ディペンダブルなITインフラを支える次世代データリポジトリ技術

セザーリ デュブニッキー・クリスチャン ウングルヌ・クリスチャン トールグ

要　旨

企業の基幹業務の電子化の進展とともに、データ保管の重要性が高まっています。NECではテープの代わりにディスクへデータを保管する次世代データリポジトリシステムを開発しました。本システムはディスクの特長である高性能/取扱いやすさを備えながらも、テープ並みのコストを実現しています。そのキーとなる「重複排除技術」「耐障害技術」「スケーラブル技術/自律管理技術」について紹介します。

キーワード

●ストレージ　●データリポジトリ　●重複排除　●超耐障害　●スケーラブル　●バックアップ　●ディザスタリカバリ　●分散格納

1. はじめに

基幹業務の電子化が進むにつれ、電子データの重要性がますます高まってきています。これらの電子データは基幹業務を継続する上で大変重要なものであり、万一の場合に備えてテープやディスクにバックアップをとってデータを保護しています。しかしながら、従来は性能/コストパフォーマンス/使いやすさを満たすようなデータ保護方法がありませんでした。

本稿では、上記を満たし、データの増加とともにシステムを拡張することが可能なデータ保護方法としての次世代データリポジトリシステム技術について紹介します。

2. 背景および従来手法の課題

基幹業務で使用するような重要なデータを保護するために、通常ではアプリケーションサーバや主要なストレージ上のデータを週末一晩かけてテープにバックアップするという手法がとられています。典型的な利用形態としては週末にフルバックアップをとり、月曜から金曜の間は増分バックアップまたは差分バックアップをとる方法が用いられています。データを復元する必要があるときは、フルバックアップをリストアした上で、必要なところまで差分バックアップを適用していきます。このように、テープによるバックアップはストレージ管理者に大きな負担がかかっていました。

一方、近年のデータの爆発的な増大や、データを使うシステムの稼働時間の長時間化に伴い、バックアップに使える時間はますます短くなってきています。このため、テープを用いた従来の手法では、要求された時間内にバックアップが完了しないケースが増えてきています。

このような理由から、バックアップ業務にディスクを導入する企業が増えてきています。ディスクを利用することで次のような利点が生まれます。

① 短い時間でバックアップが完了する
② 保存されたデータはランダムアクセスできるのでリストアしやすい
③ 保存されたデータが正しく書かれているかをチェックしやすい

しかしながら、ディスクはテープに比べて高価なため、多くの企業においてテープへ格納するまでの一時格納用として使用されています。これはディスクの利便性とテープの持つコスト面での利点を組み合わせるためですが、管理面では複雑度が増し、障害に陥る可能性も高くなってしまいます。このため、双方の利点を持ちながらも管理の簡単なバックアップシステム/データリポジトリシステムが望まれていました。

3. 次世代データリポジトリシステム

ここでは、次世代データリポジトリシステムのコンセプトとそのアーキテクチャについて紹介します。

3.1 システムコンセプト

次世代データリポジトリシステムは、次のようなコンセプトのもとに開発を進めてきました。

① ディスク方式の特長である高性能性と管理の容易性とを持つこと
② ディスクを使いながらもテープ装置並みのコストパフォーマンスを持つこと
③ 複数部品の同時障害でもデータを失わない高信頼性を持つこと
④ 小さな構成でスタートし、データの増大とともにシステムを拡張できるスケーラビリティを持つこと

3.2 システムアーキテクチャ

前述のコンセプトを実現する次世代データリポジトリシステムのアーキテクチャを**図1**に示します。

従来のストレージは、RAID (Redundant Array of Independent Disks) を構成するディスク群とそれらを制御するコントローラの対からできており、コントローラは専用LSIか処理能力の高くないCPUで構成された自由度が比較的低い専用装置として作られてきました。それに対して、本システムは複数のサーバ(ノード)から構成されており、専用コントローラを用いた従来のストレージに比べて高い処理能力と柔軟性を備えています。

このようなアーキテクチャを採用することにより、本システムは次のような特徴を備えています。

① テープを使わずディスクのみを用いることにより、高性能性と管理の容易性を実現
② ノードが持つCPUを利用してデータの重複排除を行うことによりテープ装置並みのコストパフォーマンスを実現
③ ノードが持つCPUを利用してRAIDを超える耐障害性を実現
④ ノードのCPUにより高い自律管理機能を持ち、システムダウンさせずにノードの追加や切り離しを行うほか、それに伴う自律データ再配置やスケーラビリティを実現

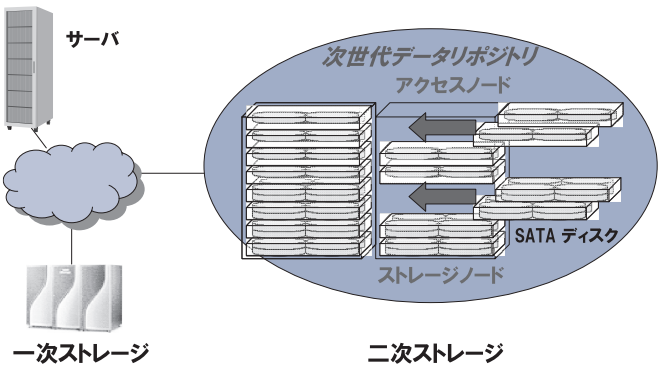

図1　次世代データリポジトリシステムのアーキテクチャ

4. 次世代データリポジトリシステムを支える技術

続いて上記の特長を実現するコア技術について紹介します。

4.1 重複排除技術

本システムでは、データが重複するかどうかをチェックし、すでにシステムに書かれたデータと重複していればシステムに書き込まないことにより、データの格納効率を大幅に高めています(**図2**)。

動作は次のようになります。まず、格納すべきデータがサーバから送られてくると、データの内容を適切なブロックに分割し、ハッシュキーを割り当てます。次に、各ブロックがすでに書き込まれたデータと重複しているかどうかをハッシュキーによりチェックします。もし重複していればそのデータを書き込

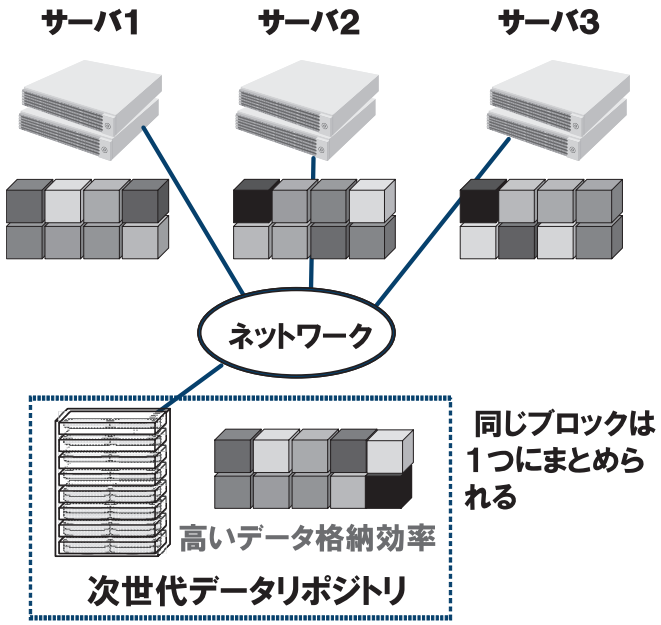

図2　重複排除技術による格納効率の大幅な向上

む代わりにすでにあるデータへポインタを張ります。重複していなければそのデータは決められた規則によって格納されます。

フルバックアップを複数セットとる場合、データの大半が重複しているため、この重複排除技術によりディスクの格納効率を大幅に高めることができます。その結果システムの容量当たりのコストをテープ並みに抑えることが可能となっています。

4.2 耐障害技術

前述のように、データをブロックに区切って同じブロックを複数のデータが共有する場合、1つのブロックが喪失した場合でもその影響は広範囲にわたってしまいます。そこで本システムでは、RAIDのパリティよりもさらに強力なデータ修復機構を採用しました(図3)。

ブロックを複数のフラグメントデータに分割し、冗長データを付加した上でストレージノードに分散格納します。万一ストレージノードに障害が生じ、複数のフラグメントデータが欠落したとしても、それが冗長フラグメントの数以下であれば強力なデータ修復機能により、元のブロックを復元します。

たとえば、ブロックを12個のフラグメントに分割するとともに、8個の冗長フラグメントを生成するとします。これら合計20個のフラグメントはストレージノードに分散して格納されます。ここでストレージノードのディスクに障害が起こるかストレージノードがダウンするかして複数のフラグメントが読み出せなくなったとしても、それが8個以内であればシステムは元のブロックデータを復元することが可能となっています。

また、ディレクトリ情報も冗長分散格納されているため、一部のディレクトリ情報に障害が起きても全体としては動き続けることが可能となっています。

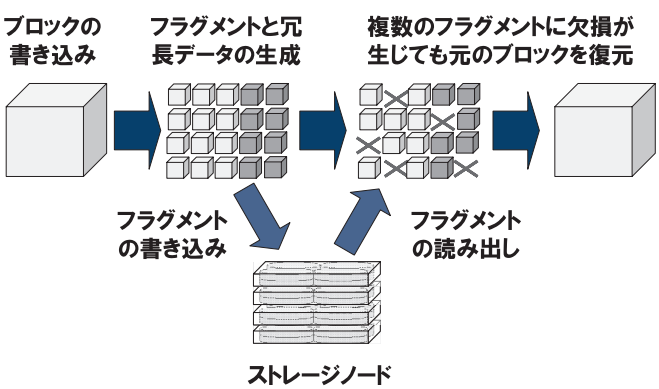

図3　強力なデータ修復機構によるデータ保護

4.3 スケーラブル技術/自律管理技術

データの増加に伴ってデータ容量や処理スピードが不足した場合、本システムではノードを追加することで解決できます。処理スピードに不足がある場合は主にアクセスノードを、データ容量に不足がある場合はストレージノードを追加します。

重複排除に関しては、ノードをまたがって重複排除するグローバル重複排除機能を備えています。これにより、データが増大して新しいノードを追加した場合でも、新しいノードを含めたすべてのストレージノードで重複排除のチェックが行われるため、データの格納効率が落ちることはありません。

ノードの追加に際しては、システムが新規ノードを自動的に認識し、オンラインでシステム内に組み込みます。システムに組み込まれた後は、システム内の負荷の偏りを自動的にバランスさせます。また、システムは障害ノードあるいはアクセス不能なノードを自動的に認識し、他のノード上にデータを再構築します。

5. おわりに

データを使いやすく安全に格納する次世代データリポジトリシステムおよびそのコア技術について紹介してきました。本システムでは、ディスクの特長である高性能/取扱いやすさをテープ並みのコストで実現することが可能となっています。また、自律管理機能により、管理の手間を大幅に削減できます。このように、本次世代データリポジトリシステムは、本当の意味での自律型グリッドストレージを様々な革新的独自技術を通して実現した試みであるといえます。

今後も革新技術を通して、ユーザにとって価値あるストレージを提供していきます。

執筆者プロフィール

セザーリ デュブニッキー
NEC Laboratories America

クリスチャン ウングルヌ
NEC Laboratories America

クリスチャン トールグ
NEC Laboratories America