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　 ６ 月 ２０ 日現在

機関番号：１４３０３
研究種目：基盤研究(C)（一般）
研究期間：2020～2022
課題番号：２０Ｋ１１７３４

研究課題名（和文）自律協調型エッジAIシステムの構成方式に関する研究

研究課題名（英文）A Study on the Configuration of Autonomous Cooperative Edge AI Systems

研究代表者
布目　淳（NUNOME, Atsushi）

京都工芸繊維大学・情報工学・人間科学系・准教授

研究者番号：６０３３５３２０

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：処理の低遅延性を特長とするエッジAI処理は、自動運転車両や配送ドローンなどへの応用が期待され、研究開発が進められている。エッジAI処理は極めて低遅延の学習推論成果を得られる反面、性能の劣る端末単体だけでの処理になることから、学習推論精度が劣化してしまうという問題を抱えている。本研究では、この課題を解決するために、複数のエッジデバイスを協調動作させる自律協調型エッジAIシステムを提案し、その基礎技術を開発した。提案方式により、個々の性能が劣るデバイスでも有機的に連携動作させることで、全体として高い性能を引き出せる基盤技術の一つを確立した。

研究成果の学術的意義や社会的意義
実行時に管理情報を交換しながらネットワーク的に近い位置のデバイスを動的にクラスタリングし、クラスタ間で自律的に処理を移送する制御が十分にシステム全体の潜在性能を引き出せることを示せた。これにより、エッジAI処理環境の新しい構成方式を示すことができた。本研究の成果は、複雑化するシステムを低オーバヘッドで自律協調制御する基盤技術になり得ることから、これまでこうした自律協調制御の適用が難しかったような他の分野に対しても応用が可能である。

研究成果の概要（英文）：Edge AI processing, which is characterized by low latency, is expected to be applied to self-driving vehicles and delivery drones, and is the subject of ongoing research and development. While edge AI processing can produce extremely low latency learning inference results, it has the problem of degrading learning inference accuracy because the processing is performed only on a single device with inferior performance. To solve this problem, we proposed an autonomous cooperative edge AI system in which multiple edge devices operate in a cooperative manner, and developed the basic technology for this system. With the proposed method, we have established one of the fundamental technologies that can bring out high performance as a whole by allowing devices with inferior individual performance to work together in an organic manner.

研究分野：情報工学

キーワード：計算機システム　エッジコンピューティング　分散協調処理　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

(1)　近年では大量の計算処理を必要とする AI 技術が一般化しつつあり、様々な応用分野で AI の利用が進んでいる。特にユーザに近い位置で使用されるスマートフォンや IoT 機器のようなものに対しても実装が進められてきている。

(2)　これまでは膨大な計算資源を有するサーバ上で学習モデルを構築し、それをネットワーク経由で利用するクラウド AI が主流であった。しかしこの構成ではネットワーク上を大量のデータが移動することになり、またネットワークでの遅延時間が大きくなることが課題となっている。今後の実用化が期待される自動運転車両や配送ドローンといった、自律制御が必要となるような分野では、この遅延時間の大きさは致命的である。そこで端末（エッジ）側で学習・推論処理を行うエッジ AI が注目されるようになってきた。エッジ AI ではネットワーク遅延を大幅に削減することが可能になる。

(3)　しかしエッジ AI では使用できる計算能力・記憶容量が厳しく制限されるため、クラウド AI では可能であった処理が十分に行えない可能性が高い。エッジ端末の高速化やクラウドネットワークの低遅延化を行ってこうした課題の解消を図ることは、不均衡だった要素技術のバランスが一時的に改善されるだけであり、課題が軽減されるに過ぎない。このため、課題の根本的な解決にはまったく新しいアーキテクチャを導入する必要があった。

２．研究の目的

(1)　ハードウェア資源の単純な増強は、課題の根本的な解決にはつながらないため、本研究では「エッジデバイスの動的クラスタリング」により、性能の見劣りするデバイス群を協調動作させる。これにより、遅延時間と処理性能の両面においてその改善を図る。

(2)　エッジデバイスの特性上、デバイス位置が固定されず、デバイス間の接続状況が時々刻々と変化することが想定される。そのため、デバイス相互の認識や状態管理を動的に行う必要があり、制御が複雑化する。本研究ではこの制御を人間（管理者）の介入（チューニング作業）に依存することなく、デバイス間の自律的な情報交換によって行う方式を開発する。

(3)　これらの技術的課題をクリアすることで、エッジデバイス間で局所的な連携を図るという新しい構成「自律協調型エッジ AI システム」を実現することを目的とする。このシステムは、近隣のデバイスを有機的に連携させることにより、低遅延性を維持しながら、単体デバイス以上の実効性能を獲得しようとするものである。

３．研究の方法

(1)　デバイス間で交換すべき情報を抽出し、その情報交換の頻度および１回の交換で移動する情報量を明らかにする。これをもとに、デバイス群の動的クラスタリング方式および低オーバヘッド通信方式の設計を行う。

(2)　特定のデバイス構成に限定することを避けるために、統一したインタフェースを提供するシステムソフトウェアの構成方式を検討する。

(3)　デバイス間の低オーバヘッド通信を支援するハードウェア機構を設計する。

(4)　大規模な評価用環境をシミュレータ上に構築する。シミュレータに与える基本的なパラメータとしては、実機あるいはコードフラグメントから実測した値を用いる。

(5)　エッジ AI 処理における学習処理段階では大量のデータに対するアクセスが発生する。この高頻度アクセスに耐えるストレージシステムとして、分散ストレージシステムを考える。エッジデバイス間で分散ストレージシステムを構成することで、デバイス間に散在するストレージ領域を有効活用し、単体のストレージでは処理できないような規模のアクセスにも対応できるようにする。

４．研究成果

(1)　デバイス群の動的クラスタリングを行うための指標として、デバイス間で交換する管理情報を決定した。送信すべき管理情報の粒度とネットワークの低位レイヤにおける情報交換粒度の相違に着目し、管理情報の基本単位を小さく設計した。この基本単位を複数同時に送信することにより、ネットワーク低位レイヤでの実効送信効率を向上させた。同時に、16 バイト程度の短い固定長データとしたことで、ハードウェアによるアクセラレーションを容易にしている。

(2) 情報交換の頻度としては、定期的に送信される生存確認パケット(heart beat)を利用することで、一定間隔で管理情報が通知されるようにした。この生存確認はWi-Fiなどのネットワークレベルで行うものから上位プロトコルのレベルで行われるものまで複数存在し、対応するレベルで管理情報を交換するための機構を実装する必要がある。デバイスの現在の負荷と実効処理能力が管理情報の重要な構成要素となるため、その情報を把握している基本ソフトウェア（オペレーティングシステム）から管理情報交換機構への通信を必要とする。下位のネットワークレベルで行う場合は様々な種類のノード間通信を利用して管理情報を付加することが可能になる反面、ネットワークデバイスの極めてハードウェア寄りの位置での実装が必要になり、柔軟性が損なわれる点が課題である。一方で、より上位のプロトコルレベルで管理情報の付加を行う場合では基本ソフトウェアの拡張で対応できるため、実装コストの低減と応用に対する柔軟性の向上が見込める。これを踏まえて、最終的にオペレーティングシステムに対する拡張機能として実装することとし、TCP/IPレベルの汎用通信に対し、広く管理情報を付加できる方式とした。

(3) 動的クラスタリングの手法として、当初はデバイスの実効性能に基づき、静的に設定した閾値からクラスタを構成するようにした。この方式では、負荷が高まったデバイスの実効性能に対して、20%の性能差が見込めるデバイスを別のクラスタとして認識するようにした。負荷の高まりに応じて上位性能のクラスタへ負荷を移送することで、システム全体の実効性能を維持する。

(4) (3)で示した手法では、閾値設定を 10%刻みで変化させた場合のシミュレーション結果をもとに、多くのパラメータ構成で良好な結果が得られた閾値として 20%という値を決定した。しかし、閾値の最適値は環境によって変化し、また、実行中の負荷変動によりその最適値が変化することが分かった。そこで閾値を動的に調整することで、実行環境に適した閾値へ調整を行う機構を開発した（引用文献①）。この方式では閾値の初期値が十分に調整されていない場合であっても、実行中に最適値へ近付ける制御を行うため、閾値の事前調整に伴う作業が不要になる。図 1 は従来方式において静的に最適な閾値を設定した場合の実行時間と、提案方式で様々な初期値から実行を開始し最長となる（最も不利な）実行時間を比べ実行時間比で示したものである。縦軸が実行時間比、横軸がサーバ数を示しており、従来方式での実行時間を 1 としている。図 1 中の N_c はクライアント数を表している。閾値の初期値が最適に設定されていない場合においても、実行中の調整操作によって実行時間を短縮することができている。図 1 のように、クラスタ構成を動的に制御する提案方式では静的に設定した最適値を使用する方式に比べて最大で 20%程度の性能向上を得ることができた。

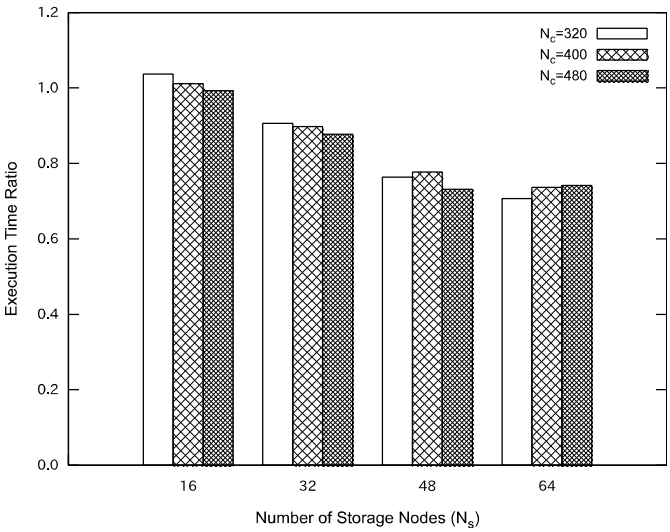

図 1 実行時間比

(5) システムの構成が大規模になるに従い、ノード間の距離にも不均質性が生じる。これは特に、計算処理や作業負荷を移送する際に移送先を選定する処理に影響を与えるため、不均質性を考慮したノード管理が必要になる。より大規模な環境に対応するために、管理情報の収集方式に変更を加え、ノード間の通信遅延時間を考慮する方式を開発した（引用文献②）。図 2 は大規模環境に対応する方式と従来方式の実行時間を比較したものである。縦軸が従来方式での実行時間を 1 とした実行時間比、横軸がネットワークスイッチ数（ネットワーク規模）、S がサーバ数、C がクライアント数をそれぞれ示

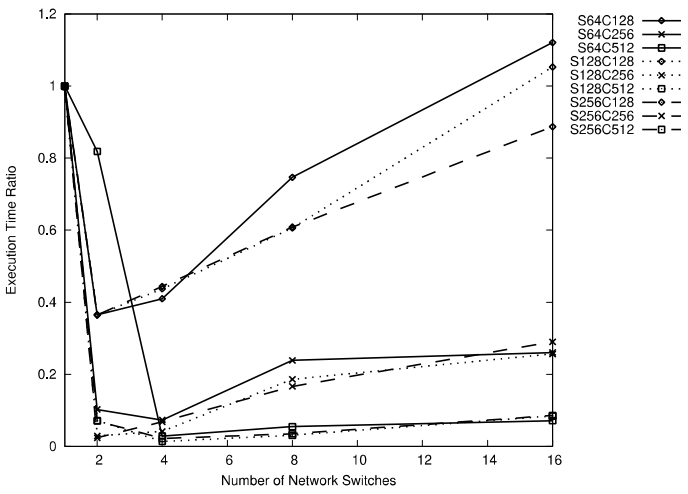

図 2 大規模環境に対応する方式における実行時間比

している。評価環境では、ネットワークスイッチの台数によってノード間の通信遅延時間に差が生じるようにしている。図 2 に示すように、ネットワークの規模を拡大し、情報交換に大きな遅延時間を伴うような環境においても、従来方式より実行時間を大きく削減できる方式を開発できた。提案方式が大きな通信遅延時間のあるような実行環境にも十分対応できることが示された。

(6) 多様なデバイスから構築されるエッジ AI 環境においては、個々のデバイスの実効性能差やデバイス間のネットワーク遅延の差が複雑化することが予想される。こうした環境においてデバイスを静的にクラスタ化してしまうと、実際のデバイスの利用状況を十分に反映することができず、計算機資源の有効利用が困難になる。オーバヘッドを伴う動的クラスタ化においては、過度の制御がかえって実行性能を悪化させることもあり、実装には注意が必要であった。本研究によって、小さなオーバヘッドで管理情報を交換し、それをもとに実効性能が近いデバイスを動的にクラスタ化する手法が確立できた。また、ネットワーク遅延が一定ではないような大規模環境においても、管理情報とその処理方式を工夫することで、自律的に処理を移送し、システム全体の実効性能を向上させることができた。本研究では、実行時に管理情報を交換しながらデバイスを動的にクラスタリングし、クラスタ間で自律的に処理を移送する制御方式を提案し、この方式が十分にシステム全体の潜在性能を引き出せることを示した。この成果はエッジ AI 処理の自律協調制御の実装に大きな意義を与えたと言える。特にネットワーク遅延時間の大きさを考慮し、ネットワーク的に近い位置でクラスタを構成する方式は、エッジ AI の有利性を活かしたまま弱点を大きく軽減できる重要な技術であると言える。さらに本研究の成果は、複雑化するシステムを低オーバヘッドで自律協調制御する基盤技術になり得ることから、将来的にはこれまでこうした自律協調制御の適用が難しかったような分野に対しても応用が可能であると考えられる。

＜引用文献＞

① Atsushi Nunome and Hiroaki Hirata, Adaptive Parameter Tuning for Constructing Storage Tiers in an Autonomous Distributed Storage System, International Journal of Networked and Distributed Computing (IJNDC), Vol.10, issue 1-2, 2022, pp. 1-10.

② Atsushi Nunome and Hiroaki Hirata, Enhancing the Performance of an Autonomous Distributed Storage System in a Large-Scale Network, In Proceedings of the 23rd ACIS International Summer Virtual Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD 2022-Summer), 2022, pp. 87-94.

５．主な発表論文等

〔雑誌論文〕　計5件（うち査読付論文　5件／うち国際共著　0件／うちオープンアクセス　1件）

１．著者名 Atsushi Nunome、Hiroaki Hirata	４．巻 -
２．論文標題 An Adaptive Tiering Scheme for an Autonomous Distributed Storage System	５．発行年 2021年
３．雑誌名 Proceedings of the 8th International Virtual Conference on Applied Computing & Information Technology (ACIT 2021)	６．最初と最後の頁 62～68
掲載論文のDOI（デジタルオブジェクト識別子） 10.1145/3468081.3471124	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 Hiroaki Hirata、Atsushi Nunome	４．巻 -
２．論文標題 Reducing the Repairing Penalty on Misspeculation in Thread-Level Speculation	５．発行年 2021年
３．雑誌名 Proceedings of the 8th International Virtual Conference on Applied Computing & Information Technology (ACIT 2021)	６．最初と最後の頁 39～45
掲載論文のDOI（デジタルオブジェクト識別子） 10.1145/3468081.3471120	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 Atsushi Nunome、Hiroaki Hirata	４．巻 -
２．論文標題 Enhancing the Performance of an Autonomous Distributed Storage System in a Large-Scale Network	５．発行年 2022年
３．雑誌名 Proceedings of the 23rd ACIS International Summer Virtual Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD 2022-Summer)	６．最初と最後の頁 87～94
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/SNPD-Summer57817.2022.00023	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 Hiroaki Hirata、Atsushi Nunome	４．巻 -
２．論文標題 Parallel Binary Search Tree Construction Inspired by Thread-Level Speculation	５．発行年 2022年
３．雑誌名 Proceedings of the 23rd ACIS International Summer Virtual Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD 2022-Summer)	６．最初と最後の頁 74～81
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/SNPD-Summer57817.2022.00021	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 Atsushi Nunome、Hiroaki Hirata	４．巻 10
２．論文標題 Adaptive Parameter Tuning for Constructing Storage Tiers in an Autonomous Distributed Storage System	５．発行年 2022年
３．雑誌名 International Journal of Networked and Distributed Computing	６．最初と最後の頁 1〜10
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s44227-022-00004-3	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 －

〔学会発表〕　計4件（うち招待講演　0件／うち国際学会　4件）

１．発表者名 Atsushi Nunome、Hiroaki Hirata
２．発表標題 An Adaptive Tiering Scheme for an Autonomous Distributed Storage System
３．学会等名 8th ACIS International Virtual Conference on Applied Computing & Information Technology (ACIT 2021)（国際学会）
４．発表年 2021年

１．発表者名 Hiroaki Hirata、Atsushi Nunome
２．発表標題 Reducing the Repairing Penalty on Misspeculation in Thread-Level Speculation
３．学会等名 8th ACIS International Virtual Conference on Applied Computing & Information Technology (ACIT 2021)（国際学会）
４．発表年 2021年

１．発表者名 Atsushi Nunome、Hiroaki Hirata
２．発表標題 Enhancing the Performance of an Autonomous Distributed Storage System in a Large-Scale Network
３．学会等名 23rd ACIS International Summer Virtual Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD 2022-Summer)（国際学会）
４．発表年 2022年

１．発表者名
Hiroaki Hirata、Atsushi Nunome

２．発表標題
Parallel Binary Search Tree Construction Inspired by Thread-Level Speculation

３．学会等名
23rd ACIS International Summer Virtual Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD 2022-Summer)（国際学会）

４．発表年
2022年

〔図書〕　計0件

〔産業財産権〕

〔その他〕
京都工芸繊維大学情報工学課程・専攻コンピュータシステム研究室ホームページ
https://www.ark.is.kit.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
研究分担者	平田　博章 (HIRATA Hiroaki) (90273549)	京都工芸繊維大学・情報工学・人間科学系・教授 (14303)	

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