様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　６ 年　 ５ 月 ３１ 日現在

機関番号：１２６０１
研究種目：挑戦的研究（萌芽）
研究期間：2020～2023
課題番号：２０Ｋ２１７８７

研究課題名（和文）シミュレーションと機械学習の協調による予測に基づいた動的負荷分散手法の開発

研究課題名（英文）Development of a dynamic load balancing method based on prediction by cooperative use of simulation and machine learning

研究代表者

下川辺　隆史（Shimokawabe, Takashi）

東京大学・情報基盤センター・准教授

研究者番号：４０６３６０４９

交付決定額（研究期間全体）：（直接経費）　　5,000,000円

研究成果の概要（和文）：近年、GPU計算では、GPU計算と相性の良い、高精度が必要な領域を局所的に高精細にできる適合細分化格子法が注目されている。本研究では、機械学習によりシミュレーションの「未来」の結果を予測し、それに基づいた動的負荷分散する手法を開発することを目標とし、深層学習による流体シミュレーションの予測、計算量や通信量に基づいた最適な領域分割方法の構築、AMR法フレームワークの高度化とそれを用いた格子ボルツマン法の構築を実現した。機械学習によりシミュレーションの予測が有用であるという知見が得られた。

研究成果の学術的意義や社会的意義
格子計算はスパコンを利用する代表的なアプリケーションで、局所的に高精細な大規模計算を実現させる意義は大きい。米国エネルギー省は、AMR法は所謂「エクサスケール」でのマルチスケール問題解決の鍵となる技術と位置付けている。本研究では、機械学習という全く異なるアプローチで数値計算結果を予測する。本研究の目標は予測に基づいた動的負荷分散の実現であるが、近似的ではあるが超高速な予測が可能である機械学習は計算科学分野の様々な要素技術で従来手法を凌駕する可能性を秘めており、本研究でその有用性を示した意義は大きい。

研究成果の概要（英文）：Recently, adaptive mesh refinement (AMR), which is well suited for GPU computation, has been attracting attention because it can locally refine regions where high accuracy is required. This research aims to develop a method to predict "future" results of simulations by machine learning and to dynamically balance the load of the simulations based on these predictions. We have realized the prediction of fluid simulations by deep learning, the construction of optimal domain decomposition methods based on the amount of computation and communication, the improvement of the AMR method framework, and applied it to the lattice Boltzmann method. We have found machine learning to be helpful in predicting simulations.

研究分野：格子法に基づいた大規模物理計算

キーワード：ステンシル計算　高性能計算　機械学習　適合細分化格子法　高生産フレームワーク　動的負荷分散

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

流体計算などの格子に基づく計算は、GPU スパコンの代表的なアプリケーションである。GPU は元々は画像処理用のプロセッサであったが、消費電力当たりの演算性能が高いため、日米欧のトップレベルのスパコンに搭載されている。効率的に計算資源を使うため、近年、GPU 計算では、GPU 計算と相性の良い、高精度が必要な領域を局所的に高精細にできる適合細分化格子法(Adaptive MeshRefinement; AMR 法)が注目されている。複数 GPU による AMR 法では、局所的に解像度が刻一刻と変化するため、GPU 間で計算負荷を動的に均等にすることが必須である。格子の解像度ごとに計算量、通信量が異なる AMR 法では、領域の再分割の評価に多くの計算が必要である。さらに、高精細領域が頻繁に変形・移動する大規模計算では、常に GPU 間の負荷を均等に保つために、頻繁に多量の GPU 間のデータ移動が生じて、大きなボトルネットとなり課題となっている。

２．研究の目的

本研究では、大規模 GPU スパコンで高性能な AMR 計算を実現するため、機械学習によりシミュレーションの「未来」の結果を予測し、それに基づいた動的負荷分散する手法を開発する。従来手法は、当該時刻の結果のみ利用して負荷分散を行なうが、本研究は、機械学習による「未来」の予測も利用した負荷分散を行う独創的で革新的な試みである。機械学習の予測は近似的ではあるが、シミュレーション実行とは別に、超高速に複数時刻の予測を行えるため、負荷分散のための領域の再分割の評価に多くの計算を行うことが可能となり、時系列変化を考慮した精度の高い評価を行える。これにより AMR 法の負荷分散を格段に高度化することを実現する。

３．研究の方法

本研究では、機械学習によりシミュレーションの「未来」の結果を予測し、それに基づいた動的負荷分散する手法を開発するため、(1)機械学習による流体計算の予測と高精細が必要な局所領域の特定と移動の推定手法の開発を行う。流体シミュレーション結果の予測には深層学習を用いる。次に、(2)予測に基づいた動的負荷分散手法の開発を行う。また、計算量や通信量を評価し、最適な領域分割を決定する。これらを開発した後、(3) 構築中の AMR 法フレームワークを完成させ、これに開発手法を導入し、実アプリケーションへの適用を行する。これを通して、機械学習による予測に基づいた動的負荷分散手法が有効であることを明らかにする。

４．研究成果

本研究では、深層学習による流体シミュレーションの予測、計算量や通信量に基づいた最適な領域分割方法の構築、AMR 法フレームワークの高度化とそれを用いた格子ボルツマン法の構築を実現した。流体計算などの格子に基づく計算では、高精度が必要な領域をより高精細な格子で計算できる AMR 法がマルチスケール問題解決の鍵となる技術として注目されている。本研究では、GPU スパコンで従来と比較して高性能な AMR 計算を実現するため、機械学習によりシミュレーションの予測が有用であるという知見が得られた。

研究を進めるにつれ、当初想定していたより、大規模な領域に対する精度が高い予測手法の構築が必要であると考え、そこに注力し研究を進めた。開発した手法で負荷分散を効率化することを直接的に示すことができなかったものの、機械学習を用いることで高速なシミュレーションの予測自体は可能であることを示し、それが動的負荷分散手法に適用できるという有益な知見が得られた。また、本研究では、従来の差分計算だけでなく、近年、多く用いられるようになった格子ボルツマン法の計算にも対応したフレームワークへと高度化を行った。これによりより多くの流体アプリケーションへの適用が可能となった。流体以外の様々なアプリケーションへの適用は今後の課題である。

以下では主な研究成果について説明する。

(1) 深層学習による流体シミュレーションの予測手法の開発

機械学習のうち深層学習を用いることで、規模の大きい流体シミュレーションの 「未来」の結果を予測する手法を開発した。本手法では、格子ボルツマン法 (Lattice Boltzmann Method; LBM) による 100 ステップごとのシミュレーション結果の 3 フレームを用いて、続く 3 フレーム（100 ステップ間隔）を予測する。

本手法の学習で用いたデータセットについて説明する。まず 1024×1024 の領域に 1 つまたは 2 つの柱体を配置し、LBM により流体が流れる計算を行う。柱体の種類として円柱と三角柱を用い、計算領域へランダムな大きさのものをランダムな位置へ配置する。複数タイムステップの 64×64 の領域の計算結果をネットワークモデルへの入力として、それに続く複数タイムステップの 64×64 の領域の計算結果を予測できるようにする。データセットを作成するため、24 セットの LBM シミュレーションを実行し、そこから、あるタイムステップの前後で入力とする 100 ステップ間隔の 3 フレームと、それに続く 100 ステップ間隔の 3 フレームを保存する。後者の 3 フレームをニューラルネットワークモデルは予測する。計算領域 1024×1024 に対して、物体の配

置は考慮せずに、一部領域が重複することも許し、機械的に 200 個の 64×64 の領域を切り出す。データ拡張およびデータセットが流体の流れる方向に依存しないように、LBM シミュレーションの結果を回転や反転させたデータもデータセットに加えることで、最終的には学習に 65､021､評価に 27､867 用いた。

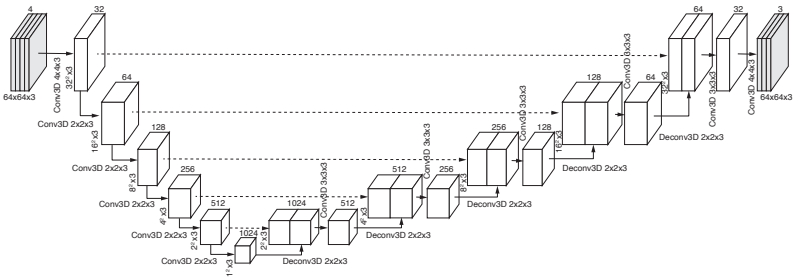

図 1　流体シミュレーションを予測するニューラルネットワーク構造

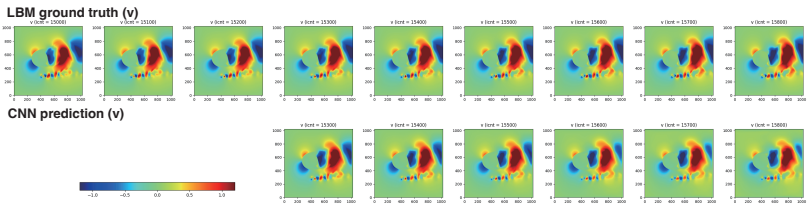

図 2　流体シミュレーションの予測結果

図 1 に本手法で用いたニューラルネットワークの構造を示す。本手法ではニューラルネットワークの入力として、64×64 の領域全域の流体の密度､速度場と､物体形状を表す符号付き距離関数の 3 フレーム分を与える。出力として、流体の密度、速度場の 3 フレーム分の予測を得る。この出力は入力データに続くタイムステップとなるようにする。この手法で用いたネットワークは Encoder-decoder モデルを基盤としており大きく前半と後半の二つの構造で構成されており、前半は複数の畳み込み層からなり、後半は複数の逆畳み込み層からなる。流体の時間発展を予測するため、空間 2 次元に加え、時間軸を 3 次元目とした 3 次元の畳み込み層と逆畳み込み層を用いたネットワーク構造とする。予測精度を上げるために、畳み込み層とそれに対応する逆畳み込み層の間に U-net などで用いられるスキップ接続を導入している。

本手法の学習には､東京大学情報基盤センターに設置された Wisteria/BDEC-01 スーパーコンピュータシステムのうち、データ・学習ノード群（Aquarius（アクエリアス））を用いる。 1 ノードあたり 8 台の NVIDIA A100 Tensor コア GPU を搭載している。 損失関数としては、それぞれの格子点の密度と速度場、さらに予測精度を上げるため、それらの空間勾配の真の値と予測値の平均二乗誤差を用いる。学習では 64×64 の領域を対象としているが、大規模なシミュレーション結果の予測を行うためには、任意のサイズの計算領域に対してその計算結果を予測できる必要がある。そこで、予測では、計算領域に対してパッチ的にニューラルネットワークによる推論を適用することで、計算領域全域の予測を可能とする。 ニューラルネットワークで得られた予測結果を入力とし、次の 3 フレームの予測を行う。これを繰り返すことで、予測を時間的に進め､「未来」の予測を行うことを可能とした。

図 2 は LBM によるある計算セットの 15000、15100、15200 タイムステップの 3 フレームを入力として、ニューラルネットワークによる予測を空間的・時間的に繰り返し適用することで、これに続く 3 フレームを予測し､それをさらに入力として、その先の 3 フレームを予測した結果である。すなわち、15300 から 15800 タイムステップの間で 100 ステップごとの予測結果が得られる。図で、上から y 方向の速度の LBM 計算による結果とそのニューラルネットワークによる予測値である。100 ステップごとに 1 つのフレームのシミュレーション結果しか用いていないが、このニューラルネットワークによって LBM によるシミュレーション結果をよく予測できていることがわかる。

(2) 計算量や通信量に基づいた最適な領域分割

複数 GPU での埋め込み境界法を伴った LBM の計算を行う前に最適な領域分割数や GPU の割り当て方を予測するために計算の性能モデルを構築した。ここでは計算機システムのノード内通信性能とノード間通信性能, また 1GPU での通信がない LBM のカーネル実行時間を用いて性能モデルを構築できるとする。ただし, 通信性能に関しては LBM のシミュレーション時の実際の通 信性能を測定するものではなく, 計算機の通信バンド幅の測定を用いる。

図 3 に複数 GPU での LBM の計算のモデルを示す。t_2 から $t_{c,9}$ までを目的変数として､図に示した説明変数から線形回帰できるものとした。

Wisteria/BDEC-01 のデータ・学習ノード群（Aquarius（アクエリアス））の 32GPU を用い、竹とんぼの飛翔シミュレーションを分割数や GPU

図 3　複数 GPU での LBM 計算のモデル化

配置を変えながら行い、性能を測定し、これと性能モデルによる予測値を比較した。領域を yz 方向に 1×32, 2×16, 4×8 分割をし，それぞれに対して以下の通信量および GPU 配置で測定を行った。YZallocation は Y 方向から、ZYallocation は Z 方向から GPU を配置することを表している。通信量の削減では、格子ボルツマン法の速度分布関数の方向などを考慮し、削減を行う。通信量の隠蔽とは、GPU 計算で GPU 間の通信を隠蔽することを表す。

(A) 通信量の削減も隠蔽もなく，GPU 配置は YZallocation
(B) 通信量の削減も隠蔽もなく，GPU 配置は ZYallocation
(C) 通信量の削減を行い，GPU 配置は YZallocation
(D) 通信量の削減を行い，GPU 配置は ZYallocation
(E) 通信量の削減と隠蔽を z 方向に対して行い、GPU 配置は YZallocation
(F) 通信量の削減と隠蔽を y 方向に対して行い、GPU 配置は ZYallocation

表 1 に(A)から(F)での性能の測定値、表 2 に図 3 に示した性能モデルによる予測値を表す。予測値には誤差は生じているものの、それぞれの実装によって得られる性能の傾向は良く表現しており、この方法が有効であることがわかる。ここでは直交格子上の計算に対して性能モデルを構築しているが、AMR 法を適用した場合も本質的には同じであり、この方法論が有用であるという知見が得られた。

表 1 (A)から(F)における性能の測定値（単位は $\times 10^3$ MLUPS）

分割数	(A)	(B)	(C)	(D)	(E)	(F)
1×32	13.6	13.6	22.7	22.7	27.3	22.8
2×16	14.8	15.6	23.2	23.0	26.2	23.7
4×8	16.3	14.6	24.4	22.1	26.8	25.1

表 2 (A)から(F)における性能モデルによる予測値（単位は $\times 10^3$ MLUPS）

分割数	(A)	(B)	(C)	(D)	(E)	(F)
1×32	16.4	16.4	23.8	23.8	29.3	23.8
2×16	18.9	17.6	25.6	24.1	28.5	24.9
4×8	18.5	18.5	24.5	24.6	26.9	26.3

(3) AMR 法フレームワークの高度化とそれを用いた格子ボルツマン法の構築

本研究では、開発中の AMR 法フレームワークを実アプリケーションへ適用することを目的とする。従来の差分計算だけでなく、近年、多く用いられるようになった LBM 計算にも対応したフレームワークへと高度化を行った。

格子ボルツマン法は、時間更新幅が格子幅の定数倍に固定される計算手法であるために、解像度ごとに時間ステップ幅が異なり、時間方向にも物理量の補間が必要となる。本フレームワークでは、これまでは、ある時刻では全ての解像度でデータを保持していることを前提としていた。図 4 に AMR を適用した格子ボルツマン法で必要となる解像度間と時間ステップ間の補間関係を示す。図は 3 解像度の場合の例である。

本研究では、まず特定の条件を満たす AMR レベル（解像度）の格子ブロックに対してステンシル計算関数や境界領域の交換を適用できるよう、機構を導入し、ステンシル計算関数および境界領域の交換において対応した。特定のレベルのみを実行するのではなく、レベルの範囲を指定して実行できるようにしたことで、GPU カーネルの実行回数を最小回数として、実行性能の向上に寄与している。次に、時間方向にも物理量の補間が必要となるため、これに対応した格子ブロックの境界領域の交換手法を導入した。また、時間方向の物理量の補間の際には解像度の変更、すなわち空間方向の補間も必要となるため（図 4 参照）、格子ボルツマン計算へも対応できるよう任意の補間関数を指定できるように拡張した。これによって、AMR 法フレームワークを用いた格子ボルツマン法の構築を実現した。

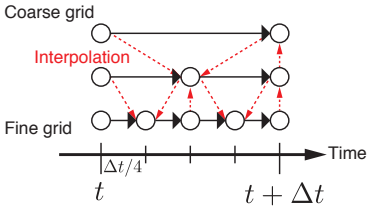

図 4 格子ボルツマン法における異なる解像度間での補間関係

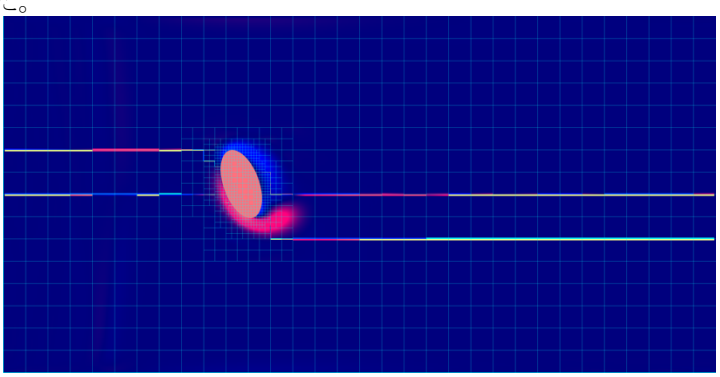

図 5 AMR フレームワークを適用した格子ボルツマン計算 図では 4 つの領域に負荷分散されている

５．主な発表論文等

〔雑誌論文〕　計4件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 畠山 昂，下川辺 隆史	４．巻 2023-HPC-188
２．論文標題 複数GPUでの埋め込み境界-格子ボルツマン法の計算の最適化と性能モデルの構築	５．発行年 2023年
３．雑誌名 研究報告ハイパフォーマンスコンピューティング(HPC)	６．最初と最後の頁 1-9
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Asahi Yuuichi、Hatayama Sora、Shimokawabe Takashi、Onodera Naoyuki、Hasegawa Yuta、Idomura Yasuhiro	４．巻 －
２．論文標題 AMR-Net: Convolutional Neural Networks for Multi-resolution Steady Flow Prediction	５．発行年 2021年
３．雑誌名 The 2nd Workshop on Artificial Intelligence and Machine Learning for Scientific Applications, IEEE Cluster 2021	６．最初と最後の頁 686 - 691
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/Cluster48925.2021.00102	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 畑山そら，下川辺隆史，小野寺直幸	４．巻 2020-HPC-175
２．論文標題 深層学習と境界交換を用いた複数領域にまたがる定常流のシミュレーション結果の予測	５．発行年 2020年
３．雑誌名 研究報告ハイパフォーマンスコンピューティング（HPC）	６．最初と最後の頁 １ - ７
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 長谷川敦，下川辺隆史	４．巻 2020-HPC-177
２．論文標題 深層学習による混相流の時間発展シミュレーション結果の予測手法の検討	５．発行年 2020年
３．雑誌名 研究報告ハイパフォーマンスコンピューティング（HPC）	６．最初と最後の頁 １ - ７
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計19件（うち招待講演　3件／うち国際学会　8件）

１．発表者名 佐久間 大我、下川辺 隆史、大森 拓郎
２．発表標題 oneAPIを用いた様々なデバイス上でのステンシル計算の実装
３．学会等名 第28回計算工学講演会
４．発表年 2023年

１．発表者名 Ziheng Yuan, Takashi Shimokawabe
２．発表標題 Accelerating lattice Boltzmann method simulation with GPU computation using C++ standard language parallelism
３．学会等名 第28回計算工学講演会
４．発表年 2023年

１．発表者名 Ziheng Yuan, Takashi Shimokawabe
２．発表標題 Accelerating lattice Boltzmann method with GPU and C++ standard parallelization
３．学会等名 10th International Congress on Industrial and Applied Mathematics（国際学会）
４．発表年 2023年

１．発表者名 下川辺隆史
２．発表標題 深層学習を用いたシミュレーション結果を予測する代理モデル開発の取り組み
３．学会等名 第7回HPCものづくり統合ワークショップ（招待講演）
４．発表年 2023年

１．発表者名
Ziheng Yuan and Takashi Shimokawabe

２．発表標題
Accelerating Lattice Boltzmann method with C++ standard language parallel algorithm

３．学会等名
International Conference on High Performance Computing in Asia-Pacific Region (HPCAsia) 2024（国際学会）

４．発表年
2024年

１．発表者名
畠山 昂，下川辺 隆史

２．発表標題
複数GPUを用いる際の埋め込み境界-格子ボルツマン法の性能向上

３．学会等名
第27回計算工学講演会

４．発表年
2022年

１．発表者名
大森 拓郎，下川辺 隆史，朝比 祐一

２．発表標題
OpenMP Offloadingを用いたGPUでの格子ボルツマン法実行における性能評価

３．学会等名
第27回計算工学講演会

４．発表年
2022年

１．発表者名
Takuro Omori, Takashi Shimokawabe

２．発表標題
Performance Optimization Of Lattice Boltzmann Method On A64FX

３．学会等名
15th World Congress on Computational Mechanics & 8th Asian Pacific Congress on Computational Mechanics（国際学会）

４．発表年
2022年

１．発表者名
Akira Hatakeyama, Takashi Shimokawabe

２．発表標題
Performance improvement of immersed boundary-lattice Boltzmann method on multiple GPUs

３．学会等名
15th World Congress on Computational Mechanics & 8th Asian Pacific Congress on Computational Mechanics（国際学会）

４．発表年
2022年

１．発表者名
下川辺 隆史

２．発表標題
深層学習による流体シミュレーション結果の予測

３．学会等名
第35回計算力学講演会（招待講演）

４．発表年
2022年

１．発表者名
鈴木翔太，下川辺隆史

２．発表標題
格子ボルツマン法に基づくGPUを用いた音響解析

３．学会等名
第26回計算工学講演会

４．発表年
2021年

１．発表者名
鈴木 翔太，下川辺 隆史

２．発表標題
埋め込み境界法を適用した格子ボルツマン法に基づく3次元音響解析

３．学会等名
オープンCAEシンポジウム2021

４．発表年
2021年

１．発表者名
Shota Suzuki, Takashi Shimokawabe

２．発表標題
Acoustic simulation using lattice Boltzmann method by multi-GPU parallel computing

３．学会等名
International Conference on High Performance Computing in Asia-Pacific Region (HPCAsia) 2022（国際学会）

４．発表年
2022年

１．発表者名
鈴木 翔太，下川辺 隆史

２．発表標題
格子ボルツマン法によるインピーダンス境界を用いた音響解析手法の構築

３．学会等名
日本音響学会 2022年春季研究発表会

４．発表年
2022年

１．発表者名
Akira Hatakeyama, Takashi Shimokawabe

２．発表標題
Multi-GPU computing of moving boundary flow using lattice Boltzmann method

３．学会等名
International Conference on High Performance Computing in Asia-Pacific Region (HPCAsia) 2022（国際学会）

４．発表年
2022年

１．発表者名
下川辺隆史

２．発表標題
深層学習による流体シミュレーション結果予測

３．学会等名
第41回計算数理工学フォーラム（招待講演）

４．発表年
2022年

１．発表者名
畑山そら，下川辺隆史，小野寺直幸

２．発表標題
畳み込みニューラルネットワークと境界交換を用いた複数領域にまたがる定常流のシミュレーション結果の予測

３．学会等名
第25回計算工学講演会

４．発表年
2020年

１．発表者名
Sora Hatayama, Takashi Shimokawabe and Naoyuki Onodera

２．発表標題
Steady Flow Prediction across Multiple Regions using Deep Learning and Boundary Exchange

３．学会等名
3rd International Conference on Computational Engineering and Science for Safety and Environmental Problems（国際学会）

４．発表年
2020年

１．発表者名
Takashi Shimokawabe and Naoyuki Onodera

２．発表標題
High-Resolution Simulations using an AMR Framework on GPU Supercomputers

３．学会等名
3rd International Conference on Computational Engineering and Science for Safety and Environmental Problems（国際学会）

４．発表年
2020年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関