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 １７ 日現在

機関番号：１３９０４
研究種目：若手研究(B)
研究期間：2016～2018
課題番号：１６Ｋ１６０３５

研究課題名（和文）高い柔軟性を有したクラウドBFTレプリケーションシステムの実現

研究課題名（英文）Development of Elastic Cloud BFT Replication System

研究代表者
中村　純哉（Nakamura, Junya）

豊橋技術科学大学・情報メディア基盤センター・助教

研究者番号：６０７３９７４６

交付決定額（研究期間全体）：（直接経費）　　1,600,000円

研究成果の概要（和文）：Byzantine Fault Tolerant（BFT）レプリケーションは，状態機械として定義される一般的なサービスに対して，ビザンチン故障と呼ばれる故障時の動作に一切の仮定をおかない最も強力な故障モデルに対する故障耐性を実現できる．本研究課題ではBFTレプリケーションについて，Amazon Web Servicesに代表されるクラウドサービスでも柔軟にレプリケーションを実現するための手法を提案した．提案手法は，レプリケーションを構成する際に問題となるレプリカ配置について可能な全ての配置案を網羅的に評価することで，利用者の目的に最もあったレプリカ配置の選択を可能にする．

研究成果の学術的意義や社会的意義
本研究課題では，BFTレプリケーションをクラウドサービス上で構築する際に重要となるレプリカ配置の決定問題について取り組んだ．BFTレプリケーションをクラウドサービスのような地理的に分散した環境で効率的に実現するための手法についてはこれまでいくつか提案されているものの，それらは既に決まっているレプリカ配置においてレプリカの役割を入れ替えることで性能を向上させるものだった．提案手法はそれらとは異なり，レプリカ配置そのものを決定する段階で利用できるという点が学術的な新規性である．

研究成果の概要（英文）：Byzantine Fault Tolerant (BFT) replication can realize fault tolerance for general services defined as state machines against the strongest fault model, called Byzantine fault, which does not make any assumptions about fault behavior. In this research subject, we propose a method to realize an efficient and elastic BFT replication on cloud services such as Amazon Web Services. The proposed method makes it possible to select the best replica deployment for a user's purpose by exhaustively evaluating all possible replica deployments.

研究分野：分散システム

キーワード：BFTレプリケーション　パブリッククラウド　ビザンチン合意　ビザンチン故障　分散システム　分散アルゴリズム　クラウドコンピューティング

様　式　Ｃ－１９，Ｆ－１９－１，Ｚ－１９，ＣＫ－１９（共通）

１．研究開始当初の背景

インターネット技術の発展によって，オンラインバンキングやオンライントレーディングなど，重要で価値を持つサービスが提供されるようになった．しかしこれらのサービスは，攻撃者にとっても魅力的であることから，日々クラッカーからの攻撃にさらされている．これらの攻撃は分散システムの分野では**ビザンチン故障**としてモデル化される．サーバクライアントモデルで提供されるサービスにおいて，ビザンチン故障から被害を防ぐための技術として，**Byzantine Fault Tolerance** (BFT) [1]というレプリケーション手法がある．そこではオリジナルのサービスを複数のレプリカに複製し，すべてのレプリカが同じ処理を行う（図 1）．全レプリカが常に同じ状態を維持し続けることで，たとえ一部のレプリカが故障してもその影響をマスクし，安定したサービスの提供を続けることができる．

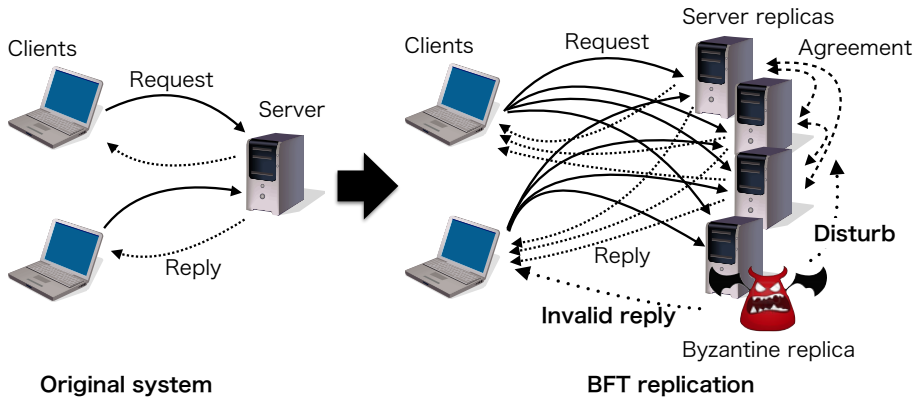

図 1 BFT レプリケーション

近年の仮想化技術の発展は，クラウドという新しい計算機の実行環境を出現させた．Amazon Web Services や Google Compute Engine などに代表されるクラウドインフラストラクチャでは，ユーザの要望に応じて自由に仮想計算機や仮想ネットワークを設置し，利用することができる．また，仮想計算機を動作させたまま他のデータセンターに移動させるマイグレーションなど，従来の物理計算機やネットワークでは行うことのできなかった，**新しい柔軟な計算機システムの運用が可能**となった．

BFT レプリケーションをクラウド上で構築することで，従来よりも高い柔軟性を持ったレプリケーションが実現可能になる．特に，**レプリカ数やネットワーク構成を攻撃負荷などに応じて柔軟に変更することは，レプリケーションの運用コストを大きく削減する効果がある**．クラウドインフラストラクチャでは仮想計算機の実行時間や通信量に応じた課金形態が主であり，過剰なレプリカ数や不要な通信は，余計な運用コストを発生するからである．しかしながら，**既存の BFT レプリケーションアルゴリズムはレプリケーション構成を実行中に変更することができない**という課題がある．

参考文献

1. M. Castro and B. Liskov. Practical byzantine fault tolerance and proactive recovery. ACM Transactions on Computer Systems, 20(4):398–461, 2002.

２．研究の目的

本研究計画では，クラウドインフラストラクチャが持つ特徴的な機能を活用するために**動的な構成変更に対応した BFT レプリケーションシステムの実現**を目標とする．本研究計画が想定する再構成可能な BFT レプリケーションシステムを，図 2 に示す．図 1 とは異なり，各レプリカはさまざまなクラウドインフラストラクチャの仮想化ホスト上で動く仮想計算機で実行される．各クラウドインフラストラクチャには Cloud Controller が存在し，クラウドインフラストラクチャ上で動作する仮想計算機を管理・制御する統一的なインタフェイスを提供する．Reconfiguration Trigger は Cloud Controller を通じてレプリカが動作する仮想計算機の状態を監視し，必要に応じて Replication Manager に再構成を要求する．Replication Manager は，Cloud Controller を通じてレプリカの管理を行う．また Reconfiguration Trigger の要求に応じて，1. レプリカの生成，2. レプリカの削除，3. レプリカの移動を Cloud Controller に指示する．

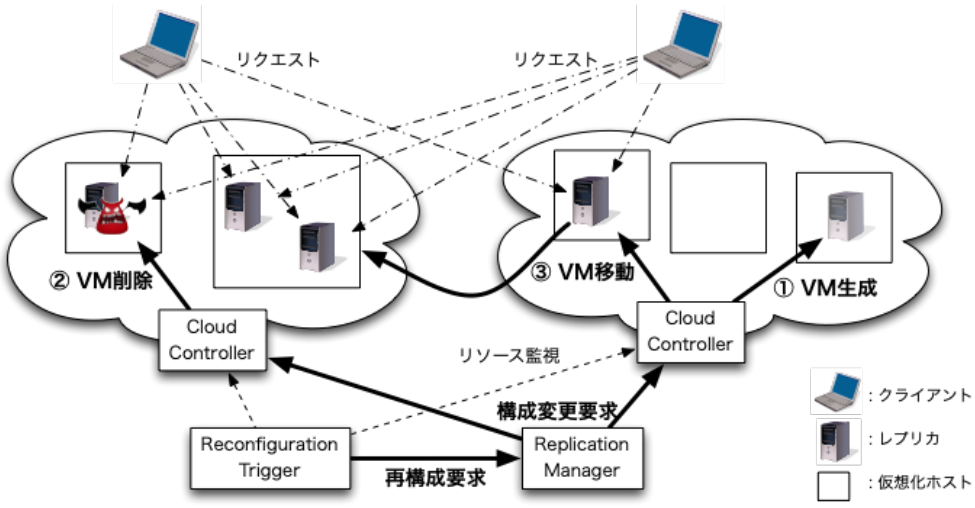

図 2 本研究計画で実現する BFT レプリケーションシステム

本研究計画では図 2 の構成要素のうち，**Replication Manager と各レプリカが実行する BFT レプリケーションアルゴリズムの実現を目標とする**．BFT レプリケーションアルゴリズムは，既存のアルゴリズムを修正し，**レプリカの追加や削除をレプリケーションを中断せずにできるように拡張する**．これにより，クラウド環境が備える特徴を活かした，柔軟な BFT レプリケーションが実現可能となる．

３．研究の方法

本研究課題は当初，次の方針で研究を進める計画だった．まず，(1) 分散アルゴリズムの問題として BFT レプリケーションの再構成を定義し，それを解くことのできるアルゴリズムを考案する．次に，(2) 設計したアルゴリズムをクラウドインフラストラクチャ上にプログラムとして実装し，既存手法との評価実験を行う．

しかし実際に研究を進める過程で次の問題が見つかった．

- そもそもパブリッククラウドを対象にして BFT レプリケーションを構築する場合には，レプリカを配置できるリージョン数が多いため，レプリカをどこに配置するべきなのかわからないこと．
- 研究計画時には判明していなかったが，当初の研究課題によく似たテーマが研究採択時に既に行われており，当初の計画通りに進めてもインパクトが弱いと予想されたこと

そのため当初の研究計画を修正し，パブリッククラウドにおいてレプリケーションを構築する際にまず必要となる，レプリカ配置の決定を容易に行うための手法の考案を目指した．

４．研究成果

本研究課題では BFT レプリケーションについて，Amazon Web Services に代表されるクラウドサービスでも柔軟にレプリケーションを実現するための手法を提案した．提案手法は，レプリケーションを構成する際に問題となるレプリカ配置について可能な全ての配置案を網羅的に評価することで，利用者の目的に最もあったレプリカ配置の選択を可能にする．BFT レプリケーションをクラウドサービスのような地理的に分散した環境で効率的に実現するための手法についてはこれまでいくつか提案されているものの，それらは既に決まっているレプリカ配置においてレプリカの役割を入れ替えることで性能を向上させるものだった．提案手法はそれらとは異なり，レプリカ配置そのものを決定する段階で利用できるという点が学術的な新規性である．提案手法は，レプリカ配置の良さを数値化する評価関数を差し替えることで，利用者の様々な目的に対応できる．評価関数の一例として，レプリケーションのレイテンシ（リクエスト送信から実行結果の受信までの時間）を最適化するケースを想定して，2 つの評価関数を設計した．これらの評価関数は Amazon Web Services 上に数千ものレプリケーションを実際に構築して網羅的なレプリカ配置の評価を行い，計測したレイテンシによって算出した最適なレプリカ配置と設計した評価関数で求めたレプリカ配置に高い相関があることを確認した．図 3 に実験結果の一部を示す．

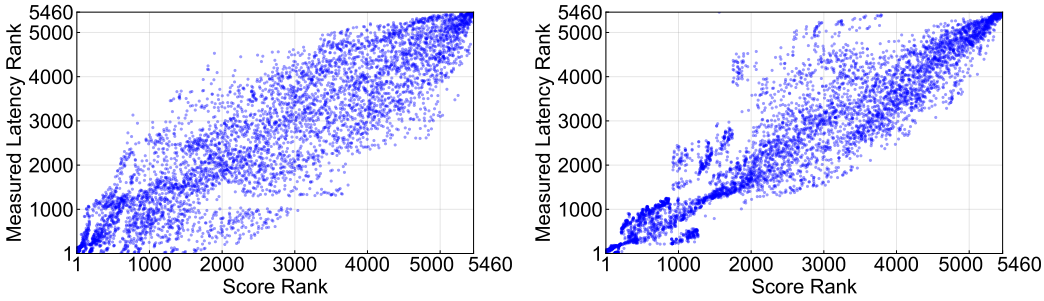

図 3 Amazon Web Services 上で実施した評価実験結果の例

その他，レプリケーションされたシステム間で送受信されたメッセージとそのときのサービス状態を協調して効率的に記録することで高い信頼性を実現するスナップショットアルゴリズムや，システム間のメッセージを効率的に配送するためのルーティングアルゴリズムを考案した．

５．主な発表論文等

〔雑誌論文〕（計 2 件）

① Yonghwan Kim, Masahiro Shibata, Yuichi Sudo, Junya Nakamura, Yoshiaki Katayama, and

Toshimitsu Masuzawa. “A Self-Stabilizing Algorithm for Constructing an ST-Reachable Directed Acyclic Graph When |S|<=2 and |T|<=2”. In Proceedings of the 39th IEEE International Conference on Distributed Computing Systems (ICDCS), (to appear), 2019. 査読あり.

② Yonghwan Kim, Junya Nakamura, Yoshiaki Katayama, and Toshimitsu Masuzawa. “A Cooperative Partial Snapshot Algorithm for Checkpoint-Rollback Recovery of Large-Scale and Dynamic Distributed Systems”. In Proceedings of the 6th International Symposium on Computing and Networking Workshops (CANDARW), pp. 285-291, 2018, doi:10.1109/CANDARW.2018.00060. 査読あり.

〔学会発表〕(計 4 件)

① 沼倉 正太, 中村 純哉, 大村 廉. RTT と SMR の通信パターンに基づく広域 SMR におけるレプリカ配置の決定手法の提案. ユビキタス・ウェアラブルワークショップ 2018 予稿集. p. 14, 2018.

② 沼倉 正太, 中村 純哉, 大村 廉. RTT と通信パターンに基づく広域 SMR におけるレプリカ配置の決定手法の提案. 第 14 回情報科学ワークショップ 予稿集, pp. 252-257, 2018.

③ 沼倉 正太, 中村 純哉, 大村 廉. パブリッククラウド上の広域 State Machine Replication のためのリージョン選択手法の提案. 信学技報, vol. 118, no. 166, DC2018-14, pp. 7-12, 2018.

④ 沼倉 正太, 中村 純哉, 大村 廉. パブリッククラウドにおける広域 State Machine Replication の特性評価. 第 80 回情報処理学会全国大会講演論文集. vol. 2018, no. 1, pp. 181-182, 2018.