C_020

大規模ディザスタリカバリシステム向け非同期リモートコピーの研究

Asynchronous Remote Copy for Very Large Disaster Recovery System

出口 彰† 二瀬 健太† 荒川 敬史† 山本 康友†
Akira DEGUCHI Kenta NINOSE Hiroshi ARAKAWA Yasutomo YAMAMOTO

1. はじめに

地震やテロによる企業情報システムの停止は、ビジネスに大きな影響を与える。このような背景から、SEC(米国証券取引委員会)は金融機関に対し、遠隔地にバックアップサイトを設け、災害時に、バックアップサイトで迅速に業務再開すること(ディザスタリカバリ)を求める指針を発表した[1]。これを受け、金融機関を中心として、ディザスタリカバリへの対応が進められている。このディザスタリカバリを支援する機能として、ストレージでは、非同期リモートコピー機能を提供してきた。

近年、金融機関などでのシステム大規模化に伴い、多くの記憶領域や、データ処理能力が必要となり、複数台のストレージを使用する構成が一般的になりつつある。このため、複数台対複数台ストレージ間での非同期リモートコピーによりディザスタリカバリシステムを構築するニーズが高まっている。しかし、災害時に副サイトで正常に業務再開するには正・副ストレージ間のデータ更新順序を同一に保つ必要があるにも関わらず、従来は1台対1台の構成でしかデータ更新順序が保証できなかった。

金融機関などの大規模システムがメインフレーム(MF)で構成されることが多いことから、本研究では、MFを対象とした複数台対複数台ストレージ間の非同期リモートコピーにおいて、データ更新順序を保つ方式を検討した。

2. 非同期リモートコピーと問題点

2.1 非同期リモートコピー

本節では、ストレージで従来から提供してきた1台対1台構成の非同期リモートコピーについて説明する。

非同期リモートコピーとは、業務サイトの正ストレージにホストから書き込まれるデータを、ライト要求とは非同期にバックアップサイトの副ストレージへ転送する機能であり、サイト間の距離を長距離にしても、ライト要求のレスポンスタイムへの影響が小さいという特徴を持つ。

災害発生時に、バックアップサイトで副ストレージを用いた業務再開を可能とするために、非同期リモートコピーでは、副ストレージでのデータ更新順序を、正ストレージと同一に保つように制御している。

以下に、従来の非同期リモートコピーの動作を説明する[2]。

(1)ホストが正ストレージに対してライトを発行する。

(2)正ストレージは、ライトを受領するとシーケンシャルな番号(シーケンス番号)をライトに対して割り当てる。次に、シーケンス番号とライトデータなどからなる転送データを作成するとともに、ホストに対してライト完了を報告する。

(3)正ストレージは、複数個の転送データを、副ストレージへ並列転送する。

(4)副ストレージは、前回最後にボリュームに書き込んだ転送データのシーケンス番号からシーケンス番号が連続している部分の転送データを特定し、それら転送データのシーケンス番号の最大値を決定する。

(5)最後に、副ストレージは(4)で決定したシーケンス番号までの転送データをシーケンス番号順にボリュームに書き込む。

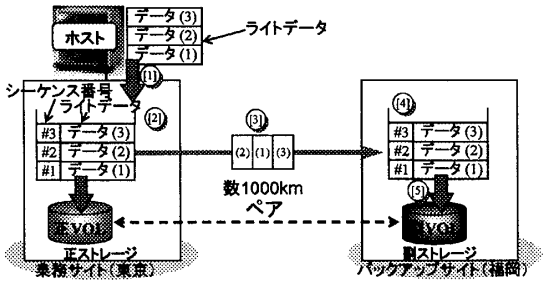

図1：非同期リモートコピーの動作概要

2.2 課題と目的

複数台対複数台の構成では、ホストからのライト要求が、複数台の正ストレージへ発行されたとしても、これらライト要求の順序を保ち、副ストレージへデータを書き込む必要がある。

複数台対複数台の構成に従来の非同期リモートコピーを単純に適用した場合、あるライトAに後続するライトBが先行するライトAとは異なる正ストレージへ発行されると、ライトBがライトAを追い越し副ストレージへ書き込まれるという状況が容易に起こる。

これを防止するためには、副ストレージ側で副ストレージ間のデータ更新順序を意識して転送データをボリュームに書き込む必要がある。本研究は、複数台対複数台ストレージ間の非同期リモートコピーにおけるデータ更新順序保証方式の確立を目的とする。

3. 方式提案

3.1 要件

複数台対複数台構成における1台の正ストレージと、その正ストレージに対応する副ストレージの間では、従来の非同期リモートコピー技術により順序を保証することができる。このため、副ストレージ側で、副ストレージ間の順序を制御することを、データ更新順序保証方式検討の基本方針とする。複数台対複数台ストレージ間の非同期リモートコピーの要件は次の二つである。

(1)副ストレージ側で、複数台の副ストレージ間のデータ更新順序を保つ。

(2)ストレージ台数に応じた性能を発揮する。

要件(1)を満たすためには、副ストレージ側で、複数台の副ストレージに転送される転送データの順序関係を把握するための機能(機能1)と、順序関係を把握した転送データを、複数台の副ストレージ間で順序を保ちボリュームへ書き込む機能(機能2)が必要となる。

また、要件(2)を満たすために、上記の2機能の処理を、各ストレージが並列に実行する必要がある。

† (株) 日立製作所 システム開発研究所
Systems Development Lab., Hitachi, Ltd.

次節以降、機能 1、機能 2 それぞれの実現方式を述べる。

3.2 機能 1 の実現方式

複数台の副ストレージ間で、正ストレージから転送される転送データの順序関係を把握するために、MF ホストがライト要求とともに正ストレージへ送信するタイムスタンプを利用する。具体的には、MF ホストがライト要求に付与するタイムスタンプを、副ストレージへ転送する転送データに格納する。これにより、副ストレージ側で、複数台の副ストレージ間での転送データの順序関係を把握することが可能となる。この処理は、転送データにタイムスタンプを格納すること以外は、1 台対 1 台構成の非同期リモートコピーと同じである。よって、各正ストレージが並列に実施可能な処理であり、性能への影響はないといえる。

3.3 機能 2 の実現方式

本節では、順序保証を優先した完全順序保証方式と、性能を優先した書き込み可能時刻指示方式を提案する。

3.3.1 完全順序保証方式

マスタ副ストレージが全転送データの中から最も古い転送データを特定し、当該転送データの書き込みを実施する「完全順序保証方式」を説明する。具体的には、マスタ副ストレージが、各副ストレージの一番古い転送データのタイムスタンプを収集し、その中から更に一番古いタイムスタンプを決定する。そして、当該タイムスタンプを持つ転送データの書き込みを、最古の転送データを持つ副ストレージへ指示する。本方式では、一つずつ転送データをボリュームに書き込むため、任意の時点でデータ更新順序を保証している。その反面、副ストレージでの転送データ書き込みが直列化されてしまう。

3.3.2 書き込み可能時刻指示方式

本節では、各副ストレージが並列に転送データを書き込む「書き込み可能時刻指示方式」について説明する。

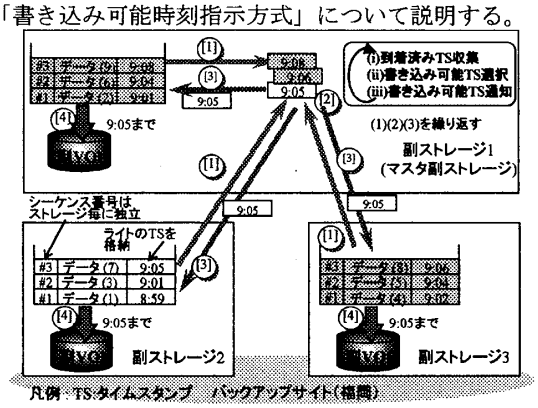

図 2：書き込み可能時刻指示方式の動作概要

バックアップサイトのマスタ副ストレージが、自身も含め、全ての副ストレージから、「副ストレージへ到着済みであり、かつ、シーケンス番号が連続している転送データ列のタイムスタンプの最大値」を収集する(図 2 動作(1))。そして、収集したタイムスタンプのうちから最も古いタイムスタンプ(書き込み可能タイムスタンプ)を選択し(図 2 動作(2))、全ての副ストレージへ通知することで、書き込み可能タイムスタンプまでの転送データ書き込みを許可する(図 2 動作(3))。各副ストレージは、マスタ副ストレージから指示されている書き込み可能タイムスタンプまでの転送データをボリュームへ書き込む(図 2 動作(4))。

書き込み可能時刻指示方式では、各副ストレージの転送データ書き込みが並列に実施できるというメリットがある。その反面、各副ストレージが並列に転送データを書き込むため、各副ストレージが書き込み可能タイムスタンプまでの転送データ書き込みを完了した時点でしかデータ更新順序を保証しない。

4. 検証

完全順序保証方式と書き込み可能時刻指示方式を 3.1 節に述べた要件に基づき比較する。表 1 に比較結果をまとめ、以降に説明する。

表 1：比較結果

要件 / 方式	データ更新順序	台数に応じた性能
完全順序保証方式	○(常に保証)	×(1 台構成時と同等の性能)
書き込み可能時刻指示方式	○(副ストレージを使用する時に保証した状態を提供する)	○(台数に応じた性能を発揮)

完全順序保証方式は、常にデータ更新順序を保証する。しかし、転送データの書き込みが直列化されるためストレージ台数に比例した性能は発揮できない。

書き込み可能時刻指示方式は、各副ストレージの転送データ書き込みを並列に実行するためストレージ台数に比例した性能を発揮できる。また、データ更新順序は、副ストレージを使用するときにデータ更新順序を保証した状態とするため実用上問題ない。これを、業務サイト被災時の制御方法を例に取り説明する。業務サイト被災などが発生し、ホストから副ストレージを使用する要求を受け付けると、マスタ副ストレージが図 2 動作(3)(4)(5)を実行することで、副ストレージのボリュームを、更新順序が保証されている状態にする。具体的には、動作(3)(4)で、副ストレージへ連続して到着している転送データのタイムスタンプ(書き込み終了タイムスタンプと呼ぶ)を決定し、動作(5)で書き込み終了タイムスタンプまでの転送データ書き込みを指示する。各副ストレージが、指示された書き込み終了タイムスタンプまでの転送データ書き込みを確実に完了すると、複数台の副ストレージは書き込み終了タイムスタンプまでの転送データを一つずつボリュームに書き込んだ場合と同じ状態となり、各副ストレージ間のデータ更新順序を保証しているといえる。

以上から、書き込み可能時刻指示方式の方が実用上優れている。

5. まとめ

本稿では、複数台対複数台ストレージ間の非同期リモートコピーのデータ更新順序保証方式を提案した。これにより、複数台のストレージを使用する大規模なディザスタリカバリシステムを構築することが可能となった。

参考文献

[1]Securities and Exchange Commission, “Interagency Paper on Sound Practices to Strengthen the Resilience of the U.S. Financial System”, Feb 2002.

[2]Hitachi Data Systems, “Hitachi Universal Replicator Advanced Technology”, Sep 2004.