情報処理学会論文誌　コンピューティングシステム　Vol.7　No.2　1–10 (May 2014)

ディザスタリカバリに向けた非同期リモートコピー構成資源算出方式

田口 雄一[1,a)]　山本 政行[1]

受付日 2013年10月18日, 採録日 2014年2月26日

概要：企業情報システムにおいては，広域災害にあっても業務を再開可能とするディザスタリカバリの重要性が高まっている．重要度の高いデータは，被災した時刻と，同時刻から遡って復旧可能である最終時刻との差であるリカバリポイントをできるだけ短く設定することが求められる．その目標値である RPO（Recover Point Objective）達成にあたってはリアルタイムでデータを遠隔地に複製する必要があり，ストレージシステムの備える非同期リモートコピー機能が有用である．本研究では，ディザスタリカバリシステム稼動開始前の設計段階で，正副サイト間の回線帯域と，一時的なバッファであるジャーナル記憶領域の容量を過不足なくかつ高精度で見積もることが可能な非同期リモートコピー構成資源算出方式を提案する．提案方式は，システム構成とその処理手順をモデル化し，ジャーナル領域のデータ蓄積量を予測する．さらにその時系列推移から各時刻におけるリカバリポイントを計算し，RPO の達成可否を評価することで，同構成の妥当性を判定する．実際にデータセンタで測定された書き込みデータ量の実績値を用いたシミュレーションの結果，従来手法を用いた場合と比べて，RPO を満たしながらも回線帯域を 70%削減できることを確認した．

キーワード：ストレージ，ディザスタリカバリ，遠隔バックアップ，システムサイジング

Asynchronous Remote Copy System Resource Sizing for Disaster Recovery

Yuichi Taguchi[1,a)]　Masayuki Yamamoto[1]

Received: October 18, 2013, Accepted: February 26, 2014

Abstract: The enterprise companies implement disaster recovery system in order to improve their IT system availability. A disaster recovery requirement is usually defined by RPO, Recovery Point Objective. RPO represents a tolerable period in which data might be lost due to a major incident. In order to achieve a short term RPO, a remote copy function provided by storage system realizes continuous data protection. This research proposes an asynchronous remote copy resource design method that calculates adaptive size of system resource, especially a bandwidth and a journal volume capacity without the overs and shorts in advance to launch a disaster recovery. In this method, a configuration and processes of the asynchronous remote copy system is defined by an evaluation model. This model makes it possible to estimate data amount stored on journal volumes and recovery points. The validity of system configuration can be judged by a reference of estimated recovery points. A simulation which referred actual datacenter records verified that this method reduces 70% of network bandwidth in comparison to legacy steps while achieving RPO.

Keywords: storage system, disaster recovery, remote backup, system sizing

1. はじめに

1.1 ディザスタリカバリの重要性

今日，情報システムは企業活動にとって不可欠な基盤で

[1] 株式会社日立製作所横浜研究所
Yokohama Research Laboratory, Hitachi Ltd., Yokohama, Kanagawa 244–0817, Japan
[a)] yuichi.taguchi.nh@hitachi.com

ある．企業情報システムは日々大量のデータを生成し，その蓄積量は増加の一途をたどっている [1]．昨今では企業が保有する膨大なデータから新たな知識や情報を発見しようとする試みが多くなされているように，データそのものが価値ある資産と認識されるようになっている [2], [3]．また企業では，ひとたびデータ消失が起これば事業機会を損失するだけでなく，顧客からの信頼や社会的信用を失うといった重大なリスクが認知されている [4]．このようにデータ保護は企業経営にとって重要課題の 1 つであり，様々な対策が講じられている [5], [6], [7], [8]．

データ保護には重要度に応じていくつかのレベルがある．重要なデータについては，一般的なストレージの冗長化 [9] やバックアップにとどまらず，多発するテロや火災のような拠点規模の損害，さらには自然災害による広域被災への対策が求められる．こうした大規模災害においてもデータを保護し，業務継続を可能とするために，ディザスタリカバリシステムが有用である．ディザスタリカバリシステムは，距離を隔てた 2 つ以上の拠点にデータを複製し，冗長化しておくことで，ある拠点で障害発生した状況にあっても，代替システムで継続稼働しようとするものである．

1.2　ディザスタリカバリ目標指標

ディザスタリカバリシステムの構築にあたっては，業務システムの可用性やデータの重要度に応じて適切に目標値を定める必要がある [7]．この目標として一般に用いられる指標が RPO（Recovery Point Objective）と RTO（Recovery Time Objective）である．RPO は障害や被災が発生した時刻と，同時刻から遡ってデータを復旧可能とする時刻との差を目標値とする指標である．また RTO は被災発生から復旧までの時間を目標値とする指標である．

RPO はデータ消失のリスクを一定時間内に抑えるために定義される．一例として，**表 1** の #2 にあげるように，RPO が 60 秒であれば，被災時刻から遡って 60 秒前までに記録されたすべてのデータを復旧可能とすることが求められる．言い換えれば被災時刻から遡って 60 秒未満に記録されたデータの消失を許容する目標設定でもある．

したがって，RPO は業務システムやデータの重要度が高ければ高いほど短く設定される．たとえば金融業で扱われるデータはその完全性や正確性が何よりも重視されるため，RPO を短く設定することが求められる．

表 1　RPO（Recovery Point Objective）設定の例
Table 1　Example of Recovery Point Objective.

#	RPO	意味
1	0 sec	障害発生時刻までに記録されたすべてのデータを復旧可能とする目標設定
2	60 sec	障害発生時刻から遡って 60 秒前までに記録されたデータを復旧可能とする目標設定
3	1 week	障害発生時刻から遡って 1 週間前までに記録されたデータを復旧可能とする目標設定

1.3　研究の目的

RPO を最小化するために，ほぼリアルタイムにデータを遠隔地に複製する技術として，ストレージシステムによるリモートコピー機能がある [10]．

ストレージシステムの備えるリモートコピー機能は，ストレージに定義された記憶領域であるボリュームを単位として，書き込まれたデータを順次，対となるボリュームに複製する．リモートコピー機能には，同期方式と非同期方式があるが，同期方式は，遠隔地に配置されたストレージの書き込み完了報告を待たなければならないため，サーバへの応答性能がサイト間の距離に起因する転送遅延により低下する可能性がある．そのため，一般に遠隔地との間のディザスタリカバリには非同期方式が用いられる．

一方，非同期方式の場合，被災時にデータの一部を消失する可能性がある．データ消失の可能性を低減させるためには，十分な量のシステム資源を用意すればよいが，コスト増の要因になる．したがって，ストレージシステムにおける非同期リモートコピー方式の実現においては，データ消失リスクの低減と資源最適化の両立が課題となる．この課題を解決するため，本論文では，ディザスタリカバリシステムの稼働開始前の設計段階で，資源量を過不足なくかつ高精度で見積もることが可能な非同期リモートコピー構成資源算出方式を提案し，その有効性を検証する．

2.　リモートコピーシステムの概要と課題

2.1　リモートコピーシステムの概略構成

図 1 にリモートコピーシステムの概略構成を示す．リモートコピーシステムは，業務システムが稼働する正サイトと代替システムが稼働する副サイトで構成される．正サ

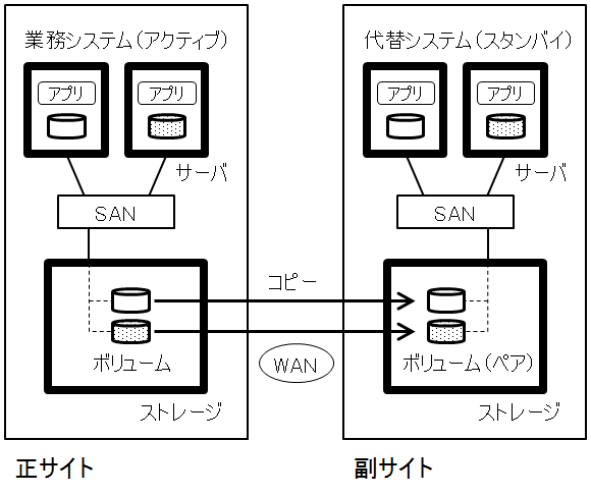

図 1　リモートコピーシステム構成
Fig. 1　Remote copy system architecture.

イト，副サイトは広域災害にあっても少なくともいずれか一方が存続するように，一定以上の距離を隔てたロケーションに設置される．

正サイトに設置されたデータセンタで業務システムが稼働する．業務システムはデータを生成，参照するアプリケーションを実行するサーバ群と，それらのデータを格納，保管するストレージで構成する．業務システムで利用されるストレージにはいくつかの種類があるが，ここではストレージエリアネットワーク（SAN）を介して複数のサーバから共有されるブロックストレージを例として取り上げる．すなわち，複数サーバで共有されたストレージに，多数のアプリケーションで生成されたデータが保管される構成となる．

2.2　非同期リモートコピー方式

非同期リモートコピー方式では，リモートコピーシステムの正サイトでサーバからストレージへの書き込みが発生すると，同ストレージでの書き込み処理完了後，即座にサーバへ完了通知を返し，書き込まれたデータを一時記憶領域であるジャーナル領域に格納する．ジャーナル領域に記録されたデータは FIFO 処理に従い，書き込みの古い順に副サイトストレージへ転送される [11], [12]．このためサーバ側で記録完了を確認したステータスであっても，副サイトへの転送を終えていないデータがジャーナル領域に残存する可能性がある．この状態で正サイトが被災した場合，未転送データは消失する．したがって，非同期方式では，目標や要件に応じてそのリスクを適切に制御する必要がある．次節ではリスク制御に向けた課題を述べる．

2.3　非同期リモートコピー構成設計の課題

非同期リモートコピーシステムの構成設計にあたっては，RPO を達成するために，正サイトストレージへの書き込みデータを RPO で定義された目標時間内に副サイトへ転送する性能を担保する構成を設計することが求められる．性能を担保するには，通信回線やバッファ容量などの IT 資源を十分に準備する必要があるが，一般的にコスト増の要因にもなりうる．

そこで，リモートコピーシステムを必要最小限の IT 資源で実現することが望ましいが，従来は設計後に運用状況に応じて適宜構成を見直す手法を採用していたため [13], [14], [15]，複雑な運用管理作業や作業中のシステム一時停止などの問題が発生していた．したがって，運用開始前の初期構成時に RPO を満たす最適な資源量を算出することが課題となる．

2.4　解決方針

上述の課題を解決するために，以下の 3 つの方針をとる．

(1)　シミュレーションのための評価モデルの策定

運用開始前の算出を可能にするため，既設システムの書き込み量を入力としたシミュレーションに基づき，リモートコピー導入後のリカバリポイントを予測するアプローチをとる．そのために実際の挙動を模擬するリモートコピーシステム評価モデルを開発する．

(2)　リカバリポイントの予測

RPO を満たすかどうかを判定するため，リカバリポイントを予測する．リカバリポイントとは副サイトにおける，現在時刻とその復旧可能時刻との差に相当する．その実現にあたり，ある想定のリモートコピーシステム構成を対象に，各時刻における未転送データ量を計算する．未転送データ量はジャーナル領域のデータ蓄積量に一致し，書き込みデータ量と転送性能のギャップにより計算できる．このジャーナル蓄積データ量からリカバリポイントを予測する．この処理では，システムへの書き込みデータ量を入力としてリカバリポイントを算出する点に特徴がある．一般に書き込みデータ量は監視しやすいパラメータであるだけでなく，すべての転送データに書き込み時刻のタイムスタンプを記録するといった独自実装が不要であるため，より汎用的なシステム構成や機器に適用可能である．

(3)　最適資源量の算出

リカバリポイントの予測を用い，同構成が RPO を達成するかどうか判定する．次に RPO を達成する様々な構成を対象にコスト計算を行い，同コストが最小となる構成を導出する．

上述の解決方針に基づき，運用開始前にリモートコピーシステムの最適な資源量を算出する．次章では，その算出方式を説明する．

3.　非同期リモートコピー構成資源算出方式

3.1　非同期リモートコピー構成算出モデル

本研究では，回線帯域やジャーナル領域容量といった IT 資源の設計が適正であるかどうか評価するための非同期リモートコピー構成算出モデルを開発した．同モデルは正副ストレージ間の非同期リモートコピー構成とそのコピー処理過程を抽象化して表現したものであり，ある書き込み負荷が生じた場合の挙動を計算によって予測するために用いる．特に正サイトストレージに対してある書き込みを発生させた場合のリカバリポイントを算出し，同構成が RPO を達成するかどうかを評価する．この評価を様々に設計された構成で再帰的に繰り返すことで，RPO を満たし，かつ資源量が最小となるリモートコピーシステム構成を見つけることが本研究の目論見である．

ストレージならびにリモートコピーシステムの構成要素は多々あるが，本研究ではリカバリポイント算出に必要な部位に絞り，抽象化したモデルによってシステム構成を表現する．**図 2** に示すように，まず正サイトに設置された

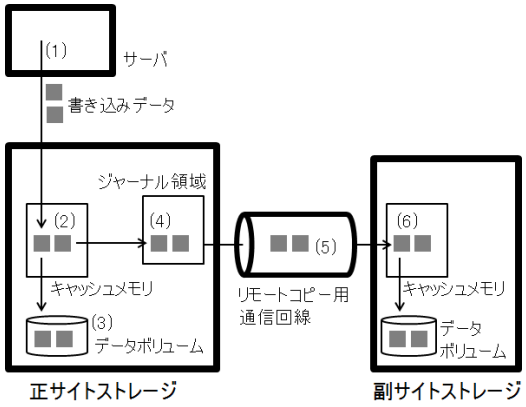

図 2　リモートコピー構成算出モデル
Fig. 2　A configuration sizing model for remote copy system.

正サイトストレージと副サイトに設置された副サイトストレージ，正副サイト間を接続するリモートコピー用通信回線を設ける．さらに各ストレージは書き込まれたデータを一時的に格納するキャッシュメモリと，リモートコピー専用のバッファであるジャーナル領域を有する [11], [12]．前者は揮発性メモリ，後者は揮発性メモリもしくはハードディスクなどの記録媒体による実装を想定する．データボリュームはデータの実体を保存する記録媒体である．

次に同モデルにおける非同期リモートコピーの処理手順を定義する．非同期リモートコピー運用において，正サイトストレージは書き込まれたデータ（処理（1））をキャッシュメモリに一時格納し，その時点でサーバに完了通知を返す（処理（2））．その後，キャッシュメモリに格納した内容をデータボリュームに記録する（処理（3））．さらにこれらの処理とは非同期で副サイトへの転送処理を実行する．転送データはジャーナル領域に時系列で格納された後（処理（4）），書き込み時刻の古い順にリモートコピー用通信回線を介して副サイトストレージに送られる（処理（5））．副サイトストレージでは受領したデータを，キャッシュメモリを介してデータボリュームに記録する（処理（6））．

すなわち同モデルにおいて，同データが正サイトストレージに書き込まれた時刻（処理（2））と，転送されたデータが副サイトに記録された時刻（処理（6））との差がリカバリポイントに相当する．**図 3** に示すように，たとえば正サイト内のサーバで生成されたあるデータが 12 時 00 分にストレージに書き込まれ，ジャーナル領域での滞留や長距離転送遅延を経て 12 時 05 分に副サイトストレージの記録を完了したケースでは，同 12 時 05 分時点のリカバリポイントは『5 分』である．このとき，仮に同 12 時 05 分に正サイトが被災した場合，12 時 00 分より後に記録されたすべてのデータは未転送であるため消失する．その際，事前に設定された RPO が 5 分より長ければ想定内であるためデータ消失も許容されるが，5 分未満であれば当初の RPO を達成できなかったことになる．

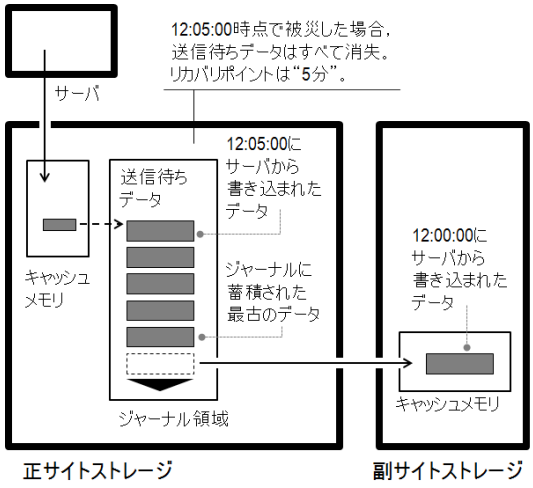

図 3　ジャーナル処理の例
Fig. 3　An example of journal processing.

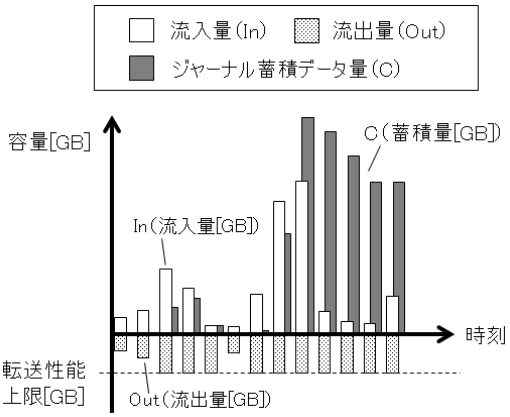

図 4　ジャーナル蓄積データ量の時系列推移例
Fig. 4　An example of journal data amount transition.

このように評価を行うことで，RPO を達成する構成であるかどうか判定することが可能となる．次節では同モデルを用いたリカバリポイント計算方法を定式化する．

3.2　非同期リモートコピー構成資源算出方法

3.2.1　ジャーナル蓄積データ量算出方法

各時刻におけるリカバリポイントは，ジャーナル領域に蓄積されるデータ量から計算することが可能である．まず，ジャーナル蓄積データ量を算出する様子を**図 4** に示す．

時刻 T における正サイトストレージへの書き込みデータ量，すなわち流入量を In_T，正サイトストレージから副サイトストレージへの転送データ量すなわち正サイトストレージからの流出量を Out_T と表記すると，ジャーナル蓄積データ量 C_T は以下の式で算出することができる．

$$C_T = C_{T-1} + In_T - Out_T \tag{1}$$

同式はすなわち，書き込みデータ量 In_T に対して，リモートコピー転送性能に依存する流出量 Out_T が不足する場合に蓄積量が増加することを意味している．ここで

C_{T-1} は時刻 T より 1 つ前の時刻におけるジャーナル蓄積データ量を表す．このようにジャーナル蓄積データ量 C もまた時系列で算出される．

流出量 Out_T は転送対象である書き込みデータ量 In_T とジャーナル蓄積データ量 C_{T-1} の和に対して転送性能が不足する場合，そのボトルネック箇所が上限となる．前述の算出モデルではハードディスクや揮発性メモリで構成されるジャーナル領域の入出力性能 P_{JNL} と，リモートコピー用通信回線性能 P_{LINE} の 2 つのパラメータがボトルネック箇所となる可能性がある．すなわち流出量 Out_T は以下のように定式化される．通信回線性能 P_{LINE} は回線帯域で表現すればよい．

$$Out_T = \min\{In_T + C_{T-1}, P_{JNL}, P_{LINE}\} \quad (2)$$

なお本計算にあたっては，キャッシュの滞留時間が入力データの時間間隔と比べて十分に短いことを前提とした．図 2 に述べたとおり，リモートコピーシステムは，キャッシュメモリに一時格納されたデータを一定時間おきにジャーナル記憶領域へ書き出す．この処理時間間隔がキャッシュによる遅延時間の最大値 $T_{DESTAGE\_INTERVAL}$ となる．ただしリモートコピーシステムは書き込まれたデータをできるだけ遅延させず，即座にジャーナル領域へ書き出そうとするため，$T_{DESTAGE\_INTERVAL}$ は入力される時系列データの時間間隔と比べて十分に短いと想定できる．たとえば流入量 In が，一般的なシステム監視のサンプリング間隔である数秒刻みで測定された時系列データであれば，リカバリポイントも同じく数秒刻みの計算となる．これに対して $T_{DESTAGE\_INTERVAL}$ はストレージ内部処理であるためミリ秒オーダの動作が可能である．数秒を単位とする計算過程において，ミリ秒単位の遅延の影響は無視して問題ない．以上の理由により，本提案ではキャッシュの影響について取り扱わず，流入量 In_T で表されるストレージへの書き込み発生時刻 T が，ジャーナル領域への書き込み時刻と一致することを前提としたモデルを開発する．

また流出量 Out_T の計算にあたっては，データ圧縮処理を考慮することでより精度を高められる．圧縮処理では正サイトストレージに滞留している転送対象データ中に同一ブロックへの書き込みがあれば，すべてのデータを送信するのではなく，最新のデータのみを転送し，転送量を削減する [17]．同処理を施すことで，実質的にはジャーナル領域入出力性能 P_{JNL} および通信性能 P_{LINE} の上限を上回る量のデータを転送することが可能となる．平均的な圧縮比を Z とすると，時刻 T における流出量 Out_T は以下の式で算出される．

$$Out_T = \min\{In_T + C_{T-1}, \frac{P_{JNL}}{Z}, \frac{P_{LINE}}{Z}\} \quad (3)$$

たとえば圧縮比 Z を 0.6 とした場合，転送対象データは 60%削減された状態で転送されるため，実効的な性能限界はそれぞれ圧縮比 0.6 で除算した値に一致する．なおここでは圧縮比 Z を定数とした例を示したが，もし既設システムの監視結果から各時刻 T における圧縮比 Z_T が既知であれば，同 Z_T を分母とすることでさらに高精度の予測を行えることは明らかである．

以上の計算では，ストレージシステムへの流入量を入力とし，ジャーナルへの蓄積データ量，すなわち要求されるジャーナル領域容量を出力とするアプローチを採用した．別のアプローチとして，ジャーナル領域容量を事前に設計値として定義し，これを上限として各時刻のジャーナル蓄積データ量を算出する手法がある．後者では，ジャーナル領域が満杯となり，それ以上蓄積できなくなる状況を考察する必要がある．このときストレージシステムはサーバからの書き込みを抑制する流入制限処理を実行するため，流入量 In_T の一部を後ろの時刻に持ち越す計算を行わなければならない．後者のアプローチは実際のストレージの挙動をより正確に再現しているといえるが，一方で流入制限に起因した，特にサーバ上のファイルシステムやアプリケーションのタイムアウトへの影響など，別の課題を考察する必要が生じる．

本研究は，既設システムで測定された，稼動実績のある書き込み負荷が発生した状況での最適資源量算出を目論む．そのため，流入量 In を変数とせず，入力値として用いる前者のアプローチを採用した．

また上記の計算式ではジャーナル領域入出力性能 P_{JNL} と通信回線性能 P_{LINE} を静的な値と仮定した．これは性能のゆらぎを予測することが困難であるための措置だが，たとえば時間帯によるスループットの低下や計画保守のための中断などが既知であれば，P_{JNL} および P_{LINE} それぞれも時系列データとして用意すればよい．流出量 Out_T の計算において，P_{JNL}，P_{LINE} もまた時刻 T の値を適用することで性能変動の影響を反映させることができる．

3.2.2　リカバリポイント算出方法

次に，時刻 T においてジャーナル領域に滞留する未転送データのうち，最も古いデータの書き込み時刻が，副サイトにおける復旧可能時刻に近似することに着目する．ジャーナル蓄積データのうち，最も古いデータ自体は未転送であるため復旧不可能だが，その直前の時刻までに書き込まれたデータは副サイトに転送済みであり，復旧可能であると想定できる．図 3 の例では 12:00:00 に書き込まれたデータが，ジャーナルに蓄積された最も古いデータの直前に書き込まれた，復旧可能データに該当する．このようにリカバリポイントを計算するためには，ジャーナル蓄積データのうち最古のデータを発見し，その書き込み時刻の直前をリカバリポイントと見なせばよい．

ジャーナル領域中，最も古いデータの発見には，ジャーナル蓄積データ量 C と流入量 In を参照する．時刻 T におけるジャーナル蓄積データ量が C_T であれば，同時点に蓄

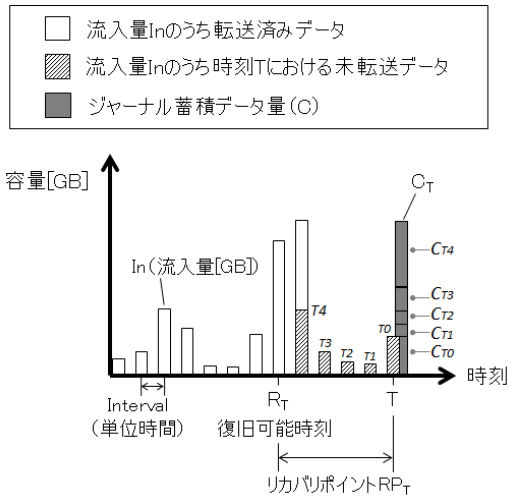

図 5　リカバリポイント算出の考え方
Fig. 5　Recovery point calculation approach.

積されたジャーナルのうち最古のデータが書き込まれた時刻は，同時刻から遡って流入量 In を累加した値が C_T に達した時刻である．またその直前が復旧可能時刻 R_T となる．すなわち，時系列の書き込みデータ量をヒストグラムで表現した**図 5** における斜線部の累加値が C_T に達する時刻（図中時刻 T_4）を特定する．さらにその直前の時刻にあたる R_T までが副サイトにおける復旧可能時刻であり，そのリカバリポイント RP_T は時刻 T と R_T の差に一致する．これらの手続きは以下のように定式化できる．

$$RP_T = T - R_T = \begin{cases} 0 & C_T = 0 \\ (n+1) \times Interval & C_T > 0 \end{cases} \quad (4)$$

ただし n は以下を満たす最小の整数である．

$$C_T \leq \sum_{i=0}^{n} In_{T-i} \quad (5)$$

RP_T は時刻 T におけるリカバリポイントを表す．$Interval$ は時系列推移の単位時間であり，前述の書き込みデータ量 In のサンプリング間隔に相当する．また n は In を時刻 T から遡って累加した値が同時刻におけるジャーナル蓄積データ量 C_T に達した時点までの累加回数であり，その 1 つ前の時刻が復旧可能時刻 R_T に一致する．

以上の手続きに従って算出したリカバリポイント RP_T の時系列推移を検証することで，各時刻において RPO を達成しているか，また回線帯域やジャーナル領域の容量といった IT 資源量に過不足がないかを評価し，最適なリモートコピー構成を導出できるようになる．

3.3　資源量算出方法

本節ではリカバリポイントの計算結果を用いたシステム構成評価と，その評価結果に基づく資源量算出方法を提案する．まず評価対象とする期間を定義し，同期間中すべての時刻において RPO を達成し，かつジャーナル領域容量と回線帯域のコストが最小となる構成を導出とすることを目論む．**図 6** にシステム構成評価の例を示す．

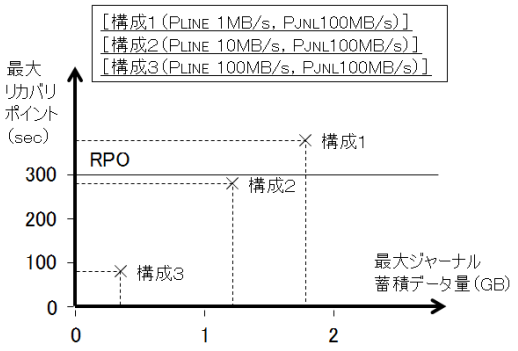

図 6　システム構成評価の例
Fig. 6　An example of configuration evaluation.

図 6 では回線帯域を変数とした 3 種類の構成を対象に，評価期間中の最大リカバリポイント RP_T と最大ジャーナル蓄積データ量 C_T の計算結果をそれぞれ例示した．

構成 1 は回線帯域不足に起因して RPO を達成できず，データ消失リスク設計が不適切であるケースに該当する．構成 2 は評価期間を通じてリカバリポイントが RPO を達成できる事例を表す．このとき最大ジャーナル蓄積データ量に一致する容量のジャーナル記憶領域を設けることで，ピーク時にもバッファが有効に作用する．構成 3 はリカバリポイントの最大値が RPO より十分に低く抑えられており目標を達成しているが，回線帯域が過剰であり，無駄なコストが生じるケースと考えられる．

以上の例にあげたように，回線帯域を変数とした多様な構成に対して本評価方式を適用し，適正構成を導出することで，リモートコピーシステムを十分な性能かつ低コストで構築することが可能となる．すなわち，以下の式で表される総コストが最小になるシステム構成を採用すればよい．

$$Cost = \min(Cost_{JNL} + Cost_{LINE}) \quad (6)$$

$Cost_{JNL}$ と $Cost_{LINE}$ はそれぞれジャーナル領域の所有コストと通信回線の運用コストを表す．図 6 で述べた手順で導出した RPO を達成する構成について，それぞれのコストを見積もり，その和が最小となる構成を最適と見なす．

4.　評価

4.1　実データを用いた実用性検証

ストレージへの書き込みを模擬するデータに Cello99 [16] を用いたリカバリポイント計算結果を**図 7** に示す．Cello99 は特定の企業内データセンタにおいて実際に発生したリード・ライトデータ量の一般公開情報であり，本方式の実用性を検証するには十分なデータである．本検証ではこのうち，あるデータボリュームに対する書き込みが局所的に増加した事象に着目し，その前後を含めた時間帯を評価期間

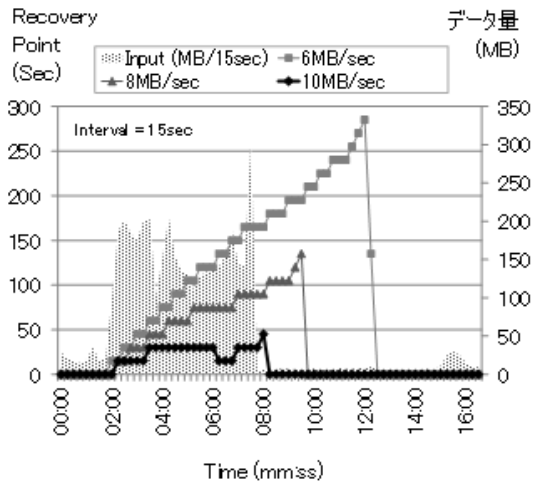

図 7　リカバリポイント時系列推移計算結果
Fig. 7　A recovery point simulation result.

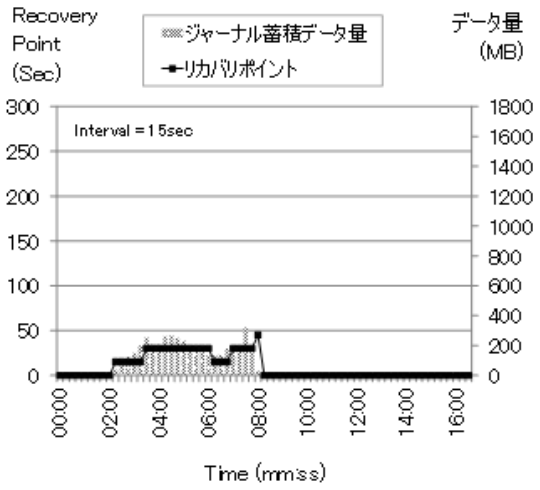

図 8　システム構成評価結果（10 MB/sec の場合）
Fig. 8　A configuration evaluation (case of 10 MB/sec).

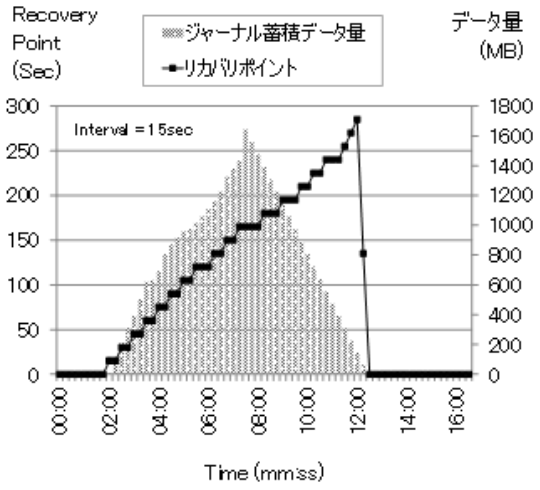

図 9　システム構成評価結果（6 MB/sec の場合）
Fig. 9　A configuration evaluation (case of 6 MB/sec).

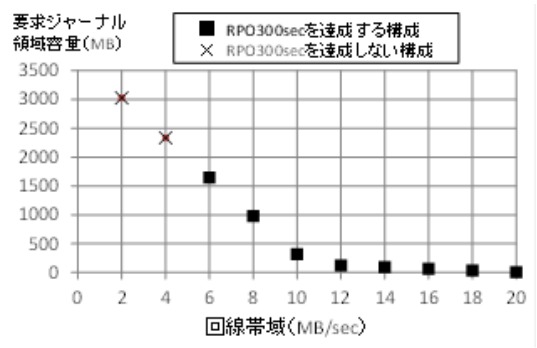

図 10　リモートコピー構成資源量算出結果
Fig. 10　Remote copy system resource sizing simulation.

として検証を行った．このような事象はジャーナル蓄積データ量が増加する要因であるため，方式評価に有効である．

なお本方式の適用にあたっては，1 週間，1 カ月といった既設システムの測定期間を評価し，期間を通じた最大ジャーナル蓄積データ量と最大リカバリポイントを予測することが有用である．

図 7 はそのデータボリュームへの書き込みを 15 秒刻みの時系列で集計し，リモートコピー用通信回線帯域を 3 通りに変動させたそれぞれの条件におけるリカバリポイント（主軸）を表す．本研究の計算手法を適用し，回線帯域を 6 MB/sec，8 MB/sec，10 MB/sec と仮定した場合のリカバリポイントの最大値はそれぞれ 285 sec，135 sec，45 sec と算出された．この評価により，たとえば同ボリュームの RPO が 45 sec であれば，少なくとも 10 MB/sec の回線帯域を，RPO が 300 sec であれば少なくとも 6 MB/sec の回線帯域を必要とするという結果が得られた．

さらに帯域 10 MB/sec，6 MB/sec それぞれの構成におけるリカバリポイントとジャーナル蓄積データ量の時系列推移を**図 8**，**図 9** に示す．

図 8 に示すとおり 10 MB/sec の帯域設計時には少なくとも 319 MB のジャーナル領域容量を設ける必要がある．同様に図 9 に示すように 6 MB/sec の帯域設計時には 1,645 MB のジャーナル領域を設ける必要がある．

回線帯域を変化させたその他の構成において同様の評価を実施した結果を**図 10** に示す．前述の回線帯域 10 MB/sec，6 MB/sec に加え，その他構成における回線帯域の増減を横軸に，それらの構成において本論文の資源量算出方式を適用した結果得られた要求ジャーナル領域容量を縦軸に表現する．さらに RPO を 300 sec とした場合に，同目標値を達成する構成は■で，達成しない構成は×で図中にプロットする．

図 10 に示した構成のうち，RPO を達成し，資源量のコストが最小となる構成を選択することにより，本研究の目的が達成される．そこで本検証では，ジャーナル領域に

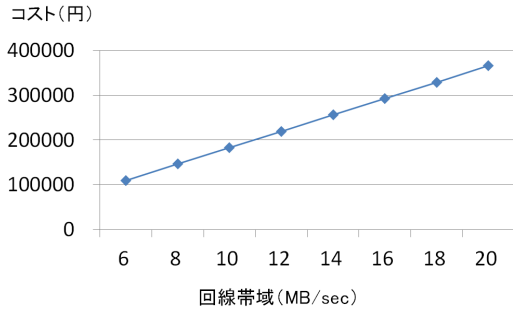

図 11　コスト試算結果
Fig. 11　A cost simulation result.

磁気記憶媒体（HDD）[*1]，通信回線に専用線[*2]を適用したケースを想定したコスト試算を行った．

図 11 は，3 カ月分の回線コスト $Cost_{LINE}$ にジャーナル領域コスト $Cost_{JNL}$ を加算した値を縦軸で表す．この結果，本ケースでは回線コストが支配的であり，帯域が最小となる 6 MB/sec の構成が最も低コストとなることが分かった．図 10 の結果とあわせることで，RPO を達成し，かつ最小コストとなる資源量（回線 6 MB/sec，ジャーナル領域容量 1,645 MB）を導出することが可能となった．

なお本検証では回線コスト $Cost_{LINE}$ が支配的であり，記憶媒体に HDD より高価な SSD や DRAM を適用した想定においても結果は同様であった．

また本評価では単一ボリュームを対象とした模擬データによるシミュレーションを行ったが，ディザスタリカバリ対象すべてのボリュームへの書き込みデータ量を合算したうえで，システム構成を設計すればよいことは明らかである．

さらに，既設システムで測定されなかった大量書き込みの発生や，通信品質の一時的な劣化といった例外ケースについても，流入量 In_T を増量する，あるいは通信性能 P_{LINE} を引き下げるといった調整により柔軟に対応できる．

5.　クラウドへの活用

本章では，提案方式の汎用性について，近年導入が進むクラウドサービスへの活用方法を通して考察する．

(1)　既存システム–クラウド間ディザスタリカバリ

この活用方法では，企業 IT システムによって生成されたデータをクラウドに転送する．転送先にクラウドを用いる場合，図 1 に示したような正サイトと副サイトで対となるストレージを設けることができるとは限らない．

本提案方式は前述の算出モデルを改編することで，ストレージのリモートコピー機能を用いないディザスタリカバリシステムにも適用可能である．

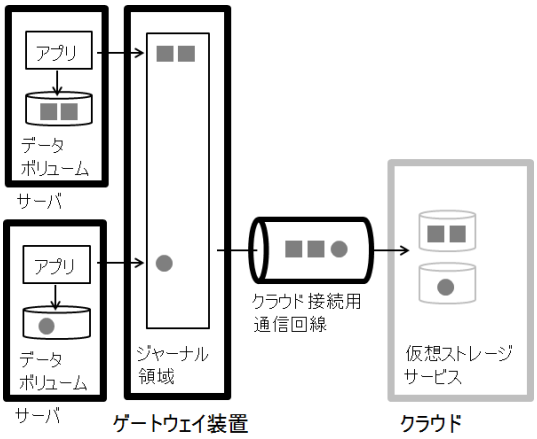

図 12　クラウド活用構成評価モデル
Fig. 12　A configuration evaluation model of cloud service.

本モデルでは，たとえばデータベースなどのアプリケーションがデータを生成・更新し，ゲートウェイ装置を通じてクラウドへ発送する構成を想定する（**図 12**）．ゲートウェイ装置には，ストレージサービスに接続するための専用アプライアンス，または一般的なキャッシングプロキシを利用する．

本モデルを用いた構成評価では，ゲートウェイ装置のジャーナル領域が図 2 の正サイトストレージにおけるジャーナル領域に該当し，3.2 節にあげたリカバリポイント計算方法をそのまま流用できる．このとき，流入量（In）には各サーバが更新する転送対象データ量の時系列推移を，流出量（Out）の上限には同システムとクラウドを接続する通信回線性能を用いる．

(2)　クラウド–クラウド間ディザスタリカバリ

本提案方式は，OpenStack[*3]などのクラウド管理ソフトウェアを利用したクラウドシステム間のディザスタリカバリにおける資源量算出にも活用できる．

これらのクラウドシステムでは，生成された仮想サーバ，仮想インスタンスの存続中にのみデータを保持するローカルストレージ領域のほかに，永続的なデータ保管を目的としたブロックストレージサービスを利用することが可能である．同サービスを提供するストレージには，OpenStack における Cinder などの仮想ストレージシステムによる実装が知られており，それらは 1.3 節に述べた外付けストレージシステムの適用が可能である．このような構成のディザスタリカバリシステムであれば，本提案方式を活用してコストを適正化できる．

6.　関連研究

従来は書き込み量のピークにあわせてシステム性能を設計することが一般的であった．たとえば図 7 に示す評価

[*1] SAS HDD, 10000 rpm, 600 GB モデル, 72,450 円＝120 円/GB と想定.
[*2] NTT 専用線サービス 42 Mbps，月額 32,000 円＝月額 6,095 円/MB/s と想定.
[*3] OpenStack は，米国における OpenStack, LLC の登録商標です.

期間における書き込み量のピークは約 300 MB/15 sec，すなわち 20 MB/sec である．本研究以前には RPO にかかわらず，このピークにあわせて 20 MB/sec の回線帯域を設けることが通例であった．これに対し，提案方式により，RPO が 300 sec であれば 6 MB/sec の帯域で足りることを構築前に算出することができる．その結果，従来と比較して 70%の回線帯域削減が可能となった．

また，従来はディザスタリカバリシステム構築前にリカバリポイントを予測するのではなく，構築後にシステム稼働状況を分析し，資源量の過不足に応じて構成を見直す方法がとられてきた．一方，提案方式は，構築前に最適資源量を見積もることができるため，従来方法よりもシステム監視や構成変更などの運用管理コストを低減させることができる可能性がある．

リカバリポイントを算出する方式について，従来稼働中システムのリカバリポイントを監視するために，転送対象データに付与されるシーケンス番号を用いてバッファ滞留時間を計測する方式が提案されている [15]．また，リカバリポイントを容易に特定できるように，正サイトストレージ側で付与される転送データのタイムスタンプを記録する方法も提案されている [11]．従来方式に比べ，本方式は，シーケンス番号やタイムスタンプを用いることなく，ストレージへのデータ流入量のみでリカバリポイントを予測できる．したがって，提案方式は従来よりも少ない情報量で，高精度な予測が可能になる．

ディザスタリカバリシステムのコスト低減に向けた取り組みには，転送データを圧縮するアプローチもある [17], [18]．重複排除の適用によりデータ量を削減する手法 [18] については，本提案の構成評価モデルを，副サイトでデータを復元するためのメタデータを転送するように改変することで応用可能である．したがって，データ圧縮と資源量予測を併用したさらなるコスト低減効果を見込める．

別の手法では，転送データ中に存在する同一ブロックへの書き込みを検出し，最新データのみを転送することでも転送データ量を削減できる [17]．こうした圧縮処理を施すディザスタリカバリシステムにおいても，稼働開始前の資源量予測が有効であることは明らかである．したがって，データ圧縮に加えて本提案による高精度な予測を適用することで，さらなるコスト低減が可能である．

7. おわりに

本論文では，ディザスタリカバリシステムの構成設計にあたり，過不足ない適量の通信回線帯域とジャーナル記憶領域容量を計算する手法を提案した．同計算にあたってはシステム構成とその処理手順をモデル化することで，ジャーナル領域におけるデータ蓄積量を算出可能とした．さらにその時系列推移から各時刻におけるリカバリポイントを計算し，ディザスタリカバリの性能目標である RPO を達成するかどうか評価することで，同構成の妥当性を判定する手法を考案した．さらに実績値によるシミュレーションを用いた評価実験により，その有効性を示した．

提案方式により，システム導入前の構成設計時にデータ消失リスクとシステムコストを適切に制御することが可能となり，要件に応じたディザスタリカバリシステムを構築できる．

謝辞　本論文の執筆にあたり，匿名査読者諸氏から数多くの有益なコメントをいただいた．ここに感謝の意を表する．

参考文献

[1] Gantz, J. and Reinsel, D.: The Digital Universe Decade – Are You Ready?, *IDC - IVIEW* (2010).
[2] LaValle, S., Lesser, E., Shockley, R., Hopkins, M.S. and Kruschwitz, N.: Big Data, Analytics and the Path From Insights to Value, *MIT Sloan Management Review* (2011).
[3] Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. and Byers, A.H.: *Big data: The next frontier for innovation, competition, and productivity*, McKinsey Global Institute (2011).
[4] 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).
[5] Toigo, J.W.: *Disaster Recovery Planning*, Principle Hall (2003).
[6] 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).
[7] 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).
[8] Rudolph, C.G.: Business continuation planning/disaster recovery, *IEEE Communications Magazine*, Vol.28, Issue 6, pp.25–28 (1990).
[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*, pp.109–116, ACM Press (1988).
[10] 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).
[11] Shulman, R.R.: Disaster Recovery Issues and Solutions, *HDS White Paper* (2004).
[12] Hitachi Data Systems: Hitachi Virtual Storage Platform, Hitachi Universal Replicator User Guide, *HDS White Paper* (2010).
[13] 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).
[14] 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).
[15] 江丸裕教，高井昌彰，原　純一：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，情報処理学会研究報告，Vol.2010-EVA-31, No.1 (2010).
[16] Mengzhi, W., Kinman, A., Anastassia, A., Anthony, B., Christos, F. and Gregory, G.: Storage Device Perfor-

mance Prediction with CART Models, *IEEE/ACM International Symposium* (2004).
[17] Hugo, P., Stephen, M., Mike, F., Dave, H., Steve, K. and Shane, O.: SnapMirror: File System Based Asynchronous Mirroring for Disaster Recovery, *Proc. USENIX Conference on File and Storage Technologies* (2002).
[18] Philip, S., Mark, H., Grant, W. and Windsor, H.: WAN Optimized Replication of Backup Datasets Using Stream-Informed Delta Compression, *Proc. USENIX Conference on File and Storage Technologies* (2012).
[19] 丸山直子，田口雄一，山本政行：ディザスタリカバリシステムにおけるストレージリモートコピー構成評価モデルの提案，情報処理学会第 70 回全国大会講演論文集，“4-539”–“4-540” (2008).

田口 雄一　（正会員）

1973 年生．1995 年早稲田大学理工学部情報学科卒業．1997 年同大学大学院理工学研究科情報科学専修修士課程修了．同年株式会社日立製作所入社．ストレージソリューションの研究に従事．

山本 政行　（正会員）

1971 年生．1994 年京都大学工学部情報工学科卒業．1996 年同大学大学院工学研究科情報工学専攻修士課程修了．同年株式会社日立製作所入社．ストレージシステムの研究開発に従事．