagai, Performance evaluation of quantum-inspired machine and quantum simulator, ICIAM2023 (10th International Congress on Industrial and Applied Mathematics), Waseda University, Tokyo, Japan, (2023/8/23).

[8] S. Morisaki, Test case prioritization studies in software engineering, 数値計算とソフトウェア生産性に関するワークショップ(WNCSP2023), (2023/8/30).

[9] O. Marques(+), ECP Overview, 数値計算とソフトウェア生産性に関するワークショップ(WNCSP2023), (2023/8/30).

[10] M. Morishita, O. Marques (+), T. Katagiri, Auto-tuning of Numerical Library by GPTune, 数値計算とソフトウェア生産性に関するワークショップ(WNCSP2023), (2023/8/30).

[11] T. Katagiri, Autotuning for Sparse Iterative Solvers and Quantum Computing Related Technology, 数値計算とソフトウェア生産性に関するワークショップ(WNCSP2023), (2023/8/30).

[12] T. Hoshino, Acceleration of Hierarchical Matrix Library HACApK, 数値計算とソフトウェア生産性に関するワークショップ(WNCSP2023), (2023/8/30).

[13] M. Morishita, O. Marques (+), Y. Liu, T. Katagiri, Experimenting with GPTune for Optimizing Linear Algebra Computations, Computing Sciences Summer Program 2023, Lawrence Berkeley National Laboratory, (2023/8/8). A Poster Presentation.

[14] F.-N. Hwang (+), "Parallel multiscale finite element method with CFD applications," Advances in Computational Mechanics (ACM 2023), (2023/10/22).

[15] T. Katagiri, "Auto-Tuning for Quantum Computing Related Technology on Supercomputers", MS53 Aspects of Software Engineering and Extreme Scale Computing, SIAM Conference on Parallel Processing for Scientific Computing (PP24), (2024/3/7).

[16] I. Luthfi (+) and I. and F.-N. Hwang (+), "Unveiling Nonlinear Wave Propagation in Finite Gratings: A Multiscale Finite Element Approach", 2024 Conference on Advanced Topics and Auto Tuning in High-Performance Scientific Computing (ATAT24), (2024/3/22).

[17] T. Katagiri, M. Morishita, "Adaptation of Auto-tuning for Quantum Annealer", 2024 Conference on Advanced Topics and Auto Tuning in High-Performance Scientific Computing (ATAT24), (2024/3/23).

[18] M. Morishita, O. Marques (+), Y. Liu, T. Katagiri, "Auto-tuning of ScaLAPACK by GPTune", 2024 Conference on Advanced Topics and Auto Tuning in High-Performance Scientific Computing (ATAT24), (2024/3/23).

(4) 国内会議発表（査読なし）

[19] 望月祐志，坂倉耕太，中野達也，土居英男,奥脇弘次，秋澤和輝，北原駿，太刀野雄

介, 松岡壮太, 小沢拓, 大島聡史, 片桐孝洋:「富岳」を利用した大規模なフラグメント分子軌道計算について,（キーノート依頼講演）, 第 28 回計算工学講演会, つくば (2023/5/31).

[20] 杉崎研司, Prasannaa V. S., 大島聡史, 片桐孝洋, 森野慎也, 望月祐志, Sahoo B. K., Das B. P.,「不老」Type II 上で cuQuantum 量子シミュレータを用いた相対論的量子化学計算の事例, 第 28 回計算工学講演会, つくば (2023/6/1).

[21] 福原諒河, 森下誠, 片桐孝洋, 河合直聡, 星野哲也, 永井亨, CMOS アニーリングにおけるサポートベクターマシンの性能評価, 2023 年並列／分散／協調処理に関する『函館』サマー・ワークショップ (SWoPP2023), 2023-HPC-190, pp.1-6 (2023/8/2).

[22] 水木直也, 福原諒河, 森下誠, 河合直聡, 片桐孝洋, 星野哲也, 永井亨, SVM による誤差を含むクラス分類における CMOS アニーリングマシンの性能評価, 情報処理学会第 86 回全国大会, pp.1-2 (2024/3/15).

[23] 樫村寛大, 森崎修司, 片桐孝洋, 河合直聡, 永井亨, 星野哲也, LAPACK を用いた固有値計算におけるテストシーケンスの最適化, 情報処理学会第 86 回全国大会, pp.1-2 (2024/3/16).

[24] M. Morishita, T. Katagiri, O. Marques (+), Y. Leu, T. Hoshino, T. Nagai, M. Kawai, Performance Evaluation of GPTune for Parameter Auto-Tuning, 第 193 回ハイパフォーマンスコンピューティング研究発表会, 情報処理学会研究報告, 2024-HPC-194, pp.1-6 (2024).

(5) 公開したライブラリなど

(6) その他（特許, プレスリリース, 著書等）