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　３ 年　 ５ 月　 １ 日現在

機関番号：１２６０１
研究種目：若手研究
研究期間：2018～2020
課題番号：１８Ｋ１８０２１

研究課題名（和文）高バンド幅と大容量を両立するアクセスパターン適応型ハイブリッドメインメモリ

研究課題名（英文）Exploiting High-Bandwidth and Large-Capacity on Hybrid Main Memories through Pattern-Aware Optimization

研究代表者
有間　英志（Arima, Eishi）

東京大学・情報基盤センター・特任助教

研究者番号：５０７８０６９９

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：本研究では、スーパーコンピュータ等のHPCシステムを対象とし、特に、近年普及が進んでいる、ハイブリッド型のメモリシステムに焦点を当て、それに適した全く新しいデータ転送技術や電力制御手法の研究開発を行った。前者のデータ転送については、Pattern-aware Stagingと呼ぶ概念を新たに導入し、これを実現するためのソフトウェア技術を提案した。後者の電力制御についても、Footprint-aware Power Cappingと呼ぶ概念を導入した上で、ソフトウェアフレームワークを提案した。これらの研究成果は共にHPC分野で著名な国際会議であるISC HPC 2020にて採択された。

研究成果の学術的意義や社会的意義
現代社会ではその活動の大部分を様々な計算機システムに委ねており、その高性能化や高効率化は極めて重要である。特に科学技術計算等に用いられるスーパーコンピュータ等のHPCシステムでは、高性能化及び高電力効率化への要望は格段に大きい。本研究では、特にハイブリッドメモリシステムと呼ばれる、HPCシステムにて近年普及し始めてきたハードウェア構成を対象とし、それに適した全く新しいデータ転送・電力制御手法を開発することで、その要望に応えている。新メモリデバイス技術の将来的な価格低下等により、当ハードウェア構成はHPC分野に留まらず、幅広く利用される事が予想され、その波及効果も大いに期待できる。

研究成果の概要（英文）：In this project, we targeted HPC (High-Performance Computing) systems, including supercomputers, in particular, those composed of recently emerging hybrid memory systems that consist of multiple different memory devices. Especially, we focused on data transfer and power management techniques specifically tailored for the hardware architecture. For the data transfer optimization, we proposed a novel concept called ``Pattern-aware Staging'' and developed a software technique based on it. As for the power management, we proposed a software framework in accordance with our newly introduced concept called ``Footprint-aware Power Capping''. Both of these works were published in ISC HPC 2020, which is a very well-known venue in HPC area.

研究分野：計算機システム

キーワード：ハイブリッドメモリシステム　データマネジメント　電力マネジメント　計算機システム　高性能計算

様　式　C－１９、F－１９－１、Z－１９（共通）

１．研究開始当初の背景

スーパーコンピュータから組み込みシステムに至るまで、あらゆる計算機システムの高性能化は、人間社会に新しい知見・サービスをもたらす上で極めて重要である。しかし、半導体微細化に基づく従来の高性能化の限界が指摘されており、今後は三次元積層、不揮発性メモリを含む様々な新デバイス技術に根ざした高性能化が重要となる。本研究では特に図 1 に示す様な、特性の異なる複数のメモリデバイス技術によって構成されたハイブリッドメモリシステムを搭載する HPC システム(HPC: High-Performance Computing、高性能計算)に焦点を当て、その上での(1)「データ転送最適化技術」及び(2)「電力割り当て最適化技術」に関する研究を行い、その高性能化や高効率化を目指した。

２．研究の目的

前述の通り、本研究の目的は、ハイブリッドメモリシステムを搭載する HPC システムの高性能化・高効率化であり、特にデータ転送や電力制御の最適化によってこれを行う。特に、当該メモリシステムにおいて、これらの最適化を行う上で、アプリケーションのメモリアクセスパターンや問題サイズに着目すべきであることは予備評価によって確認済みである。これに基づき、データ転送技術については(1)Pattern-aware Staging、電力割り当て最適化については(2)Footprint-aware Power Capping と呼ぶ新しい概念を導入し、これら概念に従った、全く新しいソフトウェアによる制御技術を開発する。

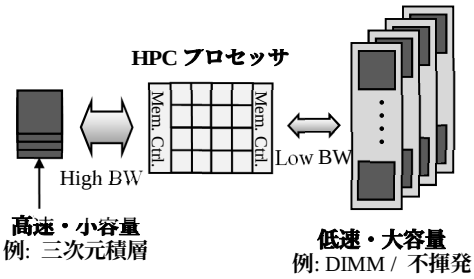

図 1: ハイブリッドメモリシステムの例

３．研究の方法

(1) Pattern-aware Staging

データ転送最適化技術については、以下手順の通りに研究開発を行った。

① まずは、シンセティックなコードを用いて、アクセスパターンがメモリ性能に与える影響の調査を行った。具体的には、高速(高バンド幅)小容量メモリ及び低速(低バンド幅)大容量メモリの両者を対象とし、GB サイズのデータチャンクに対して、ストリーミング及び不規則なメモリアクセスを同回数行い、その際の実行時間や実効バンド幅を計測した。結果として、(A)「いずれのアクセスについても、アプリケーションに十分な並列性があれば、高速小容量メモリの利用によって大幅な性能向上が見込める(バンド幅性能の違いによる)」や(B)「いずれのメモリについても不規則アクセスによる性能低下は顕著であり、それによって、メモリ間のデータ転送によるオーバーヘッドは相対的に小さくなる」といった事実が観測によって明らかになった。この評価結果の一部は図 2 に示す通りであり、これが図に示す様なステージングアクセスのモチベーションになっている。

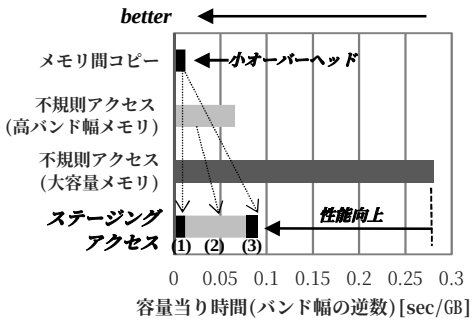

図2: 予備評価とステージング

② 上記の観測結果に基づき、Pattern-aware Staging と呼ぶデータ制御の概念を提唱した。ここでは、アプリケーション側でデータを GB オーダーの複数チャンクに分けて順次アクセスする様な状況を想定している。その際にデータは全て低速大容量のメモリに存在する様な状況を想定する。本研究では、その際に必要に応じてチャンクを高速小容量メモリ内のバッファに移動させてアクセスする。提案手法では、各々のチャンクに対してデータアクセスのパターンを実行時に解析し、その解析結果に基づき、このバッファへの転送を行うかどうかを決定する。即ち、性能向上が得られると予想される場合のみ高速小容量メモリへのデータ転送を行う。このデータ転送制御手法を Pattern-aware Staging と呼んでおり、その Block Diagram を図 3 に示す。

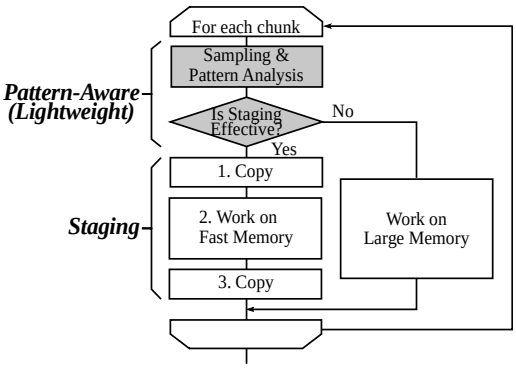

図3: 提案手法のBlock Diagram

③ 上記、実行時のアクセスパターン解析を行うためのソフトウェアによるアドレスサンプリング手法を考案した。具体的には、ソフトウェアによるヘルパースレッドプリフェッチと

呼ぶ従来から存在する手法を元に、アクセスパターン解析用に拡張したものである。ヘルパースレッドプリフェッチでは、元のコードに対してメモリアクセス及びその実行パス上にある命令だけを抜き出して別スレッドで実行する。ただし、メモリアクセスはプリフェッチ命令に変換される。本研究では、プリフェッチは行わず、そのプリフェッチアドレスだけを用い、これをパターン識別器に入力して解析を行う。

④ ブルームフィルタを用いたパターン識別手法を提案した。このパターン識別については、履歴ベースの物を利用している。具体的には図 4 に示す通り、サンプリングしたアドレスのストライドやページアドレス(サンプリングアドレスの上位ビット)を入力として用いており、これを別々の履歴に登録し、そのヒット率を調べている。ストライド履歴については、ヒット率が低ければ、不規則なアクセスをしていると言える(例えば、ストライドアクセスでは 1 エントリのみが毎回アクセスされる)。一方、ページアドレス履歴については、ヒット率が低ければ、疎なアクセスをしていると言える。この規則性と疎密性をアクセスの特徴量として利用し、この履歴のヒット率を使ってアクセスパターンの特徴を分析する。その際、履歴のアクセス速度やメモリ消費は、提案手法のオーバーヘッドに直結する。そこで、このオーバーヘッドを劇的に抑えるための手法として、ブルームフィルタを用いた履歴の実装を考案した。ベンチマークを用いて様々なアクセスパターンを検証した結果、スレッド当たり 1K から 2K のサンプリング数で十分に識別に足る精度が得られ、時間オーバーヘッドはメモリ間のデータコピー時間の 0.025～0.040%と非常に小さく、さらにメモリ消費については、スレッド辺り 256B 程度で済むことが分かった。

図 4: 履歴を用いたパターン識別

⑤ 上記特徴量を用いた意思決定のための評価関数の設計を行った。具体的には性能向上が得られるかどうかを利得とオーバーヘッドを比較することで行っており、この利得部分をパターン依存項とスケーリング係数(アクセス回数等の関数)の積として表現し、このパターン依存項を上記規則性と疎密性の線形和として表現した。この評価関数の係数を調整した上で、様々なアクセスパターンを用いて識別精度を求めた。結果として、十分実用に足る識別精度が得られることが分かり、特にエラーについては、その殆どがエラーの影響が小さい場合に起こることが分かった(即ち、データ転送をする場合としない場合で性能が殆ど変わらない場合である)。

⑥ 提案手法の効果を、様々な HPC カーネルを用いて Knights Landing ベースのシステム上で評価を行った。比較対象としては、低速大容量メモリのみを用いる場合、Numactl コマンドに Preferred オプションを付与して実行した場合、高速小容量メモリをハードウェアキャッシュモードで用いる場合である。評価の結果、低速大容量メモリのみを用いる場合と比較して最大で 3 倍程度、平均で 1.9 倍程度の性能向上が得られ、提案手法の次に性能が高いハードウェアキャッシュモードと比較して、最大で 41%の性能向上が得られた。特にハードウェアキャッシュモードについては、データトラフィックの増大が顕著であり、提案手法はこれと比較して平均で 36%のトラフィック削減を実現しており、結果として、大幅なメモリ電力削減に繋がっている事が容易に予想できる。その他、ハイブリッドメモリシステム向けの API に関する先行研究はいくつかあり、これらに対してのいくつかの定性的な優位性は明らかである。例えば、これらの先行研究は総じて、アプリケーションのプロファイルを用いた手法であるが、それ故にインプット依存性に対処するのが難しい。その一方で提案手法では、実行時にアクセスパターンを識別する点で一線を画しており、これは、プロファイルを用いる必要がない点で優位である。

(2) Footprint-aware Power Capping

電力割り当て最適化技術については、以下の通りに研究開発を行った。

① まずはシンセティックなコードを用いて、性能ボトルネックと問題サイズの関係性について調査を行った。具体的には、演算密度(Flops/Bytes 比)が可変となるシンプルなループを用意し、配列サイズ及び演算密度を変化させて性能を評価し、それを元にルーフラインモデルに基づくプロットを行った(図 5 に示す)。ここでは、高速小容量メモリはキャッシュとして利用する環境を想定する。結果として、ルーフラインの斜線部分(メモリによる性能リミットを示す)は問題サイズを大きくすればする程、下へ移動するため(低速大容量

メモリがより頻繁にアクセスされるため)、ルーフラインの概形は問題サイズによって大きく異なることが分かった。即ち、問題サイズが小さい場合には、CPU や高速小容量メモリがボトルネックとなる場合でも、問題サイズを大きくすれば、低速大容量メモリがボトルネックとなる領域が存在することが分かった。コンポーネント間の電力割り当て最適化においては、ボトルネックを特定し、当該コンポーネントに対してより多くの電力を割り当てる事が重要となるため、このボトルネックの変化を考慮した電力割り当ては、ハイブリッドメモリシステムを搭載する計算機上では極めて重要である。

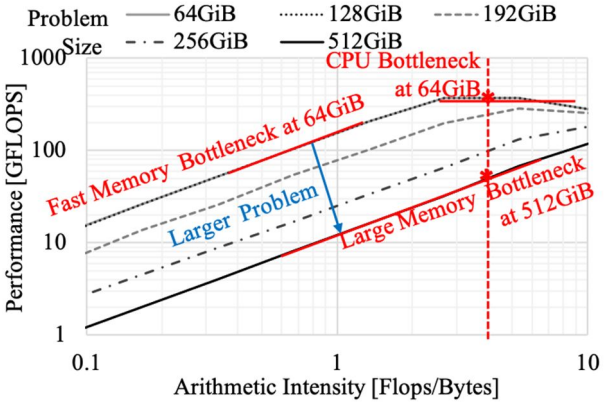

図5: ルーフラインプロットと性能ボトルネック

② 上記観測結果を元に Footprint-aware Power Capping と呼ぶ新概念を提案し、そのポテンシャルについての予備評価を行った。当概念は、上述の通り、アプリケーションの特性及びその問題サイズに応じて各コンポーネントへの電力割り当てを最適化するというものである。本研究については、図 6 の様な電力制御を想定し、スケジューラからノード電力予算(P_{total})が与えられた場合に、計算ノード内の各コンポーネントへの電力割り当てを最適化する。予備評価では、取り得る全組み合わせ試した上で、最適なものを選択しており、即ち、組み合わせの数だけアプリケーションの実行を繰り返している。これを各ベンチマークプログラム+問題サイズの組み合わせに対して行っており、一例として miniFE の場合の結果を図 7 に示す。この様に同一のアプリケーションであっても、最適な電力割り当ては、問題サイズに応じて大きく変化する場合が存在する事が分かった。

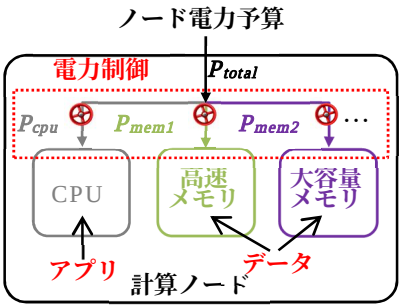

図 6: 想定する電力制御

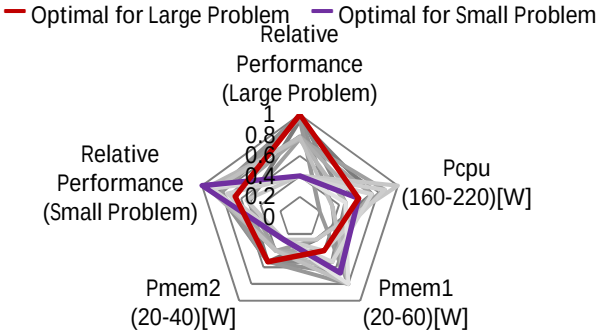

図 7: miniFE 実行時における問題サイズに応じた最適電力割り当ての変化(P_{total}=260W)

③ 上記予備評価結果を元に、Footprint-aware Power Capping の実現のための、問題の定式化及び性能モデリングを行った。性能モデルは広く用いられている、線形モデルを用いて表現し、これは基底関数ベクタと係数ベクタの内積で表現される。基底関数は、アプリケーション特徴量群の関数であり、例えば、演算密度やデータサイズを引数に含む。一方で、係数ベクタは電力割り当て組み合わせの関数とみなす。即ち、同一のアプリケーション+問題サイズの組み合わせであっても、電力割り当てに応じてモデル係数は変化し、その性能への影響は、対応する基底関数の値(特徴量を引数とする)に応じて変化する。

④ 当該モデルに従い、電力割り当て最適化フレームワークを考案した。本フレームワークはアプリケーションレベルの電力割り当て最適化であり、対象カーネルの前後にライブラリ関数コールを配置してこれを行う。ライブラリの内側では、アプリケーションの特徴量を用いて性能モデルに基づいて電力割り当てを最適化する。より具体的には図 8 に示す通りである。提案フレームワークはオフライン部分とオンライン部分に分割することができる。前者については、モデル係数の調整

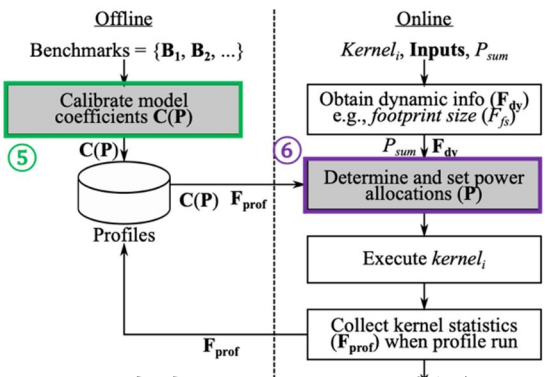

図 8: 提案フレームワークの概要

のためのベンチマークを用いたトレーニングであり、システム毎に行うものである。一方で後者は、電力割り当て最適化を行うためのアルゴリズムや、アプリケーションの静的な特徴量の抽出のためのテストランから成る(初回実行時のみ必要)。

⑤ 効率的なモデル係数調整手法を考案した。具体的には、上述のモデル係数のトレーニングについては、電力割り当ての組み合わせ毎に行う必要があり、これは組み合わせの数が増えた場合には非現実となる。そこで、ここでは、少ない電力組み合わせのみトレーニングを実施して係数を導出し、その後、係数の補完手法(特に評価では線形補完を利用)によって、トレーニングを行っていない組み合わせの係数を推定する手法を採った。

⑥ さらに、ノード電力予算、特徴量群、モデル係数ベクタが与えられた場合に、性能モデルを用いてコンポーネントへの電力割り当てを効率良く導出するためのアルゴリズムについても提案した。具体的には、ヒルクライミングに基づくヒューリスティックな手法を採っており、全探索と比較して少ないオーバーヘッドでの電力配分の決定が可能である。我々の環境では、1μs 程度での割り当ての決定が可能である。

⑦ 上記提案手法について、ハイブリッドメモリを搭載するシステム上にて評価を行った。具体的なシステムのスペック、モデルの基底関数や統計の取得手法等ソフトウェア側の設定、さらには利用ベンチマークの詳細等については、我々が出版している論文を参照されたい。結果として、提案手法によって、最適電力割り当てに近い電力割り当てができることが分かっており、例えば、合計電力予算を 300[W]と設定した場合には、最適の場合と比較して、93%/96%の性能/電力効率となる結果が得られ、さらに、合計電力予算を変化させても同等の結果が得られることが分かった。

４．研究成果

上記、Pattern-aware Staging 及び Footprint-aware Power Capping について、いずれも、HPC 分野で著名な国際会議である ISC HPC 2020 に論文が採択されており、アピール力の高い成果が得られた。当該国際会議は COVID-19 の影響によりオンラインで実施されたものの、参加者が５千人近くにのぼる等、世界各国の HPC 関係者に対して提案技術を知らせる事ができる絶好の機会である。また、論文はオープンアクセスとしており、技術解説動画も国際会議のウェブサイトから閲覧する事ができる(2021 年 4 月現在)。

また、現時点ではハイブリッドメモリシステムの採用は一部の HPC システムに限られているものの、今後新規メモリデバイス技術の低価格化が進めば、より広い HPC システムへの採用が考えられるだけでなく、一般のサーバコンピュータから組み込みシステムに至る、よりコスト制約の厳しいシステムへの展開が予想される。その様なシステムにおいても、データ転送の最適化や電力制御の最適化は広く重要な課題であり、提案アイディアの核となる部分をこれらのシステムへ応用していくことは可能であり、それ故に波及効果も大きいと言える。

また、個々の提案技術について今後の展開は以下の通りである。まず、Pattern-aware Staging については、現時点では HPC カーネルを用いた Proof-of-Concept が済んだ段階であり、今後はコンパイラによる自動化に展開していく必要がある。それができれば、より複雑な HPC アプリケーションやその他の領域のアプリケーションへの展開も容易となる。一方で、電力評価については、現時点ではノード内電力最適化に留まっているため、システム全体を司るジョブスケジューラとの連携についても考えていく余地がある。それができれば、HPC ベンチマークだけでなく、実アプリケーションを含むより現実的な環境での評価も可能となる。さらには、メインメモリだけでなく GPU や FPGA を含むよりヘテロジニアスな環境への展開も視野に入れた研究開発も今後の HPC システムを勘案すると重要である。

さらには、本研究での提案技術はその他の研究の方向性も生み出している。例えば、これまでの HPC システムでは、１計算ノードを１ジョブが占有して利用することが一般的であったが、計算ノードのリッチ化及びヘテロジニアス化が進むにつれて、複数ジョブを同時に動かすコスケジューリングの有効性も認知され始めている。本研究で考えてきた、問題サイズやアクセスパターンによるインパクトはその際にも存在し、資源配分やジョブスケジューリングではこれらを考慮して行う必要がある。これは、当研究代表者の新しい研究課題提案のモチベーションとなっており、実際に科研費の研究課題として採択されている(20K19766)。また、ソフトウェアによるデータ転送最適化のみならず、ハードウェアキャッシュによるデータ配置制御最適化にも取り組んでおり、こちらも国際会議にて論文が採択されている。

５．主な発表論文等

〔雑誌論文〕　計3件（うち査読付論文　3件／うち国際共著　2件／うちオープンアクセス　2件）

１．著者名 Eishi Arima, Martin Schulz	４．巻 －
２．論文標題 Pattern-Aware Staging for Hybrid Memory Systems	５．発行年 2020年
３．雑誌名 High Performance Computing: 35th International Conference, ISC High Performance 2020	６．最初と最後の頁 474～495
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-3-030-50743-5_24	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 Eishi Arima, Toshihiro Hanawa, Carsten Trinitis, Martin Schulz	４．巻 －
２．論文標題 Footprint-Aware Power Capping for Hybrid Memory Based Systems	５．発行年 2020年
３．雑誌名 High Performance Computing: 35th International Conference, ISC High Performance 2020	６．最初と最後の頁 347～369
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-3-030-50743-5_18	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 Eishi Arima	４．巻 －
２．論文標題 Classification-Based Unified Cache Replacement via Partitioned Victim Address History	５．発行年 2020年
３．雑誌名 2020 23rd Euromicro Conference on Digital System Design (DSD)	６．最初と最後の頁 101～108
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/DSD51259.2020.00027	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 －

〔学会発表〕　計6件（うち招待講演　1件／うち国際学会　4件）

１．発表者名 Eishi Arima
２．発表標題 Optimizations for Computing Systems with Emerging Memory Technologies
３．学会等名 Asia Pacific Society for Computing and Information Technology (APSCIT)（招待講演）（国際学会）
４．発表年 2019年

１．発表者名
Eishi Arima, and Carsten Trinitis

２．発表標題
A Case for Co-scheduling for Hybrid Memory Based Systems

３．学会等名
International Conference on Parallel Processing (ICPP)（国際学会）

４．発表年
2019年

１．発表者名
Eishi Arima, Toshihiro Hanawa, Carsten Trinitis, and Martin Schulz

２．発表標題
A Case for Power Shifting on Hybrid Memory Based Systems

３．学会等名
SWoPP

４．発表年
2019年

１．発表者名
Eishi Arima, Toshihiro Hanawa, Martin Schulz

２．発表標題
Toward Footprint-Aware Power Shifting for Hybrid Memory Systems

３．学会等名
ICPP2018（国際学会）

４．発表年
2018年

１．発表者名
Eishi Arima, Martin Schulz

２．発表標題
A Pattern Aware Optimization for Hybrid Main Memories

３．学会等名
HotSPA2019

４．発表年
2019年

１．発表者名 Eishi Arima
２．発表標題 SW/HW Optimizations for Next-Generation Supercomputing
３．学会等名 SC'20 (ITC U. Tokyo Booth)（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			
ドイツ	Technical University of Munich			