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　５ 年　 ５ 月 ２３ 日現在

機関番号：１４４０１
研究種目：挑戦的研究（萌芽）
研究期間：2020～2022
課題番号：２０Ｋ２１７９４

研究課題名（和文）トロピカル代数系に基づく超並列計算理論の構築と応用

研究課題名（英文）Principles and Practice of Massively-Parallel Computing Based on Tropical Algebra

研究代表者
伊野　文彦（Ino, Fumihiko）

大阪大学・大学院情報科学研究科・教授

研究者番号：９０３４６１７２

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）：トロピカル代数系と呼ばれる新しい代数系で記述した組合せ最適化問題を対象として，GPU（Graphics Processing Unit）などの計算アクセラレータ上で高速に解く手法を検討した．具体的には，トロピカル代数系に特有の最適化技術に加え，数万個のスレッドが動作するGPU上で応用を加速するためのデータ圧縮・解凍手法を開発した．これらの有効性を，全点対最短経路探索や量子回路シミュレーションなどの実用的なGPU応用を用いて評価した．

研究成果の学術的意義や社会的意義
人工知能技術の隆盛に象徴されるように，GPUによる計算の高速化は技術のブレークスルーに不可欠な手段として定着している．トロピカル代数系に特有の最適化技術は，道路網やSNSだけでなく，生命情報科学における生体配列の解析に対して貢献でき，適用範囲は広い．また，ライブラリとして実現したデータ圧縮技術は，煩雑なGPUプログラミングの労力を軽減でき，超並列計算機による研究開発の敷居を低下できるものと期待される．

研究成果の概要（英文）：We tried to develop a fast method for solving combinatorial optimization problems with an emerging algebraic system, called tropical algebraic system, on computational accelerators, such as the graphics processing unit (GPU). In more detail, we developed not only an optimization technique specific to tropical algebraic system but also a data compression/decompression method for accelerating GPU applications executed on thousands of threads. Their effectiveness was evaluated with practical GPU applications, such as all-pairs shortest path search and quantum circuit simulation.

研究分野：高性能計算

キーワード：トロピカル代数　最適化問題　高速化　アクセラレータ

１．研究開始当初の背景

近年，生命情報科学分野などにおいて頻出する組合せ最適化問題に対して，GPU（Graphics Processing Unit）による高速な解法が提案されている．組合せ最適化問題に限らず，計算時間の長い応用を GPU 上で加速する研究は，これまでに多様な分野で取り組みがなされている．人工知能技術の隆盛に象徴されるように，データ量が爆発的に増加している現代社会において，GPU による計算の高速化は技術のブレークスルーに不可欠な手段として定着している．

一方，並列計算機を構成するハードウェアは多層化が進み，並列プログラミングに要する労力は増加の一途である．例えば，メモリ遅延を隠蔽できる GPU の登場は，CPU に対して 10 倍の加速を実現したが，GPU 特有の設計と記述が必要であり，超並列処理を実現するための敷居は一層高くなっている．

この敷居を除去するために，コンパイラに対する指示文を逐次プログラムに追加するだけで GPU プログラムを自動的に生成する試み OpenACC がある．しかし，処理できるデータの規模や達成できる性能に関して制約がある．このように，計算アクセラレータ上で高い性能を達成できる，簡潔なプログラム記述は実現できていないのが現状である．また，そのプログラム記述を，計算原理の根底を司る代数系を含めて構築する試みは見当たらない．

２．研究の目的

本研究の目的は，エクサスケール計算時代に向けて，トロピカル代数系で記述できる組合せ最適化問題を，GPU などの計算アクセラレータ上で高速に解くことである．その実現のために，トロピカル代数系に特有の最適化技術を開発し，その計算基盤を最先端の超並列アーキテクチャである GPU 上に展開する．また，具体的な応用として，生命情報科学分野などで頻出する組合せ最適化問題を高速に解く GPU 計算手法を開発する．

３．研究の方法

これまで理論の体系化に留まっていたトロピカル代数系を GPU 上に新たに導入し，その演算規則に基づくプログラミング方法論を展開した．また，簡素な記述によりプログラミング労力を軽減するだけでなく，組合せ最適化問題をはじめとする計算量の大きな問題に対する解法を実現した．具体的には，以下に挙げる研究課題の解決に取り組んだ．

(1) トロピカル代数系に特有の最適化技術

まず，全点対最短経路探索問題がトロピカル代数系上の行列積で解けることから[1]，通常の行列積における加算演算子および乗算演算子を，それぞれ通常の最大値演算子および加算演算子に置換することで，通常の行列積をトロピカル代数系上の行列積に置換した場合の性能を GPU 上で計測した．実装には，GPU 向けの並列処理プログラミングモデルである CUDA（Compute Unified Device Architecture）を用いた．また，トロピカル代数系上の行列積においては，通常の行列積と同様に，タイリングなどのキャッシュ最適化技術が高い性能を達成するために有効であることを確認した．ここで，タイリングとは，行列を小さな部分行列（タイル）に分割し，タイルごとに計算を進めることでデータ参照の局所性を高める手法である（図 1）．図 1 の場合，左端の出力行列におけるタイル(i,j)を計算するために，i 行目のタイル行および j 列目のタイル列をタイル単位で積和すればよい．

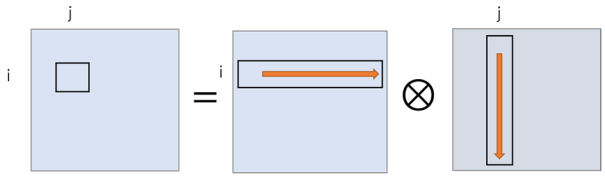

図 1　行列積における参照パターン

$$\begin{bmatrix} \begin{bmatrix} \infty & \infty \\ \infty & \infty \end{bmatrix} & \begin{bmatrix} 2 & 1 \\ 0 & 3 \end{bmatrix} \\ \begin{bmatrix} 0 & 2 \\ 1 & 2 \end{bmatrix} & \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \end{bmatrix} \quad \rightarrow \quad \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$$

計算対象の行列　　属性表

図 2　行列とその属性表

次に，トロピカル代数系上の演算子の性質に着目し，計算の一部を省略する最適化技術を開発した．着目した点は，積の吸収元ならびに和の吸収元であり，それぞれ $x \otimes \infty = \infty$ ならびに $x \oplus 0 = 0$ で表せる．すなわち，計算結果が x の値に依存しない．したがって，本技術は行列を構成する要素の値が 0 である場合，その要素に関連するデータ参照ならびに計算を省く．さらに，無限大の値を含む演算に対しては，タイル単位で計算を省く．これらの省略を実現するために，本技術は対象となる行列内の値を前処理により確認し，タイルごとに属性を定める．タイル内のすべての要素の値が 0 あるいは無限大である場合，タイルの属性を -1 あるいは 1 とし，それ以外は 0 とする（図 2）．この属性を表に格納し，属性表をもとにタイルの計算時にタイル単位でデータ参照ならびに計算を省く．この最適化技術を，計算粒度の細かいレベルから粗いレベルまで GPU 上に実現し，全点対最短経路探索問題における改善効果を評価した．

(2) プログラミング労力を軽減するデータ圧縮ライブラリ

具体的な応用を模索する過程において，CPU・GPU 間のデータ転送量を削減することが応用全体の性能最適化のために不可欠であることを再確認できた．そこで，GPU 応用の高速化を目的として，GPU 上で非可逆データ圧縮・解凍を実現する並列手法を設計し，プログラミング労力を軽減できるライブラリとして開発した（図 3）．提案手法は，CPU 向けの非可逆圧縮手法 FPC [2,3]を拡張したものであり，数万個の GPU スレッドを用いて細粒度の並列性を活用できる．一般に，データの先頭から順に処理を施す圧縮アルゴリズムは逐次制約が強く，FPC も単純には並列化できない．そこで，圧縮対象となるデータをチャンクに細分割し，異なるチャンクを独立に圧縮することにより，チャンクの数だけの並列性を確保する．さらに，同一チャンク内の異なる要素を並列に圧縮できるように，圧縮の基準となる参照要素をチャンク内で先頭要素に統一し，さらなる並列性を活用する（図 4）．また，圧縮データをメモリに書き込む際に，異なる GPU スレッド間の衝突を回避できるように，GPU スレッドに対するタスクの割り当てを工夫する．最後に，入力データのためのメモリ領域を作業領域として再利用することにより，圧縮のためのメモリ消費量ならびに書き込み量を削減し，高速化を図る．これらの工夫を組み込んだライブラリを CUDA により実装し，倍精度浮動小数点数データに対する圧縮率や精度を評価するとともに，既存の圧縮手法 cuZFP [4]に対する優位性を示した．

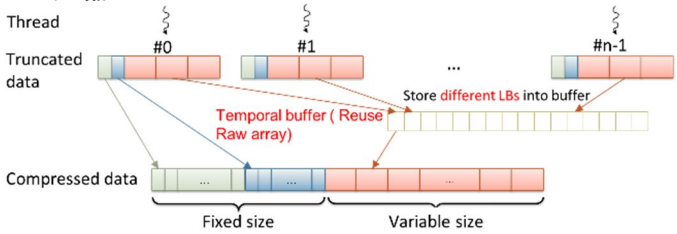

図 3　非可逆データ圧縮手法の概要

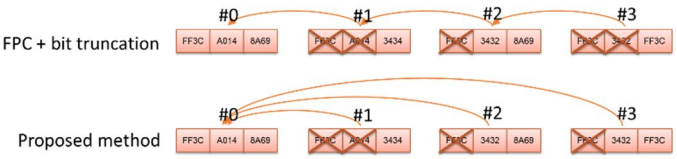

図 4　チャンク内の並列圧縮

4．研究成果

3．で挙げた研究課題の各々について，研究成果を以下にまとめる．

(1) トロピカル代数系に特有の最適化技術

吸収元の性質を基にした計算の省略に加え，タイリングによるキャッシュ最適化を施した行列積の実効性能を NVIDIA RTX A6000 上で評価した．実験に用いた行列は 2 種類であり，乱数で生成したダミーデータならびに 9th DIMACS Implementation Challenge — Shortest Paths [5]にて使用されたベンチマークデータである．前者は，行列要素の分布を変えながら実効性能の特性を調べるために用いた．後者は，米国ニューヨーク，サンフランシスコやコロラド州の道路網をグラフとして表現した実用的なデータである．

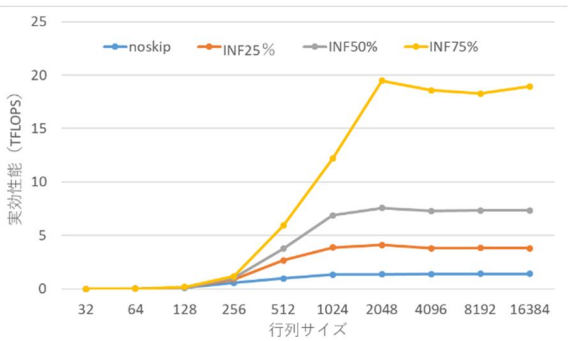

図 5　積の吸収元による改善効果

図 5 は，無限大の値を持つ要素の割合を 0%（noskip）から 75%まで変えながら，実効性能を計測した結果である．行列サイズは32 × 32～16,384 × 16,384とした．なお，積の吸収元のみを用い，和の吸収元による最適化は施していない．行列サイズならびに無限大の割合とともに，改善の度合いが増加した．本技術により計算を削減できない 0%の結果と比べて，最大で 12 倍の性能向上が得られた．一方，小さな行列に対しては，前処理に起因するオーバヘッドが原因で，改善効果は確認できなかった．

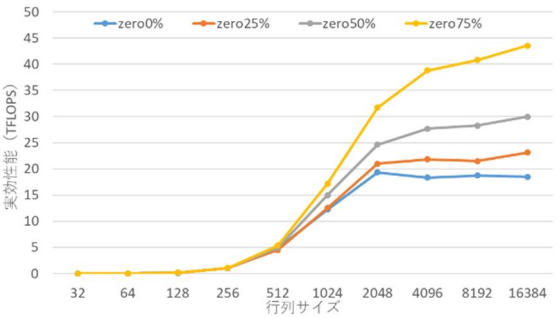

図 6　積と和の吸収元による改善効果

次に，無限大の値を持つ要素の割合を 75%に固定し，残りの要素に関して 0 の値を持つ要素の割合を 0%から 75%まで変えながら，積の吸収元に加え，和の吸収元による最適化を施した（図 6）．0 の値の持つ要素の割合が 75%のとき，積の吸収元のみを用いた場合と比べ，さらに 2 倍強の性能向上を達成した．また，図 5 と同様に，小さな行列に対しては，前処理が原因で性能向上を確認できなかった．しかし，クリーネ閉包の計算など，同じ行列に対する積算を反復する場合には，行列サイズに関わらず本技術は有用であると予想される．また，吸収元だけでなく，

中立元（和における無限大，積における 0）に着目した最適化により，さらなる高速化が期待できる．

さらに，3 つの道路網データを用いて本技術の実効性能を計測した（図 7）．これらのデータは，無限大の値の割合が 99%を占め，大半の計算を省略できる．したがって，本技術による改善効果は大きく，すべてのデータに対して 20 倍程度の速度向上を得ることができ，このときの実効性能は 60 TFLOPS 相当であった．このように，実データには無限大の値を多く含むものがあり，本技術は有用だといえる．

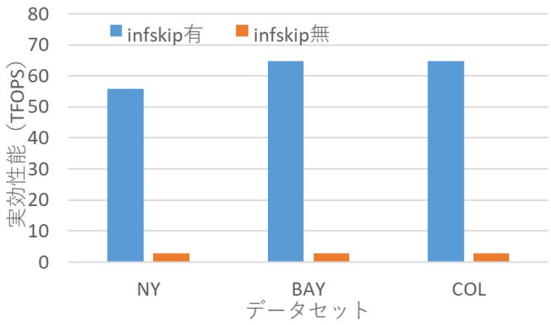

図 7　道路網データに対する実効性能

(2) プログラミング労力を軽減するデータ圧縮ライブラリ

提案手法を評価するために，既存手法に対する速度向上率ならびに圧縮率を計測した．評価に用いた応用は，量子回路シミュレータ Qiskit [6]であり，高速フーリエ変換，量子位相推定や量子もつれを実現する量子回路を用意した．本報告では高速フーリエ変換の結果を示す．Qiskit が扱うデータは，実部および虚部からなる虚数であり，メモリ上に倍精度浮動小数点数として存在する．実験に用いた GPU は，NVIDIA Tesla V100 である．

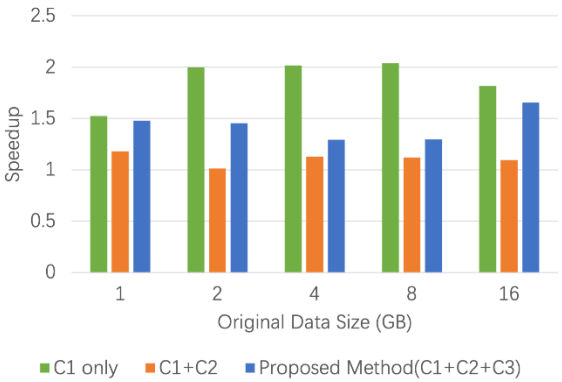

図 8　提案手法による速度向上率

提案手法の核となる 3 つの工夫を組み合わせ，データサイズを 1GB から 16GB まで変えながら，各々が寄与した改善の度合いを調べた（図 8）．図中の速度向上率は，既存手法 FPC [2,3]の GPU 実装を基準としている．細粒度の並列性（C1）を活用することにより，平均して 1.9 倍の速度向上を得た．また，タスクの割り当て（C2）を工夫して書き込み時の衝突を回避することにより，さらに 1.1 倍の速度向上が得られ，作業バッファの再利用（C3）により，さらに 1.7 倍の速度向上が得られた．したがって，主に C1 および C3 が性能改善に寄与していた．

次に，提案手法ならびに既存手法 cuZFP [4]により得られた圧縮率を調べた（図 9）．cuZFP の圧縮率が 2 倍に留まる一方，提案手法は 8 倍の圧縮率を得られた．提案手法が高い圧縮率を得られた理由として，量子回路シミュレータの扱うデータの特徴が挙げられる．このデータは，互いに異なる値を持つ実部と虚部が交互に出現する．しかし，同一チャンク内にそれぞれ同じような値が多く存在し，その多くは 0 に近い値であったため，高い圧縮率が得られた．

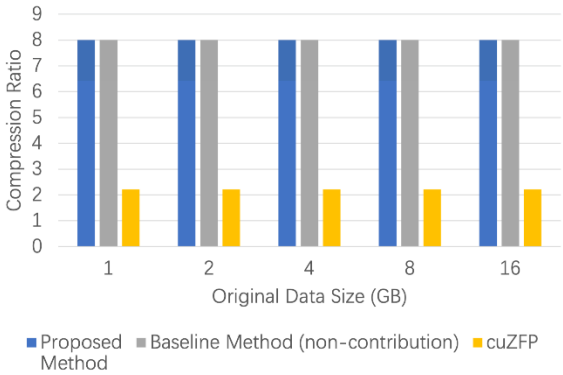

図 9　提案手法と既存手法の圧縮率

参考文献

[1] M. Mohri, “Semiring frameworks and algorithms for shortest-distance problems,” Journal of Automata, Languages and Combinatorics, 2002.
[2] X.-C. Wu, S. Di, E. M. Dasgupta, F. Cappello, H. Finkel, Y. Alexeev, and F. T. Chong, “Full-state quantum circuit simulation by using data compression,” in Proc. SC 2019, 24 pages.
[3] M. Burtscher and P. Ratanaworabhan, “FPC: A high-speed compressor for double precision floating-point data,” IEEE Trans. Computers, vol. 58, no. 1, pp. 18–31, 2008.
[4] M. Larsen, “cuzfp,” https://github.com/LLNL/zfp/tree/develop/src/cuda_zfp, 2019.
[5] http://www.diag.uniroma1.it/ ~challenge9/, 2022.
[6] Qiskit Development Team, “Qiskit 0.31.0,” https://qiskit.org/documentation/#, 2022.

５．主な発表論文等

〔雑誌論文〕　計6件（うち査読付論文　6件／うち国際共著　4件／うちオープンアクセス　1件）

１．著者名 Jingcheng Shen, Linbo Long, Xin Deng, Masao Okita, Fumihiko Ino	４．巻 79
２．論文標題 A compression-based memory-efficient optimization for out-of-core GPU stencil computation	５．発行年 2023年
３．雑誌名 The Journal of Supercomputing	６．最初と最後の頁 11055-11077
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s11227-023-05103-8	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 該当する

１．著者名 Ruiyun Zhu, Yuji Misaki, Marcus Wallden, and Fumihiko Ino	４．巻 24
２．論文標題 Cache-aware volume rendering methods with dynamic data reorganization	５．発行年 2021年
３．雑誌名 Journal of Visualization	６．最初と最後の頁 275-288
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s12650-020-00712-4	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 Marcus Wallden, Masao Okita, Fumihiko Ino, Dimitris Drikakis, and Ioannis Kokkinakis	４．巻 14
２．論文標題 Accelerating In-Transit Co-Processing for Scientific Simulations Using Region-Based Data-Driven Analysis	５．発行年 2021年
３．雑誌名 Algorithms	６．最初と最後の頁 ー
掲載論文のDOI（デジタルオブジェクト識別子） 10.3390/a14050154	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 Yuechao Lu, Yasuyuki Matsushita, and Fumihiko Ino	４．巻 E103-D
２．論文標題 Block Randomized Singular Value Decomposition on GPUs	５．発行年 2020年
３．雑誌名 IEICE Transactions on Information and Systems	６．最初と最後の頁 1949-1959
掲載論文のDOI（デジタルオブジェクト識別子） 10.1587/transinf.2019EDP7265	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 Yuechao Lu, Ichitaro Yamazaki, Fumihiko Ino, Yasuyuki Matsushita, Stanimire Tomov, and Jack Dongarra	４．巻 32
２．論文標題 Reducing the Amount of Out-of-Core Data Access for GPU-Accelerated Randomized SVD	５．発行年 2020年
３．雑誌名 Concurrency and Computation: Practice and Experience	６．最初と最後の頁 e5754
掲載論文のDOI（デジタルオブジェクト識別子） 10.1002/cpe.5754	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 該当する

１．著者名 Jingcheng Shen, Fumihiko Ino, Albert Farres, and Mauricio Hanzich	４．巻 E103-D
２．論文標題 A Data-Centric Directive-Based Framework to Accelerate Out-of-Core Stencil Computation on a GPU	５．発行年 2020年
３．雑誌名 IEICE Transactions on Information and Systems	６．最初と最後の頁 2421-2434
掲載論文のDOI（デジタルオブジェクト識別子） 10.1587/transinf.2020PAP0014	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 該当する

〔学会発表〕　計16件（うち招待講演　1件／うち国際学会　7件）

１．発表者名 Hirotoshi Yamada, Masao Okita, Fumihiko Ino
２．発表標題 Accelerating Imbalanced Many-to-Many Communication with Systematic Delay Insertion
３．学会等名 23rd International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT 2022)（国際学会）
４．発表年 2022年

１．発表者名 Yanchen Li, Qingzhong Ai, and Fumihiko Ino
２．発表標題 A One-Shot Reparameterization Method for Reducing the Loss of Tile Pruning on DNNs
３．学会等名 IEEE World Congress on Computational Intelligence (WCCI 2022)（国際学会）
４．発表年 2022年

１．発表者名
青山昂生，置田真生，伊野文彦

２．発表標題
量子回路の反復シミュレーションにおける実行パスの集約による重複計算の排除

３．学会等名
情報処理学会量子ソフトウェア研究会

４．発表年
2022年

１．発表者名
西村佳，置田真生，伊野文彦

２．発表標題
Forループの並列化可能性の判定における穴埋め言語学習の応用

３．学会等名
情報処理学会量子ハイパフォーマンスコンピューティング研究会

４．発表年
2022年

１．発表者名
羽田遼音，置田真生，伊野文彦

２．発表標題
ニューラルアーキテクチャ探索におけるガウス過程回帰の精度向上のためのバギング手法

３．学会等名
電子情報通信学会情報論的学習理論と機械学習研究会

４．発表年
2022年

１．発表者名
Jingcheng Shen, Yifan Wu, Masao Okita, and Fumihiko Ino

２．発表標題
Accelerating GPU-Based Out-of-Core Stencil Computation with On-the-Fly Compression

３．学会等名
22nd International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT 2021)（国際学会）

４．発表年
2021年

１．発表者名
Yifan Wu, Jingcheng Shen, Masao Okita, and Fumihiko Ino

２．発表標題
Accelerating a Lossy Compression Method with Fine-Grained Parallelism on a GPU

３．学会等名
12th International Symposium on Parallel Architectures, Algorithms and Programming (PAAP 2021)（国際学会）

４．発表年
2021年

１．発表者名
Fumihiko Ino

２．発表標題
A Directive-based Approach for Accelerating Large-scale Scientific Applications on the GPU

３．学会等名
4th International Conference on Artificial Intelligence and Big Data (ICAIBD 2021)（招待講演）（国際学会）

４．発表年
2021年

１．発表者名
坂本慎，置田真生，伊野文彦

２．発表標題
ニューロンの球面クラスタリングによる深層ニューラルネットワークモデル圧縮

３．学会等名
電子情報通信学会ニューロコンピューティング研究会

４．発表年
2021年

１．発表者名
寺西勇裕，置田真生，伊野文彦

２．発表標題
汎用神経回路シミュレータNESTのGPUによる高速化の検討

３．学会等名
電子情報通信学会2022総合大会

４．発表年
2022年

１．発表者名
Jingcheng Shen, Changzeng Fu, Xiangtian Deng, and Fumihiko Ino

２．発表標題
A Study on Training Story Generation Models Based on Event Representations

３．学会等名
3rd International Conference on Artificial Intelligence and Big Data (ICAIBD 2020)（国際学会）

４．発表年
2020年

１．発表者名
Jingcheng Shen, Jie Mei, Marcus Wallden, and Fumihiko Ino

２．発表標題
Integrating GPU Support for FreeSurfer with OpenACC

３．学会等名
6th IEEE International Conference on Computer and Communications (ICCC 2020)（国際学会）

４．発表年
2020年

１．発表者名
井上達博，置田真生，伊野文彦

２．発表標題
Apache Sparkにおける再計算の暗黙的な省略を考慮した性能予測

３．学会等名
情報処理学会システムソフトウェアとオペレーティング・システム研究会

４．発表年
2020年

１．発表者名
水津大樹，沈靖程，伊野文彦

２．発表標題
マルチGPU環境において大規模計算を加速するためのディレクティブ記述手法

３．学会等名
情報処理学会ハイパフォーマンスコンピューティング研究会

４．発表年
2020年

１．発表者名
Ruiyun Zhu and Fumihiko Ino

２．発表標題
A Hybrid Sampling Strategy for Improving the Accuracy of Image Classification with less Data

３．学会等名
電子情報通信学会パターン認識・メディア理解研究会

４．発表年
2020年

１．発表者名
山田広俊，置田真生，伊野文彦

２．発表標題
MPIプログラムにおける遅延挿入による不規則な多対多通信の効率化

３．学会等名
情報処理学会ハイパフォーマンスコンピューティング研究会

４．発表年
2020年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

大阪大学　大学院情報科学研究科　並列処理工学講座
http://www-ppl.ist.osaka-u.ac.jp/

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			
中国	重慶郵電大学			
キプロス	University of Nicosia			
英国	University of Strathclyde			