様式Ｆ－１９

科学研究費助成事業（学術研究助成基金助成金）研究成果報告書

平成 25 年 6 月 7 日現在

機関番号：１２６０５
研究種目：若手研究（B）
研究期間：2011～2012
課題番号：23700040
研究課題名（和文）OSカーネル障害からの早期回復手法に関する研究

研究課題名（英文）Fast Recovery from Operating System Kernel Failures

研究代表者
山田 浩史（YAMADA HIROSHI）
東京農工大学・大学院工学研究院･准教授
研究者番号：00571572

研究成果の概要（和文）：
本研究の目的はオペレーティングシステム(OS)カーネルのバグによる障害の影響を緩和して，その上で動作するサービスの可用性低下を抑えることにある．従来研究から得た知見を元に，カーネルダウンからの高速回復手法の確立を狙う．本手法を確立することで，バグによる障害が生じた際のサービスの可用性低下を抑えたり，代替機設置の管理・電力コストの軽減といったメリットが期待できる．本研究において，OS カーネルの障害から高速に復旧する機構を設計し，実際に実験を行って提案方式を定量的に評価した．その結果，復旧のためのダウンタイムを最大で約 90%以上削減できることがわかった．加えて，Solid-State Drive(SSD)を用いることでさらなる高速化が見込めること，ならびに本方式を拡張することでソフトウェアアップデートに伴うダウンタイムをも削減できること，Graphic Processing Unit(GPU)を用いることで仮想マシン移送の高速化が見込めるという知見を得た．

研究成果の概要（英文）：
The goal of this particular research is to improve availability of computer systems under operating system kernel failures. This research focuses on a software mechanism of quickly recovering from kernel failures. This approach offers several benefits such as improving service availability, reducing maintenance costs of physical hosts, reducing power-consumption. The research outcomes are as follows; 1.) A mechanism to quickly recover from kernel failures are designed and implemented. 2.) The mechanism reduces downtime for a conventional OS recovery method by at most 90%. 3.) Solid-State Drives can accelerate execution of the proposed mechanism, an extension enables the mechanism to shorten downtime caused by software update, leveraging Graphic Processing Units(GPUs) accelerates virtual machine live migration, which move a virtual machine to another physical host without any data loss, due to their parallel processing feature.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
交付決定額	3,000,000	900,000	3,900,000

研究分野： システムソフトウェア
科研費の分科・細目： 情報学･ソフトウェア
キーワード： オペレーティングシステム，ディペンダビリティ，システムソフトウェア

１．研究開始当初の背景

オペレーティングシステム(OS)カーネルの大規模化・複雑化に伴って，アプリケーションのみならず OS カーネル内のバグがシステム全体の可用性を脅かすようになってきた．OS カーネルがダウンすると，その上で稼働するアプリケーションすべてがダウンすることを意味し，ユーザに提供するサービスの可用性を著しく低下させてしまう．特に，ショッピングサイトや SLA(Service Level Agreement)を保証するサイトにおいては，サービスの停止による被害は深刻となる．テスト段階で OS カーネルからバグを根絶することは困難であり，たとえば，Linux のソースコードは数百万行からなっており，毎日のようにバグ修正レポートが報告されている．テストをかいくぐるバグはハイゼンバグやコンカレンシバグと呼ばれ，再現性が低く，修正が非常に困難であり時間を要する．そのため，これらのバグによる OS カーネルのダウンを想定した運用が求められ，OS カーネルがダウンした場合，いち早くサービスを再開することが重要となる．

従来取り組んだ研究において， OS カーネルのバグによる障害から高速に回復する手法を提案したが，適用環境が限定的であり，未だ大きな効果を得るまでには至っていない．従来手法は，テストをかいくぐるバグはメモリの状態のみを破壊することが多い点に着目して，メモリの内容を高速に健全な状態にする手法である．具体的には，仮想マシン技術のスナップショット技術を応用している．本手法を用いると，カーネルバグが生じてから復旧するまでの時間を，システムの再起動時間と比べた場合，限定的ながら最大で 84 % 削減することに成功した．また，従来研究から，1) 仮想マシンにメモリを大量に割り当てると復旧に時間がかかる，2) デーモンプログラムの使用するファイルの名前が変更されると，正しくサービスを再開することができないといった問題点が浮き彫りとなっている．これらの問題点を解決することで，OS カーネルのバグから高速に復旧するための有用な基礎技術として期待できる．

２．研究の目的

本研究の目的は，OS カーネルのバグによる障害を緩和して，その上で動作するサービスの可用性低下を抑制することにある．従来研究から得た知見を元に，カーネルダウンから高速に回復するための手法の確立を狙う．本研究の手法を確立することで，1) バグによる障害が生じた際のサービスの可用性低下を抑制できる，2) 障害復旧の高速化による代替機設置の電力コストの軽減，3) サービスの可用性低下抑制による修正パッチの入念なテストといった多くのメリットが期待できる．

本研究では上述の問題点を解決する 2 つの機構を設計・実装して，従来機構に組み込み，徹底した実験を行う．1 つめは，スナップショットを復元する際に，仮想マシンが使用しているメモリ量を少なくして使用している領域のみを復元する機構である．2 つめは，運用時におけるデーモンプログラムが使用するファイルの名前変更を監視し，スナップショットへ安全に戻れるか否かを判断する機構である．これらと従来機構を組み合わせることで，OS カーネルがダウンした際にも，サービスをより速くより安全に再開することができる．

仮想マシンのメモリ削減機構，ならびにファイル名変更監視機構を設計，オープンソースの仮想マシンモニタである Xen 上に実装する．そして，高性能なサーバ計算機を用いて，インターネット環境を想定した実験を行う．Linux や JBoss といった実システムを稼働させて，多角的かつ精密な有効性の検証を行い，さらなる知見の取得を狙う．

３．研究の方法

初年度には，仮想マシンのメモリ使用量削減機構，ならびにファイル名変更監視機構の設計，および実装を行った．設計する段階で，アプリケーションや OS カーネル，仮想マシンモニタに施す変更部分を明確にした．OS カーネルに変更を施す際には，作業効率を良くするために，仮想マシン環境を構築して実装を行った．仮想マシンモニタの開発環境は現状では十分に整備されていないため，仮想マシンモニタの改変には多くの時間を割いた．また，実際に広く使用されているサーバプログラムである Apache や実際に広く利用されている OS カーネルである Linux，仮想マシンモニタ Xen を用いた実験を次年度に予定していたため，本年度は実装を完了するまでにとどめ，入念に動作検証を行った．そのために，クライアント・サーバプログラムを動作させるための計算機を購入した．ここで，時間をかけて動作検証を入念に行っておくことで，翌年度に行う実験時において，異常な挙動が生じた際の原因の切り分けを迅速に行えると考えている．また，設計段階において Solid State Drives(SSD)にスナップショットを保存することで，さらなる高速化を得られるという感触が得られたため，SSD を購入した．SSD はその構成によって性能が大きく変わるため，異なる構成をもつ SSD を複数購入した．

最終年度には，プロトタイプの評価実験を行った．Apache や Linux といった実際に利

用されているソフトウェアを利用することで，既存のソフトウェアとの親和性を確認し，これらを用いて前年度に実装を行ったプロトタイプの評価実験を行った．当初実験用計算機を購入予定であったが，前年度に購入した計算機を使用することで十分な実験を行えることがわかったため，計算機を使いまして評価実験を行った．OS カーネルの障害を再現するために，フォールトインジェクションと呼ばれる方法を用いて意図的に OS カーネルにフォールトを挿入し，わざと障害を発生させた．この際，単純な動作をするアプリケーションを稼働させたときには，OS カーネルを復旧する際に生じるダウンタイムを最大で 90%削減した．Apache を稼働させたときにもほぼ同等の効果があることも確認できた．実験を実施していく上で，スナップショットの取得と復元を並列実行することでより高速にスナップショットを取得可能になるという感触を得た．そこで，実験計算機を購入する予定であった費用を高い並列性を持つ Graphic Processing Unit(GPU)を備えた計算機を購入した．この実験用計算機を使って，基礎的な実験を開始し，本課題で得た知見を活かし研究テーマの発展させた．

４．研究成果

本研究を通じて，OS 障害が生じた際に高速に復旧する機構の構成およびその効果を明らかにした．具体的には，仮想マシンのメモリ使用量削減機構，ならびにファイル名変更監視機構である．また，両機構を Linux や Xen に実装を行い定量的な評価を行った．実験を行ったところ，カーネル復旧に要する時間を最大で約 90%削減できることがわかった．また，Apache やオークションサイトを模したベンチマークソフトウェアである RUBiS に対しても適用可能であることがわかった．さらに，フォールトインジェクションを用いてカーネル障害を意図的に発生さた実験では，提案方式がすべての障害から回復できることがわかった．

また，Solid State Drive(SSD)を用いることでさらなる高速化が見込めることがわかった．スナップショットをディスク装置に保存するため，ディスク自身が高速化すれば必然的にスナップショット復元も高速化する．本方式はハードディスクに特化した方式では無いため，SSD にも適用可能である．SSD は爆発的に普及し始めているため，SSD の使用は現実的であると思われる．

本方式を拡張することで，ソフトウェアのアップデートに伴うダウンタイムも短縮できることがわかった．スナップショットの復元を応用するアイディアを基礎に，再起動後のスナップショットを作成する．具体的には，現在サービスを提供している仮想マシンとは別に，同じ状態を持つ仮想マシンを動的に生成する．生成した仮想マシン上で再起動を行い，再起動が完了したらスナップショットを取得し，その仮想マシンを破棄する．取得したスナップショットを復元することで，再起動後の状態を作り出すことができる．本知見は計画当初は想定しておらず，発展的に本課題の研究テーマを拡張することができた．今後は詳細に設計を行い，具体的にシステムを実装していく予定である．

またスナップショット操作について考察していく内に，Graphic Processing Unit (GPU) を用いることで仮想マシン移送を高速化できる見込みがあることがわかった．仮想マシン移送とは仮想マシンを稼働させたまま別の物理ホストに移すことを指す．仮想マシン移送では，メモリや CPU のレジスタ値といった仮想マシンの状態を移送先の物理ホストにコピーを行う．いわば別ホストにスナップショットを作成する．この作業を GPU によって並列処理することで仮想マシン移送に要する時間を飛躍的に小さくすることができると考えられる．仮想マシン移送の処理を GPU 上で実行する機構の詳細設計を開始し，こちらの機構も実装していく予定である．

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 3 件）

[1] Takeshi Yoshimura, Hiroshi Yamada, Kenji Kono, Using Fault Injection to Analyze the Scope of Error Propagation in Linux, IPSJ Trans. on Advanced Computing Systems, Vol.6, No.2, pp.1-10, 査読有, 2013

[2] Hiroshi Yamada and Kenji Kono: Traveling Forward in Time to Newer Operating Systems using ShadowReboot, Proc. of the 9th ACM Conf. on Virtual Execution Environments (VEE'13), pp.121-130, 査読有, 2013

[3] Kazuya Yamakita, Hiroshi Yamada, Kenji Kono: Lightweight Recovery from Kernel Failures using Phase-based Reboot, IPSJ Trans. on Advanced Computing Systems, Vol.5, No.2, pp.121-132, 査読有, 2012

〔学会発表〕（計 1 件）

[1]吉村剛，山田浩史，河野健二：Linux カーネルにおけるエラー伝搬，第 121 回システムソフトウェアとオペレーティング・システム研究会，2012.5.7～5.8、那覇市 IT 創造館

６．研究組織

(1)研究代表者

山田浩史（YAMADA HIROSHI）
東京農工大学・大学院工学研究院・准教授
研究者番号：00571572