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　 ６ 月　 ２ 日現在

機関番号：１２６０８
研究種目：研究活動スタート支援
研究期間：2022～2023
課題番号：２２Ｋ２１２８４

研究課題名（和文）複数階層インメモリコンピューティングとGNNを中心とした機械学習への応用

研究課題名（英文）Multi-Layer In-Memory Computing and Its Application for GNNs

研究代表者
藤木　大地（Fujiki, Daichi）

東京工業大学・科学技術創成研究院・准教授

研究者番号：６０９６３２５４

交付決定額（研究期間全体）：（直接経費）　　2,200,000円

研究成果の概要（和文）：本研究は、複数階層インメモリコンピューティング（MLIMP）を用いた計算機のアーキテクチャ上の課題解決に取り組んだ。異種メモリ間のトレードオフの積極的活用のため、GNNなどの機械学習やデータ並列型ワークロードを対象に、タスクスケジューリングと性能予測手法を考案した。

また、MLIMPがメモリ中心型計算における局所性活用課題を解決する可能性を見出した。単一階層では困難だったメモリアクセスに関する局所性活用を、メモリ階層間の性能差を利用したインメモリ計算によって実現した。入力・出力コヒーレンスを保証する「ビュー」という概念を導入し、ビューの再利用を可能にするキャッシュのプロトコル拡張を定義した。

研究成果の学術的意義や社会的意義
本研究の学術的意義は、複数階層インメモリコンピューティングという新しい計算機アーキテクチャの提案と、その上で機械学習やデータ並列処理を効率化するスケジューリング・性能予測手法を提案した点にある。また、コヒーレンスに関する課題を新たに発見し、有用なプロトコル拡張を定義したことは高く評価されている。

社会的意義としては、本研究の成果がGNNなどのデータインテンシブな計算の高速化に繋がり、創薬や金融分野におけるシミュレーションの効率化に貢献できる点が挙げられる。また、インメモリコンピューティングの汎用化は、将来的に様々な計算処理の高速化に繋がり、省エネルギーな社会の実現にも貢献する可能性がある。

研究成果の概要（英文）：This research addressed architectural challenges in multi-layer in-memory computing (MLIMP) systems. To actively leverage the trade-offs between heterogeneous memories in the memory hierarchy, we devised task scheduling and performance prediction methods targeting machine learning workloads such as GNN and data-parallel workloads.

Furthermore, we discovered the potential of MLIMP to solve locality utilization issues in memory-centric computing. We achieved memory access locality exploitation, which was difficult in single-level in-memory computing systems, by utilizing multi-layer in-memory computing that leverages performance tradeoffs between memory hierarchies. We introduced the concept of "view" to guarantee input/output coherence and defined a cache coherence protocol extension to enable view reuse.

研究分野：計算機アーキテクチャ

キーワード：計算機アーキテクチャ　インメモリ計算　メモリアーキテクチャ　キャッシュ

様　式　Ｃ－１９、Ｆ－１９－１（共通）

１．研究開始当初の背景

半導体の集積度の持続的な向上の結果、最新のサーバーではデータ移動で消費されるエネルギーが演算器で消費されるエネルギーの 40 倍と極端に大きくなっており、大量のデータをその移動を最小化した上で処理する技術が求められている。Processing in Memory（PIM）はメモリに搭載された論理回路あるいはメモリそのものを使用して計算を行う技術であり、システム全体（主記憶等含む）の 90%を超えるダイエリアを占有しているメモリを効率的に計算資源に転用することが可能となる。

PIM は SRAM、DRAM、及び不揮発性メモリ NVM など様々なメモリ基盤を対象に研究が行われており、それぞれ異なる計算特性を示している。そもそも、メモリ間の性能・容量トレードオフは従来よりメモリ階層として計算機に組み込まれており、例えば容量は小さいが速度の速い SRAM をキャッシュとして利用し、メモリアクセスの空間的・時間的局所性によってメモリ帯域・レイテンシを向上させる等の工夫がなされてきた。本研究では、単一階層のメモリを対象に研究されてきた PIM の汎用性向上のため、複数のメモリ階層に渡って実装された複数階層 PIM（Multi-Layer In-Memory Processing, MLIMP）を提案し、その有用性及びシステム上の課題を研究するものである。

２．研究の目的

複数メモリ階層が協調し計算する MLIMP では、ヘテロジーニアスな計算資源がメモリ階層上に分散している。本研究では、アプリケーションが MLIMP の恩恵を受けうるアプリケーションを調査し、MLIMP アーキテクチャの十分な動機付けを行う。先行研究では、異なるメモリ基盤が異なる計算特性を擁すことが示されており(例えば､高密度なアナログ計算が可能(NVM)や、高速なビット計算が可能（SRAM）など)、MLIMP における性能利得の最大化は、各アプリケーションに最適なメモリ階層を見つけることはもちろん、分散した計算資源を効率的に協調動作させた場合に得られると考えられる。そのため、計算強度やワーキングデータセットが不定なワークロード、特に複雑性のある機械学習手法（例えばグラフニューラルネットワーク（GNN）および従来アプリケーションのマルチプログラミング）における協調動作のための課題を解決する。本研究では、メモリで実行されるタスクのパフォーマンス予測とスケジューリング手法について、様々な入力にロバストに対応するための手法を研究するほか、PIM の入出力に対し、協調設計が容易となるようなメモリアクセスセマンティクスの研究を行う。

３．研究の方法

本研究では、まず、多層インメモリ計算技術（MLIMP）の実現に向けて、単一視点で作り上げられてきた PIM/IMC のフレームワークを拡張する。次に、コアカーネルのマッピング及びタスクスケジューリング手法・性能予測手法の検討をする。最後に、GNN 等を対象に、計算機シミュレーションによりその性能及びエネルギー効率を検証する。

４．研究成果

主要な研究成果１：複数階層インメモリ計算に向けたスケジューリング手法の提案

MLIMP のヘテロジーニアスなメモリ内計算資源の活用のため、本研究では、まずどのようなアプリケーションが多層インメモリ計算の恩恵を受けうるかの調査を行った。その結果として上がった、メモリインテンシブ・計算インテンシブな要求の両方を持つグラフニューラルネットワーク（GNN）及びマルチプログラミングを想定した多層インメモリ計算アーキテクチャの検討した。また、主要となる疎行列積演算カーネルなどについて、どのようにメモリ上で計算を行ったら性能とコストの良いトレードオフが取れるか検討し、その方法を提案した。

本研究では､多くの機械学習手法の根底にある一般行列積演算 GEMM と疎行列積演算 SpMM を PIM 高速化の主な対象とし、MLIMP が有用なアプリケーションとして特にデータサイズの不定性と計算の複雑性をもつ GNN を主軸として評価する。GNN は、例えばグラフ構造を基にした Recommendation System（User-Ad biparty graph 上での Link prediction）やタンパク質相互作用予測など、学術的・商業的に重要な推論タスクの精度を上げることが示唆されている。高

精度な GNN には mini-batching という手法が採用されており、着目するノードやリンクから k-hop までのノードを含むサブグラフ(k-hop サブグラフ）が入力される。ここで、リアルワールドなグラフはべき分布が散見され、この場合、サブグラフの大きさもべき分布となり、推論時の入力（クエリノード／リンク等）に依存しながら長く重い裾を持つ分布のサブグラフが生成される。PIM は GNN 中の種々の行列演算をこなすことができるが、メモリ容量と処理速度のトレードオフがあるために、GNN のような多様な入力データサイズのアプリケーションを一種類のメモリで処理する場合、必ず suboptimal な問題セットが発生する。MLIMP は、複数の PIM の協調による最適な実行環境を探索し、このようなケースで、従来手法では実現できなかった効率向上へと導くことができる。

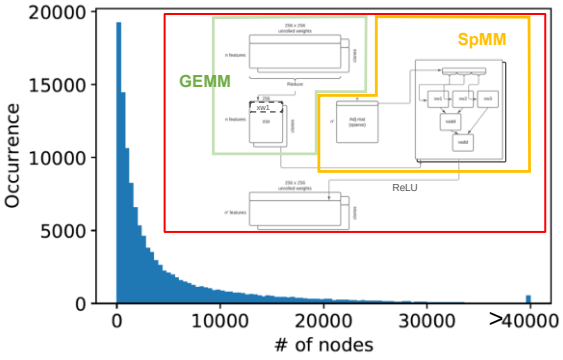

図 1　PIM による GNN 高速化（赤枠内）とサブグラフのノード数分布

研究を進めるうえで、プログラムの入力データの特性によって計算タスクのメモリへの最適マッピングが大きく影響を受けることが明らかになり、コンパイル時に判明するプログラムの特性に加え、入力から素早く特徴を抽出することが求められることが判明した（図 2)。検討をしているアプリケーションでは、入力が多種多様かつ予測困難であることから、機械学習を使ってオンザフライで入力から期待される性能を事前学習し、入力データのダイナミクスを考慮するスケジューリング手法を検討した。OS などで一般的に使用されるスケジューリング手法では不十分であり、かつ最適解を求めるには NP 困難な問題を解く必要があるが、提案手法では軽量な機械学習を用いたヒューリスティックを用いることで、十分精度の高いスケジューリングの解を求められることを示した。本研究をまとめた論文は、計算機アーキテクチャの最難関国際会議 MICRO 2022 に採録された。なお、本研究は米ミシガン大学との共同研究の成果である。

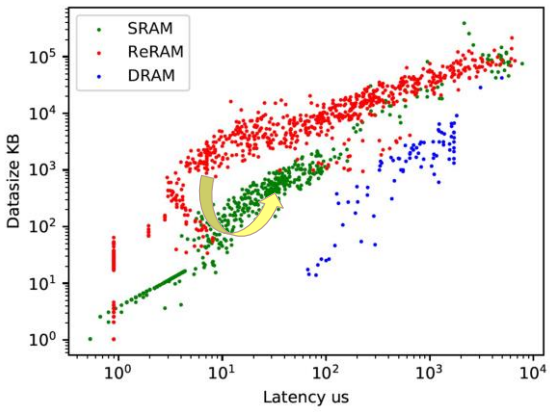

図 2　GNN のジョブサイズ／レイテンシ分布（提案手法でのスケジューリング後）

主要な研究成果 2 ：複数階層インメモリ計算の局所性利活用を可能にするキャッシュコヒーレンスプロトコルの提案

複数階層インメモリコンピューティングの研究を進める過程で、本アプローチがメモリ中心型計算に存在する局所性の利活用に関する課題を解決しうることが知見として得られた。

PIM は図 3 に示す通り、データ構造の見方を様々に変える用途に向いていることに着目し、Row-store のインメモリデータベースから解析のためのフィールドの集約をするワークロードへの応用を考えた。しかし、既存の仕組みには課題があり、集約されたデータ（ビューと呼ぶことにする）に対する内容のコヒーレンスおよび一貫性の保証のため、アクセスのたびにメインメモリなど PIM を有するメモリでビューを再生成する必要があり、入力および出力の局所性の利活用がされていなかった。しかも、コヒーレンスの保証のため、キャッシュなどの機構があえて無効化されており、一部のデータ構造では PIM で実行した方が遅延が増す結果となっていることが判明した。

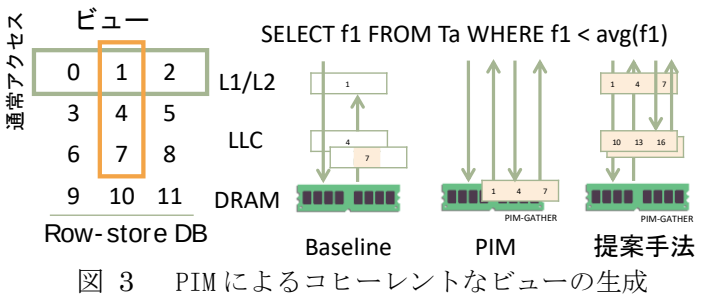

図 3　PIM によるコヒーレントなビューの生成

この課題の解決のため、PIM により生成されたビューをキャッシュ可能にするアーキテクチャである MVC を検討した。MVC はデータベースのビューの概念を参考に、メモリの内容を透過的に変形するフレームワークである。MVC の実装として、キャッシュコヒーレンスアウェアな PIM システムおよびそのための MVC キャッシュコヒーレンスプロトコルの研究・設計を行った。キャッシュ可能なビューを定義したコヒーレンスプロトコル拡張を、MESI ベースのディレクトリ型分散コヒーレンスプロトコルに実装し、モデルバリデーションツールを使用してモデル化を行い、形式検証行った。また、マイクロアーキテクチャシミュレータ ZSIM への本機能の実装し、評価を行った。

本提案により PIM により求められたビューのキャッシュ化が可能となれば、データベースはもとより、それ以外にも様々なアプリケーションで使用されているデータ構造がその恩恵を受けうることを発見した。例えば、簡単な例だと、図 *4* ような連結リストが考えられる。通常連結リストはメモリ空間上にランダムに配置されたオブジェクト同士をポインタチェイシングにより辿るが、検索に必要なキー情報のみを抽出し配列様のビューにすることで、複数回アクセス時にメモリのランダムアクセスを最小化することができる。さらに、本提案によりビューがコヒーレンスプロトコルに従うため、元のデータ構造に変更が加えられたときに、自動的にキャッシュされたビューを無効化することができる。

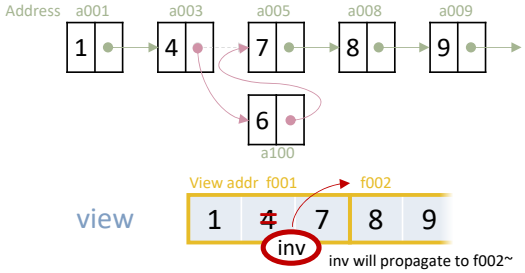

図 4　ビューの連結リストへの拡張

連結リスト以外にも、コヒーレントなビューを用いたソフトウェア最適化は、NoSQL のストレージや中間データ構造として使用されている Skiplist や AVL 木、グラフ構造などに応用することができることが判明した。簡単な Row-store インメモリデータベースのシミュレーション結果では、再利用を含むクエリに対し約 3.57 倍の性能向上が見込まれることが判明した。

本研究により、ビューの再利用が複数のメモリ階層間で可能になり、局所性の利用を可能とすることで汎用インメモリ計算の可能性を広げることができた。本研究をまとめた論文についても、計算機アーキテクチャの最難関国際会議 MICRO 2023 に採録された。

５．主な発表論文等

〔雑誌論文〕　計4件（うち査読付論文　4件／うち国際共著　2件／うちオープンアクセス　1件）

１．著者名 Khadem Alireza、Fujiki Daichi、Talati Nishil、Mahlke Scott、Das Reetuparna	４．巻 24
２．論文標題 Vector-Processing for Mobile Devices: Benchmark and Analysis	５．発行年 2023年
３．雑誌名 IEEE International Symposium on Workload Characterization (IISWC)	６．最初と最後の頁 15-27
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/IISWC59245.2023.00036	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 Fujiki Daichi	４．巻 56
２．論文標題 MVC: Enabling Fully Coherent Multi-Data-Views through the Memory Hierarchy with Processing in Memory	５．発行年 2023年
３．雑誌名 IEEE/ACM International Symposium on Microarchitecture (MICRO)	６．最初と最後の頁 800-814
掲載論文のDOI（デジタルオブジェクト識別子） 10.1145/3613424.3623784	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 Chen Yung-Chin、Ando Shimpei、Fujiki Daichi、Takamaeda-Yamazaki Shinya、Yoshioka Kentaro	４．巻 29
２．論文標題 OSA-HCIM: On-The-Fly Saliency-Aware Hybrid SRAM CIM with Dynamic Precision Configuration	５．発行年 2024年
３．雑誌名 Asia and South Pacific Design Automation Conference (ASP-DAC)	６．最初と最後の頁 539-544
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/ASP-DAC58780.2024.10473966	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 Fujiki Daichi、Khadem Alireza、Mahlke Scott、Das Reetuparna	４．巻 55
２．論文標題 Multi-Layer In-Memory Processing	５．発行年 2022年
３．雑誌名 2022 55th IEEE/ACM International Symposium on Microarchitecture (MICRO)	６．最初と最後の頁 920-936
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/MICRO56248.2022.00068	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

〔学会発表〕　計5件（うち招待講演　5件／うち国際学会　2件）

1．発表者名
藤木大地

2．発表標題
Multi-Layer In-Memory Computing

3．学会等名
FIT Top Conference Session（招待講演）

4．発表年
2023年

1．発表者名
藤木大地

2．発表標題
Data centric computing and machine learning

3．学会等名
Forest Workshop（招待講演）

4．発表年
2024年

1．発表者名
Daichi Fujiki

2．発表標題
Overcoming the Data Movement: Things Learned from Approximated Interconnects and Processing-in-Memory

3．学会等名
Green and Low Carbon Computing Summit（招待講演）（国際学会）

4．発表年
2024年

1．発表者名
藤木大地

2．発表標題
MVC: Enabling Fully Coherent Multi-Data-Views through the Memory Hierarchy with Processing in Memory

3．学会等名
FIT Top Conference Session（招待講演）

4．発表年
2024年

１．発表者名 藤木大地
２．発表標題 Towards Multi-Layer Processing-in-Memory Systems for General Applications
３．学会等名 IEEE Asian Solid-State Circuits Conference (RiSE Session)（招待講演）（国際学会）
４．発表年 2024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			
米国	University of Michigan, Ann Arbor			