

「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

(東京工業大学学術国際情報センター 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

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な技術の確立を目的としている。より具体的には、並列分散ストレージシステムを対象に、構成要素のストレージ装置をインテリジェントにすることで、ストレージシステム全体の性能向上あるいは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、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す。

これまでに、本研究のために準備した 160 台規模のシミュレータと 2 種類のプロトタイプシステムを用いて、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プライオリティ管理、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を実現・評価し、当該分野のトップレベルの国際会議や論文誌等で発表を行ってきた。また、チーム内の研究の進捗状況に関する確認と、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、チーム主催の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。

平成 18 年度は、中間評価の対応としてこれまで行ってきた研究内容をまとめると同時に、これまでの研究をさらに進め、特に先進ストレージ内で用いられる分散ディレクトリの同時実行制御、分散コミットプロトコル、広域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する方法、および XML に代表される先進ストレージ内に格納するコンテンツ処理に関する研究等において大きな成果が得られた。それらの成果に関しては、一部特許化を行うとともに、国内外の学会で発表、さらに論文誌に投稿を行った。また、自律ディスクプロトタイプによる実環境実験を行い、アクセス性能条件などの要件抽出と、対応するシステム改善のソフトウェア開発を行った。

今後も、同様にこれまでの研究を引き続き行うと同時に、プロトタイプに関しては、更なる高性能化を目指し、ハードウェアアクセラレータを含む改良を試みる。また、提案した基本技術が商用システムに取り入れられることを目標に研究を進める。

2. 研究実施内容

以下、先進ストレージシステム実現を目指した、分散ディレクトリの同時実行制御、分散コミット

プロトコル、広域ストレージネットワークにおける VPN ルータの機能を利用したストレージアクセスの性能とディペンダビリティ向上手法、ユビキタスストレージシステムの実現手法、先進ストレージ内に格納するコンテンツ処理に関する研究としての XML 処理方法、およびプロトタイプシステム改良に関して概要を述べる。

1) 分散ディレクトリの同時実行制御

自律ディスクでは、分散ディレクトリとして Fat-Btree を採用しているが、そのような Btree では複数の利用者から同時にアクセスを想定して同時実行制御が必須である。従来から、単一 Btree に対する効率的な同時実行制御手法として、B-OPT や ARIES/IM 等が提案されていたが、それらを並列 Btree に適用すると、木構造を変更する処理に対する同期コストが大きく効率が悪かった。このため、楽観的なラッチ処理をインクリメンタルに行う INC-OPT 方式を提案していたが、リスタート時のラッチのコストが大きいという問題が残っていた。

そこで、並列 Btree の構造変更可能箇所をマークすることでリスタートを軽減して同時実行制御の効率化を図る MARK-OPT 方式を提案し、160 台構成の自律ディスクシミュレータを用いてその効果を実証してきた。さらに、INC-OPT 方式や MARK-OPT 方式を含む従来の Btree 同時実行制御方式では、安定して Btree をトラバースするために、子ノードのラッチを確保してから親ノードのラッチを開放するラッチカップリング手法が用いられるが、並列 Btree におけるラッチカップリングは通信コストやラッチ保持時間の面から性能劣化の大きな要因であった。そこで、ラッチカップリングを行わないことによる問題点を解決する手法を取ることで、性能改善を行う LCFB (Latch Coupling Free parallel Btree concurrency control protocol) 手法、および LCFB 手法と Blink 手法を組み合わせ多 LCFB-Blink 手法を提案し、同様に自律ディスクシミュレータを用いてその効果を実証した。16 台のクライアントから 128 台の自律ディスクにアクセスして、従来および提案同時実行制御手法を適用した時の更新の比率を変化させた場合のスループットの推移を図 1 に示す。また、自律ディスクの台数を変化させたグラフを図 2 に示す。いずれの場合も ARIES/IM、INC-OPT、MARK-OPT、LCFB、LCFB-Blink の順に性能が良く、提案手法が優れていることがわかる。

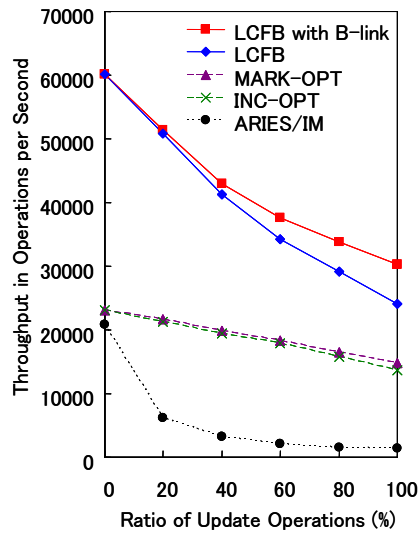


図 1 更新比率に対する各手法の違い

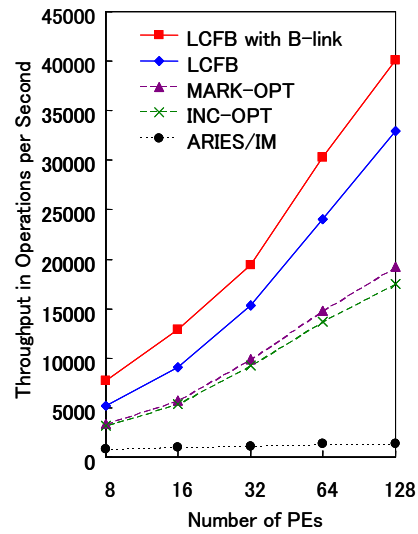


図 2 自律ディスク台数に対する違い

2) 分散コミットプロトコル

自律ディスクの Fat-Btree のトラバースでは複数のストレージ装置が関与するため、一貫性を維持するために、それらのストレージ装置間でトランザクションとして分散コミットを行う必要がある。しかし、従来の分散コミットプロトコルである 2PC (2 Phase Commit) や EP (Early Prepare) は通信コストや同期コストの関係から、性能向上のボトルネックとなる。そこで、自律ディスクのチェーン・デクラスタリングによる複製配置を有効利用した BA-1.5PC (Backup Assist 1.5 Phase Commit)プロトコルを提案した。

BA-1.5PC では、EP と同様に 2PC のプリペアコミットに対する応答を各操作の応答に暗黙的に内包すると同時に、EP のメッセージ転送前のログの強制書き込みをバックアップへの非同期 n-WAL によってコストを抑えている。非同期 n-WAL で除去した複製ストレージ間の同期は、決議フェーズに行うことで整合性を取っている。全体の処理の流れを図 3 に示す。

図 4 に、64 台の自律ディスクシミュレータを用いた場合の実験の結果を示す。実験の結果は、提案手法の BA-1.5PC は 2PC および EP と比較して性能を大幅に改善することを示している。

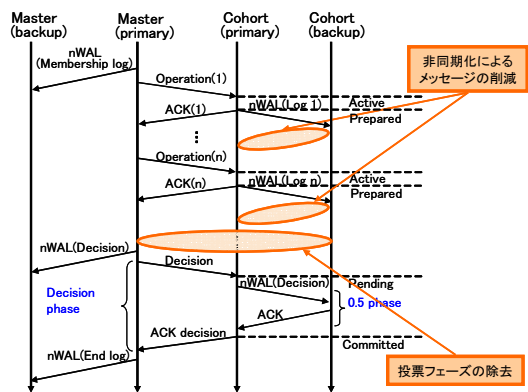


図 3 BA-1.5PC の流れ

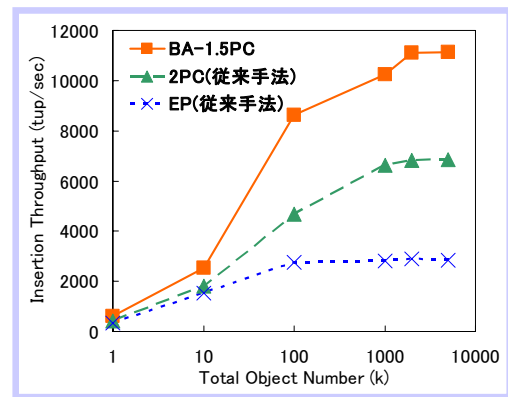


図 4 分散コミットプロトコルの性能比較

3) 広域ストレージネットワーク

現在の技術動向では、先進ストレージシステムは IP ネットワークに接続されることが前提となる。広域の IP ネットワークでストレージネットワークを構成するにはプライベート網を構成する VPN を想定する必要がある。しかし、VPN ルータを使用した場合、帯域の制限によりアクセス性能が低くなり、また信頼性の面ではネットワークの不安定要素も考慮に入れる必要が生じる。そこで、VPN ルータの機能を利用してストレージアクセスの性能とディペンダビリティ向上を目指す。例えば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。

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。この実験環境において VPN コネクションを経由したストレージアクセスのスループットを測定したところ、VPN ルータの帯域制限により、直接接続時と比較して大幅にスループットが低下した。そこで VPN ルータの複数経路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の実現を試みた。これは図 5 に示した通り、VPN ルータにより同一宛先のパケットを異なる経路に送り出すことにより、同じ宛先のストレージアクセスパケットを異なる経路経由で同時に送受信する。

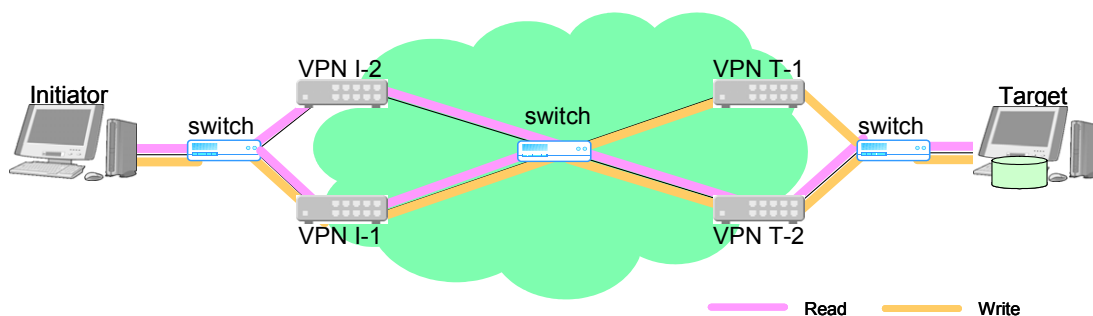


図 5 複数経路を同時に用いる iSCSI ストレージアクセス

図6は3DES暗号化性能が500MbpsのVPNルータを用いた場合における複数経路アクセス、単一経路アクセス、そして暗号化性能100MbpsのVPNルータを用いた場合の単一経路アクセスのスループットを示したものである。実験の結果、複数経路を用いたiSCSIストレージアクセスは、単一経路の場合と比較して高いスループットを達成することができた。

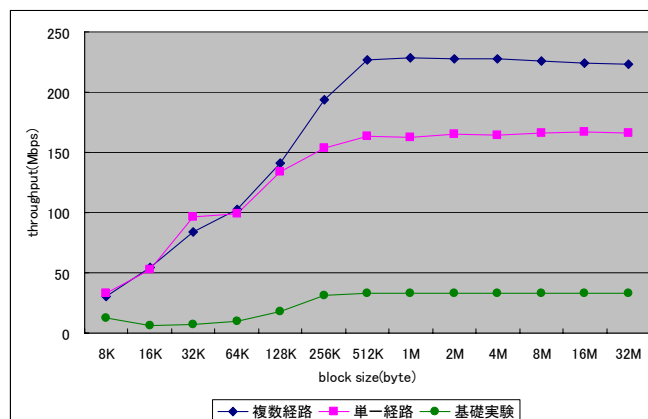


図 6 複数経路アクセスの効果

またネットワークの片方の経路に負荷がかかった状態においてストレージアクセスの実験を行ったところ、単一経路アクセスの場合と同様に複数経路アクセスにおいてもスループットが大きく低下した。この場合、負荷がかかっていない経路のみにパケットを送出すると、片方に負荷がかかった複数経路アクセスの結果よりスループットが高くなった。これは現在のiSCSIドライバの実装において、複数経路宛のパケットの送り出しがラウンドロビンで行われているため、片方の経路が遅くなるともう片方の経路へもパケットが送出されなくなってしまうためであると考えられる。従って、今後は経路の負荷や遅延状況に応じて、各経路へのパケットの送出を動的に制御するような方式の確立が望ましいと考えられる。

4) ユビキタスストレージシステム

モバイルユーザのストレージアクセスを支援するユビキタスストレージの、シミュレーションによる評価を行った。広域ネットワーク環境において、通信の影響やストレージのオーバヘッドの影響を考慮に入れ、より詳細なシミュレーションにより、従来の分散ファイルシステムであるCodaを超えるストレージアクセス性能を有することを示した。また、各ユーザが頻繁に利用するデータをモバイル機器に効率よくキャッシュするために、データに付随したタグ情報(メタデータ)を利用して各データのランキングスコアを効率よく計算するための枠組について検討した。

5) XML 処理

先進ストレージ内に格納するコンテンツ処理に関する研究として、省メモリかつ高性能なXQuery問合せ処理方式、大規模ネイティブXMLデータの分散処理とその負荷分散方式、およ

び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に関して、以下のような成果が得られた。

省メモリかつ高性能な XQuery 問合せ処理方式としては、自律ディスク内の限られた計算資源で処理可能な XQuery 問合せエンジンの研究開発を行った。XML のデータモデルの一つである Document Table Model (DTM) を二次記憶に拡張しメモリ資源の有効活用を図り、問合せ処理の特性に合わせたディスク読出しにより二次記憶のオーバーヘッドを減少させ、省メモリで動作可能な高性能 XQuery 問合せ処理を実現した。また、自律ディスクプロトタイプを利用して、その有効性をデモによって示した。

大規模ネイティブ XML データの分散処理とその負荷分散方式としては、ネイティブ XML データベースエンジンの利用を前提として、大規模な XML データを分散するための分割アルゴリズムと、問合せの分散処理法に関する研究に取り組んだ。各計算ノードの CPU コストとストレージコストをバランスしながら、動的負荷分散を行う処理方法を提案し、その有効性を PC クラスタ上で評価した。

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としては、関係データベースに写像した XML フラグメントを自律ディスクの各ノードに配置するため、コストベースの最適配置手法を開発するとともに、その有効性を実験により検証した。

6) プロトタイプシステム

実行安定性及び性能向上のため、自律ディスクソフトウェアを JAVA から C++ へ書き直すとともに、メタデータ処理の改版と実データ処理のための機能追加を行い、24Mbps HDTV 相当データのリアルタイム処理及びデータ回復処理を実現した。

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 (東工大 教授)

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」グループ

① 研究者名

植村 俊亮 (奈良先端科学技術大学院大学 教授)

②研究項目

- ・省メモリかつ高性能な XQuery 問合せ処理方式
- ・大規模ネイティブ XML データの分散処理とその負荷分散方式
- ・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
- ・ユビキタスストレージシステムの評価

(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 教授)

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)「コンテンツ対応メタデータ研究」グループ

①研究者名

上原 年博(NHK 放送技術研究所 主任研究員)

和泉 吉則(NHK 放送技術研究所 主任研究員)

②研究項目

- ・メタデータ検討
- ・コンテンツ対応検討

(5)「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) 部長付)

②研究項目

- ・自律ディスク実用実験のための改版及び機能追加実装
- ・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機検証

4. 研究成果の発表等

(1)論文発表(原著論文)

- Wenxin Liang and Haruo Yokota, SLAX: An Improved Leaf-Clustering Based Approximate XML Join Algorithm for Integrating XML Data at Subtree Classes, 情報処理学会論文誌データベース, Vol. 47, No. SIG 8 (TOD 30), pp.47-57, (2006).
- 中野真那, 小林 大, 渡邊明嗣, 横田治夫, COBALT: バージョン管理を行う並列分散ストレージにおけるアクセス負荷と記憶容量利用率の同時均衡化手法, 電子情報通信学会論文誌(D), Vol. J90-D, No.2, pp.349-358, (2007).

- 岡本拓明, 仲野亘, 小林隆志, 直井聡, 横田治夫, 岩野公司, 古井貞熙, 音声情報を統合したプレゼンテーションコンテンツ検索, 電子情報通信学会論文誌(D), Vol. J90-D, No.2, pp.209-222, (2007).
- Tomohiro Yoshihara, Dai Kobayashi, and Haruo Yokota, MARK-OPT: A Concurrency Control Protocol for Parallel B-tree Structures to Reduce the Cost of SMOs, IEICE Transactions on Information and Systems (Accepted)
- 仲野亘, 小林隆志, 勝山裕, 直井聡, 横田治夫, 講義シーン検索における検索語出現状況に基づくレーザーポインタ情報のフィルタリング, DBSJ Letters, Vol.5, No.2, pp.1-4, (2006).
- 小林大, 田口亮, 横田治夫, ストレージ複製管理のためのアクセス履歴とデータライフサイクル情報利用, DBSJ Letters, Vol.5, No.2, pp.33-36, (2006).
- 太田健介, 小林大, 小林隆志, 田口亮, 横田治夫, 柔軟なコンテンツ管理のためのルール処理へ弁別ネットワークの適用, DBSJ Letters, Vol.5, No.1, pp.1-4, (2006).
- 中野真那, 小林大, 渡邊明嗣, 横田治夫, バージョン管理を行う分散ストレージにおける偏り監視範囲分割の影響, DBSJ Letters, Vol.5, No.1, pp.105-108, (2006).
- 加藤英之, 小林隆志, 横田治夫, Web サービスベースのワークフロー管理における信頼性と負荷を考慮したスケジューリングパラメタ調整法, DBSJ Letters, Vol.5, No.1, pp.161-164, (2006).
- 城戸健太郎, 天笠俊之, 北川博之, 「PC クラスタを用いた XML データ並列処理方式の評価」, 日本データベース学会 Letters, Vol. 5, No. 2, pp. 85-88, 2006 年 9 月.
- 山元理絵, 吉原朋宏, 小林大, 小林隆志, 横田治夫, 「Web ページ推薦における推薦順位決定のための得点付け手法の比較」, DBSJ Letters, Vol.5, No.4, pp.5-8, 2007.3.
- Yuthapong Somchit, Aki Kobayashi, Katsunori Yamaoka, Sakai Yoshinori, "Dynamic Activating and Deactivating Loss Recovery Router for Live Streaming Multicast", IEICE Transactions on Communications, E89-B, No.5, pp.1534-1544, (2006)
- 中川 晋一, 木村 俊也, 三角 真, 島津 明, 山岡克式, 酒井 善則, 「患者のためのがん情報 URL リスト適正化に関する検討」, DBSJ Letters, 5, 1, pp.21-24, (2006)
- 趙剛, 小林亜樹, 酒井善則, 「適合画像の自動分類による多グループフィードバックを用いた対話型画像検索」, 映像情報メディア学会誌, Vol.60, No.12, pp.1980-1990, (2006)
- 篠宮俊輔, 山岡克式, 酒井善則, 「パケットマージ手法における出力パケットレートに基づくパケット結合方式の検討」, 電子情報通信学会論文誌 B, Vol.J90-B, No.3, pp.221-234 (2007)

(2) 特許出願

平成 18 年度特許出願: 3 件 (CREST 研究期間累積件数: 16 件)