大規模・動的分散システムの耐故障方式の研究（051107004）

Research on Fault Tolerance for Huge and Dynamic Distributed Systems

研究代表者

櫟肅之　NTT コミュニケーション科学基礎研究所

Tadashi Araragi　NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation

研究分担者

増澤利光†　佐藤雅彦††　五十嵐淳††

Masuzawa Toshimitsu†　Masahiko Sato††　Atsushi Igarashi††

†大阪大学　大学院情報科学研究科　††京都大学　大学院情報学研究科

†Graduate School of Information Science and Technology, Osaka University

††Graduate School of Informatics, Kyoto University

研究期間　平成 17 年度～平成 19 年度

本研究開発の概要

本プロジェクトでは、大規模・動的な分散システムに対応可能な耐故障方式の原理・方式を構築する。即ち、停止故障、ビザンチン故障、一時故障などに対し、プロセスの生成消滅を許す動的特性に対応できる耐故障の基本技術を開発する。また、これら故障の複合的故障に対しても、理論的計算量の導入により、各故障の質的関係を明らかにして複合故障モデルを整備するとともに、そこで要求される空間的、時間的仕様が耐故障アルゴリズムに与える影響を明らかにする。一方、実用的観点から、その方式を大規模・動的分散システムの標準的なミドルウェアへ実装し、有効性を確認する。さらに、ここで開発した耐故障方式の具体化アルゴリズムの正当性を正確に検証する方式を開発する。そのために、不定個のプロセスを扱う parameterized system や耐ビザンチン故障を念頭においた確率的暗号プロトコルの検証手法を発展させる。

Abstract

In this project, we research the theories and methods of fault tolerance for huge and dynamic distributed systems. We develop the fundamental technologies of fault tolerance against crash, Byzantine, and temporary failures in dynamic environments, where involved processes are dynamically created and eliminated. We also construct a theory of fault containment against compounded faults caused by the simultaneous occurrence of failures in a huge distributed system. This theory allows us to analyze the temporal and special relations between failures. Last, we introduce formal verification methods to check whether the proposed fault tolerance methods are correct. Here, an unbounded number of processes and probabilistic behavior of systems must be considered.

１．まえがき

現在、インターネット上でさまざまなサービスが開発・導入され、この動きは今後ますます高まると考えられる。一方、インターネットのような大規模でオープンなシステムでは、システムダウン（停止故障）やクラッカーの侵入による誤動作（ビザンチン故障）などの故障の可能性が非常に大きくなる。安心して快適なインターネットのサービスを利用するには、これら故障に対する厳密な対応が必要になる。本プロジェクトでは、インターネット上のサービスシステムを大規模・動的な分散システムとしてとらえ、そこでの耐故障の実用的な方式、理論の構築を目標とした。

２．研究内容及び成果

2.1 汎用的・実用的な耐故障方式の実現

代表的な故障であるシステムダウン、システム誤動作に対して、アプリケーションに依存しない耐故障方式として、大規模・動的な分散システムに対応できるロールバック方式と耐ビザンチン故障方式を考案した。

（1）大規模・動的分散システムのロールバック方式

分散システムの全体的な状態を定期的に記録保存し、システムダウンが生じたときに、各システムが最新の記録にもどって（ロールバック）サービスを回復する方法において、従来では、大規模・動的な環境に適応できなかった。本研究では、全体ではなく、関連のあるシステム間で部分的に状態を記録することで、大規模で、しかもユーザ/システムが動的に参加、離脱するオープンなシステムに対しても適用可能なロールバック方式を考案した。本方式は、同時に複数の故障の発生にも対応し、また、Web サービスの代表的ミドルウェア（Axis-JBoss）にも実装され、その方式の有用性が確認された。

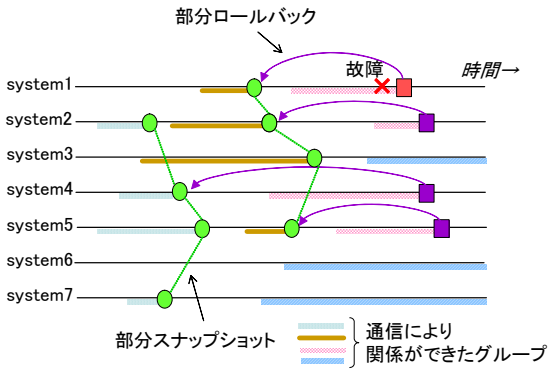

図１　大規模・動的分散システムのロールバック

（2）高速な耐ビザンチン故障方式

クラッカーの侵入などによるシステムの誤動作（ビザンチン故障）に対応する一般的な方式としてシステムレプリカ

による方式がある。これはシステムの複数のコピーを別のマシンで動かし、互いに連携して同じ動作をさせ、一つのマシンが誤動作をしてもその影響を排除するものである。このマシン間の連携ではビザンチン合意という問題を解くアルゴリズムが必要だが、従来の方式では、通信量・計算量が大きく、実用的でないと考えられていた。本方式では、リクエスト集合に対する合意アルゴリズムを新たに考案し、理論上最小のラウンド数での合意を達成し、実用的なレプリカ方式を実現した。

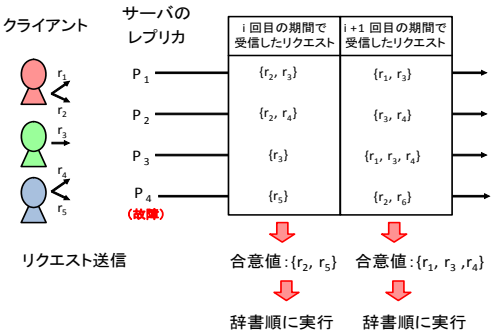

図２　レプリカ間の動作の同期

2.2 耐故障方式の性能に関する理論

大規模な分散システムでは、ある部分で障害が生じたとき、その影響範囲を最小限に抑える故障封じ込めの方式が重要になる。その封じ込め方式について、自己安定とよばれる耐一時故障のアプローチから、計算量や、定性的性能にかかわる研究を行った。

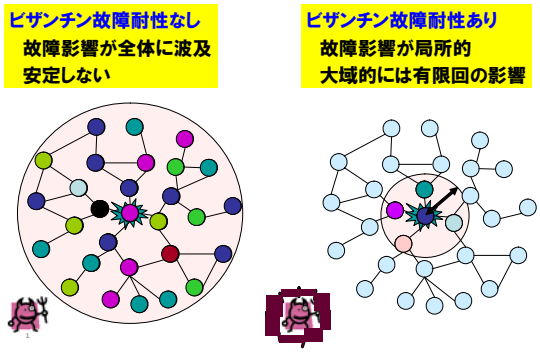

図３　故障封じ込め

（1）故障封じ込めと計算量

一般に耐故障方式の性能を測る尺度として、故障の影響範囲（空間）と故障状態から正常状態までの復帰に要するステップ数（時間）が考えられる。本研究では、まず応用上重要なリンク彩色問題、木の方向付け問題に対し、それぞれビザンチン故障の空間的封じ込め、時間的封じ込めを実現する方式を考案した。さらにこれを足がかりに、複数の故障封じ込め方式を合成して新たな故障封じ込め方式を生成する方法を考案した。また、ある特別な通信構造（リング構造）の中で実現された故障封じ込め方式を、より一般的な通信構造（木構造）上で動作する方式に変換する方法も考案した。これらの方式により、巨大で複雑なシステムに対して故障封じ込め方式を組織的に実現する一歩が築かれた。

（2）自己安定方式の安全収束

この研究では、一時故障からの復帰において、単に正常状態への復帰だけを目的にするのでなく、次善の望ましい状態（安全な状態）に迅速に移行し、その状態を保って正常状態に復帰する方法を汎用的なプロトコル（極小支配集合アルゴリズム）に対し実現した。多くのユーザを抱える大規模システムでは、このように次善の状態に速やかに移行する機能は、安全性の観点から実用面で非常に重要となる。

2.3 耐故障方式の正当性検証方式

この研究では、システムが、悪意のある侵入者から不正な操作を受けないように設計されているかを検証する方式を検討した。

（1）型理論にもとづく安全性の解析

ファイル・メモリ・ネットワークソケットなどの計算資源に対して, API によるアクセスが正しい順序で行われるかを検査するプログラム解析（資源使用解析）の研究を進めた。型システムの理論を用いて、アクセスがどの関数呼び出しの元で行われるか，といった文脈情報をもとに、不正なシステムの利用が起こらないことを効率よく検証する方法を実現した。また、Java や C#などで記述した実用システムも扱えるようにするため、利用する型システムに相互再帰的定義などの高度な表現能力を組み込み、その上での検証の正当性を証明した。

（2）セキュリティプロトコルの検証

オープンなシステムでは、セキュリティプロトコルの頑強安全性が要求される。本研究では、最も安全性レベルの強い UC 安全性を対象に、与えられたプロトコルが UC 安全であることを自動で証明する方式を検討した。UC 安全性の検証でおこなう実システムと理想システムの識別不能性の証明において、計算量仮定を用いない確率的挙動の識別不能性証明の部分を自動化し、従来、人手による安全性証明でもっとも煩雑であった部分の自動化に成功した。

３．むすび

本研究では、インターネットのような大規模・オープンな分散環境で、サイバーテロのような高度な攻撃・故障にも耐えうるロバストで効率の良い耐故障方式を実現した。また、故障の複雑度や、耐故障方式の正当性に関し、将来の実用につながる理論的成果を得た。

【誌上発表リスト】

[1] Junya Nakamura, Tadashi Araragi and Shigeru Masuyama, “Asynchronous Byzantine Request-set Agreement Algorithm for Replication”, the 1st Asian Association for Algorithms and Computation Meeting – AAAC 2008 (2008.4)

[2] Toshimitsu Masuzawa and Sebastien Tixeuil, “Bounding the impact of unbounded attacks in stabilization”, In Proc. of the 8th International Symposium on Stabilization, Safety and Security of Distributed Systems, Dallas, USA, LNCS 4280, pp. 440-453 (2006.11)

[3] Atsushi Igarashi and Mirko Viroli, “Variant path types for scalable extensibility”, In Proc. of the ACM Conference on Object-Oriented Programming, Systems, Languages, and Applications (OOPSLA 2007), Montreal, QC, pp.113-132(2007.8)

【受賞リスト】

[1] Tomoko Suzuki, Taisuke Izumi, Fukuhito Ooshita, and Toshimitsu Masuzawa, Best Paper Award, “Biologically inspired self-adaptation of mobile agent population,” (2005.8.26)

[2] 五十嵐 淳、日本 IBM 科学賞、 “オブジェクト指向言語のための型理論” (2006.11.22)