様式 C-19

科学研究費補助金研究成果報告書

平成２３年　６月　３日現在

機関番号：１４３０３
研究種目：若手研究(B)
研究期間：2009　～　2010
課題番号：21700058
研究課題名（和文）　非均質ＳＡＮ環境のための自律適応型分散ストレージシステムに関する研究
研究課題名（英文）　Development of a Distributed Storage System with Autonomous Adaptive Control for a Heterogeneous SAN Environment
研究代表者
布目　淳（NUNOME　ATSUSHI）
京都工芸繊維大学・工芸科学研究科・助教
研究者番号：60335320

研究成果の概要（和文）：非均質 SAN 環境において、データブロックのアクセス状況に応じて再配置を行う方式について研究を行った。アクセス状況の動的な変化に対応するようにデータを再配置するためには、ノード間で頻繁に利用状況を交換する必要がある。しかし、これではオーバヘッドが大きくなるという問題が生じる。これに対し、本方式ではネットワーク上に流れる小さなサイズのパケットにアクセス情報を付加することで、詳細な管理情報を低オーバヘッドで交換できる手法を開発した。

研究成果の概要（英文）：I have developed a scheme that dynamically relocate data blocks in a heterogeneous SAN (Storage Area Network) environment. In order to allocate data blocks to the most suitable storage node according to access frequency of the blocks, each node has to exchange the access information in a short period. Such extra network traffic introduces high overhead, and may cause traffic jam. In this scheme, the access information can be attached to a small-sized packet. This method can exchange fresh and detailed information about dynamic relocation of data at low overhead.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００９年度	900,000	270,000	1,170,000
２０１０年度	1,100,000	330,000	1,430,000
年度			
年度			
年度			
総　計	2,000,000	600,000	2,600,000

研究分野：情報工学
科研費の分科・細目：情報学　・　計算機システム・ネットワーク
キーワード：データストレージ、計算機システム、ハイパフォーマンス・コンピューティング

１．研究開始当初の背景

(1) 高速ネットワーク技術の発達と対応機器の低価格化により、多数のストレージ装置をネットワーク接続し、複数のコンピュータから利用することが一般的になってきている。これらのストレージ装置を独立したネットワークで接続する SAN (Storage Area Network)環境は、ストレージ領域の統合が可能になるなどの利点が多いため、様々な組織で導入が進められている。一方で、このようなストレージシステムを長期に渡って運用すると、機器の故障などによる入れ替えや機器の追加導入のために、複数世代のストレージ装置が混在する非均質な状況が生じる。この非均質 SAN 環境においては、個々のストレージ装置の特性を活かすために、管理者が装置の特性の違いを考慮した上で、データの配置を適切に決定する必要がある。しかし、デ

ータの利用状況は刻々と変化するため、最適な配置を静的に決定することは、実際には不可能に近い。そこで、ユーザ（SAN 環境の管理者）の技術レベルに期待するのではなく、より技術的な手法でこの問題を解決する必要があると考えた。

(2) データ配置を自律的に変更するためには、ノード間で協調した情報交換が必要になる。しかし、この処理を頻繁に行うと他のトラヒックに与える影響が大きいため、情報交換の頻度を抑えることと、交換する情報の内容を精選することが重要になる。そこで、重要な管理情報を特定し、それを小さなオーバヘッドで交換する必要がある。

(3) SAN の代表的なプロトコルである iSCSI は、SCSI プロトコルを IP プロトコルでカプセル化するものであり、低価格な IP ネットワークと親和性が高いことから普及が進んでいる。そこで、リモートストレージへのアクセスプロトコルとして、iSCSI を採用する。しかし、iSCSI はもともと通信粒度の異なるプロトコルをカプセル化しているために、ネットワーク上で必ずしも最高の入出力性能を引き出せるプロトコルとは言えない。このため、TCP 処理の工夫によってノード間通信の性能向上を図る研究が主流である。一方、本研究は、分散ストレージシステム全体のアクセス状況を考慮することで、各ストレージ装置のもつ特性を最大限引き出すように動的な最適化を行う。これにより、システム全体での性能向上を狙う。

(4) 本研究では、このような制御を行いながらも、そのためにかかるオーバヘッドを抑制する方式を開発する。具体的には、ネットワークパケットの「隙間」を利用することで、新たに管理情報だけを送受信することを回避する。特にギガビットイーサネットにおいては、小さなデータフレームを送信する際にキャリア拡張を行い、本来ならば不要なパディングを数百バイトも挿入する場合がある。本研究では、こうしたパディングの代替として、あるいは MTU までの空き領域に、有益な管理情報を厳選して挿入する方式を提案する。特に、iSCSI プロトコルにおいてはイニシエータとターゲットの間で交換される CDB (Command Descriptor Block)のサイズが比較的小さく、管理情報を追加する余地が大きいことが予想される。

２．研究の目的

(1) ストレージ装置へのアクセス状況に応じて、最適なデータ配置を自律的に決定する方式を開発すること。そのために必要となる管理情報についても確定する。

(2) iSCSI プロトコルでのノード間通信に(1)で決定した管理情報を付加することで、分散ストレージの自律適応制御を行うための情報を小さなオーバヘッドで交換する方式を確立すること。特に、管理情報の粒度と重要度を考慮し、交換頻度を決定する。

(3) イーサネットスイッチのようなネットワーク集線装置に、ネットワークフレームから管理情報を抽出する機能や、その管理情報を再配布する機能のような支援機構を追加することで、(2)の交換方式をハードウェアで支援することが可能かを検討すること。

(4) 管理コストを削減するために、実行する応用プログラムからヒント情報を提供してもらうことが有効かを検討し、必要であればそのための応用プログラムインタフェース(API)を策定すること。

３．研究の方法

(1) 自律適応型分散ストレージシステムの構成方式に関して詳細な設計を行った。本研究で提案するシステムでは、各ストレージ装置へのアクセス状況を定期的に交換することで、データ配置を最適化するための情報を収集する。そのため、情報交換の頻度を上げて精度を高めることと、ネットワークトラヒックが増加してしまうことがトレードオフの関係にある。このトレードオフポイントを適切に設定する方式を詳細に検討した。なお、具体的な実行環境として、SSD (Solid State Drive)と HDD (Hard Disk Drive)が混在する環境を想定した。

(2) iSCSI 環境におけるネットワークトラヒックを実際に測定し、それらの典型的なアクセスパターンを調査した。これを基に、ネットワークフレームに追加する管理情報の種類とその量、および送信するタイミングについて検討した。

(3) システムにかかるオーバヘッドをさらに削減するために、ハードウェアによる支援機構を検討した。イーサネットスイッチまたはルータなどの集線装置がネットワークに流れる管理情報を適宜キャッシュし、定期的にブロードキャストやマルチキャストのフレームに挿入する。これにより、各ノードが主体的に送信する管理情報を削減する。

(4) 動的な情報を基に最適なデータ配置を決定する手法の他に、静的な情報から再配置に関する有益な判断材料が得られるかを検討した。また、それらの情報を応用プログラム側から提供させるための応用プログラム

インタフェースを検討した。

４．研究成果

(1) データ再配置の基本方針を以下のように決定した。

- データ移動の判断はターゲット側で行うこと。
- SSD に配置されたデータへの書き込み回数がしきい値を越え、なおかつ読み出し回数がしきい値を下回る場合はドライブの寿命を考慮してデータを HDD に移動すること。
- HDD に配置されたデータへの読み出し回数がしきい値を越え、なおかつ書き込み回数がしきい値を下回る場合はデータを SSD へ移動すること。
- データの移動は移動先へのコピーの後に移動元を削除すること。
- ストレージの空き容量が少なくなった時点で、データを低速・大容量のターゲットへ移動すること。
- データの移動はブロック単位で行うこと。

(2) (1)の方針に従い、データ再配置に必要となる情報は、ターゲットデバイス側で管理することとした。ターゲットデバイスが自身に配置されたデータの利用頻度（単位時間あたりのアクセス回数）とアクセスの種類（Read/Write）を記録し、一定時間間隔でしきい値を超えていないかチェックする。ノード間で交換する管理情報の１つとして、ターゲットの IP アドレス、ストレージの容量と空き容量をストレージ情報と定義した。ストレージの容量（バイト単位）を 64 ビットで表現することとすると、このストレージ情報のサイズは 20 バイトとなる。

(3) 再配置を行うデータに関する管理情報（再配置データ情報）は、イニシエータの IP アドレス、移動元と移動先の IP アドレス、LUN および LBA とした。この場合、512 バイト長のデータブロック１個を再配置する際に要するデータ量は 536 バイトとなる。

(4) データの移動が終了した際にイニシエータに対して通知する管理情報は、移動元ターゲットの IP アドレス、LUN 及び LBA と移動先の IP アドレス、LUN、LBA とした。この移動完了通知情報のサイズは 32 バイトとなる。

(5) iSCSI プロトコルで用いられる制御パケット（表 1）のうち、確認パケット、要求パケット、準備完了パケット、実行完了パケットのデータ部分を拡張し、データ再配置に関する管理情報を格納できるようにした。これらのパケットの長さは 100 バイト程度であり、MTU サイズが 9000 バイトであるようなネットワーク環境では管理情報を追加できる余地が大きい。上記の(2)から(4)で示した管理情報をこれらの制御パケットに対し MTU（9000 バイト）まで追加する。

表 1 iSCSI における制御パケットの分類

パケット	オペコード
確認パケット	0x00, 0x20
要求パケット	0x01
準備完了パケット	0x31
実行完了パケット	0x21
実行パケット	0x05, 0x25

(6) データ再配置を行うためには、ターゲット間でお互いのストレージ利用状況を知る必要がある。そこで、iSCSI においてイニシエータとターゲットの間で定期的に送受信が行われる確認パケットを利用するものとした。ターゲットデバイスは(2)で述べたストレージ情報を確認パケットと同時に送信する。イニシエータは、受信したストレージ情報をバッファに保管する。その後、他のターゲットに確認パケットを送る際には、このバッファから取り出したストレージ情報を追加する。このようにしてターゲットのストレージ情報を間接的に周囲のターゲットに通知することで、オーバヘッドを削減する。

(7) より高速なストレージ装置へデータを移送する場合は、利用頻度が高まったデータの移送であるため、即座に行う必要がある。このため、制御パケットが生成されるタイミングでなかった場合でも、新たに独立したパケットを生成して移送を行うこととした。この移送はターゲット間で直接行う。

(8) より低速なストレージ装置へデータを移送する場合は、ネットワークトラヒックを削減するためにサイズの小さな制御パケットに付加して移送することとした。この移送はイニシエータを介して行い、再配置データ情報を追加できる制御パケットを送信するタイミングで移動元ターゲットから移動先ターゲットへ移送する。

(9) 確認パケットに管理情報を追加する場合の優先順位は、ストレージ情報、移動完了通知情報、再配置データ情報の順とした。

(10) 移送中の書き込みアクセスについては、一時的にターゲットからデバイスビジーを返して、イニシエータに再実行を行わせることで整合性を保つ。

(11) イーサネットスイッチに対して、イー

サフレームのヘッダだけでなく、ボディに特定の内容が含まれているようなフレームを解釈する機能拡張を加えることで、提案方式のハードウェア支援を行えるようにした。この拡張を行ったスイッチは、管理情報が付加された制御パケットの構成を解釈し、管理情報（特にストレージ情報）を抽出する。スイッチはこの管理情報を一定時間保持しながら、MTU サイズに満たない他の小さなイーサフレームに追加する。これにより、ターゲットやイニシエータが自発的に管理情報を送信する場合よりも多くのノードに管理情報を通知できるようにした。

(12) 提案方式では移送中のデータに対する書き込みアクセスが多発すると、性能上のペナルティが大きいため、速やかに移送を完了させる必要がある。移送対象のデータブロックが複数同時に生じた場合に処理の優先順位を決定するために、API を通してヒント情報を与えられるようにした。実行前の段階で、書き込みよりも主に読み出しが行われると予想できるデータに対しては、プログラマからの指示により、移送のスケジューリングに反映させる。また、シーケンシャルアクセスを多く行うことが予想できる場合も、API 経由で指示を行うことで、スケジューリングやブロックの先読みに利用できるようにした。

(13) 本研究で提案した分散ストレージシステムでは、高いレベルの自律適応機能を有すると同時に管理オーバヘッドを大幅に削減できるという成果が得られた。これまでも、異なる種類のストレージ装置を組み合わせて階層化ストレージを構成する提案は行われていたが、従来はネットワーク越しに動的なデータ再配置を行うにはオーバヘッドが無視できないことが懸念されており、管理者の判断でデータの配置を静的に決定する必要があった。しかし、静的にデータの配置を決定するだけではデータの実際の利用状況を反映できないため、効果は限定的であったと言わざるを得ない。今回得られた成果により、データの実際の利用状況に応じて小さなオーバヘッドで再配置を行うことが可能になり、階層化ストレージ技術のさらなる高度化が期待できる。また、動的に変化する利用状況を小さなオーバヘッドで交換できるようになったため、データの再配置だけではなく、プロセスの再配置などの分野にも応用が可能である。今後、SAN 環境で用いられるストレージ装置はさらに多様化すると考えられるため、それらの特性の違いや性能差を考慮し、適切な用途に利用できるようにするための枠組みが必要である。本研究の成果を踏まえ、システムがストレージ装置の実効性能を動的に測定し、自律的にストレージ装置を階層化するような機構を開発することで、管理者によるメンテナンスのコストを抑えながら性能を最大限に引き出せるストレージシステムに発展させることが可能であると考えられる。

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５件）

① Atsushi Nunome, Hiroaki Hirata, Masayuki Fukuzawa, and Kiyoshi Shibayama, Development of an E-learning Back-end System for Code Assessment in Elementary Programming Practice, Proceedings of the 2010 ACM SIGUCCS Fall Conference, reviewed, 2010, pp.181-186.
② 森田清隆、布目 淳、平田博章、柴山 潔、スレッドレベル並列化のためのスレッド間依存関係の分類、第９回情報科学技術フォーラム(FIT2010)講演論文集、査読有、Vol.1、2010、pp.81-86.
③ 平田博章、山田 徹、布目 淳、柴山 潔、マシン命令レベルでのプログラム実行モニタリングによる手続き呼び出し関係の正確な検知方式、第９回情報科学技術フォーラム(FIT2010)講演論文集、査読有、Vol.1、2010、pp.87-90.
④ 赤坂謙二郎、布目 淳、平田博章、柴山 潔、データベース参照のスケジューリングによる電子商取引サイトの最適化、第９回情報科学技術フォーラム(FIT2010)講演論文集、査読無、Vol.1、2010、pp.407-408.
⑤ 野間翔平、布目 淳、平田博章、柴山 潔、スタックスマッシング攻撃の正確な検出方式とその性能制約条件、電子情報通信学会技術研究報告 CPSY2009-47、査読無、Vol.109、No.319、2009、pp.25-30.

〔学会発表〕（計５件）

① Atsushi Nunome, Development of an E-learning Back-end System for Code Assessment in Elementary Programming Practice, ACM SIGUCCS Fall Conference 2010, 2010年10月26日, Norfolk, VA, USA.
② 森田清隆、スレッドレベル並列化のためのスレッド間依存関係の分類、第９回情報科学技術フォーラム(FIT2010)、2010年９月７日、九州大学伊都キャンパス.
③ 山田 徹、マシン命令レベルでのプログラム実行モニタリングによる手続き呼び出し関係の正確な検知方式、第９回情報科学技術フォーラム(FIT2010)、2010 年 ９月７日、九州大学伊都キャンパス.

④ 赤坂謙二郎、データベース参照のスケジューリングによる電子商取引サイトの最適化、第 9 回情報科学技術フォーラム(FIT2010)、2010 年 9 月 7 日、九州大学伊都キャンパス.
⑤ 野間翔平、スタックスマッシング攻撃の正確な検出方式とその性能制約条件、デザインガイア 2009 ―VLSI 設計の新しい大地―、2009 年 12 月 4 日、高知市文化プラザ.

６．研究組織

(1)研究代表者

布目　淳（NUNOME　ATSUSHI）

京都工芸繊維大学・工芸科学研究科・助教

研究者番号：60335320

(2)研究分担者

（　　）

研究者番号：

(3)連携研究者

（　　）

研究者番号：

(4)研究協力者

森田　清隆（MORITA　KIYOTAKA）

京都工芸繊維大学・工芸科学研究科・情報工学専攻

研究者番号：なし

平田　博章（HIRATA　HIROAKI）

京都工芸繊維大学・工芸科学研究科・准教授

研究者番号：90273549

山田　徹（YAMADA　TOHRU）

京都工芸繊維大学・工芸科学研究科・情報工学専攻

研究者番号：なし

赤坂　謙二郎（AKASAKA　KENJIRO）

京都工芸繊維大学・工芸科学研究科・情報工学専攻

研究者番号：なし

野間　翔平（NOMA　SHOHEI）

京都工芸繊維大学・工芸科学研究科・情報工学専攻

研究者番号：なし