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　２ 年　 ７ 月　 ６ 日現在

機関番号：６２６１５
研究種目：基盤研究(A)（一般）
研究期間：2016～2019
課題番号：１６Ｈ０１７１７

研究課題名（和文）不揮発メモリ及び小型原子時計等を前提とした分散システム技術

研究課題名（英文）Technologies for Distributed Systems with NvRAM and Atomic Clocks

研究代表者
佐藤　一郎（Satoh, Ichiro）

国立情報学研究所・情報社会相関研究系・教授

研究者番号：８０２８２８９６

交付決定額（研究期間全体）：（直接経費）　　33,400,000円

研究成果の概要（和文）：不揮発性メモリや超小型原子時計などの新しいデバイスやクラウドコンピューティングを前提とした次世代分散トランザクション技術を提案・実装・評価する。こうした新デバイスはトランザクションの実装技術は大きく変化と、性能や機能に大幅な改善をもたらす。しかし、トランザクションは枯れた技術と扱われてきており、その対応は立ち後れている。本研究では近い将来普及が予測される不揮発性メモリや超小型原子時計などの新デバイスを活かした分散トランザクションを世界に先駆けて提案・実装し、既存のトランザクションに対して性能改善していく。

研究成果の学術的意義や社会的意義
学術的には不揮発性メモリや高精度時計などの最新ハードウェアデバイスは分散システムに性能及び信頼性改善に寄与することが指摘されてきたが、その具体的な手法は明らかになっておらず、それを提示したことが本研究の意義及び貢献となる。また、実際的にはクラウドコンピューティングなどの広域に広がる分散システムのデータ管理基盤技術の提供することになる。今後、こうした最新デバイスの普及とその性能改善により、研究の有用性は高まるはずである。

研究成果の概要（英文）：We propose a next-generation distributed transaction technology with new devices, e.g., non-volatile memory and ultra-small atomic clocks, and cloud computing, and implement and evaluate it. We expected these devices to bring significant changes in transaction implementation technology and significant improvements in performance and functionality, although existing distributed transactions are seriously and essentially affected by communication latency. For example, non-volatile memory allows distributed transactions to avoid operations for persistence and atomic clocks enables messages to be ordered based on the times of their transmits in the sender side rather than the receiver side, because these messages have precise time-stamps generated from atomic clocks in the sender side.
We also evaluate our proposed methods in comparison with conventional synchronization mechanism such as two-phase commit in their consistency and performance.

研究分野：分散システム

キーワード：分散システム　不揮発記憶装置　高精度時計　トランザクション

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

情報システムのハードウェア技術は大きな変革期にあり、メニーコア化や GPU の多用というプロセッサ以外にも様々なハードウェアが登場している。この研究ではその中でも不揮発性メモリや高精度時計などの新技術を想定した分散システム技術を対象とする。例えば現在、主記憶に使われている DRAM は微細化の限界から、MRAM や FeRAM などの不揮発性メモリに置き換わると予想される。また、既存のコンピュータは水晶発振器の時計が使われ、温度変化によっては 10^{-3} のオーダーで計時に誤差が出るために、相違なコンピュータ間の協調動作における通信や動作の順序づけに時刻情報を使うことはなく、時刻情報を使った分散アルゴリズムの研究は進んでいない。しかし、現在、チップスケールの原子時計が実用化されており、各サーバーが 10^{-9} 以上の高精度の時計を内蔵する可能性もある。

一方、分散システムは耐故障性の観点から、データの複製を複数サーバーに保持することになるが、そのデータに更新があった場合、複製間の一貫性保持のコストが大きいが、あるデータに対する複数の更新処理が重なった場合、その更新順番通りに複製が更新されるとは限らないという問題があり、その解決コストは極めて大きい。また、分散データベースの場合、データベースシステムの特性上、データは不揮発性メモリに保持させる処理、つまり永続化処理が必要となるが、複製を保持するすべてサーバーにおける永続化処理はコストが大きく、トランザクションの実現コストが大きくなる。なお、これらは、分散システム及び分散データベースの特有な問題として知られており、数多くの改善策が提案されているが、既存ハードウェアを前提にする限りは、アプリケーションに依存した手法や一貫性条件の緩和以外の解決は難しいのが現実であり、新しいハードウェア技術による解決が望まれている。

２．研究の目的

分散システムのデータ管理に関わる問題、つまりデータ複製の更新性能の問題と、その更新処理におけるトランザクションの性能に関わる問題を解決する方法として超小型原子時計や不揮発性メモリなどの新しいデバイスを前提にした次世代分散データ管理技術を提案･実装･評価する。超小型原子時計により各サーバーが高精度に掲示できることが予想される。その場合、相違なコンピュータ間の協調動作における通信や動作の順序づけに時刻情報を使う余地が生まれる。ただし、前述のように時刻情報を使った分散アルゴリズムの研究は進んでいない。ここでは時刻情報を使った分散合意アルゴリズムを明らかにする。特に分散アルゴリズムの多くは、分散合意と呼ばれる複数サーバー間で同じ状態を保持する仕組みが基礎になることから、高精度時刻を前提にした分散合意アルゴリズムを設計し、その分散合意アルゴリズムの応用として複数サーバー間の複製管理を考える。具体的にはサーバー間通信における順序づけを、従前の受信順から送信順に変更するものとなる。通信遅延は予測できるとは限らない。この結果、受信側において処理の巻き戻しが起きうるが、これは後述する分散トランザクションにより解決する。

さて複製データの一貫性のためのトランザクションの巻き戻しについては、マルチバージョン(Multiple Version Concurrent Control, MVCC)トランザクション手法を導入する。MVCC は更新対象となるデータが増大するため、永続化コストが高く、HDD や SSD に代表される二次記憶装置を前提にした場合、性能劣化が大きいことが知られている。しかし、不揮発性メモリの導入により、データベースシステムではトランザクションを明示的な永続化処理を行うことなく永続されることになる。このため、ローカルなトランザクション処理については効率化が期待できる。ただし、不揮発性メモリを前提にした MVCC 手法は、海外クラウドプラットフォーム事業者を中心に研究されているが、本研究の特徴は前述の高精度時計を前提にした複製データの一貫性制御に特化していることが特徴かつ優位性となる。なお、耐故障性の観点から本研究では、3 フェーズコミットプロトコルをベースとしたものも導入していく。

３．研究の方法

二つのパートに分けて研究を進めた。①高精度時計を前提とした複製データの一貫性機構であ

り、②不揮発性メモリを想定した分散トランザクションとなる。なお、本研究は新しいデバイスを前提にするが、研究開始時点においてはデバイスメーカーにおける研究開発段階のプロトタイプであり、その入手が難しいことが予想されたことから、デバイスの基本性能などを再現する方法と、実際にデバイスを利用して評価する部分に分けて遂行した。

さて①については、従前の分散データ複製管理は、Primary Backup 手法に代表されるように一つのサーバーが複製を保持するサーバーの更新を行う方法と、アクティブレプリカ(Active Replica)と呼ばれる方式があるが、本研究では後者を対象とする。これは商用クラウドプラットフォームを含め、広域の分散システムを対象とするためである。アクティブレプリカでは複数コンピュータに対して更新を行うメッセージの全域的順序を保証、または事後に全域的順序に基づき更新する方法が前提となる。このときに問題となるのは、複数のデータ更新が重なった場合である。この場合、複製を保持するサーバーにより、データ更新の順番が相違することが起きうる。従前の方法ではデータ更新を指示したサーバーは、他のサーバーへのデータ更新を送信後に、2 フェーズコミットより更新データの到着及び実行順序の確認を行った上で、更新を指示する方法をとるために、少なとも 3 回通信が必要となり、コストが大きい。提案手法では、各サーバーが高精度な時計をもつと仮定する。

Step 1: データ更新メッセージには高精度な時計によるタイムスタンプを付与したうえで送信する

Step 2: データ複製を保持する各サーバーはデータ更新メッセージを受信すると、タイムスタンプが過去に受信・更新したデータ更新メッセージのタイムスタンプと比較する

Step 3-A: 新たに受信したデータ更新メッセージのタイムスタンプの方が新しい場合は直ちに更新を行うが、その前の状態に関するバージョンまたはログを保持する

Step 3-B: 新たに受信したデータ更新メッセージのタイムスタンプの方が古い場合は、それより新しいタイムスタンプの更新処理の巻き戻し処理を行う

さて②に対応する不揮発性メモリを前提とした分散トランザクション機構は、この巻き戻しを実現する機構として位置づけられる。既存手法と同様にアプリケーションに対して予め分離(Isolation)レベルを設定することで、巻き戻し処理を低コスト化するが、加えて不揮発性メモリの特性を活かすことになる。トランザクションのためのログ及びバージョンを不揮発性メモリに保持することになり、個々のサーバーにおいて行われるが、巻き戻し対象の更新をそのサーバー以外に通知している場合は巻き戻し範囲が広がることになる。このとき巻き戻しのための情報をログベースで管理している場合、ログを遡ることになるが、サーバーごとにログの順序が大きく異なる場合、コストが大きくなる。また、MVCC、つまりバージョンを保持する方法は分散スナップショット手法を導入することで、相違コンピュータ間で広域的スナップショットを実現して、巻き戻しのポイントが明示的に見つけられるようにした。なお、複製データの一貫性制御においては基本となる 2 フェーズコミットプロトコルに加えて、3 フェーズコミットプロトコルのひとつとなる Paxos をベースにした分散合意プロトコルを実装した。

４．研究成果

前述の①となる高精度時計を前提とした複製データの一貫性機構については、汎用トランザクションエンジンに相当するソフトウェアを設計・実装した。前述のように小型原子時計については米国製のモジュールを利用する予定でいたが、研究開始後に品質に問題が発生し、メーカーからの供給が止まり、さらに供給再開後は価格が約 3 倍となり、入手を断念することとなった。そこで当初の研究計画通りに、まずは同一コンピュータにおける各プロセスを、アクティブレプリカのサーバーとして扱い、それぞれのサーバー同士は外部サーバーを介して通信することで、各プロセスの時計の誤差がない状態を再現し、通信遅延に関しては外部サーバーとの通信において遅延をいれることで再現することができた。トランザクションエンジンについては、今後、オープンソースソフトウェアとして公表する予定となっている。このほか、原子時計に相当する方法として GPS 信号による置き換えにおいて実験も行った。

また、②の分散トランザクションについては不揮発性メモリについては複数半導体ベンダーにおいてその出荷が遅れたこともあり、大半の評価を DIMM 接続型フラッシュメモリによる実験に切り替えた。さてターゲットの不揮発性メモリであるが、入手可能な試作品（半導体ベンダーから性能評価用にサーバーベンダーに貸し出されたもの）を借りて、基本的な性能評価を行ったところ、読み込み性能に対して、書き込み性能が極端に悪いことが判明した。つまり、現時点では基本性能はフラッシュメモリと大きな差がないことになり、DIMM 型フラッシュメモリを用いた基礎評価を行った。また、分散トランザクションの実装に関しても、性能を優先するとデータをいったん DRAM などの高速かつ揮発性メモリに保持して、処理後に不揮発性メモリに書き出す方法が現実的となることが判明した。このため、既存の不揮発性メモリを前提にした実装、つまり書き込み性能が低いことを前提にした方式と、将来的に読み書きの性能差が小さくなった場合の処理方式の二種類を実装することとなった。なお、後者は当初予定通りであるが、前者に関しては、現時点では不揮発性メモリはフラッシュメモリに対して相対的に性能がよいが、容量が少ない不揮発記憶装置という扱いになり、またトランザクションエンジンに関しても永続化処理が必要となる場合がある。

研究発表に関して、論文リストにあるように本研究の初年度より積極的に国際会議などに成果を発表することができた。

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　1件／うちオープンアクセス　1件）

１．著者名 Ichiro Satoh	４．巻 110
２．論文標題 Spatial Connector: Mapping Access Control Models for Pervasive Computing and Cloud Computing	５．発行年 2017年
３．雑誌名 Procedia Computer Science	６．最初と最後の頁 174-181
掲載論文のDOI（デジタルオブジェクト識別子） https://doi.org/10.1016/j.procs.2017.06.075	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 Jingtao Sun, Ichiro Satoh	４．巻 24
２．論文標題 Theory and Implementation of an Adaptive Middleware for Ubiquitous Computing Systems	５．発行年 2016年
３．雑誌名 Journal of Information Processing	６．最初と最後の頁 878-886
掲載論文のDOI（デジタルオブジェクト識別子） http://doi.org/10.2197/ipsjjip.24.878	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 Ichiro Satoh	４．巻 94
２．論文標題 Self-Adaptively Auto-scaling for Mobile Cloud Applications	５．発行年 2016年
３．雑誌名 Procedia Computer Science	６．最初と最後の頁 9-16
掲載論文のDOI（デジタルオブジェクト識別子） https://doi.org/10.1016/j.procs.2016.08.006	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 Narito Kurata	４．巻 12
２．論文標題 Seismic Observation and Structural Health Monitoring of Buildings by Improved Sensor Device Capable of Autonomously Keeping Accurate Time Information	５．発行年 2019年
３．雑誌名 International Journal on Advances in Systems and Measurements	６．最初と最後の頁 41-50
掲載論文のDOI（デジタルオブジェクト識別子） https://www.thinkmind.org/articles/sensordevices_2018_7_20_20073.pdf	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15件（うち招待講演　1件／うち国際学会　14件）

1．発表者名 Ichiro Satoh
2．発表標題 An Approach for Recovering Distributed Systems from Disasters
3．学会等名 8th International Conference on Bioinspired Optimization Methods and Their Applications（国際学会）
4．発表年 2018年

1．発表者名 Ichiro Satoh
2．発表標題 Adaptive Deployment of Software in Smart-City.
3．学会等名 Proceedings of the 4th EAI International Conference on Smart Objects and Technologies for Social Good（国際学会）
4．発表年 2018年

1．発表者名 Ichiro Satoh
2．発表標題 An approach for testing software on networked transport robots
3．学会等名 14th IEEE International Workshop on Factory Communication Systems（国際学会）
4．発表年 2018年

1．発表者名 Ichiro Satoh
2．発表標題 Context-Aware Access Control Model for Services Provided from Cloud Computing
3．学会等名 1th International Symposium on Intelligent Distributed Computing 2017（国際学会）
4．発表年 2017年

１．発表者名
Ichiro Satoh

２．発表標題
Mobile Agent: Lost Agent Technology. Why could Mobile Agents be in Practice?

３．学会等名
15th International Conference on Practical Applications of Cyber-Physical Multi-Agent Systems（招待講演）（国際学会）

４．発表年
2017年

１．発表者名
Yuki Kano, Tatsuo Nakajima

２．発表標題
An alternative approach to blockchain mining work for making blockchain technologies fit to ubiquitous and mobile computing environments

３．学会等名
10th International Conference on Mobile Computing and Ubiquitous Network（国際学会）

４．発表年
2017年

１．発表者名
Yuki Kano, Tatsuo Nakajima

２．発表標題
A New Approach to Mining Work in Blockchain Technologies.

３．学会等名
15th International Conference on Advances in Mobile Computing & Multimedia（国際学会）

４．発表年
2017年

１．発表者名
Ichiro Satoh

２．発表標題
Access Control Model for Ambient Intelligent Services in Cloud Computing

３．学会等名
SmartCloud 2016（国際学会）

４．発表年
2016年

１．発表者名
Ichiro Satoh

２．発表標題
Toward Access Control Model for Context-Aware Services Offloaded to Cloud Computing

３．学会等名
35th IEEE Symposium on Reliable Distributed Systems Workshops（国際学会）

４．発表年
2016年

１．発表者名
Ichiro Satoh

２．発表標題
Adaptive Scaling Up/Down for Elastic Clouds

３．学会等名
10th International Symposium on Intelligent Distributed Computing（国際学会）

４．発表年
2016年

１．発表者名
Ichiro Satoh

２．発表標題
Constraint Solving-Based Automatic Generation of Mobile Agent Itineraries

３．学会等名
International Conference on Artificial Intelligence and Soft Computing（国際学会）

４．発表年
2016年

１．発表者名
Hitoshi Mitake, Hiroshi Yamada, and Tatsuo Nakajima

２．発表標題
Looking into the Peak Memory Consumption of Epoch-Based Reclamation in Scalable in-Memory Database Systems

３．学会等名
Database and Expert Systems Applications. DEXA 2019.（国際学会）

４．発表年
2019年

１．発表者名
Hitoshi Mitake, Hiroshi Yamada, and Tatsuo Nakajima

２．発表標題
A Highly Scalable Index Structure for Multicore In-Memory Database Systems

３．学会等名
International Conference on Intelligent Distributed Computing XIII. IDC 2019（国際学会）

４．発表年
2019年

１．発表者名
Yuki Yaida, Tatsuo Nakajima

２．発表標題
Experiments of Distributed Ledger Technologies Based on Global Clock Mechanisms

３．学会等名
Intelligent Distributed Computing XII. IDC 2018

４．発表年
2018年

１．発表者名
Hitoshi Mitake, Hiroshi Yamada, and Tatsuo Nakajima

２．発表標題
Analyzing The Tradeoff Between Throughput and Latency in Multicore Scalable In-Memory Database Systems

３．学会等名
th ACM SIGOPS Asia-Pacific Workshop on Systems (APSys '16)（国際学会）

４．発表年
2016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
研究分担者	倉田　成人 (Kurata Narito) (00416869)	筑波技術大学・産業技術学部・教授 (12103)	

６．研究組織（つづき）

	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	中島　達夫 (Nakajima Tatsuo) (10251977)	早稲田大学・理工学術院・教授 (32689)	