2005－DSM－36（4）
2005／3／18

ファイルサーバ仮想化環境におけるファイルデータ自律配置制御技術の開発

山川　聡　　鳥居　隆史　　梶木　善裕
NEC システムプラットフォーム研究所

概要

複雑化、多様化するコンピュータシステム環境において、コンピュータ間でのデータ共有を可能とするネットワークストレージが普及してきている。我々は、ネットワークストレージの一つである NAS（Network Attached Storage）に着目し、複数の NAS 装置を仮想的に統合し、連携動作させる技術を提案してきた。本稿では、この仮想化連携技術の一方式として、拠点に分散した NAS 装置を仮想的に統合してファイルを共有する手法と、クライアントによるファイルアクセスの際、拠点間通信ができるだけ発生しないデータの配置方法を自律的に判断する手法について提案する。本提案手法を用いることにより、拠点間でのデータ共有環境において、拠点間に大容量のネットワークを敷設しなくても、高品質のファイルアクセスサービスを提供することが可能となる。

Autonomic Data Management Control for Virtualized File Servers

Satoshi Yamakawa　Takashi Torii　Yoshihiro Kajiki
System Platforms Research Laboratories, NEC Corporation

Abstract

The network storage system realizing to share data among computers is spreading in the computer system which is complicated and diversifies. We have aimed to the NAS (Network Attached Storage) which is one of network storage, and have proposed the technologies for integrating the NAS units virtually and cooperation to make them move. In this paper, we propose the new method to share data between file servers distributed to the wide area network. The method makes it possible to manage data arrangement autonomously for reducing communications between file servers.

1．はじめに

近年、データセンターやオフィス環境において、複数のマシン間でファイルデータを共有するためのファイルサーバとして NAS（Network Attached Storage、以降ファイルサーバと NAS を同義として扱う）が急速に普及してきている。NAS は、既存のコンピューティング環境との親和性が良いことや、管理が容易であるという利点を持っており、これらの利点が普及を促進してきたといえる。このような NAS の普及伴い、NAS 装置を複数台導入するケースも増えてきており、複数の NAS 装置を連携させ、より効率的な運用や管理を行ないたいというニーズが増えてきている。これらのニーズに応えるために、我々は、複数の NAS 装置を仮想的に統合し、機能的に独立した NAS 装置を連携させる“NAS スイッチ”と呼ぶ In-band 型の仮想化装置を提案してきた[1] [2] [3]。

NAS スイッチは、NFS（Network File System）や CIFS（Common Internet File System）といった業界標準のファイルアクセスプロトコルをサポートしているサーバであれば、どのサーバも仮想化の対象とすることができることから、既に運用中の NAS 装置やファイルサーバであっても仮想化の対象として運用することができるという特徴を持っている。この NAS スイッチの仮想化機能を用いることで、サーバの

追加、削除に伴う構成変更や、データの配置場所の変更といった装置の管理作業を、ファイルアクセスサービスを停止することなく実行することができるため、管理者は容易に、システムの構成変更作業を実施することが可能となる。しかし、従来の NAS 仮想化の手法では、データ再配置に伴うデータの配置場所の指定は管理者が行なわなければならず、特に、基本性能差や距離遅延による応答性能差のある NAS 装置群を仮想化していた場合に、データをどのように配置すべきかの判断が難しかった。

我々は、このデータの配置制御に関する課題を解決するために、広域ネットワーク内に分散する NAS 装置群を仮想的に統合した環境において、NAS 装置間でのデータ再配置制御を自律的に行なわせるための手法を開発している。本手法は、クライアントによるファイルアクセスの際に、拠点間の通信をできるだけ発生させないことを目的とし、アクセス要求を最も多く発行した拠点へデータを再配置させている。本手法を採用することにより、高速、大容量のネットワークを敷設しなくても、ファイルサーバへのアクセス性を損なわずに、広域ネットワークを介したファイル共有を実現することが可能となる。

以下、本稿では、第 2 章において NAS スイッチの仮想化機能の原理とその特徴について述べた後、第 3 章にて NAS スイッチを用いて、広域ネットワークで NAS 仮想化環境を構築する手法について提案する。第 4 章では、第 3 章で提案した環境における、データの再配置制御手法について提案を行い、実験データを用いて本提案手法を検証する。

2. ファイルサーバ仮想化

本章では、ファイルサーバとして用いられている複数の NAS 装置を、仮想的に統合運用可能とする NAS スイッチの基本動作原理と、NAS スイッチが提供する付加機能について述べる。

2.1 NAS スイッチ

NAS スイッチは、NFS や CIFS プロトコルをサポートする複数の汎用の NAS 装置を、あたかも 1 台の NAS 装置であるかのように見せるための装置である。図 1 は NAS スイッチを用いた NAS 仮想化システムの基本構成を示したものであるが、NAS スイッチをクライアントと NAS 装置の間のネットワーク上に In-band 構成で組込むことにより、NAS スイッチの配下にある NAS 装置を 1 台の NAS として統合することが可能となる。また、NAS スイッチでは、NAS 装置から公開されているファイルシステム上の特定のディレクトリの配下に、他の NAS 装置から公開されているファイルシステムの特定のディレクトリをマッピングすることで、各 NAS 装置から公開されている複数のファイルシステムを 1 つのファイルシステムとしてクライアントへ提供する。このような機能を備えた NAS スイッチを導入することにより、複数の NAS 装置が設置されている環境であっても、データの格納先を意識することなく、ファイルにアクセスすることが可能となる。

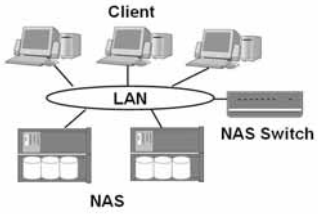

図 1:NAS スイッチによる NAS 仮想化の構成例

2.2 オンラインマイグレーション

NAS スイッチは、仮想化された名前空間内において、サービス無停止でデータを再配置する機能（我々は、この機能をオンラインマイグレーション機能と呼んでいる）を備えている。ここで言うデータ再配置とは、特定のディレクトリから別のディレクトリへデータをコピーすることではなく、ディレクトリ構成を変更せずに、データの格納先のみを NAS 装置間、もしくはファイルシステム間で移動させることである。

2.3 NAS スイッチのクラスタ化

NAS スイッチで管理されている情報は、基本的に仮想化に用いられているファイルシステムのつなぎ目に相当する部分の情報のみであり、情報量としては非常に少ない。また、この管理情報は、仮想化におけるファイルシステムの組み合わせ方が変更されない限り更新されることがない。したがって、複数の NAS スイッチで、この管理情報を共有することが容易に可能であり、同一の管理情報を持った NAS スイッチを遠隔地に分散して設置することで、広域ネ

ットワークを介したとしても仮想化された名前空間を共有することができる。また、このような広域ネットワーク環境においても、オンラインマイグレーションを実施することができ、拠点間通信を減らして応答性能を向上させるようなデータ再配置作業を、サービス無停止で実行することが可能である。

3. 広域ネットワークを介したファイル共有環境への応用

本章では、広域ネットワークを介してファイルを共有する際の問題点とその解決策を述べ、さらに NAS スイッチを用いて広域ネットワーク内に分散する NAS 装置を仮想的に統合するための手法を提案する。

3．1 広域ネットワークを介したファイル共有における問題点とその解決策

NAS 装置へアクセスするためのファイルアクセスプロトコルである NFS や CIFS は、元々、レイテンシの低いローカルネットワーク環境で利用することを想定して設計されている。したがって、レイテンシが大きい広域ネットワークを介してアクセスする場合は、ファイルの操作性能が大幅に劣化する。つまり、広域ネットワークを介したファイル共有システムで、快適なファイルアクセス環境を実現するためには、ファイルアクセスにかかるレイテンシをできるだけ小さくするような技術が必要となる。このようなレイテンシに関する課題を解決するために、ファイルデータのキャッシュ機能とファイルアクセスの排他制御機能を持った、広域ネットワークファイル共有用のアプライアンス装置を図 2 に示すように組み込んだシステムが既に提供されている[4][5]。

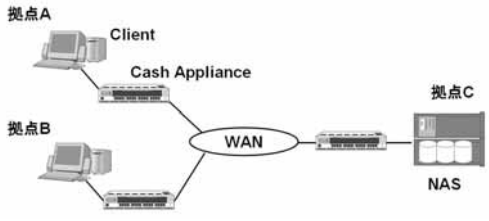

図 2：広域ネットワークを介したファイル共有システム
（ストレージ集中型システム）

文献[4][5]のいずれの方式も、一箇所に設置された NAS 装置のデータを、NAS 装置側とクライアントマシン側のエッジに設置されたアプライアンス装置を用いて、クライアント側のエッジに設置されたアプライアンス装置へデータをキャッシュさせている。これにより、ファイルアクセスにかかるレイテンシを小さくし、広域ネットワークを介したファイルアクセスの操作性を向上させている。しかし、これらシステムは、データアクセス先である NAS 装置の設置された拠点へ各拠点からアクセスが集中するストレージ集中型のシステムであることから、アクセスユーザ数やクライアント数が増加するにつれて、NAS 装置の設置拠点がボトルネックになりやすいという問題があった。

3．2 NAS スイッチの広域ネットワークへの適用

3.1 節で述べたストレージ集中型システムにおけるボトルネックの問題を解決するためには、NAS 装置自体を各拠点に分散させ、各 NAS 装置を仮想的に統合してシステム全体でデータを共有する、ストレージ分散型のストレージ仮想化システムが望まれる。

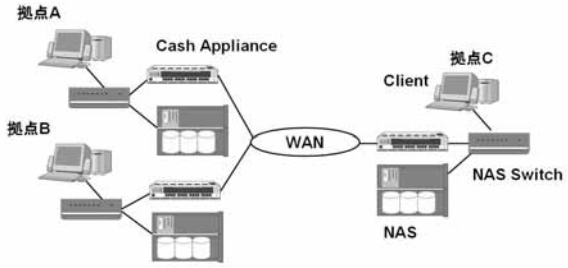

図 3：NAS スイッチを用いた広域ネットワークを介したファイル共有システム （ストレージ分散型システム）

図 3 は、広域ネットワークを介してファイルデータをキャッシュするアプライアンス装置に、分散した NAS 装置を仮想化統合する NAS スイッチと組み合わせることで、図 2 のストレージ集中型システムのボトルネック問題を解決するストレージ分散型システムである。このストレージ分散型システムでは、NAS スイッチを用いることで、各拠点に配置された NAS 装置を仮想的に統合し、各拠点のクライアントに対して、1 台の NAS 装置としてファイルアクセスサービスを提供することができる。また、各 NAS スイッチで同一となる、仮想名前空間の構成情報を共有することで、どの拠点のクライアントからでも、複数の NAS 装置を意識することなく、

同一の仮想名前空間としてアクセスすることができるようになる。さらに、一般にアクセスされる大半のファイルにはローカリティがあるという特徴を生かし、拠点ごとに、よく使われるファイルをその拠点の NAS 装置に配置することで、NAS 装置にかかる負荷や、拠点間通信によるネットワークへの負荷を、大幅に削減することが可能となる。

4. ファイルデータの自律配置制御

本章では、図 3 のように分散した NAS 装置を仮想的に統合した環境において、他拠点にあるデータへの操作に伴う拠点間通信ができるだけ発生しないように、データの配置場所を自律的に制御する手法について述べる。

4.1 ファイルデータの自律配置制御の目的

組織の変更に伴うグループやユーザの活動拠点の変更や、事業の活動フェーズの移行などにより、特定のファイル群に対するアクセス要求元の拠点が変わる場合が考えられる。このような場合、特定のファイル群に対して、従来は発生していなかった拠点間通信が頻繁に発生することとなる。図3のようなシステムにおいては、拠点間通信ができるだけ発生しないようなデータ配置を行なわなければ、限られたネットワーク資源を無駄に浪費することとなり、ファイルアクセスサービスの品質が劣化する可能がある。

このような問題に対する解決策として、拠点間に大容量のネットワークを敷設する方法が挙げられるが、システムの運用コストが大幅に上昇する可能性もあるため、全てのシステムにおいて、本解決策を適用できるとは限らない。したがって、大容量のネットワークが敷設されていない環境においても、本問題を解決するためには、ローカリティのあるファイルを、最もアクセス要求頻度の高い拠点に再配置することが望ましい。しかし、ファイルサーバの管理者が、常時アクセスパターンの状況を把握してファイルのローカリティを判断し、拠点間でのデータ再配置作業を指示することは難しく、システム内で自律的にデータを再配置させる機能が必要となる。

4.2 データ再配置制御の仕組み

ファイルのローカリティを判別するためには、各ファイルが、どの拠点からどれだけの頻度でアクセスされているかを把握する必要がある。このアクセス頻度を把握するために、全てのクライアントアクセスが NAS スイッチを経由するという特徴を用いて、各拠点に設置した NAS スイッチ内にファイルアクセスログを保存する機能設ける。さらに、管理者を介さずに自律的にデータ再配置作業を制御するために、NAS スイッチ内に、ログからアクセス頻度を算出する方法や、データの再配置手順をルール化して設定できる機能を設ける。データ再配置作業を実施するに当たっては、NAS スイッチにおいて、ディレクトリ単位でのアクセス頻度の算出と、その算出結果の拠点間での比較から、再配置すべきディレクトリを抽出し、オンラインマイグレーション機能を用いて抽出されたディレクトリのデータを再配置する。

4.3 再配置候補抽出のためのルール

データ再配置を実施するディレクトリは、以下に定義された項目に基づくルールを用いて自動的に決定される。

①アクセスログの取得ルール

クライアントから転送された READ、WRITE、RENAME 操作要求を送信元のクライアント IP アドレス、送信先の NAS 装置の IP アドレス、操作日時、操作要求先のパス名と共にログとして保存する。RENAME のログについては、アクセス頻度算出の際、パス名変更前のログとパス名変更後のログを同一パスのログとして扱うために用いられる。

②アクセス頻度の算出ルール

各拠点で取得したアクセスログのうち、自拠点のクライアントから転送された一定の期間の READ 回数、および WRITE 回数を操作要求先のファイルごとに集計する。さらに、集計されたそれぞれの回数に、操作種別によって決められた重みを掛け、それぞれを足し合わせた値をスコアとして算出する。このようにして各拠点でファイルごとに集計されたスコアをファイルのアクセス頻度として用いる。操作種別により重みを変える理由は、WRITE により大きな重みを設定することで、定常的にそのファイルにアクセスする確率が高いと推定されるデータの作成者からのアクセス頻度を高くするためである。また、一時的なデータ参照操作の集中に伴う、アクセス頻度の一時的な上昇による、余分なデータ再配置作業を排除するために、前記の手順に従って算出されたファイルのアクセス頻度に、前回、前々回

集計したアクセス頻度を一定の重みを掛けて足し合わせて最終的なアクセス頻度として用いてもよい。

③再配置を実施するディレクトリの決定ルール

②の手順により算出されたファイルごとのアクセス頻度は、そのファイルに属しているディレクトリごとに足し合わされ、ディレクトリのアクセス頻度として集計される。各拠点で算出されたディレクトリのアクセス頻度が、予め決められた閾値に達しており、かつそのディレクトリが他の拠点のファイルサーバに格納されていた場合、そのディレクトリが再配置の対象となる。また、特定のディレクトリに対し、複数の拠点で高いアクセス頻度が算出されていた場合には、特定の拠点のアクセス頻度のスコアが、全拠点で総和したアクセス頻度に対して、予め決められた一定以上の割合を占めていた場合のみ、データ再配置が実施される。

4．4 実運用環境サンプルを用いた自律再配置制御動作の検証

4. 3 節にて提案したデータ再配置制御方式の有効性を検証するため、実際に運用されているファイルサーバのアクセスログをサンプルデータとして用い、以下に示すシミュレーション環境において、本方式の有効性を検証した。

シミュレーション内容

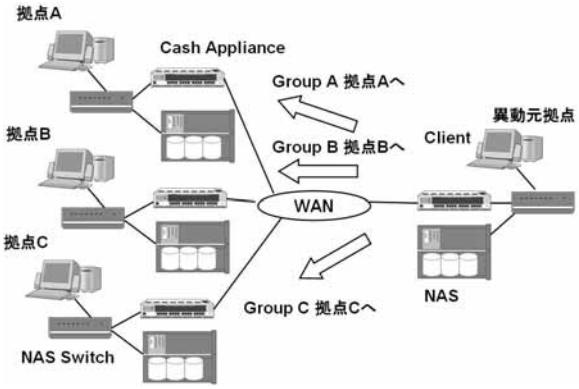

図 4: シミュレーションにおける想定環境

今回実験したシミュレーションでは、ファイルサーバを利用している 3 つの活動グループが、ファイルサーバの設置されていた拠点以外の 3 拠点に異動した場合（図 4）を想定し、自律配置制御方式に基づいて設定された一定のルールに従って、異動元の拠点から、異動先の拠点へデータ再配置を繰り返すことで、各グループが異動後の拠点に設置されたファイルサーバにどの程度アクセスすることができるかを検証した。

今回設定したルールは以下の通りである。

- ○ アクセス頻度の集計を 2 週間に 1 回実行し、集計結果を元にデータ再配置を実行する
- ○ スコア算出時の WRITE の重みを READ の 10 倍とする
- ○ 前回集計したスコアの 1/2 の値をスコア集計の際に足し合わせる
- ○ 1 日で平均 1 回の READ 要求、および 2 日で平均 1 回の WRITE 要求が発生したディレクトリがデータ再配置対象となるように閾値を設定する
- ○ 複数の拠点で高い頻度が算出され、かつ前記閾値を超えていた場合、1 つの拠点で全拠点総和の 50%以上の値となるスコアを算出した場合に、その拠点へデータが再配置される

以上のルールに従い、合計 6 週間のアクセスログのサンプルを用いて、データ再配置を 2 回実施した場合におけるシミュレーション結果を表 1 に示す。

表 1：再配置実施後の自拠点格納データへのアクセス率

	1 回目の再配置実施後	2 回目の再配置実施後
Group 1	89%	87%
Group 2	14%	48%
Group 3	0%	8%
総計	65%	70%

表 1 に示した通り、再配置の回数を重ねるごとに自拠点のデータへのアクセス率が高くなっていることが分かる。グループごと、アクセス率にばらつきが出ているが、Group 1 については、最初の 2 週間で、通常アクセスするディレクトリの大半に閾値以上のアクセスがあったことがログの状況から分かった。また、Group 3 については、ログの取得期間中に、設定した閾値を満たすほどのアクセスがなかったことが同様にして分かった。

4．5 考察

図 5 - 7 は、本シミュレーションにより、再配置が実施された 21 個のディレクトリへのグループ単位でのアクセス比率

を 2 週間ごとに示したものである。本提案における再配置制御方式では、READ 操作よりも WRITE 操作に対して、掛け合わせる重みを大きく設定することで、定常的にアクセスされる可能性の高いデータ作成者からのアクセス頻度が高くなるようにしていた。ディレクトリ 8 は、2 回目のデータ再配置のタイミングで Group 3 の拠点にデータ再配置が行なわれたが、2 回目のデータ再配置のタイミングでの集計の際には、図 6 に示す通り、Group 2 からのアクセス比率が非常に高くなっていたため、Group 2 の拠点に再配置される可能性があった。しかし、実際は Group 2 からは短期間に大量の READ アクセスが発生していただけでスコアが伸びず、逆に Group 3 からの複数の WRITE アクセスによるスコアの伸びに押されて、ディレクトリ 8 は、Group 3 の拠点に再配置されていた。今回活用したサンプルデータでは、このような現象が発生したディレクトリは、ディレクトリ 8 のみであったが、アクセス頻度算出の際、READ と WRITE に掛け合わせる重みを変化させることで、クライアントからの一時的なデータの参照操作に伴うデータ再配置を防止できることが実証された。

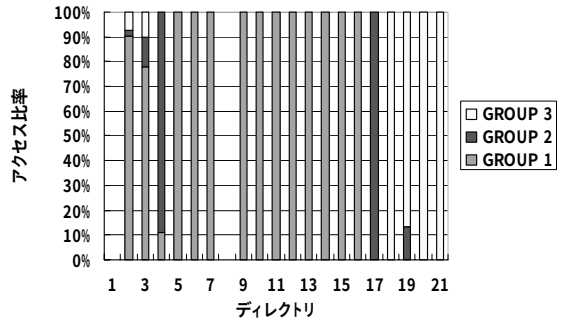

図 5: 1 - 2 週目におけるアクセス比率の状況

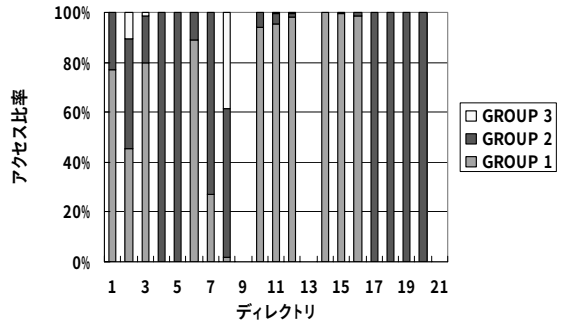

図 6: 3 - 4 週目におけるアクセス比率の状況

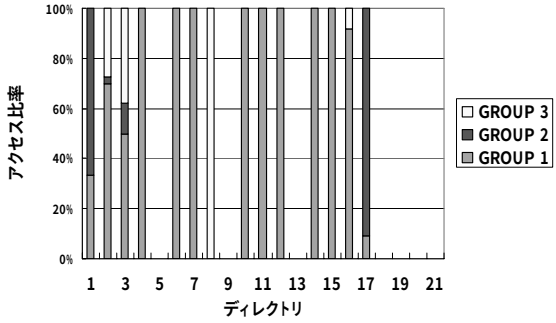

図 7: 5 – 6 週目におけるアクセス比率の状況

5. むすび

本稿では、NAS スイッチを用いて、広域ネットワークを介したファイルサーバの仮想化手法と、アクセス頻度に連動したデータ再配置制御手法について提案した。また、実際に使用されているファイルサーバのアクセスログをサンプルとして利用し、提案したデータ再配置制御方式が有効に働くことを実証した。今後は、他の実運用環境でのサンプルや、複数の設定ルールを用い、同様のシミュレーションを通して本提案方式の検証を行なうことにより、本提案方式における問題点の抽出と改善策について検討していく予定である。

参考文献

[1] 山川,石川,菊地, "NAS スイッチ:NFS サーバの仮想化技術の開発," 信学技報(CPSY), vol.102, No.275, pp13-18 (2002).

[2] 桂島,石川, "ファイルサーバ仮想化アプライアンス NAS スイッチの提案," 情処研報, 2002-DSM-26, vol.82, pp.31-36 (2002).

[3] 梶木,"ファイルサーバの仮想化技術," 信学技報(CPSY), Vol.104, No.537, pp79-84 (2004).

[4] Cisco Systems, "Cisco File Engine Series Appliances," http://www.cisco.com .

[5] Tacit Networks, "iShared Server, iShared Remote, iShared Symmetric," http://www.tacitnetworks.com .