# 審 査 の 結 果 の 要 旨

氏　　名　　バリ　ゲローフィ

インターネット上のサーバは仮想計算機上で実現されるようになった。これらサーバ群の可用性を向上させるために仮想計算機上のシステムの複製を取る手法が取られている。しかし、従来の手法では複製のためのデータ転送容量が大きく高性能ネットワークを必要としている、クライアントとTCP接続しているサーバの通信状態を効率よく複製する手法が確立していない、という問題がある。本論文では、これら問題点を解決する手法を提案、実装、評価し、その有効性を示しおり、以下のとおり８章から構成される。

第１章では、背景として従来の可用性向上技術を俯瞰し、本論文で取り組む仮想計算機を用いた可用性向上技術を明確に位置づけしている。仮想計算機上のシステムの可用性を上げるために定期的にシステムの複製（チェックポイント）を行う。システムチェックポイント時に解決されていない問題点として、i)チェックポイント時のデータ転送容量削減、ii）チェックポイントタイミング、iii)外部とTCP接続している時のTCP状態の一貫性維持があり、本論文はこれら3つの課題に対する新しい手法を提案している。

第2章は、本論文が対象とする仮想計算機の技術的背景とチェックポイント手法の基本について述べている。第３章では、チェックポイント対象であるシステムメモリ上のメモリ領域類似性を利用したデータ転送容量削減手法を提案している。メモリ領域類似性を示すために以下の実験を行っている。複数の実アプリケーションベンチマークプログラムを10分間連続して実行する。この間に100msec間隔での定期的チェックポイントにおいて直前のチェックポイントから書き変わったメモリ領域を比較すると、2Kバイトのメモリ領域で77%程度、64倍のメモリ領域で85%程度の類似性を有するという事実を示す。この事実をもとに、チェックポイント時にメモリ領域の類似性を調べるための機構として密度ハッシュ関数(Density based hash function)を用いた連想ハッシュテーブル(Content Addressable Hash Table)と類似性を持つメモリとの差分のみを転送することによりデータ転送容量を削減する手法を提案している。本提案手法の有効性を示すために、Linuxカーネルコンパイル、SPECweb、Microsoft Exchange Serverベンチマークを用いて通常チェックポイント方式と比較している。これらベンチマークの実行結果、最大で通常のチェックポイントデータ転送容量の20％を削減していることが示されている。また、Microsoft Exchange Serverベンチマークの実行結果として通常チェックポイント方式に比べて約1.8倍の性能向上を得たことを示している。

第４章では、高速チェックポイントデータ転送手法として、Infinibandネットワークが有する遠隔直接メモリアクセス(RDMA: Remote Direct Memory Access)機構を書き込み時コピー(Copy on Write)手法と組み合わせた手法を提案実装し評価している。RDMA機構を用いることによりデータ転送時のCPU介在コストが削減され、Linuxカーネルコンパイルでは高々34%の性能減でチェックポイント可能となったことを示している。

第５章では、サーバの負荷状況に応じたチェックポイントタイミング調整手法を提案している。チェックポイント後に変更されるメモリ領域はアプリケーションレベルでのメモリ変更以外にネットワークバッファ、ディスクI/Oバッファなどのカーネル内バッファのメモリ変更がある。本章では、これらメモリの変更度合いを動的に調べ、チェックポイントタイミングを決めるという新しい手法を提案している。Hadoop MR/DBベンチマークによる評価で、本手法を用いることによる5%の性能劣化だけでチェックポイント可能であることを示している。

第６章では、サーバがダウンしチェックポイント時の状態を用いて再開する場合においてもクライアントとのTCP接続状態を保つために用いられている仮想計算機内通信バッファを用いてTCP通信の高速化手法を提案している。仮想計算機上のシステムと本バッファ領域をチェックポイントし、その後バッファに格納されているパケットをクライアントに送信する。これにより、仮想計算機がダウンしてもチェックポイント時点から再開を行い、クライアントに対しては格納された通信パケットを用いてTCP通信を再開できる。さらに提案手法では、クライアントからのACKを待たずに仮想計算機内でACKを送信するため、見かけのRTT(往復遅延)が短くなり高速化に寄与する。クライアントからのACKを待たずにTCP通信経路の途中でACKを返す手法はTCP通信の高速化技法のひとつとして知られている。本論文では、TCP通信状態とともに一貫性のあるチェックポイントを実現するための技法と組み合わせて仮想計算機に実現しており、従来の研究とは異なった視点での応用であるといえる。本手法の有効性を示すため、SPECwebベンチマークを用いて比較した。本手法によりチェックポイントしないときの実行性能の90%の性能を達成したことが示されている。

第７章において、それぞれの提案手法に関連する既存研究を紹介し、提案手法の違いならびに新規性について議論し、第８章において本論文をまとめている。

このように本論文では、仮想計算機の複製を効率よく取得する手法としてメモリ領域の同一性に着目してデータ転送量を削減する手法、仮想計算機の負荷状況に依存した複製タイミング手法の2つを提案しその有効性を多くの実際に使われているアプリケーションベンチマークプログラムを用いて評価することによって示した。さらに、サーバが外部とTCP接続して通信している最中にサーバがダウンしても再実行可能とするための通信層を提案した。これら本研究の成果は、仮想計算機の高可用化に多大な貢献を行っており、仮想計算機ソフトウェアの発展に顕著な貢献をしたといえる。

よって本論文は博士（情報理工学）の学位請求論文として合格と認められる。