



metaSAN5と Scale-NAS ソリューション

エムアイシー・アソシエーツ株式会社

目次

1. 概 要	3
2. ファイルサーバーのスケールアウトシナリオと要件	3
2.1 スケールアウトシナリオ実現のための要件	3
2.2 スケールアウトシナリオのコンポーネント	4
3. CLASTOR2100-WSの基本構成と特長	4
3.1 CLASTOR2100-WSファイル共有サービス-SMB2.x	4
3.2 SMB2.xでの大容量データ転送の問題に対するソリューション	4
3.3 CLASTOR2100-WSのWindows GPTボリューム拡張	6
3.4 ネットワーク性能の拡張	6
4. CLASTOR2100-WSサーバーのスケールアウト	6
4.1. クライアントプール	7
4.2. 負荷バランサーによるファイルサーバーの IP一元化	7
5. GPTパーティション以外でのストレージの拡張 - None-Destructive Storage Expansion	8
5.1 cluStoreストレージプール機能について	8
5.2 CluStoreのポリシーベースのストレージプール管理	9
5.2.1 Fill-Upポリシー	9
5.2.2 Harmonizeポリシー	9
5.2.3 File Type ポリシー	10
5.2.4 Last Accessed ポリシー	10
6. データバックアップ	10
7. ファイルサーバーのスケールダウンとコンポーネントリプレイス	11
8. まとめ	11

1. 概要

現代のIT環境では映像、画像、PDF等の非構造化データが急激に増大しています。一方で、情報の価値はますます高まり、情報を整理し、ストレスの無い高度な情報サービスによる情報共有をIT部門は実現しなくてはなりません。その為、急激に膨張するデータを保存するファイルサーバーの重要性は増々増大していますが、一方、それを管理するIT部門では以下の様な課題に直面しています。

- ・ 急激な非構造化データの増加、ビッグデータ等への対応
- ・ 部門、担当別のストレージ管理、 空き容量管理の煩雑性
- ・ ボリューム容量、数の増大とバックアップの負担
- ・ サービス無停止でのストレージの増設と移行
- ・ 複数ベンダーのストレージ管理の煩雑性
- ・ ストレージの性能、価格の陳腐化
- ・ ストレージ予算の削減

これらの課題を解決する方法として、ファイルサーバーには必要に応じストレージ容量や、アクセス性能を追加できる拡張性が求められています。ネットワーククライアントが増えたり、データ容量が増加する場合でも、 初期投資を無駄にすること無く、 ストレージのアクセス性能を向上させることができ、データ容量の増加に従って、ストレージ容量の拡張可能なファイルサーバーのスケールアウト機能がこの課題を解決します。また、急速なIT技術革新に追従し、初期投資を陳腐化から守り、データの安全性を確保する必要があります。

現在、業界屈指のストレージベンダーからは幾つかのスケールアウトNASが販売され、インストール実績も多くなりつつあります。しかし、それらスケールアウトNASはベンダー独自のハードウェア技術やファイルシステムを使用し、何れの製品も初期の段階から高額で始まり、スケールアウトする為の費用も大きな負担になる傾向にあります。一方で、中小規模の組織や企業においては、低価格のファイルサーバーを必要に応じて購入することで、クライアントの増加、データ容量の増加に対応する傾向にあります。しかし、この様な対応ではファイルサービスの性能、信頼性の確保やバックアップもままならない状況にあり、スケールアウト可能なストレージへの投資とはほど遠い状況にあります。

この様なユーザーサイドのファイルサーバー利用状況に対し、弊社は新たなファイルサーバーのスケールアウト方法をご提案します。利用可能な技術とmetaSAN5の各コンポーネントを使用して、ファイルサーバーにストレージの仮想化、スケールアウト機能を持たせることでより低コストでスムーズなスケールアウトを実現することが可能です。

スケールアウト可能ネットワークストレージに対するユーザー要求は以下のような項目が上げられます。

1. ストレージ容量の随時追加
2. クライアント増加に追従したストレージアクセス性能の向上
3. ストレージの一元管理
4. 容易で高速なデータバックアップ
5. 低い管理維持コスト
6. 場合によりスケールダウンや、サーバー、ストレージの交換が簡単に行なえること

一般的に、以上のような要件がスケールアウトストレージに要求されます。この様なスケールアウトを実現する方法を以下にご紹介いたします。

2. ファイルサーバーのスケールアウトシナリオと要件

2.1 スケールアウトシナリオ実現のための要件

本稿では、弊社CLASTOR2100-WS ファイルサーバーと、この度発売を開始しましたmetaSAN5、更に、ファイルサーバーヘッドの仮想化を可能にするLayer 4をサポートするネットワークロードバランサーを利用して、ファイルサーバーのスケールアウトソリューションを実現させるシナリオをご紹介します。 このスケールアウトシナリオを実現させる上で、以下の様な要件を確認する必要があります。

1. ファイルサーバーを使用するクライアントが増加すること
2. クライアントに対してネットワーク経由でストレスの無いファイル共有サービスを提供し続けること
3. 必要に応じストレージエリアを増設し、その為に既存データの退避や、ストレージの初期化を必要としないこと
4. ワークグループ単位、または、全体で同一ネームスペースにアクセスすることができること
5. サーバーを一元的アドレスとして、アクセス可能なこと
6. バックアップが効率的で、低コストであること

以上の要件を満たしながらファイルサーバーのスケールアウトのシナリオを実現します。

2.2 スケールアウトシナリオのコンポーネント

1. CLASTOR2100-WSファイルサーバー：

CLASTOR2100-WSは1Uの19"ラックマウントタイプのストレージサーバーでIntel Xeon 2.3GHz 4-core CPUを搭載し、メモリーを最大48GBまで搭載することが可能です。また、使用OSとしてはCALフリーのWindows Storage Server 2008 R2を使用し、ファイルシステムとしてはWindows標準のNTFSを使用します。

2. ネットワーク：

サーバー、クライアント間はL2レベルのスイッチを使用したLANを構築します。必要に応じ、1Gb Ethernetを複数使用して、ネットワークの仮想化技術により、バンド帯域の拡張ができる必要があります。また、10Gb Ethernetも大容量のデータ転送が必要な場合の選択肢の一つです。

3. ストレージ共有ソフト metaSAN5：

今回、スケールアウトシナリオを実現する上で重要な役割をはたします。複数サーバーによるストレージ共有、ストレージの仮想化、バックアップの効率化は総てmetSAN5に含まれるコンポーネントにより実現します。

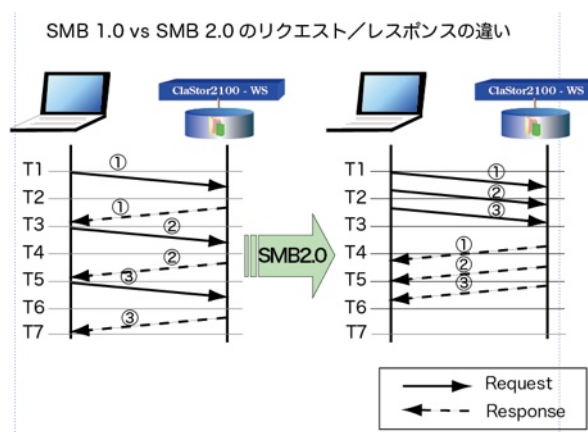
4. ロードバランサー：

複数のファイルサーバーヘッドを一元化し、クライアントから1つのIPアドレスでサーバーにアクセスし、CIFS/NFSによるファイル共有サービスを提供する必要があります。その為、ファイルサーバーのIPアドレスの仮想化と、ファイルサーバーの負荷バランスを実現するため、Webサーバー等で多く使用されるロードバランサーを使用します。

3. CLASTOR2100-WSの基本構成と特長

3.1 CLASTOR2100-WSファイル共有サービス-SMB2.x

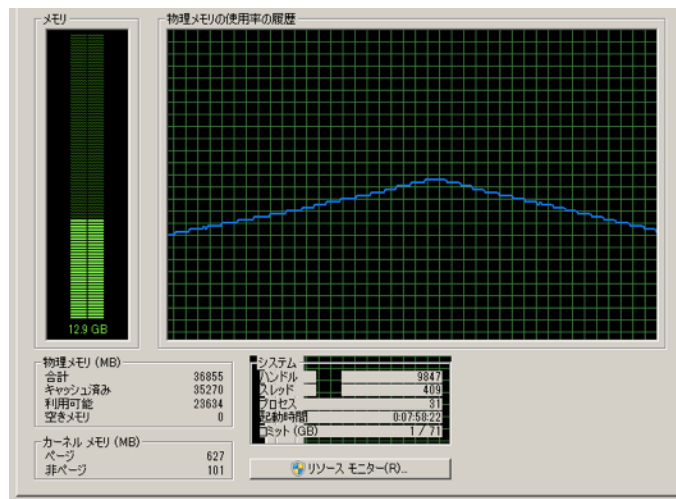
CLASTOR2100-WSはWindows Storage Server 2008 R2を搭載し、CIFS, NFS, おおび、iSCSIターゲットのサービスを提供するネットワークストレージサーバーです。CLASTOR2100-WSに実装するSMB2は、以前のSMB1.xと異なり、複数のクライアントリクエストを同時に受け付け、サービスすることが可能です。また、SMB1.xで制限されていたメモリーバッファ領域を実装メモリーの約50%まで取ることができるようになりました。その結果、SMB1.xベースのCIFSサービスに比較してSMB2.xでは数倍のアクセス性能を発揮します。ただし、SMB2.xの適用はクライアントOSがSMB2.xに対応している必要があり、クライアントワークステーションがWindows Vista以前の場合、SMB1.xでのサービスになります。



3.2 SMB2.xでの大容量データ転送の問題に対するソリューション

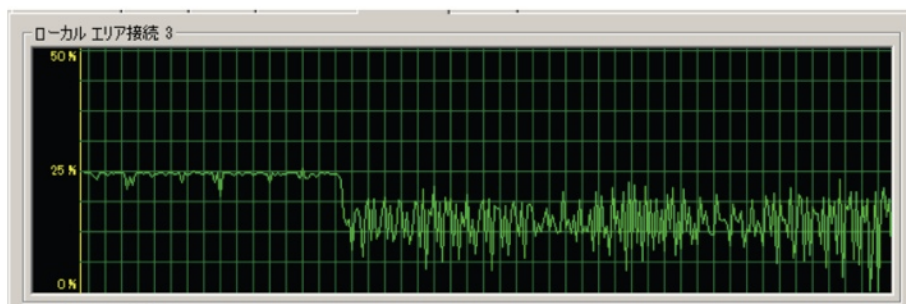
以上のように、SMB1.xとSMB2.xではそのサービス性能において大きな差がありますが、依然としてクライアントからのデータ保存は①データ転送の受付→②メモリーキャッシュに展開→③一定間隔でのストレージへの書き出しというプロセスを踏んでいます。このため、メモリーに保存されたデータはタイムリーにストレージに書き出されない限り、データがメモリーを満ち、クライアントからの書き込みのリクエストは待たされることになります。

次の図はWindows Storage server 2008 R2で3台のクライアントサーバーから512KB毎のデータ転送を同時に実行した場合の、ファイルサーバーのメモリーの消費状況と、ネットワークパフォーマンスを表しています。大容量データの同時書き込み時のメモリーキャッシュが飽和する状況と、直後からキャッシュ領域を開放する状況を表示しています。



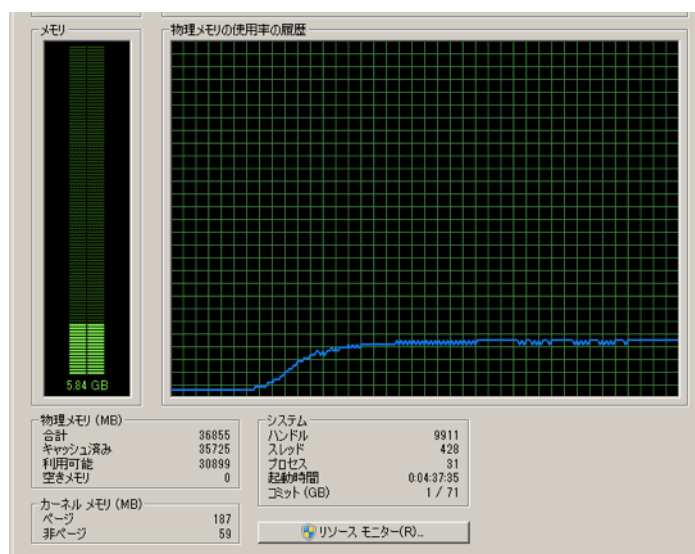
キャッシュされた書き込みデータが実メモリ容量に近づく、メモリ内データをストレージに書き出す処理が優先され、結果的にクライアントからのリクエストは待たされることになります。

下図は、上図のメモリ飽和後のキャッシュ領域の開放の間、データ転送が大幅に劣化したことをしめします。(ネットワークは10GbEx1 を使用、3台のクライアントから512KBのライトリクエストを同時に実行した際のデータです。)



以上はWindows Storage Server 2008 R2でのパフォーマンスデータですが、このサーバーにmetaSANをインストールし、前のテストと同一のデータ転送を実行した場合のメモリの使用状況を下図に表します。

metaSANをインストールした場合、大容量データ転送に際し、書き込みデータはメモリキャッシュの一定の領域まで占めますが、メモリは飽和せず、データ送信の劣化も発生しませんでした。このことから、映像、画像転送に特化したmetaSANは、大容量データをできるだけ高速にデータをストレージに書き出し、メモリに余裕を持たせるデザインであることを証明しました。

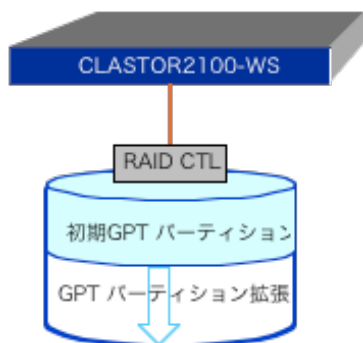


以上のことから、弊社では1台のCLUSTOR2100-WSからmetaSANを実装することをお勧めします。

3.3 CLASTOR2100-WSのWindows GPTボリューム拡張

Windows 2008 R2のファイルシステム、NTFSには1パーティション容量を最大8ZBまで拡張可能なGPT(GUID Partition Table)というパーティション方式が採用されており、ボリューム容量を随時拡張することが可能です。この機能を利用するにはRAID装置等で採用されているLogica Unit(LU)という、物理ドライブを論理ドライブとしてOSに見せる技術が必要です。

下図はCLASTOR2100-WSにRAID装置を接続し、アレイで構成されたデバイスを論理ドライブとして、Windowsに見せ、GTPパーティションとして構成したシステムを表します。



初期にGPTパーティションで構成されたボリューム領域に使用されるRAID装置の論理ドライブは、更に、新たなハードディスクを増設し、既存の論理ドライブの容量を拡張することができます。この拡張された論理ドライブの領域はWindows OSからは既存パーティションを拡張することができる新たなストレージ空間としてみることができます。Windows のDisk Managerを使用して、既存のパーティションを追加部分を含めて拡張を選択することで、瞬時にボリューム領域を拡張することが可能です。

但し、論理ドライブにストレージを追加する場合、ほぼ同一性能のドライブを使用することが経済的です。性能が異なるドライブを使用する場合は、別の論理ドライブ、別のGPTパーティションを構成します。

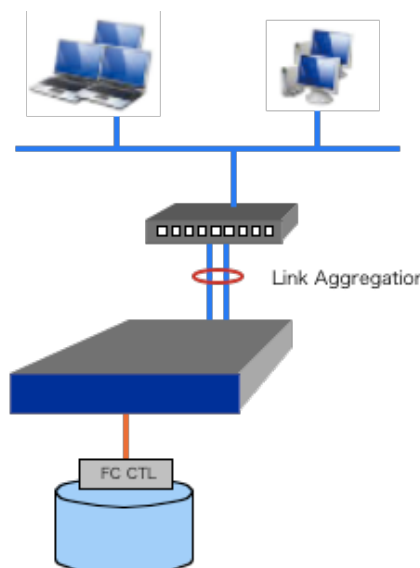
3.4 ネットワーク性能の拡張

CLASTOR2100-WSには2ポートの1GbE ネットワークの他、2ポートの10GbEネットワークが標準で装備されています。この為、これらのポートを様々な方法で利用することができます。その例として、以下にCLASTOR2100のネットワークポートの拡張に関してお紹介いたします。

ネットワークポートの動的リンクアグリゲーション(IEEE802.3ad)に対応したL2レベルのネットワークスイッチとCLASTOR2100-WSのネットワークポートをそれぞれLACP動的リンクアグリゲーションでTeaming設定することが可能です。この設定で、2本のネットワークポートは仮想的に1本の2GbEポート、または、20GbEのポートとして使用することが可能です。また、CLASTOR2100-WSにはPCI-G2の slots に追加ボードを挿入することができますので、必要に応じ、1GbE 2ポートを持つNICを追加することで、経済的に4GbEのバンド帯域の仮想IPを構成することが可能です。

右図はCLASTOR2100のネットワークポートを1つのIPにTeamingした構成を示します。

クライアントを含めたネットワーク全体に関しては、サーバーポートの大容量化だけではなく、カスケードされるスイッチ間のバンド帯域や、サーバー/スイッチ/スイッチ間の冗長化のためのSpanning Treeといったネットワーク構成を考慮する必要があります。



4. CLASTOR2100-WSサーバーのスケールアウト

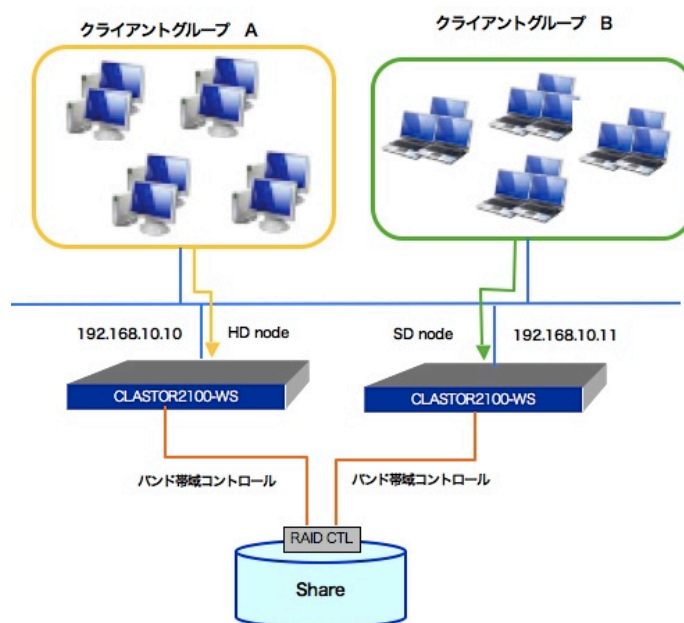
前1項の説明により、既にmetaSANをインストール済みのCLASTOR2100-WSと、SANストレージを共有するCLASTOR2100-WSを新たに追加することで、ファイルシステムへのアクセスゲートウェイを増やすことが可能です。metaSANはWindows NTFSをネイティブにサポートし、SANストレージを複数のサーバーからローカルディスクの様に、ブロックレベルでデータにアクセスすることが可能です。CLASTOR2100-WSはmetaSANを実装し、複数のCLASTORが同時にSANストレージにアクセスし、SANストレージの性能をそのままクライアントに提供することができます。

4.1. クライアントプール

多くのクライアントが同時に大容量、高性能SANストレージにアクセスすると、ネットワーク、CPU、メモリー、ストレージのポートなどのファイルサーバーリソース内のボトルネックが顕在化する傾向にあります。このような問題を解消する為に、CLUSTOR2100-WSの追加ソリューションを選択することが可能です。

metaSANはWindows Server 2008 R2のネイティブなファイルシステムであるNTFSをそのままサポートしますので、サーバーの追加時もストレージの初期化等の特別な手順や、既存CLUSTOR2100-WSのサービスを停止することなく、直ちに、ストレージ共有開始することが可能です。

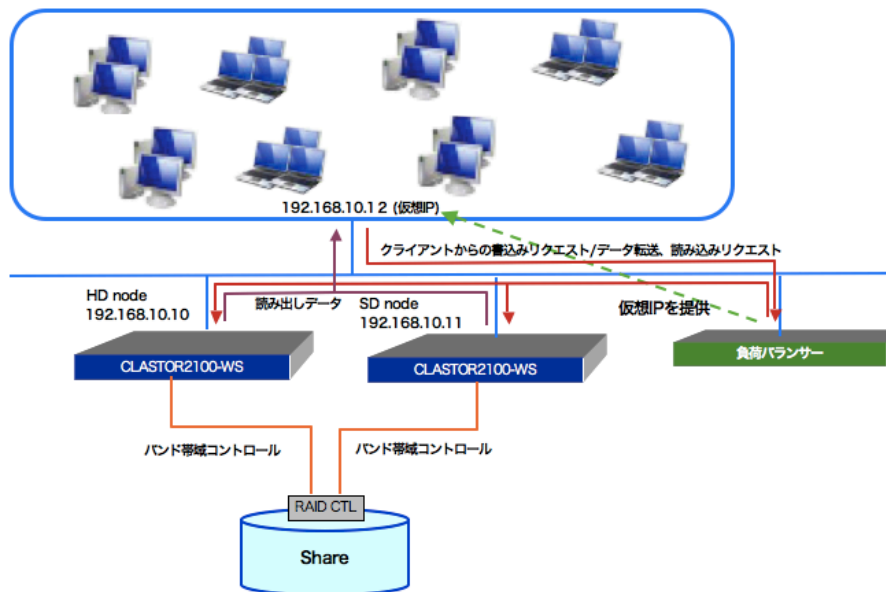
ネットワーククライアントを、SANストレージへのゲートウェイとなるCLASTOR2100-WS毎にグループ分け(クライアントプール)し、クライアントプール間で負荷バランスを行なうことが可能です。次図はクライアントプール毎に、ストレージへのアクセスするゲートとしてのCLASTOR2100-WSを配置したネットワーク構成です。クライアントプール AはHD node側からShareボリュームにアクセスし、クライアントプール BはSD node側からShareボリュームにアクセスします。プールA、プールBとも実際のユーザーアクセス負荷に合わせ、ストレージ⇔CLASTOR2100-WS間のバンド帯域を割当、ネットワーク負荷をバランスすることが可能です。



クライアントグループ別にアクセスするサーバーを分けた構成。

4.2. 負荷バランサーによるファイルサーバーの IP一元化

ネットワークストレージへの複数のパスを設け、マニュアルベースでのサーバー間負荷バランスを実現することが可能ですが、更に、負荷バランサー(Load Balancer)をネットワークに接続することでサーバー間の負荷バランスを自動化することが可能です。負荷バランサーはCLASTOR2100-WSの個別IPアドレスを一つのIPに仮想化して、クライアントに見せることができます。総てのクライアントからは、複数あるサーバーも一つの仮想IPアドレスを持つファイルサーバーとして見えます。負荷バランサーは内部処理能力により、書き込み性能で300MB/秒からそれ以上の性能を発揮します。サーバーへの負荷配分はユーザー負荷の内容により配分方法を決定することが可能です。また、必要に応じ、ファイルサーバーを負荷バランス対象の仮想IPアドレス下のノードとして追加することで、ファイルサーバー全体で大容量のネットワーク帯域を確保することができます。また、負荷バランサーに採用されるDirect Service Return(DSR)機能により、ストレージからのデータの読み出し時はバランサーを介さず、直接個々のサーバーからクライアントにデータを返すことができます。DSRを使用することで、負荷バランサーは書き込みのデータを各CLASTOR2100-WSに分配して転送するのみで、読み出しは個々のサーバーの性能をそのままクライアントが享受できます。



5. GPTパーティション以外でのストレージの拡張

- None-Destructive Storage Expansion

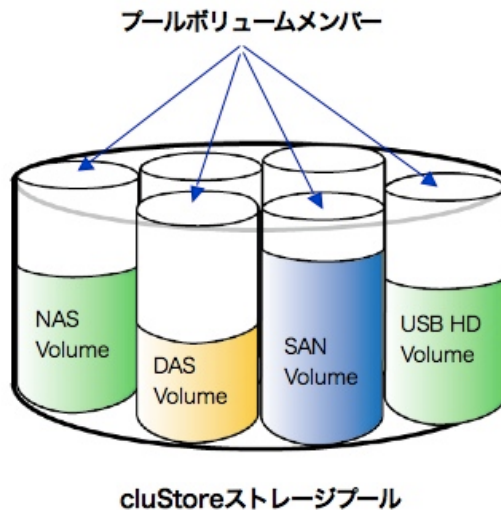
前述の 3.3項 「CLASTOR2100-WSのWindows GPTボリューム拡張」 ではGPTパーティション拡張はRAIDコントローラが構成する論理ボリュームを拡張し、GPTパーティションを拡張することで実現すると説明しました。しかし、この拡張は同一性能のドライブを使用することが現実的です。反対に、当初、15,000rpm、600GB SAS HDDを使用してアレイを構成した後、7,200rpm、3TB SASのHDDを使用してアレイを構成し、既存のネットワークストレージに追加する場合、性能の異なるストレージを一つのGPTパーティションするより、二つの別々のパーティションにした方が、性能面で確実です。

しかし、ファイルサーバーとして、ストレージを共有するには複数ボリュームより、1ボリュームの方がユーザーからも、管理上も便利です。CLASTOR2100-WSには、ストレージボリュームの仮想化オプション機能を追加することができ、複数のNTFSのファイルシステムを一つのストレージプールとして扱うことができます。この機能はmetaSANに付属するcluStoreと呼ばれるミドルウェアを使用します。

5.1 cluStoreストレージプール機能について

cluStoreストレージプール機能はCLASTOR2100-WSに直接接続するストレージの他、ファイルシステムとしてCLASTOR2100にマウントされるネットワークストレージをそのストレージプールに加えることができます。

下図はNAS、DAS、SAN、USB等各種のストレージが提供するストレージボリュームをcluStoreストレージプールで仮想化した状態をしめします。ストレージプールに新しいボリュームメンバーを追加したり、削除したりする為に、cluStoreストレージプールを初期化や、Disk Managerによる設定や、初期化等の作業は必要ありません。また、ボリュームの追加や削除の為に、CLASTOR2100-WSのファイル共有サービスを停止する必要もありません。また、ストレージプールをmetaSANの共有ストレージとすることが可能です。

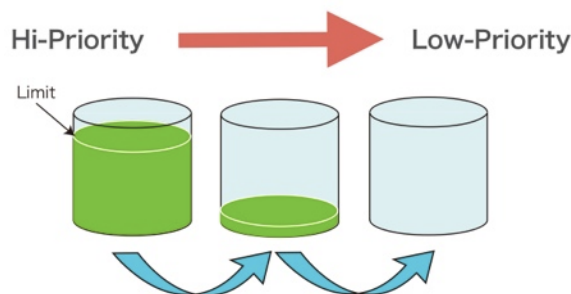


5.2 CluStoreのポリシーベースのストレージプール管理

CluStoreストレージプールにはプールを構成するメンバーのストレージボリュームに優先順位をつけることができます。この優先順位に基づき以下の4種類のポリシーでストレージプールを管理することができます。

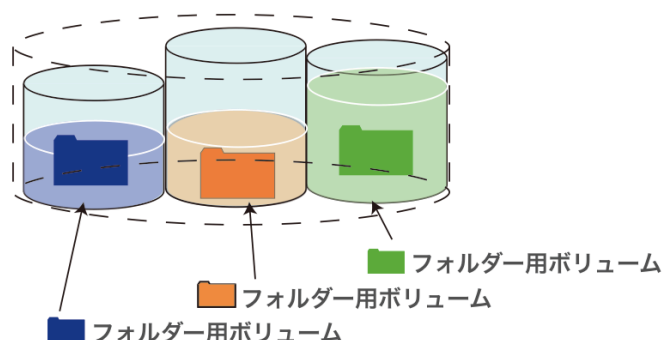
5.2.1 Fill-Upポリシー

Fill-Upポリシーにより、cluStoreはメンバーボリュームの優先順位に従い、設定されたソフトリミットまでデータを保存します。ソフトリミットに達すると、次の優先順位のメンバーボリュームにデータは保存されます。例えば、あるボリュームのデータを他のボリュームへ移行する様なデータのマイグレーションが必要な場合、データ移行元のドライブをLowプライオリティに設定し、プールに追加すると、Highプライオリティのメンバーボリュームに自動的にデータを移行させることができます。この間、移行中のデータに対してもクライアントからアクセスすることが可能です。



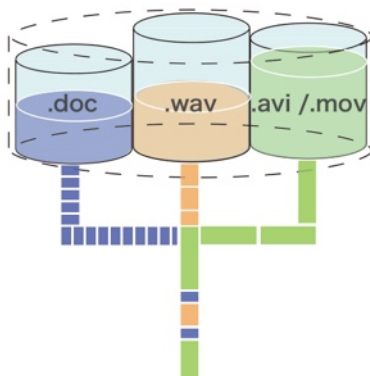
5.2.2 Harmonizeポリシー

このポリシーにより、ストレージプールに参加するメンバーボリュームに帰属していたフォルダーを移動せず、そのフォルダーのファイルはストレージプール内の同一ボリュームに書き込まれます。このメンバーボリュームをプールから取り外し、他の用途として使用する場合、その時点で、削除対象のストレージからデータ移動をする必要がありません。



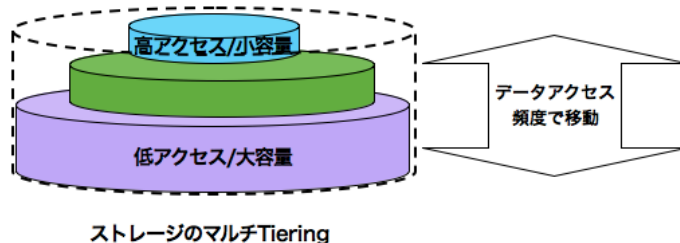
5.2.3 File Type ポリシー

このポリシーはストレージプールを構成するメンバーボリュームに対し、それぞれ帰属するファイルタイプを指定し、データのファイルタイプに応じてアロケーションを行なうことができます。例えば、このポリシーにより、高速大容量ストレージが必要なHDビデオ等のデータは、15,000rpmのドライブで構成されたメンバーボリュームに保存し、音声や、テキスト等のデータは7,200rpmのドライブで構成されたメンバーボリュームへ自動的に保存することができます。(ファイルタイプはファイルの拡張子で決定します。)



5.2.4 Last Accessed ポリシー

このポリシーによりストレージプールの階層化(Multi-Tiering)を実現します。SSDドライブ、15,000rpmドライブ、7200rpmドライブ、ネットワークドライブ等、ストレージへのアクセス性能に応じ、データアロケーションの優先度を付けることができます。データのアクセス頻度に応じ、ストレージプール内のメンバーボリューム間でデータが再配置されます。アクセス頻度の高いデータは高速なメンバーボリュームに、殆どアクセスのないデータに関しては、比較的低速な大容量メンバーストレージに保存されます。

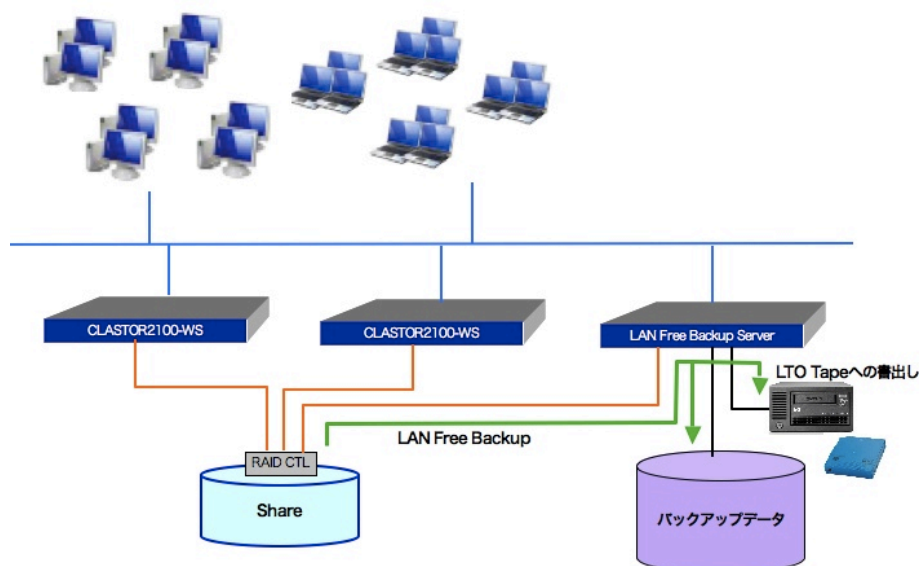


CluStoreストレージプールは以上の4つのポリシーで保存されたデータを管理することが可能です。この機能によりCLASTOR2100のストレージはサービスの停止、ストレージの初期化等を行うことなく、簡単に仮想ストレージプールを拡張したり、新旧ストレージの入れ替を行なうことが可能です。

6. データバックアップ

CLUSTOR2100-WSストレージに対するバックアップはネットワーク経由のバックアップ手法も可能ですが、弊社は効率的で、経済的なLAN-Free バックアップをお勧めします。

LAN-Freeバックアップは文字通り、LANに負荷をかけずに大容量のバックアップを実行できることを意味します。下図はその方法を示します。



右端に配置されたLAN-Free Backup ServerはネットワークストレージのデータをCLUSTOR2100-WSを介さず、直接SANボリュームをローカルディスクとしてマウントし、自分のローカルディスクのバックアップボリュームにバックアップデータを保存します。また、LTO-5のテープを使用することで、LTOFSを使用してデータの保存も可能です。ネットワークが多くのクライアントにファイル共有サービスを提供している間、バックアップ処理を高速なFC-8Gb経由で行なうことで、短時間で経済的なバックアップが実現します。バックアップはローカルディスク間で行なわれますので、ニーズとコストに応じ選択することが可能です。

7. ファイルサーバーのスケールダウンとコンポーネントリプレイス

以上でファイルサーバーのスケールアウトをご説明してきましたが、場合により、ファイルサーバーのスケールを縮小したり、古くなったサーバーや、ストレージを最新の物に交換したりする必要があります。通所、これらの作業を行なう場合、サービスの停止、および、データのマイグレーションという、時間と手間のかかる作業があり、大容量化したファイルサーバーでの作業はサービス停止とマイグレーションの為の経費がかかります。

今回ご提案したスケールアウトの方法では一つ一つのコンポーネントを組み合わせスケールアウトしていますので、簡単にスケールダウンすることが可能です。勿論、サーバーの交換はサービス停止を一切必要としません。また、仮想化されたストレージプール内のメンバーストレージも以下の手順で交換することが可能です。

1. 新規ストレージをサーバーに接続
2. 仮想ストレージプールに参加
3. 交換対象のストレージメンバーからデータの移動
4. 交換対象のストレージをサーバーから削除

以上の様に、通常ですと多くの時間とコストがかかるストレージの交換も仮想ボリュームを一切初期化することなく、シームレスに実現することが可能です。

8. まとめ

以上ファイルサーバーのスケールアウトに関しご紹介してまいりました。ユーザーにおけるスケールアウトのご要求に対し、ご紹介した種々のスケール方法で必要に応じ段階的に低コストストレージの拡張を実現してゆくことができます。以下にそれぞれの要件に対するスケールアウトの段階をまとめます。

1. ストレージ容量の随時追加 ▷ GTP拡張、または、cluStoreでの拡張
2. クライアント増加に追従したストレージアクセス性能の向上 ▷ ネットワークポートの仮想化、サーバーの多重化
3. ストレージの一元管理 ▷ metaSAN/cluStoreにより、ストレージを複数のサーバーから共有
4. サーバーの一元化、負荷バランス ▷ ネットワークロードバランサーを採用することで、IPアドレスレベルで一元化と負荷バランスを実現
5. 容易で高速なデータバックアップ ▷ metaSANを使用した LAN-Freeバックアップサーバーを建てることで、経済的で、効率的なバックアップシステムが構築可能

6. スケールダウン、コンポーネントのリプレイス ▷ 長期間使用するファイルサーバーの各コンポーネントは、総てEthernetや、FC等の標準のコミュニケーション用メディアにより結合されています。その為、一つ一つのサーバーは簡単に交換可能ですし、ストレージもデータ移動や初期化をする必要がなく、作業中でもサービスの継続が可能です。
7. 低い管理維持コスト ▷ CALフリーのWindows Storage Server 2008 R2をベースにし、Windows OSネイティブな機能を使用し、付加的にSAN共有を可能にするmetaSANや、ボリューム仮想化ソフトcluStoreを使用します。アプリケーションとの親和性が高く、将来の技術進化に追従し易く、投資の陳腐化を防ぐことが可能

現在の急激に増大する非構造化データに対し、現在多くのストレージ技術が紹介されています。一方で、今後、ハードウェアや、OS、アプリケーション等、IT分野の技術も急激に変化して行くことは議論の余地がありません。このような環境下で投資を保護し、一方で高い生産性を実現するにはできるだけオープンな技術を積極的に採用されることをお勧めいたします。弊社はWindows Server技術をコアに、OSネイティブな機能と、標準的な技術を駆使し、ユーザー要件を満たすソリューションをご提案してまいります。