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　 ８ 日現在

機関番号：１２６０１
研究種目：基盤研究(B)（一般）
研究期間：2020～2022
課題番号：２０Ｈ０４１９１

研究課題名（和文）ストレージクラスメモリを活用した高速データベースエンジンの構成法

研究課題名（英文）Software architecture of high-performance database engines for exploiting storage class memory

研究代表者
合田　和生（Goda, Kazuo）

東京大学・生産技術研究所・准教授

研究者番号：８０５７４６９９

交付決定額（研究期間全体）：（直接経費）　　13,700,000円

研究成果の概要（和文）：本研究は、ストレージクラスメモリなる新たな記憶媒体をターゲットとし、とりわけ当該媒体が従前の永続的記憶媒体に比して低レイテンシであるという特性に高次に活用することを目指し、データベースエンジンをはじめとするデータインテンシブ処理を担うシステムソフトウェアの構成法を探求するものである。入出力に掛かる新たなソフトウェア制御手法を考案し、ソフトウェア試作機を実装し、解析系関係データベース処理ならびにデータマイニング処理等を対象とする性能試験を行い、とりわけ親和性制御方式が性能向上に有意に寄与することを実験的に明らかにすることに成功した。

研究成果の学術的意義や社会的意義
ビッグデータ、DXという言葉が象徴するように、データを基軸として新たな産業的価値や社会的ソリューションを創造する機運が随所に見られ、国際的な競争が活発に進んでいる。本研究成果はそのようなデータ活用を担う中核的なソフトウェアを対象として、ストレージクラスメモリなる新たな種類の記憶媒体に着目し、基盤的なソフトウェア制御手法（特に、計算機の構成を意識した親和的入出力制御）を世界に先駆けて開発することに成功しており、新たな学術的展開に繋がる他、産業的な活用の潜在性も高く、学術・社会の両面で意義深いものと言える。

研究成果の概要（英文）：This research explores system software architecture for data-intensive processing, including database engines, to take advantage of the low-latency characteristics of storage class memory, a new persistent storage medium. We have invented software control methods related to input/outputs, implemented them into multiple data-intensive software prototypes, and performed intensive experiments with different workloads, such as analytical relational database processing and data mining processing. The experiments have successfully clarified that the developed methods, particularly an affinity control method, contribute to significant performance improvement.

研究分野：データベース

キーワード：ストレージクラスメモリ　データベースエンジン

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

半導体技術の進展により、ストレージクラスメモリ（Storage Class Memory）と称される新たな記憶デバイスが登場してきた。当該記憶デバイスは、永続性と低レイテンシ性を備え、従来、主記憶装置を構成していた DRAM（低レイテンシであるが永続的でない）、若しくは、補助記憶装置を構成していた磁気ディスクドライブおよびフラッシュメモリ（永続的であるが低レイテンシでない）の何れとも特性を異にする点が特長的であり、データベースシステムをはじめとするデータインテンシブ処理を担うシステムソフトウェアに於ける活用方法は未だ解明されていなかった。

２．研究の目的

本研究では、ストレージクラスメモリの当該特性に高度に適合することにより高速化を実現するデータベースエンジン（データベースシステムの中核的なソフトウェア）の構成法を明らかにすると共に、その有効性を実証することに挑戦し、これによって次代の新たな記憶管理アーキテクチャを確立することを目指した。

３．研究の方法

データベースエンジンがストレージクラスメモリを活用することを可能とするために、従前のデータベースエンジンに於いて問合せ処理器に対してストレージエンジンからデータを提供する役割を担っていたバッファマネージャを発展させ、異なる記憶デバイス間に跨ったデータの複製・移送を統合的に制御・調停することを実現し、この際、記憶デバイスの特性に応じた調整を可能とするためのソフトウェアモジュールを設計した。また、今日の情報システムの内部構成の非対称性（例えば、主記憶装置と演算装置の接続は NUMA (Non-Uniform Memory Access) が採用されている）を意識し、演算装置から記憶階層を俯瞰して、データインテンシブ処理の性能向上に寄与するための親和性制御手法を開発した。更に、ストレージクラスメモリは、フラッシュメモリに比して有意に低レイテンシであることが特徴的であり、異なる記憶デバイス間でデータの複製や移送を行うバッファマネージャを高速化することがシステム全体の性能向上に寄与することから、先進的なハードウェア命令を活用することにより制御の精密性と高い並列性を両立するバッファ置換アルゴリズムを開発した。

４．研究成果

ストレージクラスメモリは、遥か以前より産業界で言及されてきたが、ハードウェアとしての実装は永らくの間、限定的な試作品に留まっていた。2019 年 4 月に Intel 社が Optane DC Persistent Memory（以下、Opatne） なる記憶デバイスを発表したことにより、情報システムへの組込みに向けた研究が可能となったものの、本研究の開始時点では殆どその特性が解明されていなかった。本研究では、まず Optane を対象として、人工的なアクセス負荷を与え、その応答を観測することにより、その性能特性を実験的に明らかにするマイクロベンチマーク pmmeter を開発した（なお、2022 年 7 月に Intel 社は事実上の Optane 事業の中断を発表しているが、本報告書執筆時点に於いて、産業界を中心としてこれを置き替える他のハードウェア技術が盛んに議論されていることから、ストレージクラスメモリに関連する技術発展は今後も進展するものと期待される）。従前のプロセッサは、主記憶として利用する DRAM が揮発的であることを前提としており、メモリへのアクセスプロトコルは単一性や永続性を保証しない。しかしながら、ストレージクラスメモリは、永続的であることから、当該記憶デバイスを活用するためには、ソフトウェアの基盤的機能として、データ処理が保証しなければならない単一性や永続性を担保する必要がある。これを実現するための多様な追加的インストラクションが Intel 社のプロセッサには実装されているものの、単一性や永続性の保証に伴う性能上のオーバヘッドは明らかでなかった。pmmeter は追加的インストラクションを選択して、それに伴うオーバヘッドを計測することを可能としている。図１に、計測結果の一例（シングルスレッドによるシーケンシャルなアドレス系列に対する 64 バイトのロード及びストアのアクセススループットの計測）を示す[1]。ここでは DRAM と Optane（図中では PMEM）の双方で同じ実験を行っており、NT, +F, CLF, CLWB, CLFOPT 等はアクセスに伴う永続化のためのインストラクションの選択を意味する。この実験では、DRAM に比して Optane のスループットが 7-8 割程度低いこと、とりわけ+F（FENCE 命令の利用）により著しくスループットが低減することが分かる。マイクロベンチマークの開発により、

インストラクションの選択が性能に与える影響を定量的に把握することを実現した。

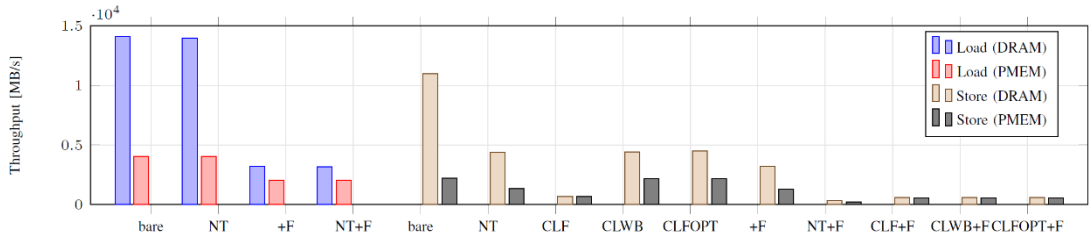

図１．マイクロベンチマークによる計測結果の一例

データベースエンジンに於けるバッファマネージャを中心として、計算機システムの物理構成を意識した親和性制御方式ならびに先端的なプロセッサが具備している排他制御命令を活用したロックフリー化バッファ管理アルゴリズムを設計し、これらを組み込んだソフトウェアモジュールの実装を進めた。また、研究代表者が過去に開発した非順序型データベースエンジンの試作器へ組み込む実装を行った。ページアクセスレベルの人工的なマイクロベンチマーク（偏りのないアクセスおよび偏りのあるアクセス）ならびに標準的な解析系デー多ベースベンチマークである TPC-H を用いた性能試験を実施した。例えば、解析系のデータベース問合せに於いては積極的にデータベースバッファを分割し、プロセッサコア間の相互作用を低減する親和制御を行うことにより、ハードウェアが潜在的に有している入出力帯域をバッファマネージャが有効に活用することができるようになる等の知見を得るに至った[2]。

同様のアプローチを並列データマイニング処理へ適用し、その有効性を検証した。図２に親和性制御方式による性能優位性に関する試験結果を示す。ここでもここでも DRAM ならびに Optane（図中では persistent memory）を比較しており、データマイニングのアルゴリズム特性とハードウェアの NUMA 特性を融合した制御方式（DPHIM）によって最大で 2 倍程度の性能向上を達成可能であることを実験的に明らかにした[3]。

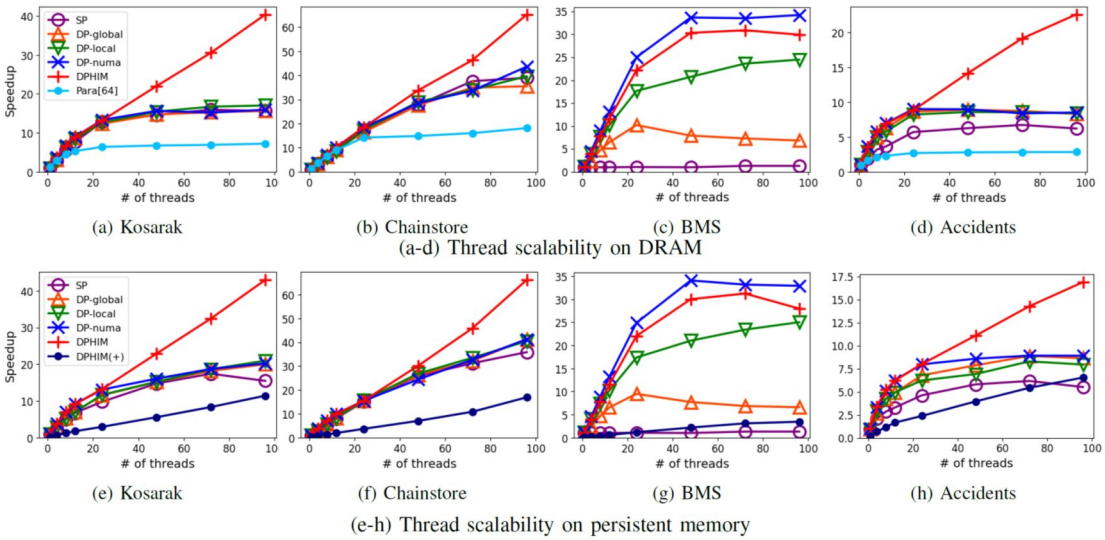

図２．データマイニング処理を対象とする性能試験の一例

＜引用文献＞

[1] H. Yoshioka et al. pmmeter: A Microbenchmark for Understanding Synchronization Cost on Persistent Memory. Proc. BigComp 2023: 326-327.

[2] T Ozawa et al. Thread-specific Database Buffer Management in Multi-core NVM Storage Environments. Proc. NVMW 2021.

[3] G. Kimura et al. Efficient Parallel Mining of High-utility Itemsets on Multicore Processors. Proc. ICDE 2023 (to appear).

５．主な発表論文等

〔雑誌論文〕　計1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　1件）

１．著者名 Yutaro Bessho, Yuto Hayamizu, Kazuo Goda, Masaru Kitsuregawa	４．巻 E105-D(5)
２．論文標題 Dynamic Fault Tolerance for Multi-Node Query Processing	５．発行年 2022年
３．雑誌名 IEICE Transactions on Information and Systems	６．最初と最後の頁 909-919
掲載論文のDOI（デジタルオブジェクト識別子） 10.1587/transinf.2021DAP0004	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26件（うち招待講演　0件／うち国際学会　6件）

１．発表者名 Mika Takata, Kazuo Goda, Masaru Kitsuregawa
２．発表標題 μ-join: Efficient Join with Versioned Dimension Tables
３．学会等名 Proceedings of the 27th International Conference on Database Systems for Advanced Applications (DASFAA 2022)（国際学会）
４．発表年 2022年

１．発表者名 Hiroki Yuasa, Kazuo Goda, Masaru Kitsuregawa
２．発表標題 Exploiting Embedded Synopsis for Exact and Approximate Query Processing
３．学会等名 Proceedings of the 33rd International Conference on Database and Expert Systems Applications (DEXA 2022)（国際学会）
４．発表年 2022年

１．発表者名 Hirotaka Yoshioka, Yuto Hayamizu, Kazuo Goda, Masaru Kitsuregawa
２．発表標題 pmmeter: A Microbenchmark for Understanding Synchronization Cost on Persistent Memory
３．学会等名 Proceedings 2023 International Conference on Big Data and Smart Computing (BigComp 2023)（国際学会）
４．発表年 2022年

１．発表者名
Genki Kimura, Yuto Hayamizu, R. Uday Kiran, Masaru Kitsuregawa, Kazuo Goda

２．発表標題
Efficient Parallel Mining of High-utility Itemsets on Multicore Processors

３．学会等名
Proceedings of the 39th IEEE International Conference on Data Engineering (ICDE 2023)（国際学会）

４．発表年
2023年

１．発表者名
吉岡弘隆，早水悠登，合田和生，喜連川優

２．発表標題
不揮発メモリを対象とする性能マイクロベンチマークpmmeterの検討と予備試験

３．学会等名
電子情報通信学会データ工学研究会，電子情報通信学会技術報告

４．発表年
2022年

１．発表者名
木村元紀，早水悠登，ラゲウダイキラン，合田和生，喜連川優．吉岡弘隆，早水悠登，合田和生，喜連川優

２．発表標題
NUMA環境に於ける高効用アイテムセットマイニングの並列実行方式の検討と予備実験

３．学会等名
電子情報通信学会データ工学研究会，電子情報通信学会技術報告

４．発表年
2022年

１．発表者名
三浦優也，小沢健史，合田和生

２．発表標題
ストレージ上のデータベースに対するCPUとGPUを併用した基本処理の実行方式に関する予備実験

３．学会等名
情報処理学会データベースシステム研究会，電子情報学会研究報告

４．発表年
2022年

１．発表者名
髙田実佳，喜連川優，合田和生

２．発表標題
シノプシス埋込みによる近似問合せ処理の試作実装と初期評価

３．学会等名
第15回データ工学と情報マネジメントに関するフォーラム／第21回日本データベース学会年次大会(DEIM2023)

４．発表年
2023年

１．発表者名
木村元紀，早水悠登，ウダイラゲ，喜連川優，合田和生

２．発表標題
再帰的演算を含む分析処理の高効率な並列実行方式の提案と有効性評価

３．学会等名
第15回データ工学と情報マネジメントに関するフォーラム／第21回日本データベース学会年次大会(DEIM2023)

４．発表年
2023年

１．発表者名
吉岡弘隆，早水悠登，合田和生，喜連川優

２．発表標題
不揮発メモリを対象とする空間索引構造の実装方式の検討と予備実験

３．学会等名
第15回データ工学と情報マネジメントに関するフォーラム／第21回日本データベース学会年次大会(DEIM2023)

４．発表年
2023年

１．発表者名
三浦優也，小沢健史，合田和生

２．発表標題
GPU直接I0を用いたデータベース問合せ処理の検討と予備実験

３．学会等名
第15回データ工学と情報マネジメントに関するフォーラム／第21回日本データベース学会年次大会(DEIM2023)

４．発表年
2023年

１．発表者名
Tsuyoshi Ozawa, Ryoji Kawamichi, Yuto Hayamizu, Kazuo Goda, Masaru Kitsuregawa

２．発表標題
Early experience of Utilizing Persistent Memory for Database Bulk Loading

３．学会等名
The 20th USENIX Conference on File and Storage Technologies (FAST2022), Refereed Work-in-Progress Presentation（国際学会）

４．発表年
2022年

１．発表者名
湯浅拓樹，合田和生，喜連川優

２．発表標題
B+木へのシノプシス埋め込みによる近似問合せとその予備的な実験

３．学会等名
電子情報通信学会データ工学研究会

４．発表年
2021年

１．発表者名
加藤滉貴，小沢健史，合田和生，喜連川優

２．発表標題
並列データベースシステムに於けるRDMAを用いたリモート入出力性能の検討

３．学会等名
電子情報通信学会データ工学研究会

４．発表年
2021年

１．発表者名
湯浅拓樹，合田和生，喜連川優

２．発表標題
既存データ構造へのシノプシス組み込みによる近似問合せ手法

３．学会等名
第14回データ工学と情報マネジメントに関するフォーラム／第20回日本データベース学会年次大会(DEIM2022)

４．発表年
2022年

１．発表者名
髙田実佳，合田和生，喜連川優

２．発表標題
世代管理されたディメンション表を対象とする結合処理の効率的実行

３．学会等名
第14回データ工学と情報マネジメントに関するフォーラム／第20回日本データベース学会年次大会(DEIM2022)

４．発表年
2022年

１．発表者名
加藤滉貴，小沢健史，合田和生，喜連川優

２．発表標題
並列データベースシステムに於けるRDMAを用いたリモート入出力性能の測定と問合せ処理への影響

３．学会等名
第14回データ工学と情報マネジメントに関するフォーラム／第20回日本データベース学会年次大会(DEIM2022)

４．発表年
2022年

１．発表者名
木村元紀，合田和生， Rage Uday Kiran，喜連川優

２．発表標題
高効用アイテムセットマイニングの高効率な並列化手法とその評価

３．学会等名
第14回データ工学と情報マネジメントに関するフォーラム／第20回日本データベース学会年次大会(DEIM2022)

４．発表年
2022年

１．発表者名
吉岡弘隆，合田和生，喜連川優

２．発表標題
不揮発性メモリ性能測定のためのマイクロベンチマークの設計と実装

３．学会等名
第14回データ工学と情報マネジメントに関するフォーラム／第20回日本データベース学会年次大会(DEIM2022)

４．発表年
2022年

１．発表者名
Tsuyoshi Ozawa, Yuto Hayamizu, Kazuo Goda, Masaru Kitsuregawa

２．発表標題
Thread-specific Database Buffer Management in Multi-core NVM Storage Environments

３．学会等名
12th Annual Non-Volatile Memories Workshop（国際学会）

４．発表年
2021年

１．発表者名
吉岡弘隆，合田和生，喜連川優

２．発表標題
不揮発メモリデバイスの性能評価のためのマイクロベンチマークに関する初期検討

３．学会等名
電子情報通信学会データ工学研究会

４．発表年
2020年

１．発表者名
吉岡弘隆，合田和生，喜連川優

２．発表標題
不揮発メモリデバイスを対象とするデータベース演算の実行コスト測定方式に関する検討

３．学会等名
電子情報通信学会データ工学研究会

４．発表年
2020年

１．発表者名
吉岡弘隆，小沢健史，合田和生，喜連川優

２．発表標題
不揮発メモリデバイスを対象とする結合演算のマルチスレッド実行性能に関する実験と一考察

３．学会等名
電子情報通信学会第13回データ工学と情報マネジメントに関するフォーラム／第19回日本データベース学会年次大会

４．発表年
2021年

１．発表者名
加藤滉貴，小沢健史，合田和生，喜連川優

２．発表標題
RDMAを用いたリモート入出力性能の実験的考察

３．学会等名
電子情報通信学会第13回データ工学と情報マネジメントに関するフォーラム／第19回日本データベース学会年次大会

４．発表年
2021年

１．発表者名
湯浅拓樹，合田和生，喜連川優

２．発表標題
B+木の付加情報を用いた近似問合せ手法の検討

３．学会等名
電子情報通信学会第13回データ工学と情報マネジメントに関するフォーラム／第19回日本データベース学会年次大会

４．発表年
2021年

１．発表者名
髙田実佳，合田和生，喜連川優

２．発表標題
世代管理されたマスタ表とトランザクション表との結合処理に関する考察

３．学会等名
電子情報通信学会第13回データ工学と情報マネジメントに関するフォーラム／第19回日本データベース学会年次大会

４．発表年
2021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