2B-01

3拠点間でデータ冗長化を自動継続する高可用システムのストレージ制御方式

Automatic Storage Replication Failover Method for High Availability across 3 Sites

須藤　梓†　　斎藤　秀雄†　　川口　智大†　　長尾　尚†

Azusa Sudo　Hideo Saito　Tomohiro Kawaguchi　Takashi Nagao

1. はじめに

地震やテロなどの災害による基幹業務の停止は，企業にとって莫大な損失に繋がる[1]．被災した場合でも業務を継続するため，遠隔の拠点間でサーバやストレージといったIT機器を冗長化する，拠点間高可用システムが導入される．さらに，特段の高可用性を要する業務では，被災拠点が復旧するまでに他拠点が被災し，システムが停止することを防止するため，3拠点以上での冗長化が求められる．

高可用性を実現する手段として，ストレージでのリモートコピー機能(RC)がある．RCは大別して，両拠点のデータを同期して更新する同期RCと，遠隔拠点のデータを非同期に更新する非同期RCの2種類がある．同期RCは，サーバへの応答性能が拠点間の距離に応じて低下するため，拠点間距離が近い場合に用いられる．一方，非同期RCは，サーバからの更新要求とは非同期にRCを行うため，拠点間距離に依存せず，拠点間距離が遠い場合に用いられる[2]．更に，同期RCの派生として，両拠点のストレージ何れにもデータの更新や参照が可能なActive-Active型同期RC(HA)が存在する[3][4]．HAは，ダウンタイムゼロで拠点間フェイルオーバが可能という特長を持つ．これらRC機能は，応答性能やデータ整合性等のシステム要件に基づいて使い分けられる．

本稿では，正サイト被災後も2拠点間の冗長化を維持する，HAと非同期RCを統合した3拠点間高可用システムの実現方式について述べる(図1)．

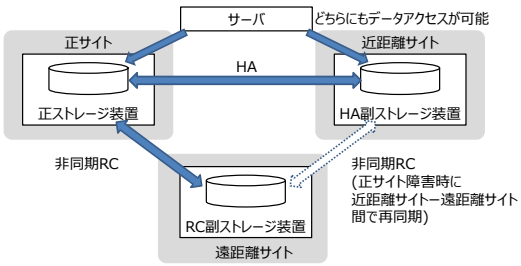


図1　3拠点間高可用システム

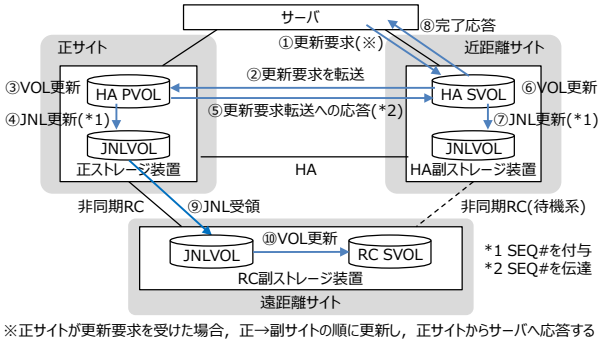
2. 3拠点間高可用システムの課題

2.1　システム構成とデータ更新処理

3拠点間高可用システムにおけるデータ更新処理の流れを示す(図2)．

正サイトと近距離サイトは，両方が更新要求を受けて互いにデータを同期する．データの更新順序を保証するため，近距離サイトで更新要求を受けた場合は，正サイトへ転送し，必ず正サイトのボリューム(HA PVOL)，近距離サイトのボリューム(HA SVOL)の順でデータを更新する．併せて，正サイトは，遠距離サイトとのデータ整合性を保証するため，更新データにデータ更新順序を示す情報(SEQ#)を組み合わせたジャーナル(JNL)を記録する．遠距離サイトは，正サイトで作成したJNLを受領し，遠距離サイトのボリューム(RC SVOL)のデータを更新する．

正サイトが被災し，近距離サイトへフェイルオーバする場合に，近距離サイトー遠距離サイト間の再同期を高速に行えるよう，正サイトのJNLに付与したSEQ#を近距離サイトへ伝達する．これにより，正サイトー近距離サイト間でJNLを同期することができ，正サイト被災時は，遠距離サイトが受領したJNLの続きを近距離サイトから受領できるようになる．



※正サイトが更新要求を受けた場合，正→副サイトの順に更新し，正サイトからサーバへ応答する

図2　3拠点間高可用システムにおけるデータ更新処理

2.2　JNL補填方式と3拠点間高可用化の課題

正サイトでのデータ更新処理中に取消処理等が発生し，データ更新が中断された場合，取得したSEQ#に対するJNLが作成されず，SEQ#が欠番となる場合がある(JNL欠損)．この状態のまま放置された場合，遠距離サイトは，正サイトの欠損JNLの到着を待ち続けるため，RC SVOLへの更新が停止する．そこで，非同期RCでは，欠損したJNLのSEQ#を作成済として扱うことでJNLを補填しSEQ#の連続性を保つ，JNL補填方式を用いる(図3)．

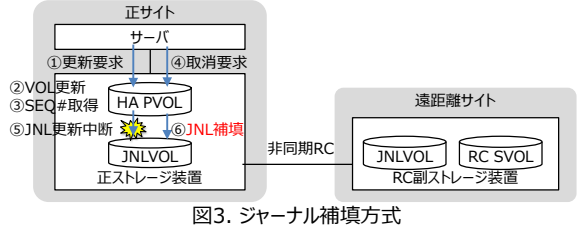


図3. ジャーナル補填方式

3拠点間高可用システムでは，正サイト被災時に，近距離サイトでも同様のJNL欠損が発生し，RC SVOLへの更新が停止するため，近距離サイトでのJNL補填が必要となる．

しかし，図4(a)に示すように，近距離サイトは，正サイトからの更新要求転送が遅延しているのか，正サイトで障害が発生しているのかを判断できないため，欠損

† (株) 日立製作所 研究開発グループ
Research & Development Group, Hitachi Ltd.

JNL を補填することができない．

また，前述の条件以外で，欠損 JNL を補填できないケースについて説明する．

説明に先立って，更新要求の分割について述べる．サーバからの更新要求が，ストレージ装置のデータ管理サイズの上限値を超えた場合は，図 2②～⑦の処理を分割して行う．分割数 N(N≧2)の場合，正サイトは，データの整合性を担保するために，N 分割の 1 番目の更新要求を受けた時点で N 個の連続した SEQ#を予約する．図 4(b)は，サーバからの更新要求が 2 分割される例である．正サイトは 1 分割目の更新要求転送を受領し(図 4(b)②)，SEQ#10,11 を予約して，SEQ#10 の JNL を作成する．

ここで，②への完了応答が近距離サイトに到達しない場合は，正サイトは，近距離サイトから 2 分割目の更新要求転送が遅延しているのか，近距離サイトで障害が発生しているのかを判断できないため，予約した SEQ#11 の欠損 JNL を補填することができない．

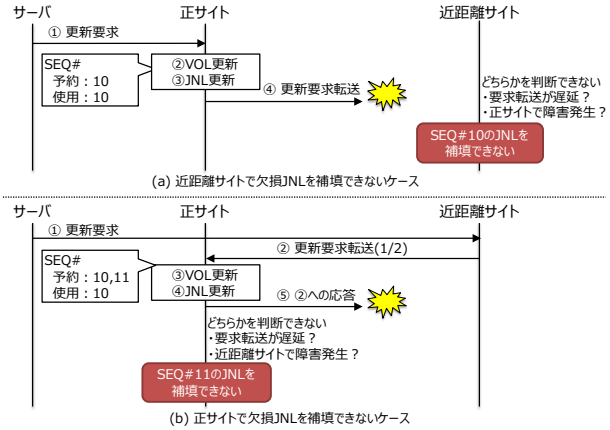


図4. 3拠点間高可用システムでの課題

3. 方式提案

3.1 方式 1：欠損 JNL タイムアウト方式

上記課題に対し，一定時間経過しても JNL が作成されない SEQ#が存在する場合は，強制的に JNL 補填を行う方式を提案する．

図 4(b)のケースを解決する本方式の流れを図 5 に示す．SEQ#11 の JNL 欠損が発生した後，この欠損状態に時限を設け，タイムアウト時に強制的に JNL 補填を行い，JNL 欠損を解消する．サーバに対しては，近距離サイトから処理失敗を応答することを期待する．これは，正サイトから近距離サイトに②の応答が返らないためである．

また，図 4(a)のように，近距離サイトに JNL 欠損が発生した場合は，近距離サイトで後続の SEQ#11 が作成されてから一定時間経過した後，同様に JNL 欠損を解消する．

図5. 方式1:欠損JNLタイムアウト方式の処理例

3.2 方式 2：補填 JNL 同期方式

方式 1 は，タイムアウトまで JNL 欠損状態となるため，JNL 補填に時間がかかるデメリットがある．そこで，正サイトで JNL 補填を行った場合は，JNL 補填が可能となる SEQ#の範囲を近距離サイトへ伝達する方式を提案する．

本方式の流れの例を図 6 に示す．本例は，正サイトが更新要求転送を受領し，SEQ#10 を予約した後で障害が発生するケースである．障害発生後，正サイトは処理中断を契機に，SEQ#10 の JNL 補填を行う(図 6④)．この後，正サイトは近距離サイトへ，JNL 補填要求と併せて JNL 補填可能な SEQ#の範囲を伝達することで，近距離サイトでも JNL を補填することが可能となる(図 6⑤⑥)．

図6. 方式2:補填JNL同期方式の処理例

4. 方式評価

提案した 2 方式が，JNL 欠損発生の原因を補完できるかを評価した．JNL 欠損の原因となる，処理中断契機を表 1 に示す．これらのうち，JNL 補填要求を伝達できるケース(#1,2)では，方式 2 の補填 JNL 同期方式により即座に JNL 欠損を補填することができる．しかし，サイト間の通信障害等，JNL 補填要求を伝達できないケース(#3,4)では，方式 2 では対応できない．これらのケースについては，方式 1 の欠損 JNL タイムアウト方式で補完する．

以上から，方式 1 及び 2 により JNL 欠損を補填することで，3 拠点間高可用システムが実現できることを確認した．

表 1. JNL 欠損発生原因

#	事象	頻度	方式 1	方式 2	方式 2 の補足
1	サーバからの処理中断要求	高 *1	○ *3	◎	
2	内部障害(データ転送障害)	中	○ *3	◎	
3	内部障害(CPU 障害)	中	○ *3	×	処理継続不能のため JNL 補填要求不可
4	正/近距離サイト間通信障害	低 *2	○ *3	×	SEQ#通知不能のため JNL 補填要求不可

*1: サーバの動作に依存するため多数発生することを想定する必要有り
*2: 通信経路が複数あるため発生頻度は低い
*3: タイムアウトまでは JNL 欠損状態となるため方式 2 より回復が遅い

5. おわりに

本稿では，3 拠点間高可用システムを実現する上で課題となる JNL 欠損に対し，欠損 JNL タイムアウト方式と補填 JNL 同期方式を提案した．本方式により，JNL 欠損を補填し，正サイト被災後も 2 拠点間冗長化を維持することが可能となった．

参考文献

[1] Eagle Rock Alliance, “Contingency Planning Research”, 1996.
[2] Hitachi Data Systems Corp., “Hitachi Virtual Storage Platform, Hitachi Universal Replicator User Guide”, 2013.
[3] 川口智大, “Markov 連鎖を用いたデータセンタ High Availability システムの信頼性評価方法”, 研究報告システムソフトウェアとオペレーティング・システム(OS), 2015-OS-133(9), pp.1-6, 2015.
[4] 長尾尚, “拠点間高可用ストレージを実現するデータ二重化方式”, 情報科学技術フォーラム講演論文集, pp.233-244, 2015.