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　　元 年　　６ 月 １３ 日現在

機関番号：３２６１２
研究種目：基盤研究(C)（一般）
研究期間：2016～2018
課題番号：１６Ｋ００１５０

研究課題名（和文）先進的デバイスの利活用による高性能データ基盤システムに関する研究

研究課題名（英文）Research on high performance data infrastructure system by utilizing advanced devices

研究代表者

川島　英之（KAWASHIMA, Hideyuki）

慶應義塾大学・環境情報学部（藤沢）・准教授

研究者番号：９０４０７１４８

交付決定額（研究期間全体）：（直接経費）　　3,500,000円

研究成果の概要（和文）：本研究では先進的デバイスを用いた高性能データ処理に関する研究を行った。そのなかに分散データベースの高性能化がある。分散データベース管理システムにおいて外部キー制約や二次索引、実体化ビューの管理を行うための高性能な処理方式としてRAMPトランザクションがある。RAMPトランザクションは隔離性を緩和することで高性能化されているがそれを先進的デバイスによって高性能化する技法は未開拓である。そこで，本研究では高性能インターコネクトであるInfiniBandを利用し、RDMAの機能を用いてRAMPトランザクションを高速化する手法を提案した。成果は情報処理学会山下記念研究賞を受賞した。

研究成果の学術的意義や社会的意義
本研究成果により、分散トランザクション処理が高性能化される。分散トランザクションはソーシャルネットワーク、ビッグサイエンスなど、膨大なデータを用いる環境において頻繁に使われる処理である。そのような処理を高性能化することは、多くのユーザにとって便益をもたらす。これが本研究の貢献である。

研究成果の概要（英文）：In this research, we researched high performance data processing using advanced devices. Among them is the improvement of distributed database performance. Read Atomic Multi-Partition (RAMP) transactions are high-performance processing methods for managing foreign key constraints, secondary indexes, and materialized views in a distributed database management system. Although RAMP transactions have been enhanced by mitigating isolation, techniques to improve them by advanced devices are unexplored. Therefore, in this research, we proposed a method to accelerate RAMP transaction using InfiniBand, which is a high performance interconnect, using the function of Remote Direct Memory Access (RDMA). The result won the Information Processing Society of Japan Yamashita Memorial Research Award.

研究分野：システムソフトウェア

キーワード：トランザクション　システムソフトウェア　データシステム

１．研究開始当初の背景

計算機の低価格化や処理能力向上により、従来は考えられなかった規模のデータが急速に生成され始めていた。例えば、Facebook 社で扱うデータ量は日々600 テラバイト増大しており、大型ハドロン衝突型加速器（LHC）が生成するデータ量は年間 30 ペタバイトと言われていた。このような規模のデータを処理するには、従来型のデータベース管理システムは適用できないとの認識があった。このため、Facebook、Amazon、Google 社などの大規模データを扱う事業者は、Warehouse-Scale Computers（WSCs）と称されるように、数千台の計算機を用いて大規模クラスタであるデータセンタを設置し、WSCs で動作する専用のビッグデータ処理基盤を構築していた。

ビッグデータ処理基盤には、処理効率化と低消費電力化のために先進的デバイスの利活用が求められていた。WSCs では計算機が数千台以上利用されるが、現状の消費電力でより多くのデータを扱えることが望まれた。究極の高性能化と省電力化を求めるスーパーコンピュータにおいては GPU や Xeon Phi などの先進的デバイスを用いることは常識となっていたが、同様にビッグデータ処理基盤においても、より一層の計算処理効率化のためにアクセラレータ、I/O 効率化のために不揮発メモリ等を用いることが、これから必要になっていくと考えられた。

２．研究の目的

本研究の目的は、新しい高性能データ基盤の構築原理を探求することだった。データ基盤のワークロードは 2 種類に大別されるため、二種類の基盤技術研究を行うことが目的だった。第一のワークロードは分析処理（例：TPC-H）であり、その具体例にはリレーショナル結合演算や MapReduce だった。第二のワークロードはトランザクション処理（例：TPC-C）であり、その具体例には所謂 DBMS があった。

(1) 分析処理に関する研究（結合演算・MapReduce）

分析処理には様々な演算が含まれるが、特に負荷が高い演算として結合演算と MapReduce が挙げられる。これらの演算をハードウェアとソフトウェアを協調動作させることで高性能化させる研究を行った。

(2) トランザクション処理に関する研究（ログ先行書込・分散トランザクション処理）

トランザクション処理における高負荷処理の 1 つにログ先行書込（Write Ahead Logging）がある。これはクラッシュ後のシステム回復に必要である。これを行うにはデータをメモリからストレージへ移動する必要がある。その性能限界を引き出すため、ハードウェアとソフトウェアを協調設計する研究を行うことは研究目的の一つだった。

分散トランザクションを高性能化する最新技術は RDMA(Remote Direct Memory Access)の利活用だった。RDMA を活用することにより、世界最高性能の分散トランザクション処理を実現する要素技術を開発することも研究目的の一つだった。

３．研究の方法

上述のように高性能ハードウェアを活用するために、分析処理基盤とトランザクション処理基盤を再設計する研究方法を採用した。

４．研究成果

(1) 並列ログ先行書込みプロトコル

本研究はフラッシュストレージをログ用のストレージデバイスとするときにふさわしい WAL プロトコルとして P-WAL を提案した。フラッシュストレージは複数のメモリチップに対して並列にアクセスすることで高い性能を発揮する。P-WAL はフラッシュストレージの特性を活用し、各ワーカが専用の領域にログを書き込む並列ログ書き込み方式を用いる。この方式により従来の直列 WAL 方式で発生する，排他制御処理とストレージ I/O にともなう性能低下問題を解決した。P-WAL をトランザクションシステム上で実装し、性能評価を行った結果、P-WAL は直列 WAL 方式に対して図 1 に示すように大幅な性能向上を示した。

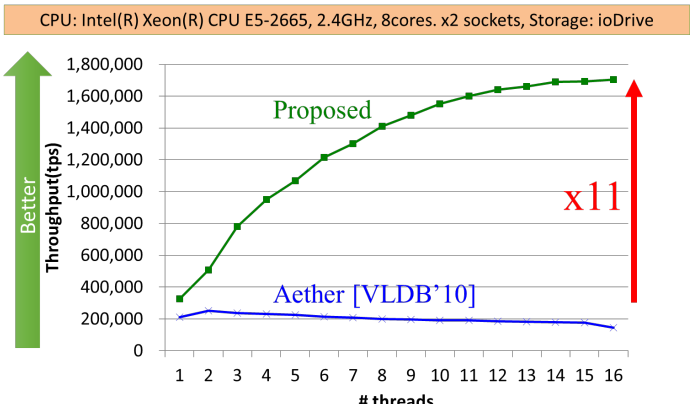

図１　P-WAL による性能向上

(2) RDMA による RAMP トランザクションの高性能化
分散データベース管理システムにおいて外部キー制約や二次索引，実体化ビューの管理を行うための高性能な処理方式として Read Atomic Multi-Partition (RAMP) トランザクションがある。RAMP トランザクションは隔離性を緩和することで高性能化されているが、それを先進的デバイスによって高性能化する技法は未開拓だった。そこで、本研究では高性能インターコネクトである InfiniBand を利用し、Remote Direct Memory Access (RDMA) の機能を用いて RAMP トランザクションを高速化する手法を提案した。まず、RDMA-Write による GET/PUT オペレーションの高速化手法として GET+/PUT+方式を提案した。続いて、RDMA-Read によるさらなる GET オペレーションの高速化手法として GET*方式を提案した。提案手法の評価のため、プロトタイプ In-Memory Key-Value Store を実装した。Yahoo! Cloud Serving Benchmark を用いた実験において、従来方式と比べて高速化を達成した。

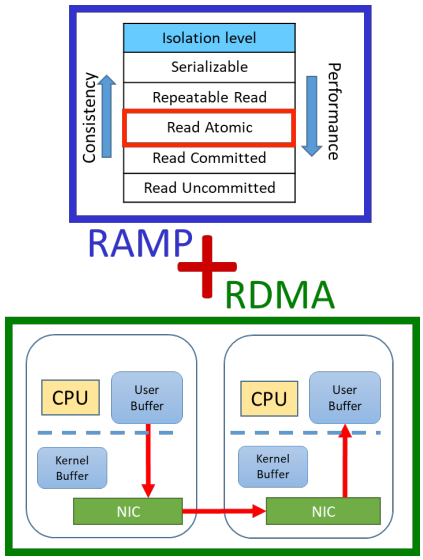

図 2 RAMP + RDMA

(3) Shuffling の高性能化
本研究では、偏ったデータによって引き起こされる MapReduce シャッフリングの問題に対処するために設計された 3 つのメモリ内シャッフリング方法を提案し、検証した。結合シャッフルアーキテクチャ (CSA) は、対応するブロックのメタデータを含む両方のブロック、シャッフル転送の単位、およびメタブロックをシャッフルするために、単一の対全対全交換を使用する。分離シャッフルアーキテクチャ (DSA) は、メタブロックとブロックのシャッフルを分離し、それぞれのシャッフルプロセスに異なる全対全交換アルゴリズムを適用して、偏った分布におけるストラグラーの影響を軽減する。スキュー対応メタシャッフル (DSA w / SMS) を使用した分離シャッフルアーキテクチャは、各ワーカープロセスのメモリ消費量に基づいて、ブロックの適切な配置を自律的に決定する。このアプローチは、一部のワーカープロセスがノードメモリの制限を超える可能性がある、非常に偏った状況を対象とした。この研究では、InfiniBand や Intel Omni-Path などの高性能インターコネクトを採用した、プロトタイプのインメモリ MapReduce エンジンにおける 3 つのシャッフル方法の実装を評価した。実験の結果、、DSA w / SMS が極端に歪んだデータ分布に対して唯一の実行可能な解決策であることを示した。

(4) 並行性制御法 TicToc と並列ログ先行書込み法 P-WAL の統合
本研究ではメニーコアマシンと NVDIMM を用いた環境でトランザクション処理を高性能化する手法を検討した。トランザクション処理の分野では、ハードウェアの発展に伴って楽観的並行性制御や並列ログ書き込みが登場した。既存の並列ログ書き込み手法は、ログレコードに順序を与えるために共有カウンタを用いる。しかしメニーコアマシンにおいて、共有カウンタへのアクセスはボトルネックとなる。そこで、本研究は楽観的並行性制御において生成したタイムスタンプをログレコードに付与する手法を提案する。評価実験では提案手法と従来の高性能化技法のスループットを TPC-C ベンチマークを用いて評価した。その結果、NVDIMM へのログ書き込みを想定した時に提案手法は既存手法のと比較して 1.37 倍のスループットを示した。しかし、SSD へのログ書き込みを想定した時には、依然として提案手法よりも既存手法が良いスループットを示した。よって、メニーコアマシンと NVDIMM を用いた環境では、従来の高性能化技法である Early Lock Release と Group Commit は性能劣化をもたらすことがわかった。

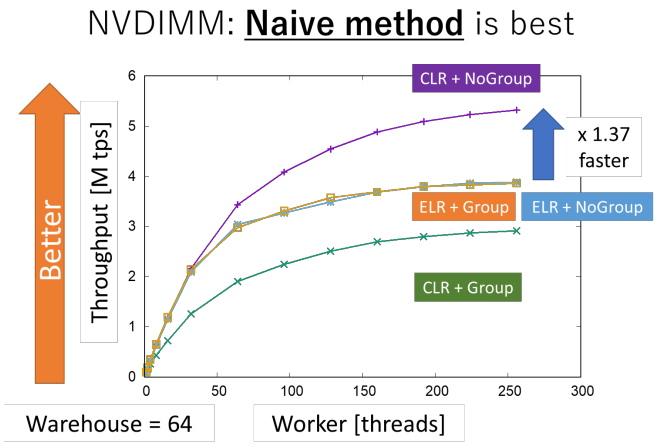

図 3 TicToc+P-WAL

５．主な発表論文等

〔雑誌論文〕（計 9 件）

1. Yasuhiro Nakamura, Hideyuki Kawashima, Osamu Tatebe. Integrating TicToc with Parallel Logging. CANDAR Workshops, pp. 105-111, 2018. 査読有.
2. Takayuki Tanabe, Hideyuki Kawashima, Osamu Tatebe, Integration of Parallel Write Ahead Logging and Cicada Concurrency Control Method, SMARTCOMP, pp. 291-296, 2018. 査読有.
3. Ryota Takizawa, Hideyuki Kawashima, Osamu Tatebe. Performing External Join Operator on PostgreSQL with Data Transfer Approach. Proceedings of the International Conference on High Performance Computing in Asia-Pacific Region, pp. 1-7. 2018. 査読有.
4. 村田直郁, 川島英之, 建部修見. RDMA の適用による RAMP トランザクション処理の高速化. 情報処理学会論文誌データベース. Vol. 10, No. 2, pp. 19-30, 2017. 査読有.
5. 神谷孝明, 川島英之, 星野喬, 建部修見. 並列ログ先行書き込み手法 P-WAL, 情報処理学会論文誌データベース. Vol. 10, No. 1, pp. 24-39, 2017. 査読有.
6. Naofumi Murata, Hideyuki Kawashima, Osamu Tatebe. Accelerating Read Atomic Multi-partition Transaction with Remote Direct Memory Access, IEEE International Conference on Big Data and Smart Computing, pp. 239-246, 2017. 査読有.
7. Kentaro Horio, Hideyuki Kawashima, Osamu Tatebe. Efficient Parallel Summation on Encrypted Database System. IEEE International Conference on Big Data and Smart Computing, pp. 178-185, 2017. 査読有.
8. Ryuya Mitsuhashi, Hideyuki Kawashima, Takahiro Nishimichi, Osamu Tatebe. Three-Dimensional Spatial Join Count exploiting CPU Optimized STR R-Tree. IEEE Big Data Conference. pp. 2938-2947, 2016. 査読有.
9. Harunobu Daikoku, Hideyuki Kawashima, Osamu Tatebe. On Exploring Efficient Shuffle Design for In-Memory MapReduce. BeyondMR workshop. 2016. 査読有.

〔学会発表〕（計 11 件）

1. 田辺敬之, 川島英之, 建部修見. 並行性制御法 Cicada の評価, 第 10 回データ工学と情報マネジメントに関するフォーラム, 2017.
2. 中村泰大, 川島英之, 建部修見. 並列 WAL を適用した TicToc の評価. xSIG, 2017.
3. 渡辺敬之, 川島英之, 建部修見. 並行実行木 Masstree における一括構築法の並列化. xSIG, 2017.
4. 梶原顕伍, 川島英之, 建部修見. Raft に基づく分散データベースにおけるデータ分割, xSIG 2017.
5. 渡辺敬之, 川島英之, 建部修見. 並行実行木 Masstree の調査, 第 9 回データ工学と情報マネジメントに関するフォーラム, 2017 年 3 月 6 日, 高山グリーンホテル（岐阜県高山市）.
6. 中村泰大, 川島英之, 建部修見. 並行実行制御手法 TicToc の調査, 第 9 回データ工学と情報マネジメントに関するフォーラム, 2017 年 3 月 6 日, 高山グリーンホテル（岐阜県高山市）
7. 梶原顕伍, 川島英之, 建部修見. 分散合意手法 Raft の調査, 第 9 回データ工学と情報マネジメントに関するフォーラム, 2017 年 3 月 6 日, 高山グリーンホテル（岐阜県高山市）
8. 渡辺敬之, 川島英之, 建部修見. 並行実行木 Masstree の一括構築法, 情報処理学会第 139 回 OS 研究会, 2017 年 3 月 2 日, アクロス福岡.
9. 中村泰大, 川島英之, 建部修見. 並行実行制御手法 TicToc と並列ログ先行書き込み手法 P-WAL の結合, 情報処理学会第 139 回 OS 研究会, 2017 年 3 月 2 日, アクロス福岡.
10. 神谷孝明, 星野喬, 川島英之, 建部修見, トランザクション処理システムのリカバリ可能性の再考, 情報処理学会第 139 回 OS 研究会, 2017 年 3 月 2 日, アクロス福岡.
11. 川島英之, 建部修見, 並列データベースシステムにおける演算子間データ配送方式, 情報処理学会第 138 回 OS 研究会, 2016 年 8 月 8 日, キッセイ文化ホール（長野県松本市）

〔図書〕（計 0 件）

〔産業財産権〕
○出願状況（計 0 件）

○取得状況（計 0 件）

〔その他〕

６．研究組織

(2)研究協力者
研究協力者氏名：佐野 健太郎
ローマ字氏名：(SANO, Kentaro)