Vol.2010-EVA-31 No.1
2010/3/8

ディザスタリカバリにおける非同期リモートコピーのリカバリポイント監視方式

江丸裕教[†,††]　高井昌彰[†]　原　純一[††]

企業情報システムにおけるデータ保護の重要性は高まる一方である．一方，自然災害やテロなどシステムの停止やデータ損失に繋がりかねないリスクは数多く存在する．一般に，被害を受けたシステムを復旧すること，また復旧させるための体制やシステムをディザスタリカバリと呼ぶ．企業情報システムにおけるディザスタリカバリを実現する上で，重要な役割を担うのがストレージシステムである．ストレージシステムが提供するディザスタリカバリ技術の一つに，データを遠隔地にミラーリングすることによってデータを保護する非同期リモートコピーがある．非同期リモートコピーでは，被災後に遠隔地においてどの時点のデータを復旧できるかを判断するリカバリポイントを監視する必要がある．本論文では，これまで実用的なリカバリポイント監視方式が存在していなかったオープンシステムにおけるConsistent Con tinuous 型の非同期リモートコピーを対象とし，①特別な作りこみが不要であること，②正サイトだけで監視が行えること，③監視精度を管理できることの3要件を満たすバッファ滞留時間計測法を提案する．また，シミュレーション環境での評価実験により提案手法の有効性を示す．

Monitoring Recovery Point for the Asynchronous Remote Copy in Disaster Recovery

Hironori Emaru,[†,††]　Yoshiaki Takai[†]　and Junichi Hara[††]

The importance of the data protection keep rising in the enterprise information systems. On the other hand, there are many risks which may cause an unexpected system halt and data loss. Therefore, planning recovery procedures in preparation for a disaster is important. This is referred to as disaster recovery. A storage system takes a major role to achieve disaster recovery. Asynchronous remote copy is one of disaster recovery technologies, and protects data by mirroring the data to a remote location. In the asynchronous remote copy, it is important to monitor the recovery point of the data which will be used in the recovery phase. In this paper, we discuss three requirements for recovery point monitoring: (1) interoperability of SAN environment, (2) monitoring without secondary site information, and (3) accuracy management. Based on the three requirements, we propose a buffer residence-time method. This method can be applied to the consistent continuous asynchronous remote copy in open systems. We show our method satisfies the three requirements by evaluations in a simulation environment.

1. はじめに

1.1 ディザスタリカバリの重要性

企業情報システムにおける無停止運用やデータ保護の重要性は，マーケットのグローバル化や，Web による 24 時間 365 日のサービス提供などを背景に高まる一方である．ところが，テロや自然災害など，企業情報システムの停止やデータ損失に繋がりかねないリスクは数多く存在する．これらのリスクを低減するには，災害や障害を起こりうるものと想定し，災害や障害の発生時に停止したシステムをいつまでにどうやって復旧するかを予め計画しておく必要がある [1]．被害を受けたシステムを復旧すること，また復旧させるための体制やシステムをディザスタリカバリと呼ぶ [2, 3]．

1.2 RPO と RTO

ディザスタリカバリでは，被災後にどの時点のデータをいつまでに復旧するかを指標として事前に定めておき，その指標が遵守されていることを監視しつつ運用することが重要である．この指標の内，どの時点のデータを復旧できるべきかを Recovery Point Objective （RPO），被災後いつまでに業務を再開できるべきかを Recovery T ime Objective（RTO）と呼ぶ [4]．RPO と RTO は一般にコストをかければかけるほど短縮することが可能であるが，かけられるコストには限界がある．そこでこれらの目標値は，通常，サービス提供者と利用者の間で，被災確率や業務停止時間に応じた損害額とシステム構築のコストを勘案して決定される [5]．

1.3 ストレージシステムとリモートコピー

企業情報システムにおけるディザスタリカバリを実現する上で重要な役割を担うのが，企業情報システムの価値の根幹をなすデータを高信頼・高性能に格納するストレージシステムである．

ストレージシステムが提供するディザスタリカバリ技術の一つに，データを遠隔地にミラーリングすることによってデータを保護し，被災時には保護されたデータを用いて遠隔地で業務を継続するリモートコピーがある [6, 7]．ただし，リモートコピーにも様々な手法があるため，全ての手法に対して汎用的に RPO を遵守しているかどうかを監視する技術，すなわち汎用的なリカバリポイント監視技術は存在しない．

そこで本論文では，リモートコピーおよびそのリカバリポイント監視技術を分類した上で，今まで有効な技術が存在していなかった，オープンシステムにおける Consistent Con tinuous 型の非同期リモートコピーに適用可能な方式であるバッファ滞留時間計測法を提案し，シミュレーション環境での評価実験によって有効性を示す．

[†]北海道大学情報基盤センター
Information Initiative Center, Hokkaido University
[††]株式会社日立製作所システム開発研究所
Systems Development Laboratory, Hitachi Ltd.

2. 本研究の位置づけ

2.1 リモートコピーとは

リモートコピーとは，業務を行っている場所（正サイト）から地理的に離れた位置にストレージシステムを備えた副サイトを用意し，業務遂行によって変化するデータを正サイトと同様に副サイトにも常にミラーリングする技術である．リモートコピーは，図 1 に示した状態遷移を取り，以下の手順で運用される．

(1) 運用開始前に，正サイトのデータを副サイト上のストレージシステムに全てコピーする．この動作を初期コピーと呼ぶ．
(2) 初期コピーの完了後，正サイトの業務を開始する．業務中に発生したライト I/O は，正サイトと副サイトの両方のストレージシステムに書き込まれる．
(3) 正サイトが被災した場合，正副入れ替えを行い，正サイトのストレージシステムと同一のデータを持つ副サイトのストレージシステムを用いて業務を再開する．
(4) 何らかの障害で，副サイトに対してライト I/O を転送できない場合には，コピー一時中断状態に移行する．障害原因を取り除いた後，再同期指示により正サイトに保持していた差分をコピーすることで，コピー状態に復帰する．

2.2 リモートコピーの分類

リモートコピーは，副サイトへのデータ転送を実際に行うコンポーネント（ホスト，ネットワーク，ストレージ）によって分類される．

ホストベースのリモートコピーは，ホスト単位でデータ転送を制御・実行するため，リモートコピーの障害がホストの業務処理に影響を与えるという問題がある．一方，ネットワークベースのリモートコピーでは，アプライアンスやネットワークスイッチが副サイトへのデータ転送を行うため，この部分の負荷集中がスケーラビリティ上の問題をもたらす．このため，企業向けシステムでは，ストレージベースのリモートコピーを用いるのが一般的である．実際，ストレージベースのリモートコピーがリモートコピー全体に占める割合は，2007 年時点で 83.7%である [8]．本論文では，ストレージベースのリモートコピーを対象とする．

2.3 同期リモートコピーと非同期リモートコピー

ストレージベースのリモートコピーには，業務ホストからライト I/O が発行された際に，副サイトでのライト I/O 書き込み完了応答を待ってサーバに応答を返す同期方式と，副サイトでのライト I/O 書き込み完了応答を待たずにサーバに応答を返す非同期方式がある．

同期方式は，常に正サイトと副サイトのデータが一致するため，被災時に失われるデータ量を最小限に抑えられるが，副サイトの書き込み時間のために業務ホストの応答性能に影響を与えてしまう．そのため，両サイトが 100km 以上の超遠距離に配置される場合や，ネットワーク品質が低い場合には使用できない．

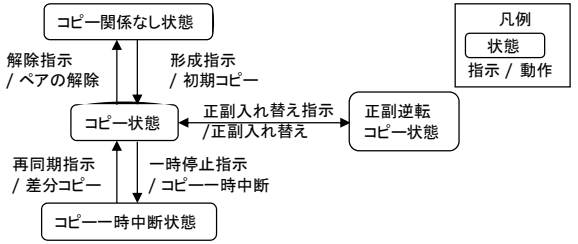

図 1 リモートコピーの状態遷移図
Fig.1 State transition diagram of the remote copy.

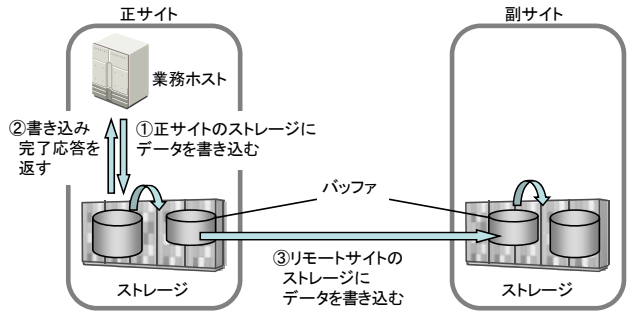

図 2 ストレージベースの非同期リモートコピー
Fig.2 Storage-based asynchronous remote copy.

一方，非同期方式では，図 2 に示すように副サイトに転送すべきライト I/O を一度正サイト内のバッファに蓄積し，正サイトへの書き込みとは非同期に副サイトに転送・書き込みを行う．本方式は，ネットワークの帯域変動や急激な業務負荷変動時にも業務ホストの応答性能に影響を与えずに済むため，サイト間の距離が超遠距離になる場合や，帯域が不安定なネットワーク回線を用いる場合に有効である．同期リモートコピーは，あるライト I/O の書き込みが正副ともに完了しない限り次のライト I/O が発行されないため，リカバリポイントの監視は基本的に不要である．したがって，リカバリポイントを監視する必要があるのは非同期リモートコピーのみである．なお，リカバリポイントを監視するのは，図 1 に示した状態のうち，コピー状態および正副逆転コピー状態の場合のみである．

2.4 ストレージベースの非同期リモートコピーにおけるリカバリポイント監視技術

ストレージベースの非同期リモートコピーは，副サイトへのデータ転送方法の観点から，Periodic 型，Continuous 型，Consistent Continuous 型に分類される [9]．

Periodic 型は業務ホストから書き込まれたライト I/O を一度正サイトのストレージシステムに保持し，定期的にまとめて副サイトに転送する．またこのタイミングで業務ホスト上では静止化処理を行う．静止化とは，アプリケーションを一時停止させてバッファをフラッシュするなどして，転送データを回復可能な整合性の取れたものにすることである．

Continuous 型は業務ホストから書き込まれたライト I/O を副サイトに常時転送するが，ライト I/O の順序性は維持されない．副サイトで回復可能な整合性を確保するため，静止化されたデータを取得可能な他の方式と組み合わせて使用されるケースが一般的である [10]．

Consistent Continuous 型は業務ホストから書き込まれたライト I/O を，順序性を維持しつつ，常時副サイトに転送する．本方式は，いつ正サイトが被災しても副サイトからの回復が可能なため，上記の 3 分類の中で最も小さいリカバリポイントを実現可能である．

一方，リカバリポイント監視技術という観点から非同期リモートコピーを分類すると，業務ホストがライト I/O にタイムスタンプを付与する仕組みを持っているメインフレームシステムと，持っていないオープンシステムに大別することができる．

表 1 は，非同期リモートコピーにおけるリカバリポイント監視技術を，データ転送方法ならびに業務ホスト種別の観点から分類したものである．

メインフレームシステムでは（表 1 の領域 A），ライト I/O のタイムスタンプを利用できるため [11), 12)]，副サイト側でタイムスタンプと現在時刻を比較することによりリカバリポイントを容易に監視することができる．一方，オープンシステムでは，データ転送方法により適用できるリカバリポイント監視技術が異なる．

Periodic 型の場合（表 1 の領域 B），例えば 5 秒間隔でライト I/O を転送しているケースでは，転送が失敗した時点でリカバリポイントが 5 秒以上になるということを意味する．したがって，転送の成功/失敗を監視することがリカバリポイントの監視に相当する．

Continuous 型（表 1 の領域 C）は，整合性を担保できる他の技術と組み合わせて使用され，静止化されたデータの取得は定期的に行われるのが一般的であるため，Periodic 型と同様な監視技術が適用できる．

Consistent Continuous 型（表 1 の領域 D）については，リカバリポイントを最小化可能なデータ転送方法であるにも関わらず，リカバリポイントを監視する有効な手段がこれまで存在していなかった．

本研究ではオープンシステムにおける Consistent C ontinuous 型の非同期リモートコピーをターゲットとする．以降，本論文ではオープンシステムにおける Consistent Continuous 型の非同期リモートコピーを非同期リモートコピーと呼ぶ．

表 1 非同期リモートコピーとリカバリポイント監視技術の分類
Table 1 Classification of the asynchronous remote copy and monitoring methods of the recovery point.

転送方法 / ホスト種別	Periodic 定期的にライトI/Oを副サイトに転送	Continuous 順序性を維持せずに常時ライトI/Oを副サイトに転送	Consistent Continuous 順序性を維持しつつ常時ライトI/Oを副サイトに転送
メインフレームシステム A	I/O中のtimestampを利用		
オープンシステム	periodicであることを利用した監視 (B)	組み合わせて使用される定期的に静止化する技術を利用 (C)	(D)

3. 非同期リモートコピーにおける書き込みタイムラグ

3.1 リカバリポイントと書き込みタイムラグ

非同期リモートコピーでは，ネットワーク障害の発生，他業務によるネットワーク帯域の圧迫，想定以上のライト I/O 発生などの様々な要因によって，業務ホストでライト I/O が発行されてから副サイトに書き込まれるまでの遅延時間に変化を生じる．本論文では，この遅延時間を書き込みタイムラグと定義する．正サイトが被災した場合，副サイトでは書き込みタイムラグ分だけデータが失われた状態で業務復旧の作業を行うことになる．すなわち非同期リモートコピーにおいては，書き込みタイムラグの監視がリカバリポイントの監視に相当する．

3.2 書き込みタイムラグ監視の要件

本節では，大規模な企業情報システムにおけるストレージシステムで最もよく用いられている Storage Area Network（SAN）[13]および SAN で使用されている Fibre Channel Protocol（FCP）[14), 15)]の環境に構築された非同期リモートコピーにおける書き込みタイムラグ監視を実現するにあたって，求められる要件を整理する．

第 1 の要件は，監視の精度を管理できることである．実際の書き込みタイムラグに対して精度の高い値が得られることも重要であるが，どのような条件で監視を行った場合に，どれだけの精度を確保できるかが予めわかっていることがより重要である．精度がわからなければ，実際に監視により得られた値の評価はできないためである．

第 2 の要件は，正サイトだけで書き込みタイムラグが監視できることである．非同期リモートコピーを用いるシステムでは，東京と大阪，ニューヨークとサンフランシスコといったように，サイト間の距離が数百 km，数千 km 離れていることは珍しくない．このような場合，正サイトと副サイトは別々の管理者によって管理されているこ

とが一般的であり，副サイト側に書き込みタイムラグの情報取得の仕組みを用意せずに，正サイトだけで監視を行いたいというニーズがある．また，正サイトは自社のデータセンタを利用するが，副サイトは他社のデータセンタを利用するといったケースのように，そもそも副サイトに情報取得の仕組みを配置できないこともある．

第 3 の要件は，システムを構築する業務ホスト，アプリケーション，ストレージシステムに対する特別な作りこみが発生しないことである．例えば，メインフレームシステムのようにI/Oにタイムスタンプを埋め込む手法をSAN環境に適用しようとすると，プロトコルレベルでの改変が必要となり，OS，ネットワークスイッチ，ストレージシステムなど影響の及ぶ範囲が大きい．また，業務ホスト上で稼動している OS やアプリケーションソフトウェアにプローブを入れることは，特に商用ソフトウェアを用いる場合，技術上もまたライセンス契約上も困難である．

3.3 バッファ滞留時間計測法

SAN 環境におけるストレージシステムの運用では，バッファは想定される書き込みに対し，ネットワークの切断など予期しない事態が発生しても数分程度はコピー一時中断状態に移行しないだけの容量を確保するのが一般的である．したがって，書き込みタイムラグの監視にはミリ秒の精度は必要なく，秒単位の精度で計測できれば十分に実用的である．

また，非同期リモートコピーにおいて，データは一度バッファに蓄積された後，副サイトからのトリガによって副サイトに転送される．副サイトからのトリガは逐次的に発行される．すなわち，あるトリガによって正サイトから転送されたデータが，副サイトに書き込まれた時点で次のトリガが発行される．そのため，副サイトやサイト間のネットワークに何らかの障害が発生して副サイト側でのデータの書き込みやサイト間のデータ転送が実行されない場合には，正サイトのバッファにデータが滞留したままとなる．一方，データが転送される Fibre Channel のレイテンシは 100km あたり 0.5msec 程度であり，サイト間距離が数千 km の長距離になっても転送時間はミリ秒のオーダである．

以上のことから，ライト I/O が正サイトのバッファに滞留している時間を書き込みタイムラグと見なすことができる．

FCP では，ファイバチャネルフレーム（FC フレーム）が通信の単位となっており，各フレームのヘッダにはシーケンス番号が埋め込まれている [16]．

本論文で提案するバッファ滞留時間計測法は，バッファの入口で取得したライトI/Oのシーケンス番号を定期的に記録しておき，バッファの出口から取得したライト I/O のシーケンス番号をバッファの入口で取得したシーケンス番号と突き合わせることによって，このライト I/O がいつ投入されたものかを求める手法である．

ここで説明のため，正サイトのバッファにおける流入曲線と流出曲線を図 3 に示す．

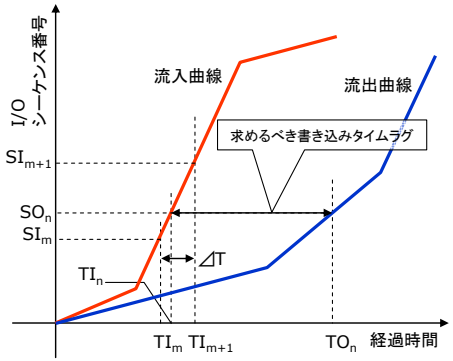

図 3　バッファ滞留時間計測法による書き込みタイムラグ
Fig.3　Time-lag of the write I/O by the buffer residence-time method.

流入曲線とは，正サイトのバッファの入口に着目し，時間経過に伴いバッファに投入されたライト I/O のシーケンス番号をプロットしたものである．流出曲線は同バッファの出口に着目し，時間経過に伴いバッファから副サイトに転送されたライト I/O のシーケンス番号をプロットしたものである．なお，実際には一定時間（⊿T）ごとの離散点であるが，本グラフはこれを線形補間している．

時刻 TO_n におけるバッファ出口から取得したライト I/O のシーケンス番号を SO_n とすると，このシーケンス番号のライト I/O がバッファに投入された時刻は，流入曲線と水平線 $y = SO_n$ との交点である．交点の時刻を TI_n とすると，時刻 TO_n における書き込みタイムラグは，$TO_n - TI_n$ である．

ここで TI_n は以下のように求めることができる．

(1) 一定時間(⊿T)ごとにバッファの入口でライト I/O のシーケンス番号を取得する．時刻 TI_m におけるシーケンス番号を SI_m として配列に保持しておく．

(2) 時刻 TO_n におけるバッファ出口の I/O のシーケンス番号を SO_n とし，$SI_m < SO_n < SI_{m+1}$ を満たす SI_m，SI_{m+1} を検索する．このとき $TI_m < TI_n < TI_{m+1}$ となり，TI_n を次式で近似する．

$$TI_n = TI_m + \Delta T\,(SO_n - SI_m) / (SI_{m+1} - SI_m)$$

明らかに TI_n が取りうる値は監視間隔である⊿T の範囲に限定されるため，本手法による誤差は最大⊿T となる．すなわち監視間隔によって監視の精度を管理することができる．また，FCP に準拠したフレームに標準で埋め込まれているシーケンス番号を正サイトでのみ取得することによって，書き込みタイムラグを求めることができる．

したがって，本手法は書き込みタイムラグ監視に求められる 3 つの要件を全て満たしている．

4. 評価

本章では，提案手法であるバッファ滞留時間計測法の有効性をシミュレーションによって検証する．

4.1 シミュレーションモデル

非同期リモートコピーの環境モデルを構築し，その上でシミュレーションを実行する．シミュレーションの条件および処理の流れは以下の通りである

(1) 複数の業務ホストからライト I/O が発行される．ライト I/O は FC フレームで 2KB 固定とする．
(2) 各業務ホストには個別のバッファが割り当てられ，各々ライト I/O を書き込む．
(3) バッファに対するライト I/O の書き込みとは非同期に，バッファから副サイトへのデータ転送を行う．

4.2 実験結果および考察

はじめに，正サイトと副サイト間のネットワーク帯域に時間変動が無い環境に適用した結果を示す．図 4 は，表 2 に示すパラメータにおける書き込みタイムラグである．横軸はシミュレーション経過時間，縦軸は書き込みタイムラグである．なおライト I/O Rate は図 6 のように時間変動するものとした．ライト I/O Rate の変動によって，書き込みタイムラグは変化するが，10 秒の監視間隔に対し，誤差（実際の書き込みタイムラグとの差）が最大 1 秒以内に抑えられていることがわかる．

表 2　シミュレーションのパラメータ
Table 2　Parameter of the simulation.

パラメータ	設定値
業務ホスト台数 16	台
ライト I/O Rate	図 5 に示すとおりに変動
バッファ容量	各業務ホスト毎に 600MB
サイト間距離	500km
ネットワーク帯域 300	MB/sec で一定
監視間隔 10	秒
副サイトにおけるレイテンシ 0.2msec	で一定
副サイトにおける書き込み速度 3200	0MB/sec で一定

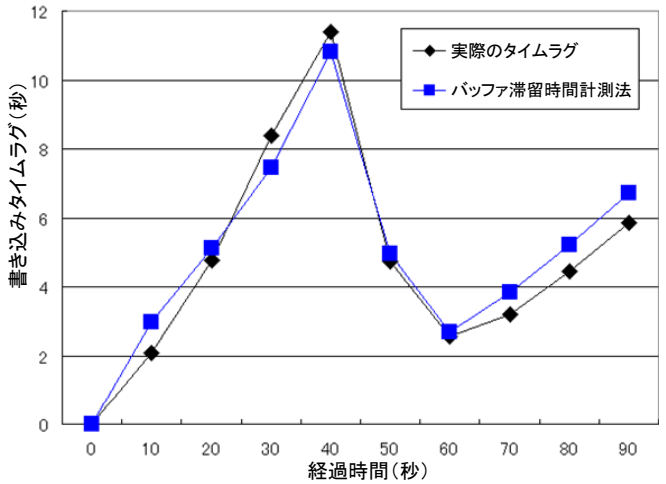

図 4　シミュレーション結果（ネットワーク帯域一定）
Fig.4　Simulation results (constant bandwidth).

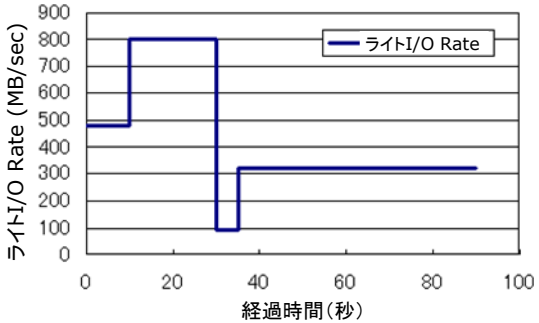

図 5　ライト I/O Rate
Fig.5　W rite I/O Rate

次に，正サイトと副サイト間のネットワーク帯域に時間変動がある環境に適用した結果を示す．本シミュレーションでは，帯域幅をシミュレーション開始から 10 秒間は 300MB/sec，10 秒後から 30 秒後まで 100MB/sec，その後は 400MB/sec とした．それ以外のパラメータは表 2 の通りである．また，ライト I/O Rate の変動は，1 つ目のシミ

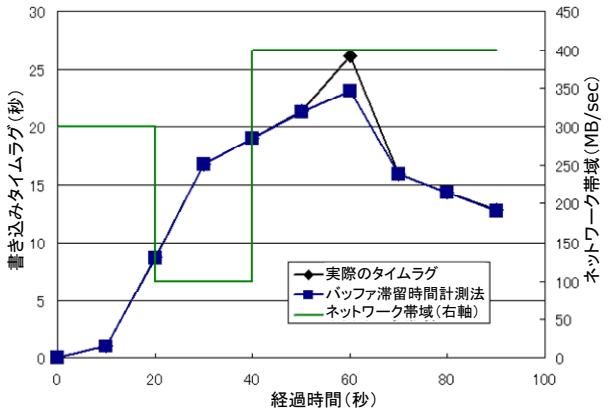

図 6　シミュレーション結果（ネットワーク帯域変動）
Fig.6　Simulation results (varying bandwidth).

ュレーション同様，図 5 に示す通りとする．図 6 にシミュレーションで得られた書き込みタイムラグの結果を，ネットワーク帯域の変化とあわせて示す．

提案手法の最大誤差は，シミュレーション開始 60 秒後の 3.0 秒であり，監視間隔の範囲に収まっている．本手法が，ネットワーク帯域の変動があるような環境においても，実際の書き込みタイムラグに対する誤差を監視間隔値以下に抑えられることがわかる．

5.　おわりに

本論文では，オープンシステムにおける Consistent Co ntinuous 型の非同期リモートコピーにおける書き込みタイムラグ監視において，監視精度の管理，正サイトだけの監視，デファクトスタンダードの利用といった 3 つの要件を示し，これら要件を満たすバッファ滞留時間計測法を提案した. さらにシミュレーションによる評価実験から，提案手法の有効性を示した．

今後の課題としては，実環境における本手法の有効性の検証が挙げられる．正サイトのみの監視で実用的な書き込みタイムラグが得られることを，実際のリモートコピー環境でも検証する必要がある．

また監視間隔の適切な設定方法にも検討の余地がある．監視精度を上げるためには監視間隔を短くすればよいが，これによってストレージシステムに対して余分な負荷をかけ，業務に影響を及ぼす可能性がある．この負荷を定量的に評価することによって，システムの規模やビジネスニーズに合わせた適切な監視間隔を設定するための指標が必要である．

参考文献

1) Rudolph, C.G..: Business continuation planning/disaster recovery, *IEEE Communications Magazine*, Vol.28, Issue 6, pp.25-28 (1990).
2) Fallara, P.: Disaster recovery planning, *IEEE Potentials,* Vol.22, Issue 5, pp.42-44 (2003).
3) Toigo, J.W.: Disaster Recovery Planning: Preparing for the Unthinkable (3rd Edition), Prentice Hall (2003).
4) Keeton, K., Santos, C., Beyer, D., Chase, J. and Wilkes, J.: Designing for Disasters, *Proc. 3rd USENIX Conference on File and Storage Technologies*, pp.59-62 (2004).
5) セキュリティ技術国際動向調査研究報告書, 財団法人日本情報処理開発協会 (2008).
6) 加藤 礼基：遠隔コピー機能による災害対策システム構築に関する考察（信頼性・拡張性），情報処理学会研究報告，Vol.2003, No.61, pp.23-28 (2003).
7) 加倉井 宏一，荻田 光一郎：災害対策システムのリニューアルにおける現実的災害対策レベルの評価，情報処理学会研究報告，Vol.2004, No.106, pp.1-6 (2004).
8) Gsoedl, J.: Replication alternatives, *Storage Magazine*, Vol.8, No.2, pp.10-18 (2009).
9) Azagury, A. C., Factor, M. E., and Micka, W. F.: Advanced functions for storage subsystems: Supporting continuous availability, *IBM Systems Journal*, Vol.42, No.2, pp.268-279 (2003).
10) Dufrasne, B., Gardte, W., Jamsek, J., et al.: *IBM System Storage DS8000: Copy Services in Open Environments*, pp.313-319, IBM Corporation (2008).
11) Dufrasne, B., Gardte, W., Jamsek, J., et al.: *IBM System Storage DS8000: Copy Services with IBM System z*, IBM Corporation (2008).
12) Gopisetty, S.: Automated planners for storage provisioning and disaster recovery, *IBM Journal of Research and Development*, Vol.52, No.4/5, pp.353-366 (2008).
13) Preston, W.C.: *Using SANs and NAS*, O'reilly (2002).
14) Sachs, M.W. and Varma, A.: Fibre Channel and related standards, *IEEE Communications Magazine*, Vol.34, Issue 8, pp.40-50 (1996).
15) Heath, J.R. and Yakutis, P.J.: High speed storage area networks using a fibre channel arbitratedloop interconnect, *IEEE Network*, Vol.14, Issue 2, pp.51-56 (2000).
16) 喜連川優：ストレージネットワーキング技術，オーム社 (2005).