平成 19 年度
実績報告

「情報社会を支える新しい高性能情報処理技術」
平成 15 年度採択研究代表者

横田　治夫

東京工業大学・教授

ディペンダブルで高性能な先進ストレージシステム

1. 研究実施の概要

本研究は、新しい情報化社会に求められるディペンダブルで高性能な先進ストレージシステムを構築するための基本的技術の確立を目的として、平成 15 年度にスタートした。そのためのアプローチとして、並列分散ストレージシステムの構成要素であるストレージ装置にインテリジェンスを持たせ、ストレージシステム全体の性能向上と信頼性向上のための管理コストを削減する方法を検討してきた。さらに、そのようなストレージに格納するコンテンツとその取り扱い方法も含め、情報社会におけるストレージシステムの新たな位置づけを目指した機能に関しても検討を行っている。

平成 19 年度は、実用化にむけて、これまで提案してきた手法をさらに発展させることを主眼に研究を進めてきた。自律的なストレージ管理に関しては、実用化の際に問題となる、管理処理によるサービスの品質低下を防ぐための手法や、容量・性能が異なるストレージ装置が混在する環境で性能を引き出すための手法に関して提案・評価を行った。また、IP ベースのストレージネットワークにおいて VPN ルータの機能を有効利用して複数コネクションを作り、性能とディペンダビリティを向上させる手法の提案と評価を行った。ストレージ内に格納するコンテンツの取り扱いに関しても、XML で記述されたデータの検索を行う XQuery を分散ストレージ装置中の環境で効率よく行う方法と分散格納方法について提案と評価を行った。また、自律ディスクプロトタイプの性能向上のための管理ソフトウェアの改善と一般に広く普及している Windows 環境クライアントからのアクセス性向上のためのインタフェース機能の開発・実装および性能評価を行った。プロトタイプシステムについては、本年度の領域シンポジウムおよび情報処理学会全国大会併設イベント「わくわく IT」においてデモンストレーションを行った。

これらの成果に関しては、国内外の学会で発表、さらに論文誌に投稿を行った。また、チームで主催する国際ワークショップである ADSS2007 を IEEE ICDIM2007 の併設ワークショップとして開催し、多数の参加者のもと、研究成果の発表とともに、有用な意見交換を行うことができた。

今後も、実用化を目標に、先進ストレージシステムのための要素技術の提案、プロトタイプ等を

用いた評価等を行うとともに、チーム全体の達成イメージの提示を行っていく。

2. 研究実施内容

情報化社会に求められるディペンダブルで高性能な先進ストレージシステムを構築するための基本的技術の確立を目的として、研究を推進してきた。そのための実験環境として、シミュレーション用の 160 台規模のブレードシステム、および本プロジェクト内で開発した大容量版と小容量多ノード版の 2 種類のプロトタイプシステムを用意した。それらを用いて、ストレージのアクセス負荷と容量バランスの両立、通常処理とデータ管理処理の優先度管理による品質保証、それらのための管理データ構造の同時実行制御等、ストレージ管理コスト削減のための自律的な機能等の検討を行っている。また、IP 接続による高機能なネットワークストレージを構成するための基礎技術に関しても研究を行っている。それらと同時に、ストレージに情報を格納するための基盤となる記述様式として XML に着目し、XML を分散ストレージ上に効率よく格納する方法、そのためのインデックス方法などに関しても研究を行ってきた。上記に関する様々な実現手法を提案・評価し、当該分野のトップレベルの国際会議や論文誌等を含め、積極的に発表を行ってきた。さらに、海外の研究動向の調査を積極的に行うとともに、チーム内の研究内容の発表と、それに対する海外のトップレベルの専門家からのコメントの収集を目的に国際ワークショップを開催し、研究の位置づけを明確にしてきた。

以下、通常処理とデータ管理処理の優先度管理による品質保証、XML の分散格納とそのための問合せ処理、ストレージネットワークの性能と信頼性向上手法、プロトタイプの管理ソフトウェアの開発、国際ワークショップの実施内容に分けて報告を行う。

1) 通常処理とデータ管理処理の優先度管理による品質保証

分散ストレージシステムにおけるストレージ装置間でのアクセス負荷均衡化のためのデータ移動は、一時的に負荷の高いストレージ装置の負荷がさらに高くなり、クライアントへのサービスとしてのデータ提供の性能が保障できないという問題がある。本研究では、クライアントへのサービスの性能保証のため、Replica-assisted Migration(RM)と呼ぶ手法を提案してきた。

RM 手法では、信頼性確保のためのデータを利用して、データ移動経路選択と複製へのアクセス回送の 2 種の複製データ利用法を提案し、それらを組み合わせることによって、データ移動時にもクライアントへの十分なサービスの提供を目指している。そのため、各ノード最大許容負荷を目標とした複製利用制御アルゴリズム、ストレージノードキャッシュを考慮した回送制御、現実的アクセスを対象とした複製利用制御を提案した。

読み出しのみ、読み書きを行う人工的アクセスパターン、大規模 WEB サーバおよびファイルサーバの 4 種のワークロードを用いた実験を行い、提案手法が性能保証に有効であることを実証した。図 1 は読み出しのみ、図 2 は読み書きを行う人工的ワークロードに対する実験結果である。

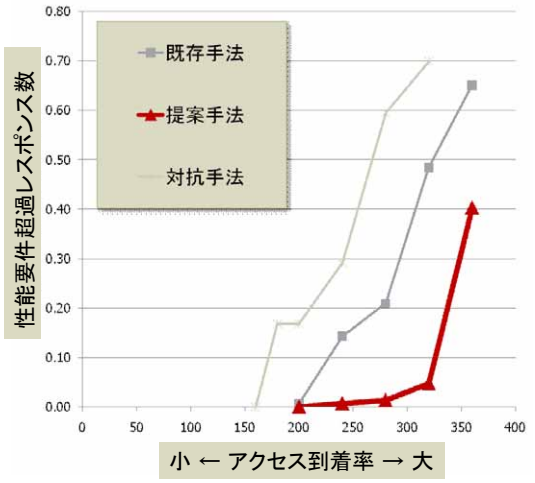


図 1　読み出しのみを行うワークロード

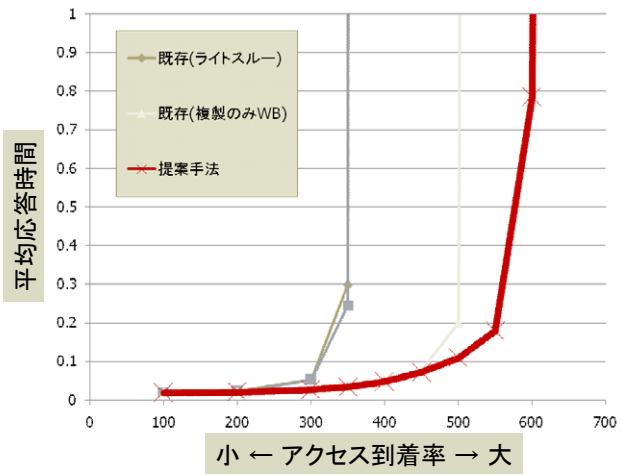


図 2　読み書きを行うワークロード

2)　XML の分散格納とそのための問合せ処理

a)　リソースの有効利用を考慮した分散 XQuery 問合せ処理方式

複数の自律ディスクとホスト計算機を連携させる効率の良い分散 XQuery 処理方式の研究を行った。これまでの分散 XQuery 問合せ処理は、リモートプロシジャコールに基づくものしか存在せず、パイプライン処理等の効率の良い処理が不可能であった。そこで、XQuery 問合せ処理に call-by-need の計算モデルを導入し、分散したオペレータ間でパイプライン処理を可能にした。また、オペレータ間のデータのフロー制御を的確に行い、計算機リソースの有効利用も可能とした。さらに、複数の計算機にまたがる演算結果を直接問合せ呼出し元に返し、計算機間の通信オーバヘッドを減らす direct result forwarding も導入した。これらを組合せることにより、従来のリモートプロシジャコールに基づく分散 XQuery 問合せよりも、最大で 22 倍高速な問合せ処理を実現した。以上は、XQuery 関数を任意の計算ノードで実行可能なよう、XQuery 言語の一部を拡張しており、ある XQuery の部分問合せを、例えばその部分問合せのデータソースを格納する自律ディスクの内部で実行させることが容易となった。今後、昨年度までの研究成果である、DTM に基づく XML データのストレージ格納方式と組合せ、自律ディスククラスタを利用した分散 XQuery 処理システムのプロトタイプを開発する予定である。

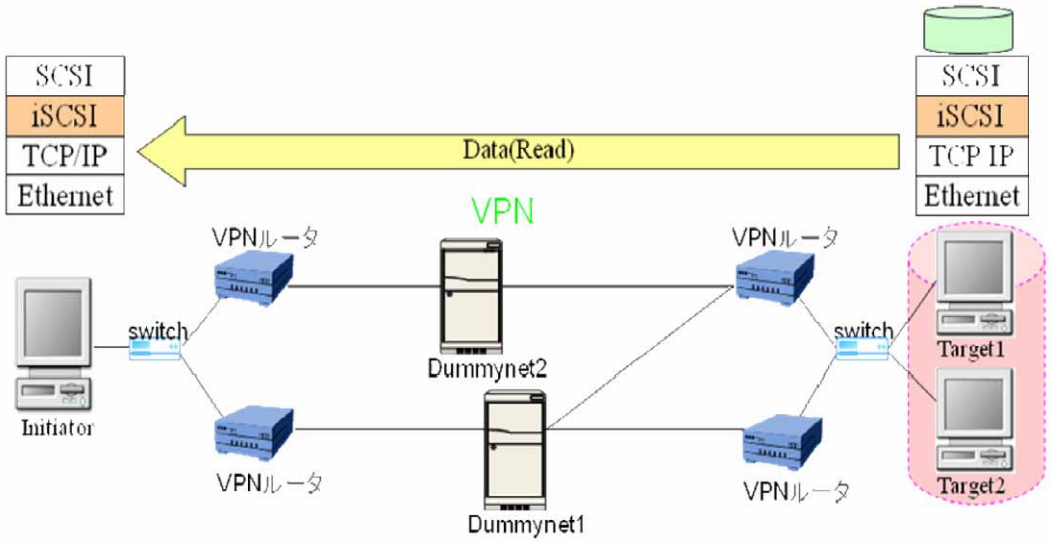
b)　自律ディスクに基づく関係データベースを利用した大規模 XML データ分散処理方式

平成 18 年度までの成果を受け、引き続き自律ディスク上に格納された XML データの分散問合せ処理方式の研究・開発を行った。今年度は、特に、XML データの問合せ処理アルゴリズムの一つである Holistic Twig Join に着目した。Holistic Twig Join は、複雑なパス問合せを一度に実行できる効率の良いアルゴリズムとして知られており、これを自律ディスク上で実行することにより、高い問合せ性能を実現することが期待されるが、我々の知る限りこの分散処理方式を検討した例はない。今年度は、1)Holistic Twig Join に適したデータ構造の自律ディスクノードへの分散方式、2)問合せワークロード、XML データの統計量などを利用した、上記データ構造の最適分割方式、の

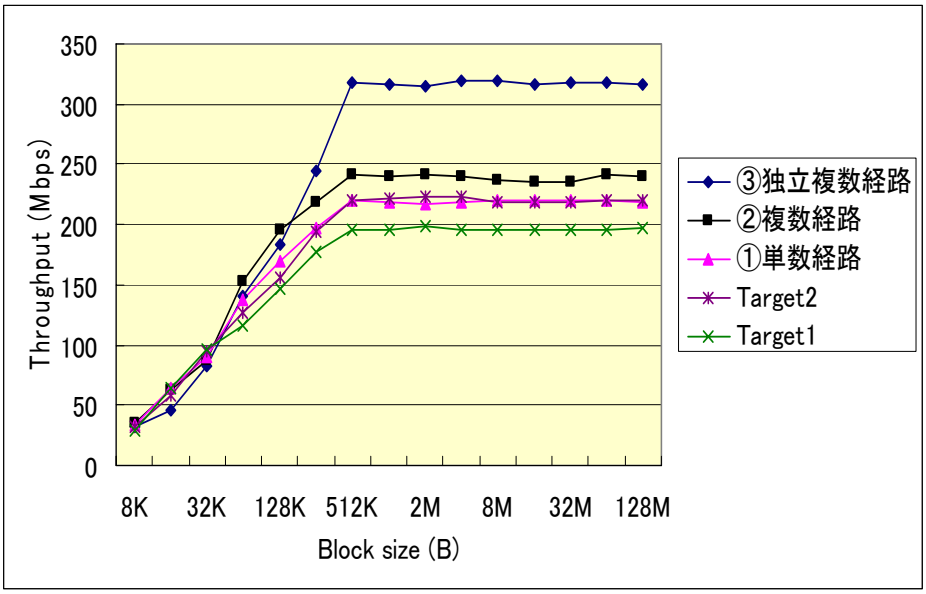
2 点を検討するとともに、実験によりその有効性を検証した。来年度は、最終的な取りまとめに向け、プロトタイプシステムの実装と、デモシステムへの組込みを行う予定である。

3) ストレージネットワークの性能と信頼性向上手法

ネットワークを用いたストレージアクセスにおける性能と信頼性向上を達成するため、遠隔データバックアップなどの用途を想定し、拠点間をVPNで接続した環境において、ストレージアクセスを行う際のネットワークの利用手法の改良を検討する。そのための実験環境として、VPN ルータ複数台を利用して構築した擬似的な広域ネットワーク環境の両端に、IP-SAN の代表的なアクセスプロトコルである iSCSI のイニシエータおよびターゲットを配置し、VPN コネクション上に iSCSI アクセスのコネクションを形成した。まず VPN ルータのマルチルーティング機能を利用して、複数経路を同時に用いる iSCSI ストレージアクセスを実現したことにより、単一の iSCSI ドライブに複数のコネクションを張ることに成功した。さらに下図に示す通り、複数のドライブを用いこれらを RAID 0 として構築したターゲットに対し、複数経路アクセスを実行して性能評価を行った。



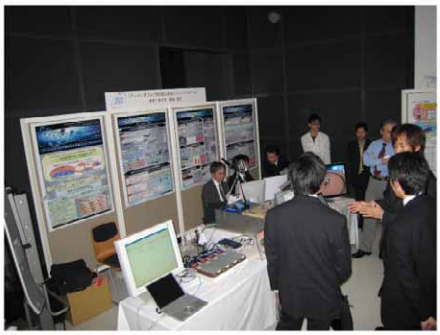
下図は 3DES 暗号化性能が 500Mbps の VPN ルータを用いた場合において、単独の Target1 と Target2 それぞれに対する単一経路アクセス、Target1 と Target2 を用い RAID 0 を構築したターゲットに対する単一経路アクセスおよび複数経路アクセス、そして参考として測定した Target1 および Target2 に対する独立複数経路アクセスの測定結果である。実験の結果、複数経路アクセスを用いたことによる性能向上を示すことができた。独立複数経路アクセスの結果は、提案方式のいわば理想値であるが、この場合のボトルネックは送信側のVPNルータであるため、VPNルータの性能向上が急速に進んでいることを考えると、提案方式は将来的に高い性能が期待できる。



4)　プロトタイプの管理ソフトウェアの開発

プロトタイプ機にて実使用環境を想定して自律ディスクソフトウェアへ以下の改版、機能追加を行った。具体的には、Windows クライアント向けインタフェース(CIFS インタフェース)の開発・実装と、CIFS 向けメタデータ処理機能の開発・実装を行った。上記改版にて評価環境を構築し性能測定を行い、アクセラレータにはメタデータ処理向上が効果的との知見を得た。

また、プロトタイプ機を用いて、10 月 12 日に開催された本年度の領域シンポジウムおよび 3 月 12 日～13 日に開催された情報処理学会全国大会併設イベント「わくわく IT」においてデモンストレーションを行った。以下の 2 枚の写真はそれぞれの様子を撮影したものである。



5)　国際ワークショップの実施内容

これまで、当研究チーム主催の国際ワークショップである ADSS（ADvanced Storage System workshop）を 2004 年と 2005 年に開催し、チーム内の研究成果の発表の場とするとともに、研究成果に対する海外のトップレベルの専門家からのコメントの収集と、海外の研究動向の調査を目的とし、チームの研究の位置づけを明確にしてきた。

ADSS2004 は、最初のトライアルとして、チームメンバーと招待講演者のみのクローズド形式とし、ワークショップの後で関連する研究所の見学も行った。ADSS2005 は、発表自体は同様にクローズド形式ではあるが、関連する国際会議（FAST2005）と同じ会場で国際会議に続く形で開催し、国際会議に引き続いて参加する専門家も多く、参加者の半数以上がメンバー外で盛況であった。

本年度開催した ADSS2007 は、これまでとは異なり一般に論文募集を行う完全なオープン形式とし、関連する国際会議である IEEE ICDIM2007 の併設ワークショップとして、各国からの 15 件の投稿の中で、口頭発表 7 件、ポスター発表 3 件という構成で開催した。出入りがあったので正確な数は把握できなかったが、ICDIM の参加者を含む 30 名以上参加者を得て、盛況に開催された。以下の 3 枚の写真は、ADSS2007 の会場の様子を撮影したものである。



3. 研究実施体制

(1)「先進ストレージ研究統轄・推進」グループ

①研究者名：横田　治夫（東京工業大学）

②研究項目

・ディペンダブルで高性能な先進ストレージシステムの研究の統括・推進

・先進ストレージシステムにおけるデータ管理機能の検討、実装、評価

・分散ディレクトリの同時実行制御手法

・分散コミットプロトコル

・セキュアストレージの実現手法

(2)「高度メディア蓄積・管理手法研究」グループ

①研究者名：宮崎　純（奈良先端科学技術大学院大学情報科学研究科）

②研究項目

・リソースの有効利用を考慮した分散 XQuery 問合せ処理方式

・自律ディスクに基づく関係データベースを利用した大規模 XML データ分散処理方式

(3)「ストレージネットワーク研究」グループ

①研究者名：小口　正人（お茶の水女子大学）

②研究項目

・先進ストレージシステムにおけるストレージネットワークの性能とディペンダビリティに関する研究

(4)「システムアーキテクチャ研究」グループ

①研究者名：太田　光彦（富士通株式会社）

②研究項目

・自律ディスク実用実験のための改版および機能追加実装

・自律ディスクプロトタイプ試作機を用いた自律ディスクの実用性検証

4. 研究成果の発表等

(1) 論文発表（原著論文）

1. 山元理絵，吉原朋宏，小林大，小林隆志，横田治夫，「アクセスログに基づく Web ページ推薦における LCS の利用とその解析」，情報処理学会論文誌データベース，Vol. 48，No.SIG 11(TOD 34), pp.38-48, 2007.6
2. Tomohiro Yoshihara, Dai Kobayashi, Haruo Yokota, ″MARK-OPT: A Coucurrency Control Protocol for Parallel B-Tree Structures to Reduce the COST of SMOs″, IEICE Transactions on Information and Systems, Vol.E90-D, No. 8, pp.1213-1224. 2007.8.
3. 並木悠太，神戸康多，小林大，横田治夫，「Fat-Btree をインデックスに用いた PostgreSQL の分散検索」，DBSJ Letters, Vol.6, No.2, pp.61-64, 2007.9.
4. 渡部 徹太郎，小林 隆志，横田 治夫，「ファイル検索に向けたアクセスログからのファイル間関連度の導出」，DBSJ Letters, Vol.6, No.2, pp.65-68, 2007.9.
5. 小林大，横田治夫，「並列ストレージにおけるデータ再配置による長期的負荷均衡化と短期的応答性能の両立」， 情報処理学会論文誌データベース，Vol. 49，No. SIG15 (TOD37), 2008.22.
6. 仲野亘，小林隆志，直井聡，横田治夫，「講義講演シーン検索におけるレーザーポインタ情報の活用法」，電子情報通信学会論文誌(D), Vol. J91-D, No.3, pp.654-666, 2008.3.
7. 油井誠，宮崎純，植村俊亮:「効率的な XQuery 処理のための DTM に基づく XML ストレージ」，情報処理学会論文誌: データベース, Vol.48，No.SIG 11 (TOD34)，pp.128-148，情報処理学会, 2007 年 6 月
8. Reyn Nakamoto, Shinsuke Nakajima, Jun Miyazaki, Shunsuke Uemura: “Tag-Based Contextual Collaborative Filtering”, IAENG International Journal of Computer Science, Volume 34, Issue 2, pp. 214-219, November 2007
9. Neila Ben Lakhal, Takashi Kobayashi, Haruo Yokota, “FENECIA: Failure Endurable

Nested-transaction based Execution of CompositeWeb Services with Incorporated State Analysis”, VLDB Journal (Accepted)

10. 小林大, 横田治夫,「負荷変動と応答性能維持を考慮した高可用並列ストレージシステムのための複製利用と更新要求制御」、情報処理学会論文誌, (採録).
11. 加藤英之, 小林隆志, 横田治夫,「OXTHAS: Web サービスベースのワークフロー管理における障害を考慮した負荷分散手法」, 電子情報通信学会論文誌(D), Vol.J91-D, No.4,(採録).

(2) 特許出願

平成 19 年度 国内特許出願件数:1 件(CREST 研究期間累積件数:17 件)