情報処理学会論文誌 Vol. 52 No. 2 464–476 (Feb. 2011)

仮想リモートレプリケーションによる3拠点ストレージシステム制御方式

牧 晋広[†1,†2] 平岩友理[†1]
今津剛行[†1] 吉永努[†2]

データ損失なく災害復旧可能な3拠点ストレージシステムは広域分散された3拠点それぞれの主，近郊，遠隔ストレージ間でデータを3重化する．3重化はリモートレプリケーションとよばれる2重化技術を用いて，主ストレージ上のデータを同一時点の状態で近郊ストレージに，過去時点の状態で遠隔ストレージに転送することで実現される．災害等による主拠点システム停止後，主ストレージと同一データが格納される近郊ストレージを用いシステム復旧する．さらに，近郊ストレージは遠隔ストレージのデータと自身のデータとを比較し，その差分を交換することで，短時間に遠隔ストレージとの間でリモートレプリケーション（デルタリモートレプリケーションとよぶ）を再構築し，データを2重化する．一般に，リモートレプリケーションの構築可否はレプリケーション構築時にストレージ間で確認される．したがって，従来の3拠点ストレージシステムでは，デルタリモートレプリケーション構築可否がシステム復旧まで確認できなかった．本稿は3拠点ストレージシステムにおいて主ストレージから近郊，遠隔ストレージへのリモートレプリケーション構築中に，デルタリモートレプリケーションを仮想的に構築しその構築可否を確認する方式を提案する．評価では，本方式の適用により，デルタリモートレプリケーション構築可否の確認が完了するまでの時間を1/39から1/143に短縮可能なことが示された．

Control Mechanism with Virtual Remote Replication on 3 Data Center Storage System

Nobuhiro Maki,[†1,†2] Yuri Hiraiwa,[†1]
Takeyuki Imazu[†1] and Tsutomu Yoshinaga[†2]

We have been studying three-datacenter storage system where data on production storage are replicated to other two storage systems (“backup storages”), and those backup storages take over replication activity with short downtime after a production storage failure. The system builds a replication between backup storages and those backup storages exchange the differential data between them so that these two backup storages form mirroring relationship. Usually, capability of forming remote replication is validated between source and target storages upon actually forming replication relationship. Therefore, the system cannot validate whether it is possible to build the replication between backup storages or not until it fully takes over the replication activity. To make higher reliability, the conventional system had to perform a recovery test that actually builds the replication between backup storages. In this paper, we provide a new mechanism that virtually builds a replication while data on production storage are replicated to backup storages. Evaluation indicates a system with our new mechanism can reduce the time to validate readiness for the replication between backup storages shorter in the range from 1/39 to 1/143 when compared to the conventional mechanism.

1. はじめに

企業における情報システムの重要性は高まる一方である．その一方で，テロ，災害等によるシステム停止やデータ損失の発生を避けることは難しく，いったんシステムが停止するとその企業は莫大な損失を被る[1)]．災害等発生後も短時間にシステム復旧を実現するために，ディザスタリカバリシステム（DRシステム）が利用される．DRシステムは，地理的に離れた複数拠点にある計算機システムのデータを複数拠点間で冗長に保持するシステムで，災害発生時は被災を逃れた拠点に保持されたデータを使用し，システムの復旧が可能となる[2)–5)]．

特に米国テロ事件以降，法規制強化等によりDRシステムへの期待が高まる中，高い確度でのシステム復旧を実現する3拠点ストレージシステムが注目されている．3拠点ストレージシステムは3拠点に配置されたストレージ間でデータを3重化するDRシステムである．3重化は2種類のデータ2重化技術，すなわち，データ無損失だが転送距離に制限のある同期リモートレプリケーションと，理論上の転送距離に制限はないがデータ損失リスクのある非同期リモートレプリケーションを組み合わせて実現する．これにより，主拠点にあるストレージ（主ストレージ）上のデータは近郊拠点のストレージ（近郊ストレージ）に同一時点の状態で，遠隔拠点のストレージ（遠隔ストレージ）に過去時点の状態でつねに複製される．

†1 株式会社日立製作所
Hitachi, Ltd.
†2 電気通信大学
The University of Electro-Communications

ここで，3 拠点ストレージシステムは主ストレージ上のデータ 3 重化の違いにより，マルチターゲット方式，カスケード方式に分類される．マルチターゲット方式は主ストレージのデータを同期リモートレプリケーションと非同期リモートレプリケーションとで複製し，2 拠点のストレージに分散配置する方式である．カスケード方式は主ストレージのデータを同期リモートレプリケーションで近郊ストレージに複製，近郊ストレージが受信したデータをさらに非同期リモートレプリケーションで遠隔ストレージに複製する方式である．いずれの方式でも 3 拠点ストレージシステムは広域にシステムを分散した構成で，主拠点被災時もデータ損失のないシステム復旧が近郊拠点から可能という，両 2 重化技術の特徴をあわせ持つ．さらに，近郊，遠隔ストレージの両方が動作可能な場合，主ストレージと同一データが格納される近郊ストレージから遠隔ストレージにリモートレプリケーションを再構築できる．これにより，3 拠点ストレージシステムは近郊拠点でのシステム復旧後に再度被災する場合にも高い確度で遠隔拠点からシステムの復旧が可能となる．

また，3 拠点ストレージシステムではシステム復旧時に再構築するリモートレプリケーションの再構築時間を短縮する研究，開発がなされている[6),10),13)]．本稿では，被災を逃れた拠点のストレージ間で短時間に再構築するリモートレプリケーションをデルタリモートレプリケーションとよぶことにする．

著者らもマルチターゲット方式を基にした 3DC-DDR: Three Data Center storage system with Differential Data Resynchronization mechanism で上記の研究，開発を行っている．3DC-DDR では，主拠点被災後に近郊ストレージが遠隔ストレージのデータと自身のデータとを比較し，差分のみを両ストレージ間で交換することで，遠隔ストレージとの間にデルタリモートレプリケーションを再構築する[8),14)]．

一般的に，リモートレプリケーション構築可否の確認処理はリモートレプリケーション構築時にストレージ間で実施される．構築可否の確認により，人手で実施されるリモートレプリケーションの構成設定や通信設定の有効性を判断する．3DC-DDR を含め 3 拠点ストレージシステムのデルタリモートレプリケーションはシステム復旧時に再構築されるため，その構築可否の確認がシステム復旧時まで実施できない．システム復旧時に被災を逃れた拠点のストレージ間で確度の高い 2 重化を実現するために，従来の 3 拠点ストレージシステムではテスト目的で，システム復旧前に実際にデルタリモートレプリケーションを再構築する必要があった．デルタリモートレプリケーションを再構築するには事前に同期，非同期の両リモートレプリケーション構築が必要である．通常，同期，非同期リモートレプリケーション構築が完了するまでには数時間から数日の時間を要し，この時間が結果として，3 拠点ストレージシステムにおけるシステム動作確認のための時間を従来の 2 拠点ストレージシステム以上に増大させていた．システム動作確認の結果，正常動作しないことが判明した場合，システムの再構築等の手続きが判明時点から発生し，さらなるシステムの本番稼働時期を遅らせることになり問題となる．以後，デルタリモートレプリケーションをテスト目的で構築することを構築テストとよぶことにする．

3 拠点ストレージシステムでの動作確認のための時間短縮を実現するべく，本稿では 3DC-DDR において同期，非同期リモートレプリケーション構築中にデルタリモートレプリケーションを仮想的に構築することで，構築テストなしに当該レプリケーション構築可否の確認を実現する仮想レプリケーション制御方式を提案する．

2. 3DC-DDR の概要

2.1 構　　成

3DC-DDR はホスト計算機とストレージからなる拠点計算機システムがストレージネットワークにより相互に接続される構成をとる（**図 1** 参照）．ストレージネットワークのプロトコルには FCP（Fibre Channel Protocol）や iSCSI（Internet Small Computer System Interface）等が使用される．各拠点には同一構成の機器を配置することで，任意の拠点からシステム復旧が可能になる．

ホスト計算機には業務プログラムと，管理プログラムが動作する．管理プログラムはシステム使用者向けの操作手段とストレージ制御手段を備える．操作手段は操作端末によるグラフィカルユーザインタフェース（Graphical User Interface: GUI）やコマンドラインインタフェース（Command Line Interface: CLI）等である．**表 1** に CLI の内容を示す．ストレージ制御手段はシステム使用者による入力に従い，制御コマンドをレプリケーション元のストレージに発行することで，リモートレプリケーションの制御や，処理状態の把握を行う．制御コマンドは制御内容と，レプリケーション種別，構成情報を含む．制御内容は，リモートレプリケーションの処理動作を規定し，開始，一時停止，再同期，削除，状態取得がある．レプリケーション種別は同期方式，非同期方式，デルタ方式のいずれかのリモートレプリケーションを規定する．構成情報は，1 か複数のレプリケーション元と，レプリケーション先ボリュームの組（レプリケーションペアとよぶ）から構成される．

ストレージは記憶装置のほかに，制御装置，バッファを備える．記憶装置はハードディスクドライブ等から論理的な記憶資源であるボリュームを提供する．制御装置はデータ送受信機能，およびリモートレプリケーション機能を備える．バッファはリモートレプリケーショ

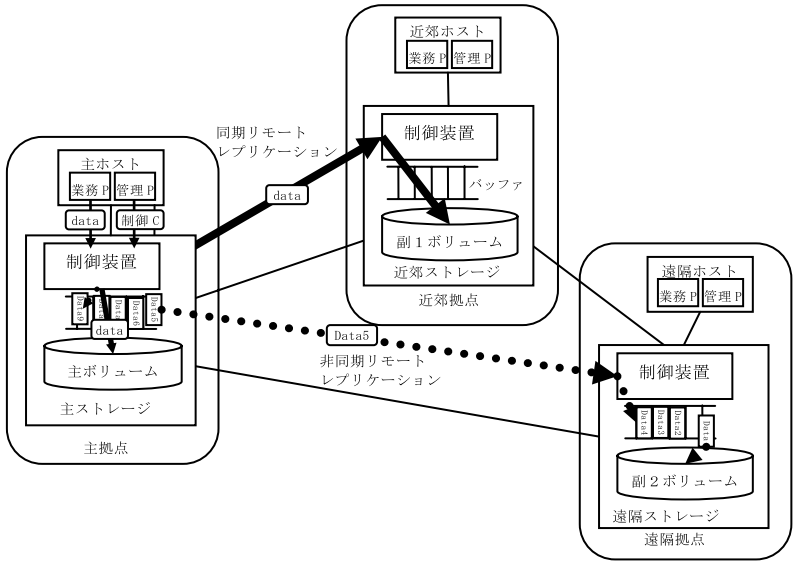

図 1　3DC-DDR システム構成
Fig. 1　3DC-DDR overview.

表 1　管理プログラムが提供するコマンドライン
Table 1　Command line interface of management program.

コマンドライン	説明
MAKE レプリケーションペア オプション	指定した 1 か複数のレプリケーションペア構築を開始する
SUSPEND レプリケーションペア オプション	指定した 1 か複数のレプリケーションペアのレプリケーション処理を一時停止する
RESYNC レプリケーションペア オプション	指定した 1 か複数のレプリケーションペアを再同期する
DELETE レプリケーションペア オプション	指定した 1 か複数のレプリケーションペアを削除する
QUERY レプリケーションペア オプション	指定した 1 か複数のレプリケーションペアの状態を取得する

ンで転送されるデータを一時保持するために使用される．また，バッファは複数のリモートレプリケーション処理に備え，複数系統用意される．

2.2　システム動作

3DC-DDR においてシステムを開始するには，主ストレージから近郊ストレージに同期リモートレプリケーションを，遠隔ストレージに非同期リモートレプリケーションを構築する．これは，システム使用者が主ホスト上の管理プログラムを用いて操作する．たとえば，同期リモートレプリケーション構築の場合は，管理プログラムに「MAKE 同期リモートレプリケーション」を指示する．近郊，遠隔ストレージへのリモートレプリケーション構築完了後，構築テストを実施する．構築テストは実際にデルタリモートレプリケーションを再構築，解除を実施する．テスト終了後，主ホストでは業務プログラムを開始する．

主拠点被災時のシステム復旧は，近郊・遠隔拠点が被災を逃れる場合と，遠隔拠点のみ被災を逃れる場合でその動作が異なる．

近郊・遠隔拠点が被災を逃れる場合，3DC-DDR は近郊拠点からシステム復旧を行う．近郊ホストは被災直前の主ストレージと同一データによるシステム復旧が可能となる．システム復旧は IP ネットワークの主拠点から近郊拠点への引き継ぎ（DNS の書き換え等）処理や，近郊ホストの復旧処理がある．近郊ホストの復旧処理は OS や業務プログラムが持つ障害復旧機能（たとえば，データベース管理システムの場合，クラッシュリカバリ機能）を利用する．システム復旧後，近郊，遠隔ストレージ間にデルタリモートレプリケーションを再構築する．デルタリモートレプリケーションを再構築することで，被災直前の主ストレージのデータで転送されていないデータに加え，システム復旧後に生成される近郊ホストによる更新データも近郊ストレージから遠隔ストレージに転送することが可能となる．

遠隔ストレージのみ被災を逃れる場合，3DC-DDR は遠隔拠点からシステム復旧を行う．遠隔ストレージには主ストレージとは異なるデータが格納されるが，過去の一時点としては完全なデータであるため，過去の一時点に遡ったシステム復旧は可能となる．

2.3　リモートレプリケーションの処理

3DC-DDR では，主ストレージが近郊，遠隔拠点の両ストレージ間と連携し，リモートレプリケーションを構築する．主ストレージは主ホストから受付ける制御コマンドに従い，リモートレプリケーション処理を実施する．リモートレプリケーションの処理はストレージ内のボリューム単位に実施される．また，リモートレプリケーション処理は同期，非同期リモートレプリケーション処理，デルタリモートレプリケーション処理がある．はじめに，同期，非同期リモートレプリケーション処理を説明する．

2.3.1　同期，非同期リモートレプリケーション処理

同期，非同期リモートレプリケーション処理は初期レプリケーションと定常レプリケーションに分類される．以下，それぞれを説明する．

(1) 初期レプリケーション

主ストレージが制御コマンドを受付けると，制御コマンド内の制御内容を解析し，制御内容が「開始」であれば，当該主ストレージは初期レプリケーション処理を開始する．初期レプリケーションとは，レプリケーション先のデータをレプリケーション元のデータに一致させる処理である．したがって，初期レプリケーション期間中は両ストレージ間のデータは一致せず，レプリケーション先のストレージからはシステム復旧が不可能となる．

初期レプリケーションが開始されると，リモートレプリケーション構築可否の確認とレプリケーションボリュームの全複製を順次実施する．

• リモートレプリケーション構築可否の確認：

リモートレプリケーションが正常に稼働するための要件（稼働要件）は，制御コマンドに指定されたレプリケーションペアが存在し，当該ストレージ間でデータ送受ができることである．そこで，3DC-DDR では，

(a) 主ストレージが制御コマンドの構成情報に規定されるレプリケーション元のボリュームが存在することを確認
(b) レプリケーション元のボリュームが存在する場合，主ストレージが当該制御コマンドをレプリケーション先のストレージに転送
(c) 当該制御コマンドを受付けたレプリケーション先ストレージが主ストレージ同様に，レプリケーション先のボリュームとして規定されたボリュームが自ストレージ内に存在することを確認

を実施する．

以上，3DC-DDR は，(a)，(c) で 3DC-DDR は制御コマンドに指定されたリモートレプリケーションの構成を，(b) の対象ストレージ間の制御コマンドの転送結果でデータ転送可否を，それぞれ検証でき，動作要件を満たすことが確認できる．

• レプリケーションボリュームの全複製：

主ストレージはレプリケーション元ボリューム上の全データを複製し，レプリケーション先のストレージに分割して転送する．レプリケーション先のストレージは受信したデータを順次格納する．主ストレージのデータ転送はレプリケーション先ストレージの応答を待つことなく，連続的に実施される．この処理により，主ストレージは自ストレージに格納されるレプリケーション対象の全データをレプリケーション先ストレージに格納されるデータと一致させる．初期レプリケーションは主ストレージにあるレプリケーション対象の全ボリュームのデータすべてを複製し終えると完了となり，主ストレージは定常レプリケーション処理を引き続き実施する．初期レプリケーション処理から定常レプリケーション処理への移行には特別な制御コマンドを必要としない．

(2) 定常レプリケーション

定常レプリケーション処理では，主ストレージは主ホストによる更新データを受付けるときのみ当該データを複製し，一方を主ストレージ内ボリュームに格納し，他方をレプリケーション先ストレージに転送する．レプリケーション先ストレージは，主ストレージが受付けた順序と同じ順序で当該データをボリュームに格納する．

定常レプリケーション処理では，レプリケーション元，先ストレージ間で送受されるデータが主ホストから発行される更新データのみになるため，初期レプリケーションの通信トラヒックに比べ，その量を大幅に削減することができる．また，リモートレプリケーション処理中に，主ホストから発行された更新データが主ストレージとレプリケーション先ストレージで同じ順序で書き込まれるため，主拠点が被災しても被災を逃れた拠点のホスト計算機は主ストレージの障害復旧と同じ手順でシステムを復旧できる．

定常レプリケーションは同期方式，非同期方式で処理内容が異なる．以下，それぞれについて説明する．

(2-1) 同期リモートレプリケーション

主ストレージは主ホストから更新データを受信するとそのデータを複製し，一方を主ボリュームに書き込み，他方を近郊ストレージに転送する．近郊ストレージはデータ受信後，そのデータを副 1 ボリュームに書き込み，応答を主ストレージに返す．主ストレージは応答受信後，主ホストに更新完了を報告する．

以上により，主ホストが更新完了の報告を受けた時点で，遠隔ストレージでは主ストレージと同一データの保持が保証される．

(2-2) 非同期リモートレプリケーション

主ストレージは主ホストから更新データを受信すると，当該更新データにシリアル番号を付与し転送フレームを作成し，その転送フレームを自ストレージ内のバッファに格納する．バッファ格納後，主ストレージは主ホストに更新完了を報告する．上記処理とは非同期に，主ストレージは通信トラヒックの状況に応じ，適切なタイミングでバッファに格納された転送フレームを取り出し，遠隔ストレージに転送する．転送フレームにはシリアル番号が付与

されているため，主ストレージは遠隔ストレージからの転送フレームの応答を待つことなく，連続的に転送フレームを転送する．ここで，遠隔ストレージは，受信した転送フレームのシリアル番号を基に転送フレームをソートし，副 2 ボリュームに古い順にデータを書き込む．これにより，遠隔ストレージは主ホストによるデータの更新順序を再現できるため，同期リモートレプリケーションのように主ストレージは転送先ストレージからの応答を待って次のデータを転送する必要がない．

以上により，主ホストは主ストレージ内のバッファに更新データ格納後，更新完了の報告を受ける．そのため，主ホストのアクセス時間はデータ転送距離による影響を理論上受けない．

2.3.2　デルタリモートレプリケーションの処理

デルタリモートレプリケーションは，同期，非同期の両リモートレプリケーションで転送されたデータの差分を使用して，近郊，遠隔ストレージ間に構築する非同期リモートレプリケーションである．

デルタリモートレプリケーションの処理は差分蓄積処理と差分再同期処理からなる．

差分蓄積処理はシステム復旧前に実施する準備処理で，主ストレージから転送された個々のデータを，識別可能にし，近郊，遠隔ストレージの双方で蓄積する処理である．差分蓄積処理は同期，非同期の両リモートレプリケーション開始後であれば，任意のタイミングで開始可能である．

差分再同期処理はシステム復旧時に実施するデルタリモートレプリケーションの再構築処理で，近郊，遠隔ストレージに蓄積されたデータの差分を特定し，特定した差分を両ストレージ間で交換することで，デルタリモートレプリケーションを再構築する．近郊，遠隔ストレージ間で差分データを交換するためには，近郊，遠隔ストレージのいずれかに，主ストレージ上のレプリケーション対象の全データが転送されている必要がある．これを実現するためには，同期，非同期，両リモートレプリケーションの初期レプリケーションが完了している必要がある．

以下，それぞれの処理内容について**図 2** を用いて説明する．

(1) 差分蓄積処理

主ストレージから転送されたデータを識別可能な状態で蓄積するために，差分蓄積処理では，同期リモートレプリケーションの処理を以下のように拡張する．

- シリアル番号付き同期転送処理：
  主ストレージは，主ホストから受信した更新データのシリアル番号の付与を，非同期リモートレプリケーションに加え，同期リモートレプリケーション用の転送フレームにも実施する．
- 同期転送データの蓄積処理：
  通常の同期リモートレプリケーション処理では，近郊ストレージは受信した転送フレームのデータを読み出し，副 1 ボリュームに書き込む．本処理ではさらに，近郊ストレージはその転送フレームをバッファに書き込む．バッファに転送フレームを格納することで，近郊ストレージは受信した直近の転送フレームを一定量蓄積する．

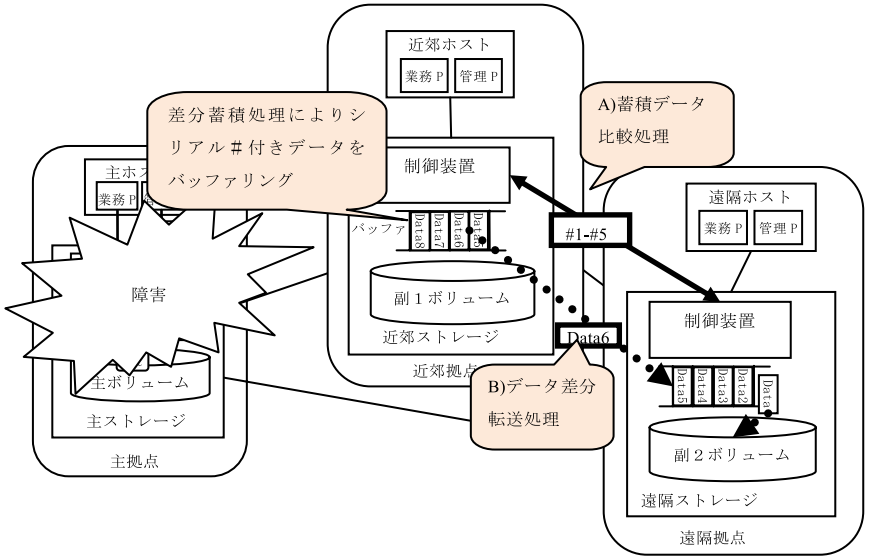

図 2　デルタリモートレプリケーション処理
Fig. 2　Process of delta-remote replication for 3DC-DDR.

以上から，同期転送データの蓄積処理により，近郊ストレージは同期リモートレプリケーションで転送された直近のデータを一定量蓄積可能になる．また，シリアル番号付き同期転送処理により，近郊ストレージはバッファに蓄積されたデータの識別が可能になる．

差分再同期処理を実施可能にするためには，データ蓄積処理において，近郊，遠隔ストレージのバッファに蓄積された転送フレームのシリアル番号がバッファ間で重複か，連続すればよい．上記条件を満たすことで，近郊，遠隔ストレージ間で主ストレージから転送され

た直近の全データがいずれかのバッファに存在することになる．これにより，両ストレージ間でデータ損失のないデルタリモートレプリケーションの再構築が可能になる．

(2) 差分再同期処理

近郊，遠隔ストレージに蓄積されたデータの差分を特定し，特定した差分を両ストレージ間で交換するために，差分再同期処理では，以下の処理を A)，B) の順で実施する．

A) 蓄積データ比較処理：
近郊ストレージは遠隔ストレージのバッファに蓄積される転送フレームのシリアル番号について最大値と最小値を取得し，自装置のバッファにあるシリアル番号と比較する．これにより，近郊ストレージは近郊，遠隔ストレージ間の差分データを特定する．

B) データ差分転送処理：
近郊ストレージは処理 A) で特定した差分データに該当する転送フレームを古いシリアル番号から順にバッファから取り出し，遠隔ストレージに非同期リモートレプリケーションにおける定常レプリケーション処理と同一の手順で転送する．

以上から，蓄積データ比較処理により，蓄積したデータの内，近郊，遠隔拠点の両ストレージ間の差分を特定することが可能となり，データ差分転送処理により，特定した差分を両ストレージ間で交換することが可能となる．

2.4 課　　題

デルタリモートレプリケーション構築可否の確認について考える．2.3.1 項で述べたとおり，デルタリモートレプリケーション構築可否の確認を実現するには，リモートレプリケーションの稼働要件を満たせばよい．すなわち，デルタリモートレプリケーション開始の制御コマンドに指定されたレプリケーションペアが存在し，近郊，遠隔ストレージ間でデータ送受ができることを，レプリケーション元である近郊ストレージが確認すればよい．

ここで，デルタリモートレプリケーションの処理を考える．差分蓄積処理は上記のとおり，同期リモートレプリケーションの拡張処理であり，制御コマンドは同期リモートレプリケーションを制御する主ストレージに発行される．一方，差分再同期処理は近郊，遠隔ストレージ間でデータを送受する処理であるため，制御コマンドはレプリケーション元の近郊ストレージに発行される．以上を考えると，近郊ストレージが最初にデルタリモートレプリケーションに関する制御コマンドを受付けるのは差分再同期処理であり，これはシステム復旧時の処理となる．そのため，近郊ストレージはシステム復旧のタイミングまで構築可否の確認ができないことになる．これを回避するためには，システム復旧前にデルタリモートレプリケーションの構築テストが必要となる．構築テストを実施するには同期，非同期リモートレプリケーションの構築が必要になるため，多大な時間が必要となる．したがって，構築テストを実施することなく，デルタリモートレプリケーション構築可否の確認をシステム復旧前に実現することが課題となる．

3. 仮想レプリケーション制御方式

構築テストを実施することなくデルタリモートレプリケーション構築可否の確認を実現するべく，著者らはシステム復旧前にデルタリモートレプリケーションを仮想的に構築する仮想レプリケーション制御方式を提案する．本章では，仮想レプリケーション制御方式の概要について簡単に述べた後，その処理方式について説明する．

3.1 概要と実現方式

仮想レプリケーション制御方式では，差分蓄積処理をレプリケーションボリュームの全複製を持たない，仮想的なデルタリモートレプリケーションの初期レプリケーションに相当する処理と位置づける．そのために，仮想レプリケーション制御方式では差分蓄積処理を同期リモートレプリケーションの拡張処理ではなく，デルタリモートレプリケーションの処理として実現する．これにより，差分蓄積処理の開始時には，デルタリモートレプリケーションのレプリケーション元である近郊ストレージが制御コマンドを受付けることになり，結果，近郊ストレージが当該制御コマンドの構成情報を用いてデルタリモートレプリケーション構築可否の確認が可能となる．さらに，差分蓄積処理と差分再同期処理が一連のデルタリモートレプリケーションに関連する処理として扱うことができるため，システム使用者が管理プログラムに指示する CLI や GUI によるリモートレプリケーションの指示対象が一致する．これは，システム使用者のユーザビリティが向上すると著者らは考える．**表 2** にシステム

表 2　システム使用者によるレプリケーション操作手順（CLI によるシステム構築（システム開始から差分再同期処理開始まで）の場合）
Table 2　Replication management operations for 3DC-DDR establishment.

方式適用なしの操作手順	方式適用ありの操作手順	説明
MAKE 同期レプリケーション	MAKE 同期レプリケーション	リモートレプリケーション構築
MAKE 非同期レプリケーション	MAKE 非同期レプリケーション	
MAKE 同期レプリケーション DELTA	MAKE デルタレプリケーション DELTA	差分蓄積処理開始
RESYNC デルタレプリケーション DELTA	RESYNC デルタレプリケーション DELTA	差分再同期処理開始

構築時の CLI による操作手順を示す．方式適用時は差分蓄積処理と差分再同期処理のレプリケーション対象がデルタリモートレプリケーションに統一されていることが分かる．

仮想レプリケーション制御方式を実現するために，著者らは制御コマンド事前受け付け処理と，仮想レプリケーションステータス生成処理を 3DC-DDR に追加した．

制御コマンド事前受付け処理は近郊ストレージでデルタリモートレプリケーション向けの制御コマンドをシステム復旧前に受付ける処理である．

仮想レプリケーションステータス生成処理は差分蓄積処理における近郊，遠隔ストレージの両バッファに蓄積されるデータ差分の状態を，仮想的なデルタリモートレプリケーションの処理状態として変換する処理である．これにより，管理プログラムは差分蓄積処理から差分再同期処理への移行可否の判定を，同期，非同期リモートレプリケーションの処理状態の把握と同様に，近郊ストレージにデルタリモートレプリケーションの状態取得のための制御コマンドを発行することで可能となる．

3.2　処 理 方 式

3.2.1　制御コマンド事前受け付け処理

制御コマンド事前受付処理で，近郊ストレージは制御コマンドを受付けると，当該制御コマンドを用いてデルタリモートレプリケーションの構築可否を検証し，差分蓄積処理を実施する．

処理フローを**図 3** に示す．制御コマンド（制御内容は開始）を受付けると，近郊ストレージは制御コマンドから構成情報を取り出し，(a) 構成情報に規定されるレプリケーション元のボリュームが自ストレージに存在することを確認し，存在する場合，(b) 当該制御コマンドをレプリケーション先である遠隔ストレージに転送する．(c) 遠隔ストレージでは主ストレージ同様に，レプリケーション先のボリュームとして規定されたボリュームが存在することを確認し，近郊ストレージにその結果を報告する．以上の (a) から (c) により，2.3.1 項リモートレプリケーション構築可否の確認で示した処理同様に，リモートレプリケーションの稼働要件を満たすことができる．

次に，近郊，遠隔ストレージ間では差分蓄積処理を開始する．差分蓄積処理の動作は上記の処理と同様である．

3.2.2　仮想レプリケーションステータス生成処理

近郊ストレージは差分蓄積処理において近郊，遠隔ストレージそれぞれのバッファに蓄積されたデータ（転送フレーム）の差分を，デルタリモートレプリケーションの処理状態に変換する．**表 3** のようにデルタリモートレプリケーションの処理状態を，差分再同期処理の開

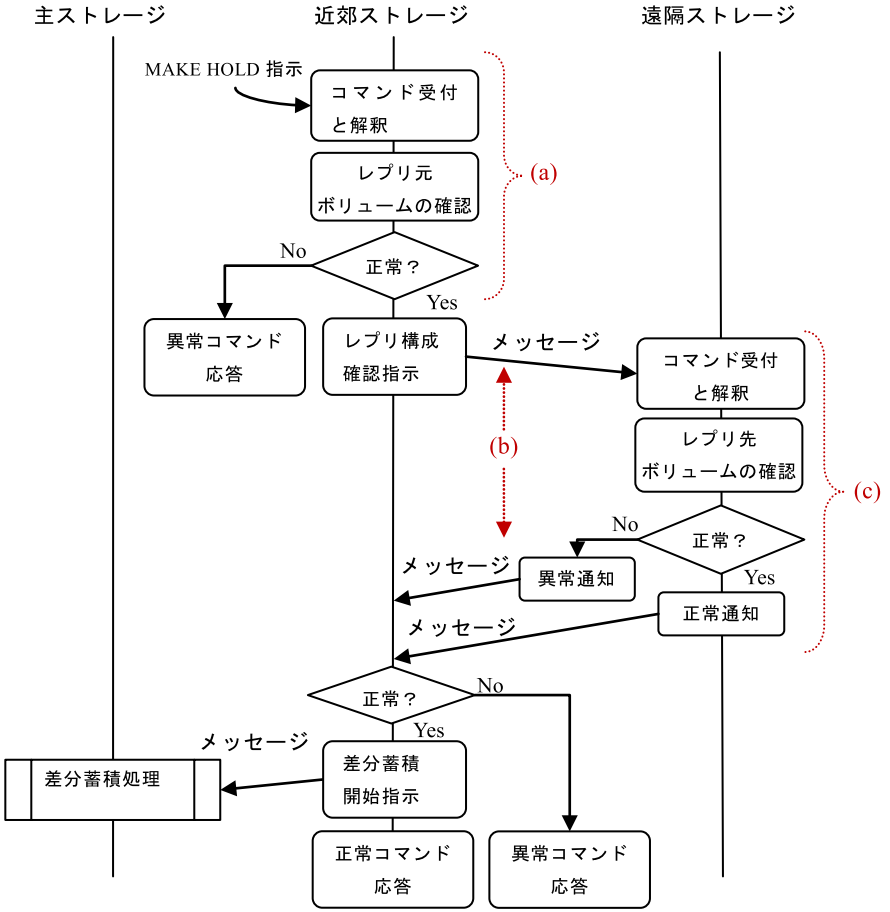

図 3　制御コマンド事前受け付け処理フロー
Fig. 3　Flow of control command pre-receive process.

始可否により 3 状態分追加定義する．追加分は HOLD TRANS，HOLD，HOLD ERROR である．表中の BUFMIN はバッファ内転送フレームのシリアル番号最小値を，BUFMAX はバッファ内転送フレームのシリアル番号最大値を示す．たとえば，BUFMIN（近郊）とは近郊ストレージのバッファにある転送フレームのシリアル番号が最小のものを指す．また，

表 3　デルタリモートレプリケーションの処理状態とその意味
Table 3　Processing states of delta remote replication and their descriptions.

処理状態	各ストレージのバッファの状況	意味
SIMPLEX	なし	デルタリモートレプリケーションが動作していない
DUPLEX	なし	デルタリモートレプリケーションが定常レプリケーション状態
DUPLEX PENDING	なし	デルタリモートレプリケーションが差分再同期中状態
PENDING	なし	デルタリモートレプリケーションが別状態に遷移中
SUSPEND	なし	デルタリモートレプリケーションが一時停止状態
HOLD TRANS（追加状態）	BUFMIN(近郊)>BUFMAX(遠隔)+1 ，もしくは少なくとも同期，非同期レプリケーションのいずれかが初期レプリケーション中	差分蓄積処理が開始され，差分再同期実施不可な状態．HOLD 状態に遷移中
HOLD（追加状態）	BUFMIN(近郊)≦BUFMAX(遠隔)+1，かつ BUFMAX(近郊)≧BUFMAX(遠隔)かつ同期，非同期両レプリケーションが定常レプリケーション	差分蓄積処理が開始され，差分再同期実施可能な状態
HOLD ERROR（追加状態）	バッファなどストレージ内ハードウェアの障害，通信障害	ストレージ，ストレージネットワークが障害状態で，差分再同期処理が実施不可能な状態

処理状態が HOLD の場合，近郊，遠隔ストレージ内のバッファに格納される転送フレームのシリアル番号は重複か連続の状態になり，差分再同期処理が可能な状態となる．

近郊ストレージは当該制御コマンドを受付けると，遠隔ストレージから遠隔ストレージ内にあるバッファの BUFMIN，BUFMAX を取得し，近郊ストレージ内の BUFMIN，BUFMAX を比較し，表 3 に基づき変換し，デルタリモートレプリケーションの処理状態とする処理を開始する．近郊ストレージでは管理プログラムからデルタリモートレプリケーションに対する制御コマンド（制御内容は状態取得）を受信したときに，上記の変換を実施しその処理状態を結果として報告する．

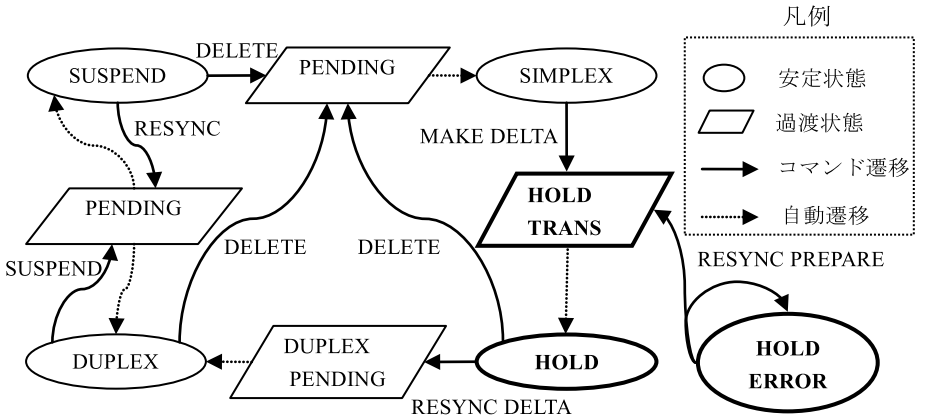

図 4　仮想レプリケーション制御の状態遷移
Fig. 4　State transition of virtual replication process.

次に，**図 4** に状態遷移を示す．デルタリモートレプリケーションが動作していない処理状態は SIMPLEX 状態となる．ここで，システム使用者が MAKE DELTA を指示することで，デルタリモートレプリケーションに対し仮想構築（差分蓄積処理の開始）が開始され，デルタリモートレプリケーションの状態は HOLD TRANS 状態を経由し HOLD 状態に遷移する．ここで，この一連の遷移は図 3 の処理フローにおけるコマンド受け付けから蓄積処理開始の処理に該当する．次に，RESYNC DELTA 指示が実施されると，デルタリモートレプリケーションが再構築され，デルタリモートレプリケーションの状態は HOLD 状態から DUPLEX PENDING 状態を経由して，DUPLEX 状態になる．HOLD ERROR 状態は任意の状態から遷移する．そのため，図では HOLD ERROR 状態への矢印を記載していない．HOLD ERROR 状態に遷移した場合，障害要因が取り除かれていれば，RESYNC PREPARE 指示により，HOLD TRANS を経由して，HOLD 状態に復帰することができる．

3.3　方式適用時のシステム動作

仮想レプリケーション制御方式を適用することで，3DC-DDR は構築テストなしにデルタリモートレプリケーション構築可否を確認できる．本方式適用後のシステムの想定運用は以下のとおりである．

はじめに 3DC-DDR では同期，非同期リモートレプリケーションを構築する．これは，システム使用者が主ホストの管理プログラムを通じ主ストレージに制御コマンドを発行する

ことで実現する．各ストレージでは初期レプリケーション処理を開始する．初期レプリケーションの完了有無によらず，この時点で，仮想レプリケーション制御が可能な状態になる．そこで，3DC-DDR では仮想的なデルタリモートレプリケーション構築を開始する．これは，システム管理者が主ホストもしくは近郊ホストの管理プログラムを通じ，近郊ストレージに制御コマンドを発行することで実現する．ここで，制御コマンドの応答が正常の場合，差分蓄積処理が開始され，デルタリモートレプリケーションの再構築が可能な状態に設定されていると判断する．次に，先に実施した同期，非同期リモートレプリケーションの初期レプリケーション完了後，業務プログラムを開始する．

主拠点被災時のシステム復旧動作は 2.2 節システム動作で示した内容と同様である．

4. 性能評価

仮想レプリケーション制御方式の有効性を検証するため，仮想レプリケーション制御方式を備えた 3DC-DDR を実機システム上に試作し，評価を行った．

4.1 評価環境

表 4 に評価環境を示す．著者らは，仮想レプリケーション制御方式をストレージ：Hitachi Universal Storage Platform[⋆1]のマイクロコードで，またリモートレプリケーションの制御や処理状態の把握手段を管理プログラム：Business Continuity Manager（BCM）上で実装した．ストレージでの実装をマイクロコードで行う理由は ASIC 等によりチップ化するよりも柔軟かつ短期間に実装できるためである．すなわち，マイクロコードで実装することで，仕様変更等が容易化され，複雑化するストレージ処理を簡単にデバッグでき開発期間の短縮が可能となる．BCM はホスト計算機上で動作し，ホスト計算機，ストレージ間のネットワークを介してストレージに制御コマンドを発行する．また，各ストレージ間のネットワークはシングルパスで構成した．

表 4　評価環境
Table 4　Experimental environment.

ホスト計算機 （主，近郊，遠隔ホスト）	ハードウェア：IBM mainframe System z9‡ OS: z/OS v1.8 管理プログラム：　Hitachi Business Continuity Manager5.1
ストレージ （主，近郊，遠隔ストレージ）	Hitachi Universal Storage Platform H65A3-5 ボリュームサイズ：2.838GB（共通） ボリュームタイプ：3390-3 キャッシュサイズ：4GB ボリュームの RAID レベル：RAID5 [9]
ストレージネットワーク （データ回線）	ホスト計算機とストレージ間：ESCON§ 17MB/s x 4 ストレージ間：Fibre Channel 2Gb/s 各ストレージ間ストレージネットワークの回線距離は全て 10m

‡ IBM System z9 は International Business Machine Corporation の登録商標です．
§ ESCON は International Business Machine Corporation の登録商標です．

⋆1 Universal Storage Platform は Hitachi Data Systems の登録商標です．

測定時間は BCM の CLI コマンド実実行時間（E-Time）を 3 回測定し，その平均値を採用した．また，測定時間は仮想レプリケーション制御方式を適用しない場合の測定結果で，レプリケーションペア数が最少の測定時間を 1 として正規化した．

4.2 実験結果および考察

4.2.1 仮想レプリケーション制御方式の効果

3DC-DDR で，仮想レプリケーション制御方式を適用しない場合と，適用する場合における同期，非同期リモートレプリケーションの構築開始からデルタリモートレプリケーション構築可否の判定が完了するまでの時間（完了時間とよぶ）を測定した．

図 5 と**表 5** に測定結果を示す．図 5 は同期，非同期，デルタリモートレプリケーションそれぞれのレプリケーションペア数を増加したときの，完了時間である．表 5 は方式適用時の測定時間を抜き出した表である．

仮想レプリケーション制御方式を適用する場合はレプリケーションペア数増に対し，ほぼ一定の完了時間を維持する．仮想レプリケーション制御方式の測定結果を取り出した表 5 を見ても，レプリケーションペア数が 24 以降はほぼ一定の時間を推移している．ペア数 16 の場合は極端に完了時間が短いが，これはストレージマイクロコードの実装による影響と考えられる．レプリケーションペア数が 20 以下の場合，制御コマンド事前受付処理における制御装置内部のデータアクセス時間が大幅に短縮されるためである．

一方，仮想レプリケーション制御方式を適用しない場合は，レプリケーションペア数が増加するのに合わせ，完了時間も増加している．測定した範囲内では仮想レプリケーション制御方式適用しない場合は適用する場合に比べ 39 倍から 143 倍の時間が必要になった．

この原因は，仮想レプリケーション制御方式を適用する場合，同期，非同期リモートレプリケーションの開始後，即座に仮想的なデルタリモートレプリケーションの再構築が可能となり，デルタリモートレプリケーション構築可否が確認できる．その一方で，方式適用がない場合は同期，非同期リモートレプリケーションの初期レプリケーションが完了するま

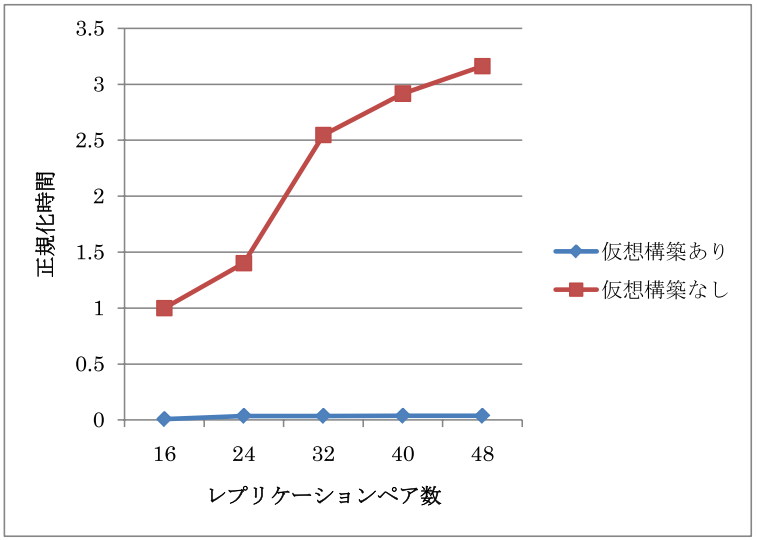

図 5　デルタリモートレプリケーション構築可否が完了するまでの時間（測定結果をレプリケーションペア数 16，方式適用なしの測定時間で正規化）
Fig. 5　Evaluation result of the time until 3DC-DDR finishes validating the behavior of delta-remote replications.

表 5　方式適用時の構築可否完了までの時間（レプリケーションペア数 16 の測定時間で正規化）
Table 5　Evaluation result of the time with proposal mechanism.

ペア数	16	24	32	40	48
完了時間	1	5.13	5.13	5.28	5.22

で，構築可否の確認のために 3DC-DDR はデルタリモートレプリケーション再構築を待つ必要があるためである．すなわち，仮想レプリケーション制御方式を適用しない場合，デルタリモートレプリケーション構築可否を確認するためには，同期，非同期リモートレプリケーションの初期レプリケーション完了を 3DC-DDR が待つ必要があるためである．一般に，初期レプリケーションはレプリケーション対象のデータ量に応じその処理時間が増加する．企業情報システムにおけるデータ量が増加傾向にある現状を考えると，仮想レプリケーション制御方式の有効性は今後もさらに高まると考える．

4.2.2　仮想レプリケーション制御失敗の影響

仮想レプリケーション制御方式により，デルタリモートレプリケーションのペアが存在し，近郊・遠隔ストレージ間が通信可能なことを保証できる．しかしながら，以下の場合，デルタリモートレプリケーションの再構築（差分再同期）が失敗になる．

- **仮想レプリケーションステータスが HOLD TRANS：**
  これは同期，非同期リモートレプリケーションの初期レプリケーションが完了していない状態や，近郊，遠隔ストレージそれぞれのバッファに十分なデータ（転送フレーム）の蓄積がない状態で起こる．バッファに十分な転送フレームがない状態は非同期リモートレプリケーションで使用されるストレージネットワークで大幅な遅延が生じる場合等で起こる．
  ここで，初期レプリケーションが完了していない状態を考える．初期レプリケーションが完了するまでは業務プログラムが動作していないため，業務への影響は少ないと考える．一方，バッファに転送フレームが十分にない場合，差分再同期処理は失敗する．しかしながら，主ストレージから近郊ストレージに対象データの転送が完了しているのであれば，近郊拠点でのシステム復旧は可能となる．また，近郊ストレージには主ストレージと完全に同一のデータが格納されている．そのため，近郊ストレージの全データを遠隔ストレージに転送することで，近郊，遠隔ストレージ間でリモートレプリケーション構築も可能となる．差分再同期処理が失敗した場合の影響を**図 6** に示す．図 6 はレプリケーションペア数を変化させた場合の近郊，遠隔ストレージ間でリモートレプリケーション構築に必要な時間を差分再同期処理の使用有無で比較した結果である．結果をみると，レプリケーションペア数が増加するに従い，影響が大きくなる．評価では，最大 86.8 倍の差が生じている．レプリケーションペア数が増加するに従って，差分再同期処理を使用しない場合の時間が増大するのは，レプリケーションペア数増大に従いデータ転送量も増大するためである．
- **仮想レプリケーションステータスが HOLD ERROR：**
  これは近郊・遠隔ストレージ間に張られたストレージネットワーク回線もしくはストレージハードウェアが障害になった場合等でおこる．この場合，障害部位の取り除きを行い，再度仮想レプリケーションステータスを取得することで，システム復旧の可否を判断することになる．ストレージハードウェアの部位によっては，近郊拠点からのシステム復旧が失敗する可能性がある．この場合，遠隔拠点からのシステム復旧となり，復旧後のシステムは最新データを失う可能性が高い．

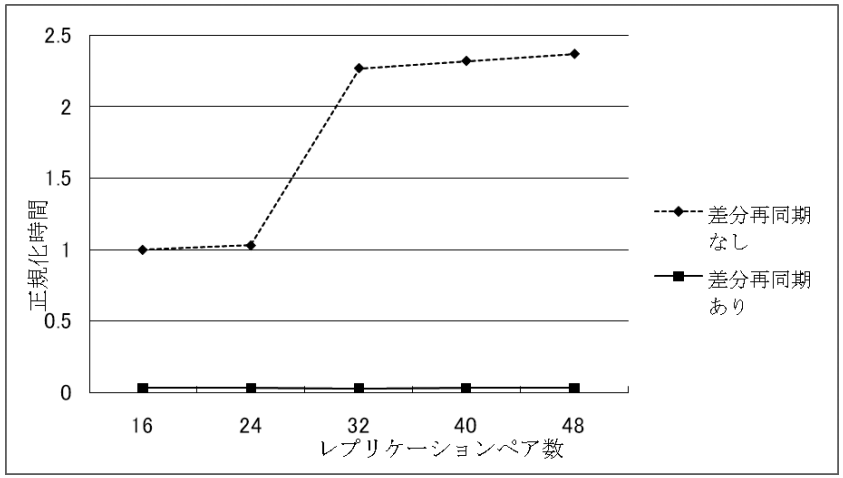

図 6　デルタリモートレプリケーションが使用できない場合の影響（測定結果をレプリケーションペア数 16，差分再同期なしの測定時間で正規化）
Fig. 6　Impact of replication establishment time without using of delta remote replication.

4.2.3　ネットワーク遅延の影響

本評価環境では，ストレージ間ネットワークに十分な距離を設定していない．ストレージネットワークのプロトコルに FCP を用いた場合，距離によるネットワーク遅延は，同期リモートレプリケーションで数ミリ秒，非同期リモートレプリケーションで数十から数百ミリ秒程度である．ここで，今回の評価での影響を考える．仮想レプリケーション制御方式適用時の測定結果では最大 10%程度影響を受ける．方式適用のない場合，測定結果に対しその影響は十分に小さく，ほぼ無視できる．仮想レプリケーションの適用がない場合，初期レプリケーションの時間が大半を占め，その時間はネットワーク遅延に対し十分に大きいためである．

以上より，今回の評価では最大で 10%程度測定結果に誤差が生じる可能性がある．しかしながら，仮想レプリケーション制御の優位性を覆すほどの影響はないと考える．

5. 関 連 研 究

上記のように，3 拠点ストレージシステムには大きく 2 種類の方式が存在する．マルチターゲット方式と，カスケード方式である．

マルチターゲット方式の研究には，本 3DC-DDR のほかに，たとえば SRDF/Star [⋆1] がある[10]．SRDF/Star も近郊，遠隔ストレージ間にデルタリモートレプリケーションを再構築する機能が備わる．SRDF/Star には本研究のような構築テストなしにデルタリモートレプリケーション構築可否を確認する機能に関する記載を見つけることはできなかったが，仮想レプリケーション制御方式の適用は可能と考える．

カスケード方式の研究には，我々が提案している 3DC-Cascade 方式[11),12)]，Metro/Global Mirror（旧 PPRC Cascading）[13)] 等がある．カスケード方式では，主ストレージからのデータ転送は近郊ストレージのみであり，主ストレージの処理負荷が低く，ホスト計算機による IO 処理性能の低下を抑止できる可能性がある．その一方で，近郊拠点が被災すると，データ複製処理の継続が不可能になるというデメリットがある．このデメリットを解決するため，Metro/Global Mirror では Incremental Resync とよばれるデルタリモートレプリケーションを，主ストレージと遠隔ストレージ間で再構築する．Incremental Resync にも，本研究のような構築テストなしに Incremental Resync の構築可否を確認する機能に関する記載を見つけることはできなかったが，仮想レプリケーション制御方式の適用は可能と考える．ただし，この場合，主ストレージ，遠隔ストレージ間に仮想的なデルタリモートレプリケーションを構築する形態での実現になる．

また，カスケード方式に類似する方式として中継バッファ方式[7)] がある．中継バッファ方式はカスケード方式同様に主ストレージから遠隔ストレージに順次データを転送するが，近郊拠点がストレージではなく，中継バッファであることがカスケード方式と異なる．中継バッファは主ストレージから転送されるデータを一時保存するバッファとそのバッファ上のデータを遠隔ストレージに非同期に再転送する機構を備える．中継バッファ方式には主拠点被災後に中継バッファから遠隔ストレージへのデータ転送を待たずにシステム復旧を実現する手法が提案されている．しかしながら，上記手法は主拠点被災前後で同一の中継バッファから遠隔ストレージにデータ転送を実施するものであり，本研究が対象とする主拠点被災後に異なるストレージ間でリモートレプリケーションを構築するものとは異なる．

6. お わ り に

著者らが 3 拠点ストレージシステムとして提案している 3DC-DDR におけるシステム動作確認のための時間短縮を実現するべく，デルタリモートレプリケーション構築可否の確認

⋆1 SRDF は EMC Corporation の登録商標です．

のために必要だった構築テストを不要にする仮想レプリケーション制御方式を提案した．

仮想レプリケーション制御方式の有効性を検証するために，本提案方式適用有無のシステムで，デルタリモートレプリケーション構築可否の確認が完了するまでの時間を実環境で測定した．測定結果，本提案方式を適用したシステムは，本提案方式を適用しないシステムに比べ，全般にわたって構築可否の確認までの時間が短いことが分かった．測定した範囲では，本提案方式適用により，本提案方式を適用しないものに比べ，測定時間を 1/39 から 1/143 に短縮できることを示した．本提案方式は 3DC-DDR に限らず，3 拠点ストレージシステムの多くに適用可能である．

以上より，3 拠点ストレージシステムに本提案方式を適用することで，3 拠点ストレージシステムでの動作確認のための時間短縮が実現できる．

参 考 文 献

1) Patterson, D.A.: A Simple Way to Estimate the Cost of Downtime, *Proc. LISA '02: USENIX 16th System Administrators Conference* (*LISA '02*), pp.185–188 (2002).
2) Toigo, J.W.: *Disaster RECOVERY Planning*, Principle Hall (2003).
3) 谷井成吉：コンピュータシステム災害復旧の対策，ダイアモンド社 (2006).
4) IBM: Advanced functions for storage subsystems: Supporting continuous availability, *IBM SYSTEM JOURNAL*, Vol.42, pp.268–279 (2003).
5) Schulman, R.R.: Disaster Recovery Issues and Solutions, HDS White Paper (2004).
6) ITCentrix (Barometrix): Three-Node Disaster Recovery Topologies, An Approach to Significantly Reduce the Risk of Permanent Data Loss, White Paper (2006).
7) 大和純一，管　真樹，菊池芳秀：広域災害に対するストレージによるデータ保護，電子情報通信学会，Vol.89, No.9 (2006901), pp.801–805 (2006).
8) HDS: Disaster Recovery Issues and Solutions, Hitachi Data Systems Corp., White Paper (2004).
9) Patterson, D.A., Gibson, G. and Katz, R.H.: A case for redundant arrays of inexpensive disks (RAID), *SIGMOD '88*: *Proc. 1988 ACM SIGMOD International Conference on Management of Data*, New York, NY, USA, ACM Press, pp.109–116 (1988).
10) EMC: Using Asynchronous Replication for Business Continuity between Two or More Sites, EMC Corp., White Paper (2004).
11) 岡田　渡，牧　晋広，宮田和久，佐藤雅英：3 拠点カスケードリモートコピー構成における主ホスト障害時のデータ可用性維持方式，第 68 回情報処理学会全国大会論文集，Vol.3, pp.367–368 (2006).
12) 牧　晋広，岡田　渡，宮田和久，佐藤雅英：マルチサイトリモートコピー制御における低ホスト負荷障害監視方式，第 68 回情報処理学会全国大会論文集，Vol.3, pp.369–370 (2006).
13) IBM: IBM System Storage DS8000: Copy Services in Open Environments, IBM Corp., Red Book (2008).
14) Maki, N., Hiraiwa, Y., Imazu, T. and Sowa, M.: A Proposal of Management Interface for Differential Data Exchange Mechanism on 3 Datacenter Storage Systems, *Proc. 6th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology* (*ECTI2009*), pp.644–647 (2009).

(平成 22 年 5 月 31 日受付)
(平成 22 年 11 月 5 日採録)

牧　晋広（正会員）

1993 年名古屋工業大学工学部電気情報工学科卒業．1995 年電気通信大学大学院情報システム学研究科情報ネットワーク学専攻修了．1998 年同大学院博士課程単位取得退学．同年（株）日立製作所入社．システム開発研究所にてストレージ管理ソフトウェアの研究開発に従事．現在同研究所主任研究員，および電気通信大学大学院情報システム学研究科情報ネットワークシステム学専攻博士後期課程在学中．

平岩　友理（正会員）

1992 年茨城大学工学部情報工学科卒業．同年（株）日立製作所入社．システム開発研究所にて大型汎用計算機およびストレージ管理ソフトウェアの研究開発に従事．現在同研究所主任研究員．

今津　剛行（正会員）

2000 年大阪市立大学工学部情報工学科卒業．2002 年同大学大学院工学研究科電子情報系専攻修士課程修了．2002 年（株）日立製作所入社．現在，同社ソフトウェア事業部にて，ストレージ管理ソフトウェアの開発に従事．

吉永　　努（正会員）

1986 年宇都宮大学工学部情報工学科卒業．1988 年同大学大学院修士課程修了．同年より宇都宮大学工学部助手．1997 年から翌年にかけて電子技術総合研究所・客員研究員．2000 年より電気通信大学大学院情報システム学研究科助教授．現在，同教授．博士（工学）．計算機システム，並列分散処理，ネットワーク・コンピューティング等に興味を持つ．電子情報通信学会，IEEE 各会員．