早稲田大学大学院理工学研究科

博 士 論 文 概 要

論 文 題 目

エンタープライズサーバにおける
ファイルシステムとインターコネク
ションネットワークに関する研究

Studies on File Systems and
Interconnection Networks for
Enterprise Servers

申 請 者

氏 名		
	保田（藤井）	淑子
	YASUDA (FUJII)	YOSHIKO

専攻・研究指導
(課程内のみ)

２００４年 １２月

エンタープライズサーバは、企業における業務システムの利便性と性能を向上するために使用される。たとえば、クラスタ型エンタープライズサーバは、複数の要素サーバをEthernetのようなローカルエリアネットワークで接続した構成をとる。一般に、各要素サーバはそれぞれ独立したハードウェアリソースを持ち、その上でOS（Operating System）が稼動し、ディスクはファイルシステムにより管理される。そのため、システムの大規模化および複雑化に伴う管理コストの増大や使い勝手の悪化が問題となっており、特にNAS（Network-Attached Storage）のようなストレージデバイスでは、台数増加に伴う管理コストの増大が深刻な問題となっている。また、マルチプロセッサ型エンタープライズサーバは、各要素プロセッサをインターコネクションネットワークで接続し、メッセージパシング等の通信プロトコルを用いて要素プロセッサ間でデータをやり取りするため、通信オーバヘッドが大きいことが問題となっている。

従来から複数NASをローカルエリアネットワークで接続し仮想的に1つのNASに見せるクラスタ型NASシステムが提案されているが、従来技術では特定のOSのみに対応するような限定された使用方法を想定していた、あるいはシステムの構成変更に伴い無効になってしまうファイルの識別子を再有効化させるための機能をクライアント計算機に追加しなければならなかった等、クラスタ化によってNASのメリットである使い勝手のよさや管理容易性が損なわれてしまうという問題があった。また、クラスタ型NASシステムでは、複数のユーザ間で共有される場合にユーザごとに使用可能なディスク容量を保証する必要があるが、ユーザのファイルが複数のNASに分散して配置されるため、ユーザごとにディスクの使用量を制限することは困難であった。さらに、クラスタ型NASシステムでは、それを構成する要素NASに障害が発生すると、そのNAS上の最新のファイルが消失してしまうという問題があった。また、従来からマルチプロセッサ型エンタープライズサーバでは、インターコネクションネットワークに関する研究が盛んに行われてきた。しかしながら、従来研究では、数値応用プログラムであってもユーザがトポロジを意識したデータマッピングを行わない場合や非数値応用プログラムのようにトポロジを意識したマッピングができないアプリケーションを実行する場合、インターコネクションネットワーク上でのデータ転送オーバヘッドの最小化が非常に困難であった。また、ブロードキャスト通信を実現するために、別系の専用ネットワークを装備する、あるいはソフトウェアによりブロードキャストアルゴリズムを記述していたが、高コストあるいは大きなオーバヘッドが問題となっていた。

本論文では、これらの問題点を解決するために、クラスタ型エンタープライズサーバにおける管理容易で使い勝手もよくかつ高信頼なファイルシステムの仮想化技術と、マルチプロセッサ型エンタープライズサーバにおける低オーバヘッドのインターコネクションネットワーク技術について論ずる。

本論文は、7章から構成される。以下に各章の概要を示す。

第1章「緒言」では、本研究の背景と目的について述べる。本章では、企業の業務システムにおけるエンタープライズサーバの位置付けについて述べるとともに、エンタープライズサーバの課題とその課題に対する従来研究の取り組み、および本研究の意義について述べる。

第2章「仮想一元化NASシステムX-NASのコンセプト」では、主にオフィスや企業部門で使

用される安価なNAS(以下エントリNASと呼ぶ)の管理容易性を維持しつつ、ディスク容量を簡単に拡張可能な仮想一元化NASシステムX-NASについて述べる。X-NASの核である仮想一元化ファイルシステムMVFSは、異種OSの稼動するクライアントに対して単一のファイルシステムビューを提供する。また、X-NASの統合管理機能により、管理者が従来の単体NASを管理するのと同じように、複数NASを簡単に管理できる。開発した様々な統合管理機能のうち自律リバランス技術は、クラスタ型NASシステムにおいて自然発生する複数ディスク間の残容量の偏りや、構成変更時に発生するディスク残容量の偏りに起因する問題を解決できる。NFSv2ベースのX-NASプロトタイプを用いて、UNIX系ファイルサーバ標準ベンチマークSPECsfs97の性能を評価した結果、8台構成のX-NASが単体のNFSサーバに比べて応答時間を10%短縮し、スループットを25%向上できることを確認した。また、自律リバランス技術の性能を評価し、ディスクのクラスタ化により自然発生するディスク残容量の偏りによる性能低下を57%、容量拡張時に発生するディスク残容量の偏りによる性能低下を200%改善できることを確認した。その結果、提案した仮想一元化ファイルシステムにより単体エントリNASの管理容易性と使い勝手のよさを有し、単体エントリNASの性能を維持できるクラスタ型NASシステムを実現できることが示された。

第3章「ディスク使用量制限機能の提案と評価」では、クラスタ型NASシステムの使い勝手と管理容易性を向上することを目的としたディスク使用量制限技術について述べる。本章では、クラスタ型NASシステムにおいてユーザ毎のディスク使用量を制限する機能XQUOTAを提案した。XQUOTAでは、容量スケーラビリティと実装容易性、耐故障性を重視し、ユーザ毎のディスク使用量情報をクラスタ型NASシステムの各要素NASで分散して管理する分散管理方式を採用した。各要素計算機で管理するユーザ毎のディスク使用量をネットワーク経由で取得するオーバヘッドを削減するため、NFSプロシージャの種類とローカルファイルシステムにおける実際のデータアクセスサイズを元にディスク使用量を概算し、見積り値を正確なディスク使用量の代わりに用いる見積りベースQuota方式を提案した。本方式の有効性を検証するため、XQUOTAをクラスタ型NASシステムX-NASのプロトタイプに実装し、評価実験を行った結果、XQUOTAがオーバヘッド削減目標10%に対して、3%以下に抑制可能であり、クラスタ型NASシステムに適したディスク使用量制限機能を実現できることが示された。

第4章「同期バックアップ機能の提案と評価」では、クラスタ型NASシステムの信頼性向上に向けた同期バックアップ技術について述べる。本章では、汎用のファイルアクセスプロトコルであるNFSを利用して、ユーザがファイルを作成及び更新するのに同期してファイルブロック単位でバックアップNAS上のファイルも作成及び更新可能なNASの同期バックアップを提案した。さらに、クラスタ型NASシステムとバックアップNAS上のデータ一致保証に伴う処理オーバヘッドを低減するため、両システム上のファイルの識別子の対応関係をメモリ上に記録するレプリケーションキャッシュ、クラスタ型NASシステムとバックアップNASへのアクセスの並列実行を可能にするレプリケーションスレッド、並列実行処理効率をより向上させる部分非同期処理機能を開発した。同期バックアップ機能の有効性を検証するため、本機能をNFSv3ベ

ースのX-NASプロトタイプに搭載しファイルサーバプログラムNetBenchおよびSPECsfs97、バックアップ用途を想定したファイルコピープログラムを用いて性能評価を行った。評価の結果、上記3つのオーバヘッド低減機能の効果により、ファイルサーバプログラムでは、同期バックアップなしX-NASの約80%の性能を達成できることを確認した。また、同期バックアップ性能の最悪ケースと考えられるファイルコピープログラムでは、同期バックアップなしX-NASの約70%の性能を達成できることを確認した。本機能の導入により、エントリNASユーザに対して、同期バックアップなしX-NASの70%から80%の性能を維持しつつ、バックアップNASの導入コストを低減して信頼性を向上できるクラスタ型NASシステムを提供できることが示された。

第5章「インターコネクションネットワークの通信性能評価」ではマルチプロセッサ型エンタープライズサーバにおけるインターコネクションネットワークの通信性能評価について述べる。本章では、代表的なインターコネクションネットワークである多段クロスバネットワーク、トーラスネットワークおよびハイパキューブネットワークにおいて、等方転送およびランダム転送を行った場合の通信性能をソフトウェアシミュレーションより定量的に評価した。等方転送は数値応用プログラムにおいてトポロジを意識せずにプロセッサ台数のみを意識してマッピングを行った場合に出現し、ランダム転送は転送先が動的に変化し予測できないようなアプリケーションで出現する。評価の結果、等方転送の場合、多段クロスバネットワークがトーラスおよびハイパキューブに比べて1/7から1/2の時間で通信できることを明らかにした。また、ランダム転送の場合、多段クロスバが他ネットワークトポロジに比べて、同等あるいは2/3の時間で転送できることを確認した。これらの結果から、マルチプロセッサ型エンタープライズサーバ上で等方転送とランダム転送を多く必要とするアプリケーションにおいてはインターコネクションネットワークトポロジとして多段クロスバネットワークが有効であることが示された。

第6章「多段クロスバネットワークにおける高速ブロードキャスト方式の提案」では、第5章で有効性が示された多段クロスバネットワークにおける高速ブロードキャスト技術について述べる。本章では、初期データやプログラムあるいはプログラムにおける演算の途中結果を複数の要素プロセッサに転送する場合に必須となる1対全通信（ブロードキャスト）を、多段クロスバネットワークで高速に実現する高速ブロードキャスト機構「シリアライズクロスバ方式」を提案した。従来、ブロードキャストは専用ネットワークにより実現するか、あるいはユーザがトポロジを意識してソフトウェアによりブロードキャストアルゴリズムを記述していた。一方、シリアライズクロスバ方式は、多段クロスバネットワークの特定のクロスバスイッチにおいて複数のブロードキャストを逐次制御することにより、専用ネットワークを装備することなく、デッドロックを回避しつつ、高速なブロードキャストを実現できる。このシリアライズクロスバ方式をマルチプロセッサ型エンタープライズサーバ日立SR2000シリーズに実装し、性能を評価した。評価の結果、高速ブロードキャスト機構を使用した場合、ソフトウェアブロードキャストに比べて2倍の性能を得ることができ、かつ、ハードウェアの限界性能の95%以上を達成できることを確認した。その結果、提案した高速ブロードキャスト技術の有効性が示された。

第7章「結言」では、本研究で得られた成果について述べる。

研 究 業 績

種 類 別	題名, 発表・発行掲載誌名, 発表・発行年月, 連名者（申請者含む）
論文 ○	保田 淑子, 川本 真一, 江端 淳, 沖津 潤, 樋口 達雄, “仮想一元化 NAS システム X-NAS の同期バックアップ機能の実現と評価”, 情報処理学会論文誌 コンピューティングシステム(ACS9), (掲載決定)
○	Y.Yasuda, S.Kawamoto, A.Ebata, J.Okitsu, T.Higuchi, N.Hamanaka, “RX-NAS: a Scalable Reliable Clustered NAS System”, 情報処理学会論文誌 Vol. 46 No. 1, 2005 年 1 月
○	保田 淑子, 沖津 潤, 川本 真一, 江端 淳, 樋口 達雄, “クラスタ型 NAS システム向きディスク使用量制限機能の提案と評価”, 情報処理学会論文誌 コンピューティングシステム Vol. 45 No. SIG11, pp.24-35, 2004 年 10 月
○	Y.Yasuda, S.Kawamoto, A.Ebata, J.Okitsu, T.Higuchi, “An Online Backup Function for a Clustered NAS System (X-NAS)”, Proceedings of 12th NASA / 21st IEEE Conference on Mass Storage System and Technologies, pp.165-170, April 2004
○	Y.Yasuda, S.Kawamoto, A.Ebata, J.Okitsu, T.Higuchi, N.Hamanaka, “Scalability of X-NAS: a Clustered NAS System”, 情報処理学会論文誌 コンピューティングシステム Vol. 44 No. SIG11, pp.68-78, 2003 年 8 月
○	Y.Yasuda, S.Kawamoto, A.Ebata, J.Okitsu, T.Higuchi, “Concept and Evaluation of X-NAS: a Highly Scalable NAS System”, Proceedings of 11th NASA / 20th IEEE Conference on Mass Storage System and Technologies, pp.216-224, April 2003
○	Y.Yasuda, H.Fujii, H.Akashi, Y.Inagami, T.Tanaka, J.Nakagoshi, H.Wada, T.Sumimoto, “Deadlock-free Fault-tolerant Routing in the Multi-dimensional Crossbar Network and Its Implementation for the Hitachi SR2201”, Proceedings of 11th IEEE International Parallel Processing Symposium, pp. 346-352, April 1997 H.Fujii, Y.Yasuda, H.Akashi, Y.Inagami, M.Koga, O.Ishihara, M.Kashiyama, H.Wada, T.Sumimoto, “Architecture and Performance of the Hitachi SR2201 Massively Parallel Processor System”, Proceedings of 11th IEEE International Parallel Processing Symposium, pp.233-241, April 1997

研 究 業 績

種 類 別	題名, 発表・発行掲載誌名, 発表・発行年月, 連名者（申請者含む）
講演 (査読付シンポジウム)	保田 淑子, 川本 真一, 江端 淳, 沖津 潤, 樋口 達雄, “仮想一元化 NAS システム X-NAS の同期バックアップ機能の実現と評価”, コンピュータシステム・シンポジウム 2004, pp.13-22, 2004 年 11 月 保田 淑子, 沖津 潤, 川本 真一, 江端 淳, 樋口 達雄, “クラスタ型 NAS システム向きディスク使用量制限機能の提案と評価”, 先進的計算基盤システムシンポジウム SACSIS2004, pp.325-334, 2004 年 5 月 Y.Yasuda, S.Kawamoto, A.Ebata, J.Okitsu, T.Higuchi, N.Hamanaka, “Scalability of X-NAS: a Clustered NAS System”, 先進的計算基盤システムシンポジウム SACSIS2003, pp.259-266, 2003 年 5 月 川本 真一, 保田 淑子, 江端 淳, 沖津 潤, 樋口 達雄, 濱中 直樹, “クラスタ型 NAS システム X-NAS における容量拡張・縮退機能の実現”, 先進的計算基盤システムシンポジウム SACSIS2003, pp.157-158, 2003 年 5 月（ポスター発表）
講演 (研究会)	保田 淑子, 川本 真一, 江端 淳, 沖津 潤, 樋口 達雄, “仮想一元化 NAS システム X-NAS における同期バックアップ機能の実現と評価”, 並列/分散/協調処理に関するサマー・ワークショップ SWoPP 2003, 2003 年 8 月 川本 真一, 保田 淑子, 江端 淳, 沖津 潤, 樋口 達雄, “仮想一元化 NAS システム X-NAS における自律容量リバランス機能の実現と評価”, 並列/分散/協調処理に関するサマー・ワークショップ SWoPP 2003, 2003 年 8 月 川本 真一, 江端 淳, 沖津 潤, 保田 淑子, 樋口 達雄, “ファイル自律配置方式を備えた仮想一元化 NAS システム X-NAS の実現と評価”, 第 14 回データ工学ワークショップ 4-B-01, 2003 年 3 月 保田 淑子, 川本 真一, 濱中 直樹, “Java アプリケーションサーバの特性解析”, 並列/分散/協調処理に関するサマー・ワークショップ SWoPP 2002, 2002 年 8 月 藤井 啓明, 保田 淑子, 明石 英也, 稲上 泰弘, 柏山 正守, 和田 英夫, 住本 勉, 河辺 峻, “並列計算機 SR2201 の方式と評価”, 並列/分散/協調処理に関するサマー・ワークショップ SWoPP 96, 1996 年 8 月 保田 淑子, 田中 輝雄, 稲上 泰弘, “スーパースカラ方式とベクトル処理方式の比較―主記憶アクセス特性に着目して―”, 並列/分散/協調処理に関するサマー・ワークショップ SWoPP 95, 1995 年 8 月 保田 淑子, 藤井 啓明, 田中 輝雄, 稲上 泰弘, “ハイパクロスバネットワークの通信性能評価”, 並列/分散/協調処理に関するサマー・ワークショップ SWoPP 93, 1993 年 8 月

研 究 業 績

種 類 別	題名, 発表・発行掲載誌名, 発表・発行年月, 連名者（申請者含む）
講演（FIT）	江端 淳, 川本 真一, 保田 淑子, 沖津 潤, 樋口 達雄, 濱中 直樹, “仮想一元化 NAS システム X-NAS の同期バックアップ実現に向けた順序制御方式の検討”, FIT 2003, 2003 年 9 月
講演（全国大会）	保田 淑子, 樋口 達雄, 濱中 直樹, “大容量サイクルベースバストレーサを用いた Web コンピューティングアプリケーションの特性解析”, 情報処理学会第 64 回全国大会, pp.1-103-104, 2002 年 3 月