重複排除の隠れたリスクの特定：

HYDRAstor はこの問題をどのように解決したか？

2006年10月

NEC

イントロダクション

重複排除技術(de－duplication）は、特にディスクベースのバックアップ製品に関して、最近大変注目されている。重複排除技術（コモナリティファクタリング（commonality factoring)、 非冗長ストレージ、重複データ削減としても知られる）は新しい、ユニークなデータのみを特定し、保存する。つまり、同じデータが既に保存されていれば、再度保存されるのではなくポインタによって参照される。実際、重複排除の効果は計り知れない。バックアップシステムは、その性質上、同じデータを何度も繰り返しバックアップする。重複排除技術の導入により、IT 部門は導入コスト削減、TCO 削減、および低回線容量環境での効果的なディザスタリカバリ目的の複製を可能にする等の効果が得られる。

第一世代のディスクベースバックアップ製品（VTL[1]等）はバックアップ操作を改善できるが、導入コストや TCO の高さが導入の障害となっていた。避けられない現実として、これらの製品は膨大な量の高価なディスクを追加で購入しなければならない。１か月分のバックアップを VTL で取ろうとすれば、実際のデータ生成量の４倍から５倍の容量の VTL を購入する必要がある(この比率は１ヶ月のデータ保持サイクルにおける三世代戦略に備えるためである。 数ヶ月分のデータ保持にはさらに多くのディスク容量が必要となる)。ディスク使用量の削減は、重複排除技術がビジネスモデルを変えるほどの変化をもたらすことのできる領域である。ディスクベースバックアップのすべての利点は、ディスク容量とこれにかかるコストが大きく削減されることによりもたらされる。

重複排除技術

バックアップはその性質上、ほとんどの部分について直近のバックアップからの変更がない場合であっても、データが繰り返しバックアップされる。重複排除アルゴリズムはこの冗長データを特定し、１回のみ保存する。データストリングが最初に一度保存されたら、その後は同じストリングについてはデータポインタにより保存場所を参照することによりバックアップされる。

> 第一世代の重複排除製品はバックアップとリストアの環境に重大なリスクをもたらした。１つの「ブロック」の喪失により、関係するすべてのバックアップイメージのリカバリが不可能になるのだ。

時が経つにつれ、重複排除技術は 20:1 のデータ削減を可能にする。つまり、バックアップ対象のデータのうち、実際にディスクに保存されるのはたった５％ということだ。例えば、重複排除を使えば、100TB のバックアップイメージがたった 5TB の物理ディスクに保存される。実際の削減率はデータの性質や IP 部門のバックアップ戦略によって異なるだろう。フルバックアップを毎日行う部門は増分バックアップを行う部門よりも大きな削減率をより早く得られるであろう。

[1] Virtual Tape Library、仮想テープライブラリ。

データ削減技術のビジネス価値には例えば以下がある。

導入コスト削減 – 必要なディスク容量が 80%削減される。データ削減技術を用いた製品は TB ベースでの大きな容量を低価格で実現。

TOO 削減 – データ削減ファクタは時間とともに増大するので、物理ディスクの使用量はデータ削減技術を使用しない製品に比べて増加の速度が抑えられる。

費用効率の高いリモートレプリケーション – 標準技術を用いたリモートデータレプリケーションはデータ量が帯域幅の制限を越えるため、多くのケースで現実的ではない。データ削減技術を用いれば、重複削除済みのユニークなデータだけが転送されるため転送データの 90-95%が削減され、WAN 経由でのレプリケーションが現実的なものとなる。

重複排除の利点は明確だが、第一世代の製品に見られる繋ぎ合わせの実装は重大なリスクをもたらす。重複排除はソフトウェアベースの圧縮のため CPU に負荷がかかり、パフォーマンスは 40%低下しうる。第一世代の重複排除製品では、処理能力は 65%も減少しうる。このペナルティにより、ディスクがテープより速いというメリットは大きく打ち消される。

しかし、第一世代の重複排除製品に関する最も深刻なリスクは、バックアップを喪失し、リカバリ不可能となるリスクである。複製データ削減率が 20:1 以上に達するため、数十、数百のバックアップイメージが 1 つのディスクブロックの復元に依存する。もしそのディスクブロックが失われたら、関連するバックアップイメージはすべて復元不可能となりうる。ディスクが機能不全となった場合の影響はもっと悪く、VTL やディスクベース装置に保存されたバックアップすべての利用に影響を与えうる。

さらに、大容量の(S)ATA ディスクドライブはリビルドに 24 時間かそれ以上かかることがあり、リビルド処理は I/O に多大な負荷をかけるため RAID グループの全ディスクに影響する。リビルド時の処理能力の大幅な低下は大規模データ修復の RTO[2]実現を妨げるだけでなくバックアップウィンドウに間に合わない結果にもつながりかねない。

RAIDで十分に保護できるか？

第一世代の VTL とディスクベース製品のほとんどはディスク障害対策として RAID を使用している。RAID5 は 1 つのパリティを使用し、1 つのディスク障害からリカバリ可能で、オーバーヘッドは 20%である。しかし RAID5 には重大な欠点がある。同じ RAID グループで 2 つのディスク障害が起こった場合、そのグループ内のデータは復元不可能である。データ回復力を高めるため、ディスクベースバックアップ製品の中には RAID 5 の代わりに RAID 6 を使うものもある。RAID 6 はダブルパリティを用いており、同一 RAID グループにおける 2 つのディスク障害からのリカバリが可能である。しかし、ストレージのオーバーヘッドは RAID 6 で 35-40%と、RAID 5 に比べ大変大きい。

RAIDは重複排除を用いない環境での保護には十分かもしれないが、重複排除の環境では不十分である。重複排除の環境では、1 つのディスクブロックが数百のバックアップイメージで使用されるデータを格納する場合がある。もしそのブロックが失われたら、すべての関連イメージがリカバリ不可能となる。さらに、典型的なストレージアレイの300GB SATAドライブのリビルドには24時間かかりうる：500GBであればもっと長くかかる。ドライブ容量はテラバイトにまで増大しており、リビルドに時間がかかれば、リビルドが完了する前に統計的には同一RAIDグループ内に第二、第三の障害が起こる可能性がある。

さらに、RAID 技術では、ディスクのリビルド時に深刻なパフォーマンス低下が起こる。RAID 5/6 グループのリビルドは I/O に大きな負荷をかけ、グループの全ディスクに影響する。リビルド時の大きなパフォーマンス低下は、同時にリストアが行われる場合、RTO を妨害しうる。長いリビルド時間およびパフォーマンスロスによりバックアップウィンドウ枠を超えてしまい、データをさらに危険な状態に

[2] Recovery Time Objective：システム復旧にかかるまでの目標時間

さらすことになる。RAID 6 の多くのアレイはダブルパリティリカバリ中に RAID に読み書きすることを推奨していない。RAID 6 で繰り返されるリカバリ処理が重大なパフォーマンス低下を引き起こすためだ。この処理を中断すればデータはリカバリ不能に陥る可能性がある。

重複排除の環境には RAID 搭載システムでは不十分:
ーデータ回復力（resiliency）の制限
ーリビルド時のパフォーマンス低下

RAID 5 同様、RAID 6 システムでは追加のパリティ計算によるライトペナルティが生じる。このペナルティは VTL やディスクアプライアンスの全体のパフォーマンスを低下させる。このペナルティは単なる 2 つのパリティ計算よりも悪い: なぜならダブルパリティ計算は標準的な CPU 上には簡単にはマップできず、表索引を必要とするからだ。[3]

理想的な解は、2 つ以上のディスク障害に耐え、ライトペナルティもなく、ディスクのリビルド時のパフォーマンス低下がなく、RAID 5 以上のディスクスペースは使わないというものだろう。NEC はこの 3 年間、まさにこのような解を求めて研究を続け、そして HYDRAstor にたどり着いた。

HYDRAstor: 次世代ディスクバックアップとリストアソリューション

HYDRAstor は、現在の VTL や disk-as-disk 装置におけるリスクや制限なしに、バックアップやリストアの抱える主要な問題を解決するために広範囲にわたり徹底的に設計された次世代のディスクベースデータ保護装置である。HYDRAstor 独自のアーキテクチャにより、シングルインスタンスで 200 MB/sec から 20,000 MB/sec をはるかに上回る総処理能力を誇る高性能バックアップおよびリストアを可能にする。

HYDRAstor のディスク容量はテラバイト（TB）からペタバイト（PB）まで簡単に無停止で拡張できる。1 システムで数ヶ月または数年分のバックアップデータを$0.90/GB 以下で保存でき、これは典型的なテープシステムよりも低コストである。このスケーラビリティと低コストは、HYDRAstor 独自のグリッドベース ストレージ アーキテクチャおよび特許出願中の DataRedux™と Distributed Resilient Data™ (DRD)技術によりもたらされる。第一世代の VTL およびディスクアプライアンスとは異なり、HYDRAstor は重複排除環境におけるデータ回復力の増大という要求を満たす唯一のソリューションである。

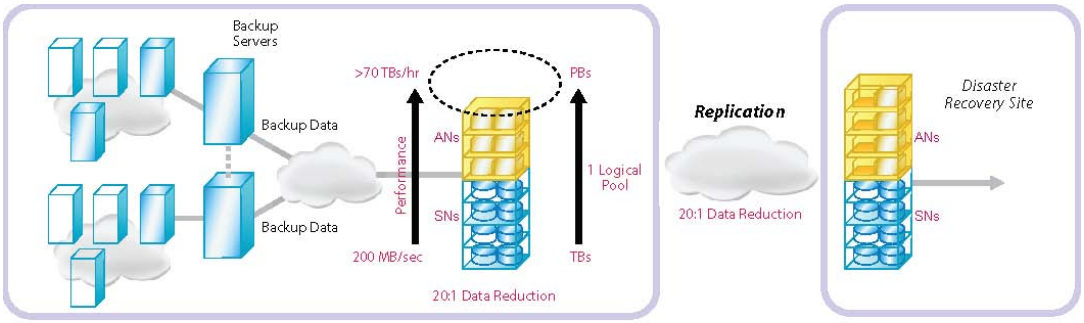

Figure 1. HYDRAstor Next-generation Disk-based Data Protection

[3] H. Peter Arvin: “The mathematics of RAID 6.”

グリッドベースストレージアーキテクチャ

HYDRAstor のグリッドベースのストレージアーキテクチャは他のディスクベースバックアップシステムとは異なる。このアーキテクチャは、HYDRAstor のインテリジェント OS と、最適な組み合わせの業界標準サーバーやディスクとのコンビネーションで、最高のデータ回復力および事実上無制限の性能と容量拡大を提供する。重複排除されたデータを危険にさらす第一世代の VTL や disk-as-disk バックアップアプライアンスとは異なり、HYDRAstor は特許出願中の DRD 技術を通じて高度なデータ回復力を実現する

ターンキーソリューションとして提供されているため、HYDRAstor の革新的なグリッドベースアーキテクチャは、ユーザが追加のリソースをどこで必要とされていようと適用できるようにすることにより、現在製品の操作上の制限を排除している。さらに、これらのリソースはおおきな労力を要する設定なしに追加でき、HYDRAstor としてのチューニングや干渉も自己調整され、自己回復力を持つシステムである。性能拡張性とストレージ容量はビジネスニーズに基づき独立して増加させることができ、1 つのコンポーネントの追加によりシステム全体の性能と容量を増加させることができる。HYDRAstor はこの拡張性を、**アクセラレータ**ノード(Accelerator Nodes)および ストレージノード(Storage Nodes)を通じて実現する。

アクセラレータノード — 性能の拡張

図 1 に示す HYDRAstor アクセラレータノード (ANs)は、コモディティなサーバーである。AN は 1 つ以上のバックアップサーバーとギガビットイーサネット(GbE)で接続され、CIFS と NFS の両方をサポートする。これらは拡張性とデータ回復力のためにクラスター化され、各 AN は 100 MB/sec 以上の処理能力を持つ。AN は、負荷を各 CPU に分散させるのみならず、どのノードも SOPF[4]にならないことを保証する密結合分散ファイルシステムと共に動作する。もし 1 つかそれ以上の AN に障害が発生したら、残りの AN がすべての負荷を引き受ける。AN は性能向上のため、無停止で拡張できる。例えば、10 の AN で 1000 MB/sec を超えるバックアップおよびリストア処理能力を提供する。

HYDRAstor 環境で必要とされる AN の数を決めるには、以下の計算式を用いる：

$$\text{AN\#} = \frac{\text{Aggregate peak data transfer in MB/sec}}{\text{100 MB/sec}}$$

ストレージノード — 容量の拡張

図 2 に示すストレージノード (SNs) もまた TB 単位のストレージ容量を持つコモディティなサーバーで構築される。SN はバックアップイメージのためのディスク容量を提供し、HYDRAstor OS により管理される専用ネットワーク経由でアクセラレータノードに接続される。HYDRAstor OS は複数 SN 間のストレージを仮想化し、どの AN からもアクセスできる 1 つの論理プールを作成する。HYDRAstor の革新的な容量の仮想化により、プロビジョニング作業が不要となる。HYDRAstor では、ストレージ管理者は LUN、ボリューム、ファイルシステムを作成したりサイズを決めたりする必要がない。HYDRAstor は、SN が追加されると、自動的に利用可能容量内での既存データの負荷バランスをとり、性能と活用の最適化を行う。HYDRAstor の分散化されたグリッドアーキテクチャーが、バックアップやリストアの性能に影響を与えることなく、これらの利点を提供する。

---

[4] a single point of failure システム上のあるコンポーネントが異常を来たすと、そのシステム全体が障害に陥ってしまうようなコンポーネントの総称

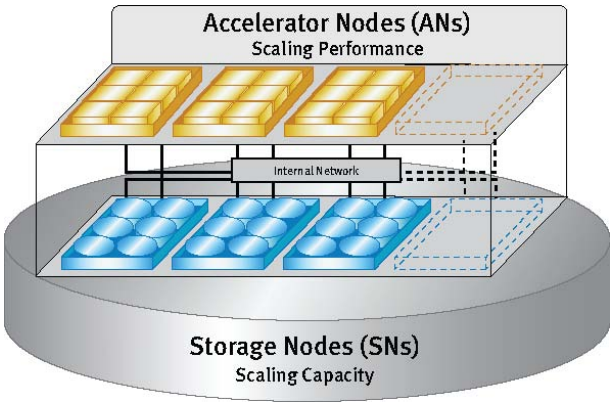

Figure 2. HYDRAstor グリッドアーキテクチャー

DataRedux テクノロジー — ストレージの効率性

HYDRAstor の DataRedux 技術は重複排除を実現するものだが、これはデータを可変長の「データチャンク」に分離するという NEC の特許出願中の技術「チャンキング(chunking)」から始まった。ハッシュ法により、そのデータチャンクが既に保存されているものかどうかを判断する。もしそのデータチャンクが既に保存されていれば、そのデータのアドレスを示すためのポインターが用いられる。もしデータが保存されていなければ、そのチャンクがディスクに保存される前に、以下で説明するデータ回復性（data resiliency）が適用される。

重複排除効果はデータチャンキングアルゴリズムにより推進される。チャンキングをまったく用いない製品や、データを固定サイズに分割する製品ではデータ削減率は低い。チャンキングをまったく用いない製品ではファイルの１バイトしか変更していなくてもファイル全体が再度保存される。データを固定サイズに分割する製品では、変更されたバイトを含むチャンクの後に続くチャンクがすべて保存される。しかし HYDRAstor では高度なアルゴリズムと可変長チャンクの利用により冗長データの発見の機会を最大化している。データに対する真の変更を含むチャンクのみがユニークなものとして特定される。これらのチャンクはディスクに保存される前に物理的に圧縮される。

重複排除と物理的な圧縮は初期のアレイベースの VTL 製品では性能ボトルネックになるが、HYDRAstor のグリッドベースアーキテクチャでは多くのノードに負荷を分散することができ、これにより性能低下とアイドルノードを防いでいる。

Distributed Resilient Data テクノロジー — 復旧の確実性

データ削減技術と共に使われた場合の RAID 5 や 6 の制限とリスクを解決するため、HYDRAstor は特許出願中の Distributed Resilient Data (DRD)技術を導入する。DRD は、３つ以上のディスク障害からのデータ保護を、RAID ライトペナルティなしに、リビルド中の性能低下を引き起こすことなく、リビルド時間を長時間化させることなく、RAID 6 よりも少ないストレージオーバーヘッドにて実現する。

HYDRAstor のデータ削減技術によりユニークなデータチャンクが特定された後、DRD はそのチャンクを、決められた数のデータフラグメントに分割する。オリジナルのデータフラグメントのデフォルト数は９である。求められるデータ回復レベルにより（デフォルトは３）、DRD はそのデータ回復力レベルに合うパリティフラグメント数を、オリジナルの９つのデータフラグメントにのみ基づいて計算する（言い換えれば、DRD はデータではなくパリティに基づきパリティフラグメント数を計算する）。故に、パリティを読んで再計算し再書き込みを行わねばならない初期の VTL および RAID を使った装置と異なり、HYDRAstor にはこの「パリティペナルティ」が発生しない。DRD は１２のフラグメント（９つのオリジナルフラグメント＋３つのパリティフラグメント）を、その構成内に存在する SN に最大限分散する。12 未満の SN しかないシステムでは、複数のフラグメントが同一の SN に保存される。この場合、

HYDRAstor はこれらのフラグメントを異なるディスクに分散させる。最小構成である 4 SN の場合、データ回復力を最大にするために、３つのフラグメントが各 SN の異なるディスクに書き込まれる。

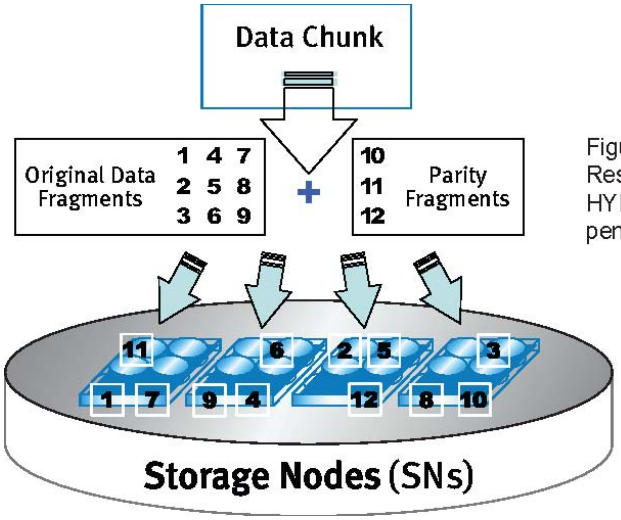

Figure 3: Distributed Resilient Data (DRD), HYDRAstor's Patent-pending Technology

デフォルトの３パリティフラグメント設定において、１つのデータチャンクは１２フラグメントのうちのどの９フラグメントからでも再生できる。つまり、同一グループで３つのフラグメントが同時に喪失しても、データの完全性は損なわれないことを意味する。HYDRAstor の最小構成である 4 SN の環境において、１つの SN 障害（５つのデータディスク）または複数 SN にまたがる３つのディスク障害が起こった場合でも全てのデータを保護できる。

１２以上の SN を含む大規模な HYDRAstor 環境では、１つのチャンクの１２のフラグメントのうち各 SN が保存するのは１つ以下である。従って、デフォルト設定で３つ以上の SN 障害からデータを守ることができる。クリティカルなデータであれば、管理者はもっとデータ回復力の高い設定にすることができる。データ回復力の設定に基づき、HYDRAstor は自動的に適切な数のオリジナルデータフラグメントを計算し、必要な数のパリティフラグメントを作成する。これらのフラグメントはデータ回復力が最大になるように常に分散化される。

HYDRAstor には書き込み性能劣化は起きない。 リビルド実行中であっても、処理は中断されず、影響も受けない。

このデータ回復力の追加により、HYDRAstor に多大なストレージオーバーヘッドが必要になると思われるかもしれない。しかしそうではない。HYDRAstor のデフォルトアーキテクチャに対するディスクスペースオーバーヘッドは 3/12 すなわち 25%であり、RAID 5 のオーバーヘッド（約 20%）よりもほんの少しだけ多く、RAID 6（約 35-40%）より少ない。一方でデータ回復力は RAID 5 よりも 300%、RAID 6 よりも 50%高いのだ。

RAID 5 や 6 と異なり、HYDRAstor はディスクのリビルド中であっても性能低下を引き起こすことなくバックアップやリストアの実行が可能である。フラグメントは十分な処理能力を持つ複数の Storage Node にまたがって分散されているため、バックアップやリストアの実行はリビルドの際にも性能低下をもたらさない。障害が発生したコンポーネントは自動的に「発見」され、バックグラウンドで自動的にリビルドが始まる(Figure 4 参照)。

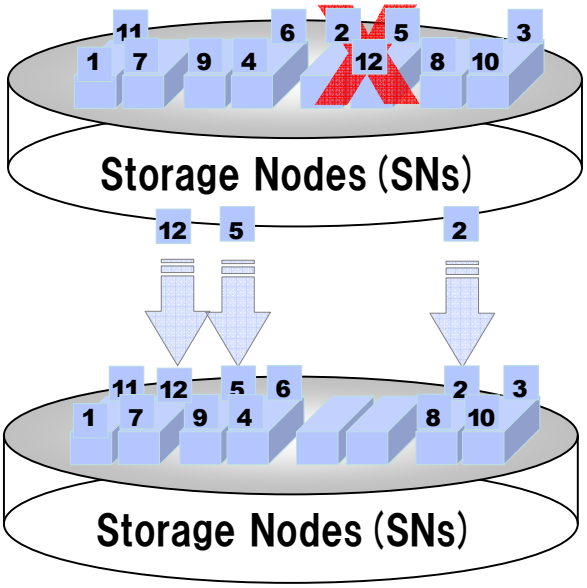

Figure 4: Data Resiliency実行中。ストレージノードに重大な障害が起こった場合、HYDRAstor は直ちに障害を発見し、リビルドし、他の運用可能なストレージノードにデータを書き込む。

No Single Point of Failure

HYDRAstor では SPOF は存在しない。
特許出願中の DRD 技術により、RAID 5 と同等のディスクスペースオーバーヘッドでありながら RAID 5 より 300%も高いデータ回復力を実現する。

第一世代の製品の固定ハッシュテーブルは、通常、複数障害点（multiple points of failure）を持つ。１つのディスクアレイをベースとしているために、どれか１つのコンポーネント（マザーボード、RAID カード等）が故障するとシステム全体の障害という結果になってしまう。HYDRAstor においてこれと同じ状態は１つの AN または１つの SN に障害が起こった場合であるが、既に示したように、この場合でも HYDRAstor 内のデータのアクセスのしやすさに影響を与えない (Figure 4 参照)。

第一世代の製品は、通常、複数障害点を持つ。第一世代製品の固定ハッシュテーブルは SPOF ともなりうる。もしハッシュテーブルが喪失したら、データはリカバリ不能になる。HYDRAstor はこのような集中的なテーブルを持たない。HYDRAstor はハッシュテーブルを、複数の SN に障害が起こった場合でもハッシュデータを喪失しないようにすべてのストレージノードに分散させる。

HYDRAstor 利点

HYDRAstor 特有の利点をまとめると以下のようになる。

低コスト– HYDRAstor'の安全なデータ削減能力により、テープと同等の価格でのディスクベースバックアップを実現できる

データ回復力– 特許出願中の Distributed Resilient Data (DRD) 技術により、利用機関は、データ喪失やリカバリ不能の心配をすることなく、安全に HYDRAstor の重複排除技術を十分

に活用できる。

動的拡張性– アクセラレータノード（ANs）とストレージノード（SNs）は利用機関のニーズを満たすよう性能と容量を上げるため、独立して無停止で追加することができる。このような、簡単で費用効率の高いインフラ調整能力を提供する製品は他にはない。

性能ペナルティなし – RAID システムとは異なり、HYDRAstor はディスクのリビルド実行中でも、バックアップやリストア操作を危険にさらす性能低下が起こらない。

SPOF なし – HYDRAstor は、冗長性の高いターンキー装置として提供され、SPOF となる集中化されたリソースを持たない。

HYDRAstor はデータ削減の利点である低い TCO と、第一世代の VTL や disk-as-disk 装置では見られない高いシステム利用可能性を提供する。エンタープライズデータセンターは、クリティカルなデータの保護に関して HYDRAstor が発揮する優れた機能を評価するであろう。