情報処理学会論文誌 Vol. 52 No. 2 446–463 (Feb. 2011)

動的帯域予約ネットワーク上での iSCSI セッション多重度制御手法の特性解析

野本義弘[†1,†2] 大崎博之[†3]
井上史斗[†3] 今瀬真[†3]

高遅延・広帯域の IP ネットワークを用いた，遠隔ストレージ間の大容量データ転送用途に，利用帯域の特性に応じて TCP コネクション数を自動調節可能な iSCSI プロトコルの TCP コネクション数制御機構 iSCSI-APT（iSCSI with Automatic Parallelism Tuning）を適用した場合の性能評価について述べる．光ネットワーク制御の高度化にともない，広帯域ネットワークを時間予約可能なサービス DCN（Dynamic Circuit Network）の提供が始まっている．筆者らは，iSCSI-APT 技術を，DCN に適用した場合のシミュレーション解析と性能評価を行った．この結果，TCP コネクション数が固定の方法に対する優位性があること，TCP コネクションごとのデータ転送能力に違いがあることが分かった．

Analysis of Automatic Parallelism Tuning Mechanism for iSCSI Protocol on a Network with Dynamic Provisioning

Yoshihiro Nomoto,[†1,†2] Hiroyuki Ohsaki,[†3]
Fumito Inoue[†3] and Makoto Imase[†3]

We evaluate the performance of a remote IP storage replication via long-fat network using iSCSI with automatic parallelism tuning (iSCSI-APT), which automatically adjusts the number of TCP connections according to the network status. With advancement of optical network control, introduction of dynamic circuit network (DCN) service which enables dynamic bandwidth reservation has been started. We analyze the performance of iSCSI-APT with and without a parallelism tuning mechanism. Through simulations, we show that the parallelism tuning mechanism significantly improves the effectiveness of iSCSI-APT, and that every TCP connection gains different amount of data transmission throughput.

1. はじめに

近年の広域・広帯域ネットワークの普及にともない，企業ユーザを中心に，ディザスタリカバリなど広域データ転送への需要が高まっている．また，学術系ネットワーク特有のアプリケーションとしても各研究拠点間での大容量データ転送などの要求がある．このため，すでに広く普及した IP（Internet Protocol）ネットワークを経由して，遠隔設置したストレージ装置との間で，大容量データを二重化する，あるいは高速転送する技術に対する需要が存在する．

データの二重化方式には，ファイルシステムにより管理されるファイルを対象とした転送方式と，データベースなどファイルシステムに依存しないブロックデータを対象とした方式がある．特に，後者には，サーバとストレージ間の接続をネットワーク化した SAN（Storage Area Network）技術が利用されることが多い．とりわけ，経済性と既存のインフラとの整合性に着目し，IP ベースの SAN である IP-SAN を構築するためのプロトコルとして，これまで DAS（Direct Attached Storage）で使われてきた SCSI プロトコルを TCP（Transmission Control Protocol）パケット内にカプセル化する iSCSI（Internet Small Computer Systems Interface）[1] が知られている．**図 1** に，iSCSI を利用した遠隔ストレージのバックアップシステム構成例を示す．図 1 は，遠隔ストレージ内に保存されたデータをバックアップサーバがローカルストレージに読み出す構成を例示している．

iSCSI は 2004 年に IETF（Internet Enginnering Task Force）で標準化されたプロトコルであり，イニシエータ機能のパーソナルコンピュータ用 OS への標準実装や，サーバ系 PC でのターゲット機能サポート，iSCSI 対応ストレージ製品の普及，および，低廉化が進んでいる．iSCSI を用いることにより，既存の SCSI アプリケーションが IP ネットワークを経由して遠隔のデバイスに接続することができる．その一方で，iSCSI は，下位レイヤに TCP プロトコルを用いることから，広域・広帯域ネットワークにおいてスループットが低下するという問題が指摘されている[2]．

†1 日本電信電話株式会社サービスインテグレーション基盤研究所
Service Integration Laboratories, NTT Corporation
†2 名古屋工業大学大学院工学研究科
Graduate School of Engineering, Nagoya Institute of Technology
†3 大阪大学大学院情報科学研究科
Graduate School of Information Science and Technology, Osaka University

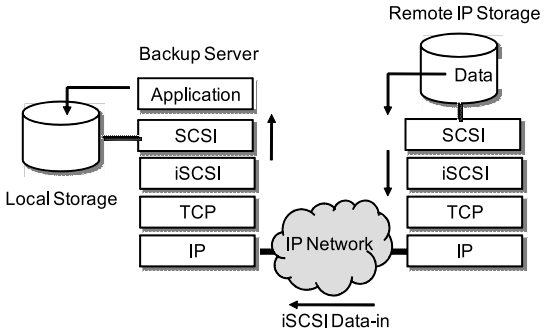

図 1　iSCSI を用いた遠隔ストレージとの二重化構成例
Fig. 1　Configuration example of remote storage replication using iSCSI.

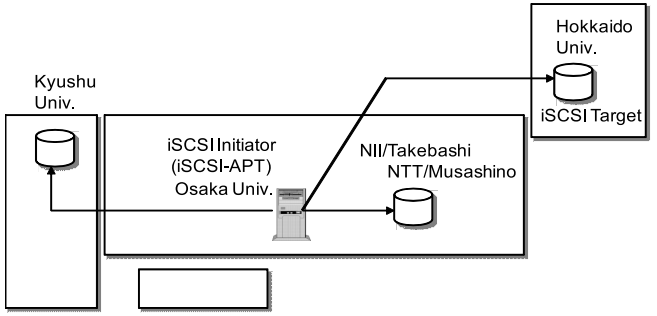

図 2　高品質遠隔バックアップ実験
Fig. 2　High performance remote backup trial.

iSCSI を用いて大容量データを高速に，連続読み込み，あるいは連続書き込みをするためには，与えられた広域・広帯域ネットワークにおいて，いかに高い iSCSI スループットを実現するかが鍵となる．多くの場合，iSCSI スループットの低下要因は，TCP そのものの機能に起因している．広域・広帯域ネットワークにおける TCP のスループット低下は，既知の問題であり，これまでさまざまな解決方法が提案されている[3),4)]．

iSCSI 以外のプロトコルを対象に，広域・広帯域ネットワークにおける TCP のスループット低下を回避する手法として，複数の TCP コネクションを利用する研究が行われてきた[5)–8)]．iSCSI では，1 本の iSCSI セッション内に，複数の TCP コネクションを用いて，データ転送を行う機能 MC/S（Multi-connections per Session）が規定されている[1)]．MC/S 機能を適切に運用すれば，iSCSI の高速化が期待できる．しかし 2 章で述べるように MC/S 機能を利用しても単純に iSCSI のスループットが向上するとは限らない．利用方法によっては，逆にスループットを低下させてしまう場合がある．複数の TCP コネクションの利用のもとで，スループット向上を実現するためには，与えられた広域・広帯域ネットワークの状況に応じて，TCP コネクション数を適切に設定しなければならない[6)]．

筆者らは，与えられた広域・広帯域ネットワークに応じて，MC/S の TCP コネクション数を自動的に調整する機構 iSCSI-APT（iSCSI with Automatic Parallelism Tuning）を提案している[9)–14)]．iSCSI-APT は，広域・広帯域ネットワークを対象に，遠隔に設置されたストレージ装置間のデータ二重化を行うなど，連続的なデータ転送を主な用途とする制御技術である．iSCSI のスループットが最大化されるよう，広域・広帯域ネットワークの遅延や実効帯域に応じて，TCP コネクション数を自動的に調整する機能を備える．加えて，iSCSI イニシエータへの付加機能であるため，MC/S 機能をサポートした市販の iSCSI ターゲット装置を対象に適用可能であるとともに，TCP の実装に依存しないことから適用範囲が広いという長所を持つ．

一方，大容量データ転送に関連するネットワーク開発の動向として，動的にエンドツーエンドパスを設定し，帯域を確保する技術 DCN（Dynamic circuit network）の開発が進められている[15),16)]．これに対応し，国立情報学研究所（NII）では学術情報ネットワーク SINET3（Science Information Network 3）において，レイヤ 1 帯域オンデマンドサービス[17)–21)]（以下，L1-BoD サービス）を 2008 年 6 月より提供中である．筆者らは，L1-BoD サービスを利用した iSCSI-APT の有効性評価を含め，学術系ネットワークに適した高速データ転送システムの研究のため，産学連携の高品質遠隔バックアップ実験（**図 2**）を実施している．本実験では，L1-BoD サービスを利用して，大阪大学に設置するバックアップサーバと，他の 3 地点（北海道大学情報基盤センター，NTT 武蔵野研究開発センター，九州大学情報基盤研究開発センター）に設置した iSCSI ストレージ装置との間で，大容量のブロックデータを転送する．バックアップサーバは，回線の予約結果（帯域と利用開始終了時刻）の情報を持たず，経路指定を行わない[⋆1]．このため，バックアップサーバは回線の開通を自動的に検知し，データ転送を開始する．回線の検知には，RFC4171 で規定されている iSNS（Internet Storage Name Service）[22)] などを利用する．回線の開通を検知後，バック

⋆1 現在，L1-BoD サーバは人手による Web アプリケーションサービスを提供中であり，サーバが自律的に予約を実現するためには，今後，Web サービスインタフェースの実現が期待される[21)]．

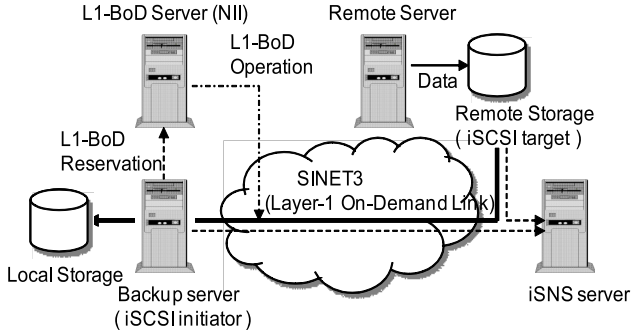

図 3　バックアップシステムの構成例
Fig. 3　Configuration example of backup system.

アップサーバはデータ転送のスループットを最大化する TCP のコネクション数を，予約で利用可能となった帯域に合わせて自動調節する．このような仕組みを技術開発の目的としている．当該実験の指向するシステム構成例を**図 3** に示す．

本論文では，L1-BoD サービスの利用を想定した iSCSI-APT の有用性について，まず，シミュレーションモデルを構成し，予備的なシミュレーションをとおして実装の方向性を明らかにする．次いで，1 本の iSCSI セッション内の複数の TCP コネクションの特性について iSCSI-APT を利用する場合と利用しない場合の比較をとおして，性能の解析を行う．

本論文の構成は以下のとおりである．2 章では，関連研究を紹介する．3 章では，iSCSI-APT のシミュレーション手法について説明する．4 章では，回線開通時と帯域変動時，それぞれの状況におけるシミュレーション結果を示し，最後に 5 章において，まとめと今後の課題について述べる．

2. 関 連 研 究

広域・広帯域ネットワークにおいて iSCSI のスループットが低下するという問題は広く知られており，これまでさまざまな解決方法が提案されてきた．広域・広帯域ネットワークにおける iSCSI の性能を評価した研究として，文献 23)–25) などが存在する．文献 23)–25) では，それぞれ実験・シミュレーション・解析によって iSCSI の性能評価が行われており，iSCSI イニシエータとターゲット間のネットワーク遅延 RTT（Round Trip Time）が増大すると，iSCSI スループットが大きく低下することが示されている．

iSCSI MC/S 機能の有効性を評価した研究として，文献 3)，26)，27) などが存在する．文献 3) では，iSCSI MC/S 機能を利用することにより，単一の TCP コネクションを用いた場合と比較して，高い iSCSI スループットを実現できることを示している．文献 3) では MC/S の多重度を 5 に固定したときの結果のみが示されている．文献 26) では MC/S の多重度が iSCSI スループットに与える影響を調査している．その結果，MC/S の TCP コネクション数を大きくすると iSCSI のスループットが増加することが示されている．文献 27) では，8.2 Gbit/s，24,000 km の高速・広帯域ネットワークを用い，64 本の TCP コネクションを用いた iSCSI の高速大容量データ転送の実験を行っている．

一方，iSCSI プロトコル以外で，複数コネクションの有効性を評価した研究は数多くあり，文献 4)，5)，28)–31) などがあげられる．たとえば，文献 4)，5) では，並列 TCP コネクションの性能評価が行われており，並列 TCP コネクション数が増加するにつれて TCP のスループットは増加するが，TCP コネクション数が多すぎると，逆に TCP のスループットの低下することが示されている．特に，文献 5) では，並列 TCP コネクション数に対して，TCP のスループットが「上に凸」の形状を形成することを示している．

つまり，並列 TCP コネクションを利用する iSCSI MC/S 利用においては，並列 TCP コネクション数が大きいほど良いということではなく，一定の TCP コネクション数を超えると iSCSI スループットが逆に低下することを意味している．以上を整理する．iSCSI スループットはネットワーク遅延（RTT）が大きくなるにつれ低下する．iSCSI のスループット低下を防ぐための一手法として iSCSI MC/S 利用は有効と考えられるが，ネットワーク環境に応じて，コネクション数を適切に調整することが重要である．

本論文では，シミュレーションにより TCP コネクション数制御手法の特性解析を行う．シミュレーションには，米国 DARPA の研究プロジェクト VINT（Visual InterNet Testbed）の研究成果である NS-2[32] を使用する．NS-2 を利用した iSCSI プロトコルに関するシミュレーションは，文献 24)–26)，33)，34) などで行われている．文献 24) を例外として，これらの文献はすべて，iSCSI を適用する回線のスペックのみに着目した，TCP コネクション数が固定の場合の評価である．文献 24) では，回線スペックに加え，ストレージ装置のスペックをも対象とする．文献 24) の対象とするストレージ構成は単体のストレージ装置ではなく，iSCSI インタフェースとファイバーチャネルインタフェースのゲートウェイ装置で中継接続されるストレージアレイとなっている．このため，ストレージ構成の内部構造まで考察したキューイングモデルを採用し，詳細なストレージ特性まで考慮した評価を行っている．

本論文は，文献 24)–26)，33)，34) と異なり，MC/S で利用する TCP コネクション数を動的に制御するモデルを対象としていること，さらに，回線のスペックだけでなく，ストレージアクセス遅延特性も包含した評価モデルを対象としていることに特徴がある．このような詳細な iSCSI プロトコルの MC/S 機能のシミュレーション評価は，筆者らの知る限りでは行われていない．

3. シミュレーション手法

3.1 iSCSI MC/S 機能

図 4 に MC/S の構成を示す．MC/S は転送するブロックデータを，複数の TCP コネクションを利用して並列に転送する機能である．利用できる TCP コネクション数の上限値 `MaxConnections` は，iSCSI のセッションパラメータとして格納されている．`MaxConnections` 値の決め方は，iSCSI セッションのログイン時（セッション確立時）に，iSCSI イニシエータおよびターゲット間で交渉され，より小さい値が適用される．

図 5 は，iSCSI セッション内での iSCSI イニシエータとターゲット間の iSCSI コマンドの基本的なサイクル（読み出しの場合）を示している．MC/S では，iSCSI Read から iSCSI Data-in 完了までのサイクルが，同一の TCP コネクション内で完結すればよく[1)]，後続の iSCSI Read は，すでに確立中のいずれの TCP コネクションも利用することができる．したがって，図 5 左側の iSCSI read cycle と右側の iSCSI read cycle は，異なった TCP コネクションで並列に送信することも可能である．なお，図中では，ネットワーク遅延を RTT，iSCSI ターゲット内のアクセス遅延を Storage I/O delay と表記している．

3.2 iSCSI-APT シミュレーションモデル

次に，筆者らが提案している TCP コネクション数の自動制御機構 iSCSI-APT の動作概要について**図 6** を用いて説明する．

iSCSI-APT は，チャンク（chunk）と呼ぶ小さな転送データ単位ごとの iSCSI スループット G を計測しながら（図中の (1)），次のチャンク転送に用いるコネクション数 N をフィードバック制御（図中の (2)）により最適化（図中の (3)）する機構である．

スループット G の測定から，iSCSI スループットを最大化するための TCP コネクション数 N を探索するアルゴリズムは，文献 6) で提案されているアルゴリズム（黄金探索法）を適用する．課題となるのは，iSCSI-APT におけるチャンクサイズの定義方法である．**図 7** を用いて，定義の考え方を説明する．図 7 では，TCP コネクション数 N それぞれで，独

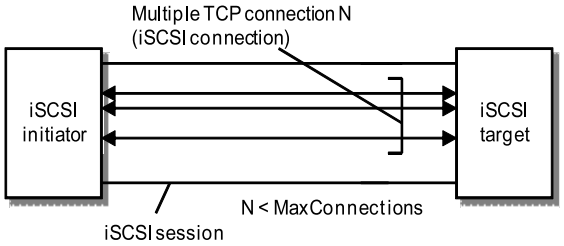

図 4　iSCSI MC/S の概要
Fig. 4　Overview of iSCSI MC/S.

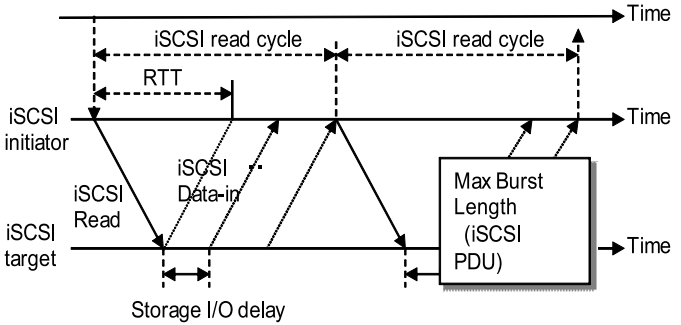

図 5　iSCSI データ転送（リード）シーケンスモデル
Fig. 5　Read sequence model of iSCSI data transfer.

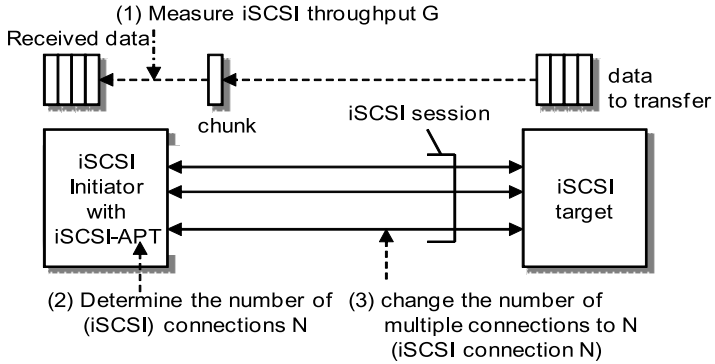

図 6　iSCSI-APT の動作概要（イニシエータのリード時）
Fig. 6　Overview of iSCSI-APT processing (read access by initiator).

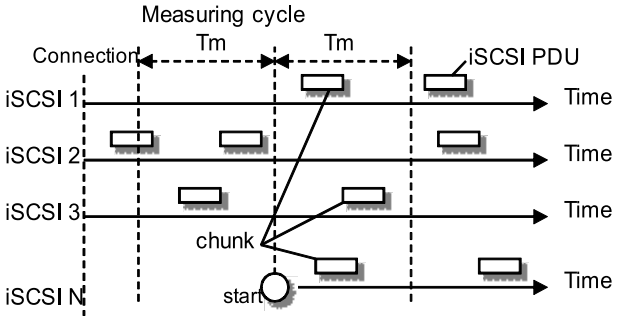

図 7　iSCSI-APT におけるチャンクサイズの定義
Fig. 7　Definition of chunk size in iSCSI-APT.

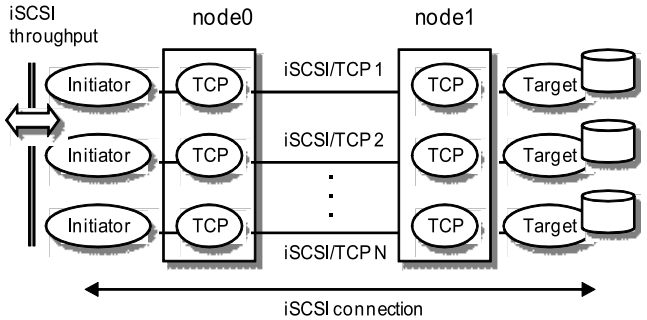

図 8　NS-2 を用いた iSCSI コネクション多重モデル
Fig. 8　Multiple iSCSI connection model using NS-2.

立に iSCSI Protocol data unit（PDU）の転送が行われている．図 7 は，スループット G の算出に必要な，測定時間 T_m と，測定時間 T_m 内にデータ転送を完了した iSCSI PDU の総和などのパラメータを示している．後者の iSCSI PDU の総和がチャンクサイズとなり，iSCSI-APT のチャンクサイズは iSCSI PDU の整数倍となる．

スループット G の計測に際して，測定時間 T_m を固定する方式と，通過する所定の数の iSCSI PDU を基準に，不定の T_m を適用する，チャンク固定の方式の 2 つに大別できる．

図 7 では，右側の計測時間 T_m 内に通過（右側に移動）した iSCSI PDU は 3 個であり，チャンクサイズは iSCSI PDU 3 個分のサイズとなっている．

iSCSI PDU のサイズは，iSCSI パラメータ `MaxBurstLength` が上限値となる．シミュレーションでは，iSCSI PDU サイズを最も転送効率の高い `MaxBurstLength` に設定する．

本論文では，特性解析のためのシミュレーターに NS-2 を使用し，iSCSI イニシエータとターゲットのシミュレーションモデルは，文献 33) の手法に基づいて構成した．文献 33) では，複数のスペックの異なる回線をシーケンシャルに接続したネットワーク構成を対象に iSCSI 層処理を OTcl（Object Tool Command Language）で記述することで，iSCSI スループットを算出するシンプルなシミュレーションモデルを提示している．本論文では，TCP 層に NS-2 の FullTCP（tahoe version）を使用する．

iSCSI イニシエータとターゲットはそれぞれの TCP エージェントが 1 対 1 で接続されるため，TCP コネクションを N 本使用する iSCSI セッションでは，各々 N 個の TCP エージェントを配置することになる．

本論文で用いる iSCSI コネクション多重モデルを**図 8** に示す．ノード 0（イニシエータ）とノード 1（ターゲット/ストレージ装置）が直結され，それぞれに N 個の TCP（tahoe version）エージェントが存在する．すなわち，本シミュレーションでは iSCSI セッションが 1 本，iSCSI コネクションが N 本存在することになる．このとき，iSCSI スループット G の計測点は，すべてのイニシエータエージェントの左側（上位層）となる．

本モデルは，iSCSI セッションを 1 本と見なすことも，N 本と見なすことも可能であるため，TCP コネクションを 1 本使用する iSCSI セッション N 本のモデルと等価となる．また，文献 5)，6) の対象とする GridFTP[35] と，文献 7)，8) の対象とするネットワークブロックデバイス（NBD[36]）による TCP マルチコネクションを用いた並列データ転送モデルとの違いは，TCP 上位層プロトコルの相違に起因する PDU 構成単位の違いと，PDU 入出力制御の違いである．このため，前者に対応した構成単位の修正と，後者に基づく制御アルゴリズムの差し替えによって，汎用モデル化が可能である．

本モデルでは，TCP 層の上位である iSCSI 層は OTcl で記述する．複数コネクションでは当該シーケンスが N 本並列に動作する．ここで，図 5 の Storage I/O delay は，

Storage I/O delay ($T_{\mathrm{I/O}}$) = 位置決め時間 + 回転待ち時間 + 転送時間

である．位置決め時間と回転待ち時間の平均値は，ディスクユニットの回転数により自動的に決まり，5,400 回転/分の場合，5.6 ms，7,200 回転/分の場合，4.3 ms となる．

OTcl で記述された iSCSI 層は，上位層からの読み出し（あるいは，書き込み）命令を iSCSI PDU 単位の読み出し（あるいは，書き込み）命令として，iSCSI ターゲットに対して発行する．iSCSI ターゲットは，iSCSI PDU 単位で iSCSI イニシエータにデータを転送する．この一連の動作は，同一の iSCSI コネクションで完結しなければならない[1]．

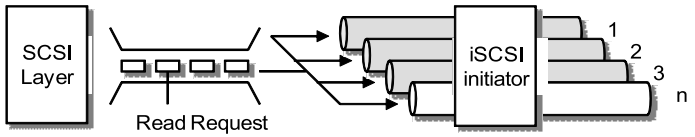

図 9　iSCSI コマンドの TCP コネクションへの振り分け（読み出し要求）
Fig. 9　iSCSI command distribution for TCP connection.

接続された複数の iSCSI コネクションに対して，SCSI データをどのように振り分けるか（**図 9**）は，iSCSI イニシエータの実装に依存する．多くの実装では，ラウンドロビンで処理される例が多い．その場合，特定の TCP コネクションでのデータ転送が遅れると他のコネクションに空きがあっても，使用できない状態が起こりうる．そのため，本シミュレーションでは，iSCSI スループットを最大化する目的で，連続する iSCSI Read 要求を，1 回目のみ各 iSCSI コネクションに，順に振り分ける．その後，iSCSI Data-in 受信の完了したコネクションに，順に振り分ける処理を行う．

3.3　予備的なシミュレーションとその結果

ここでは，スループット G の計測方法を決定するための予備的なシミュレーションを行う．iSCSI-APT を適用する L1-BoD サービスでは現在，最短 15 分の帯域予約が可能であるため帯域予約時刻での過渡状態を考慮すると，実効的に 12 分以内にデータ転送を完了する利用例がありうる．このような利用例に対しても iSCSI-APT が有効であるためには，オーバヘッドにあたる iSCSI-APT の処理過程を数分以内に完了する必要がある．本節では，上記の処理過程に要する時間として 60 s を評価基準としてシミュレーション評価を行う．

ところで，iSCSI-APT アルゴリズムの根幹である，スループット G の計測方法には，チャンクサイズを固定する方法と，計測時間を固定する 2 つの方法が考えられる．前者は，図 7 において，スループット G を計測するサイクルを，通過する所定の iSCSI PDU 数とする方式であり，後者は，図 7 において，一定の時間間隔（図中では T_m）に通過する iSCSI PDU 数を計測する方式である．後者については，時間制御であるため，後述する 4 章でのシミュレーションに際して，処理過程短縮の観点を含め，評価することは容易である．このため，予備的なシミュレーションの目的としては，前者の方式に対してのみ，基準時間に応じた iSCSI-APT の運用が可能か否かを評価するものとする．その評価結果により，前者の手法の 4 章での採否を決定する．

本シュミレーションに使用するパラメータを**表 1** に示す．

表 1 では，ネットワーク帯域として，L1-BoD の最大値である 1 Gbit/s を適用する一方で，TCP ソケットバッファサイズには，L1-BoD で利用可能な最も狭い帯域 150 Mbit/s を想定した 64 kbyte を適用する．この意図は以下のとおりである．L1-BoD での帯域予約に際して，必ずしも所望の帯域を確保できず，代替として狭帯域を利用する場合がある．このような可能性に対しては，カーネル資源の有効活用の観点から，TCP ソケットバッファサイズを小さめ（狭帯域用）に設定しておき，所望の帯域が確保できた場合でも，当該 TCP ソケットバッファサイズのまま，コネクション数の自動調整でスループットを確保するような運用方法が実装上有効と考えられるためである．

表 1　シミュレーションパラメータ
Table 1　Simulation parameters.

ネットワーク帯域	1	[Gbit/s]
RTT	10	[ms]
iSCSI Max Burst Length	1	[Mbyte]
iSCSI PDU Length	1	[Mbyte]
Storage I/O delay（$T_{\mathrm{I/O}}$）	0	[ms]
TCP コネクション数 N の初期値	4	
ネットワークリンクのキュー長	500	[packet]
ネットワークリンクのキュー管理方式	DropTail（FIFO）	
TCP ソケットバッファサイズ	64	[kbyte]

また，遠隔ストレージとのデータ転送には，書き込みと読み出しの 2 つのパターンが存在するが，両者の違いは，図 7 における iSCSI PDU の移動方向が左方向か，右方向かに帰着し，iSCSI-APT の制御手法に影響しないため，本節でのシミュレーションでは読み出しの評価のみを行う．その結果，遠隔ストレージにおいて iSCSI Read 到着後，iSCSI Data-in が発行されるまでの処理遅延が発生するが，iSCSI PDU が数 100 個単位の計測時間（単位 s）に影響しないため，ここでは無視するものとする．

iSCSI のデータ転送は，連続のデータ転送，かつ，固定長（iSCSI PDU）単位のデータ転送である．iSCSI スループットの高速化の観点からは，iSCSI PDU を最大値の `MaxBurstLength` に設定することが望ましい．

iSCSI PDU が最大値のもとで，チャンクサイズを固定値として計測時間をシミュレーションしたのが，**図 10** である．図中の x100 は，チャンクサイズを iSCSI PDU の 100 倍（100 個分）のサイズに固定化することを意味する．いい換えると，x100 のグラフ（縦軸）は，iSCSI PDU 100 個が通過するのに要する時間を示す．図 10 から，iSCSI-APT の処理段階で，実効的に使用する TCP コネクション数（横軸）が 10 以下の過程では，計測時間

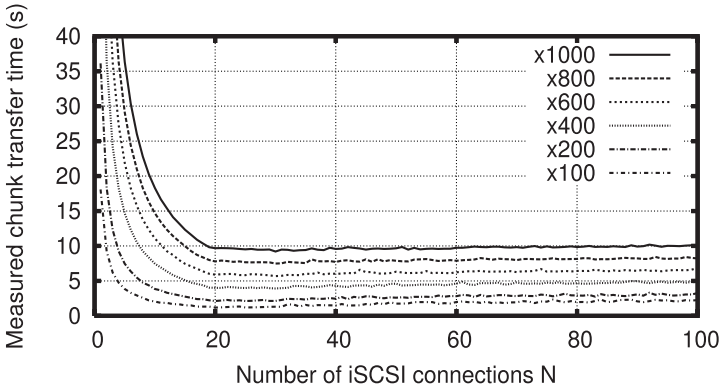

図 10　チャンクサイズ固定時の計測時間
Fig. 10　Measurement time when chunk size is fixed.

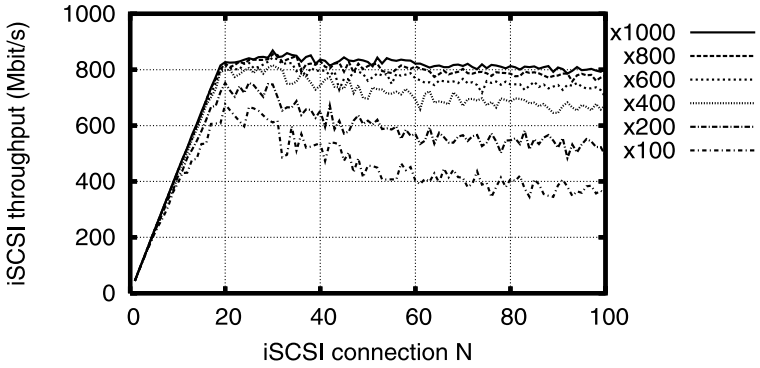

図 11　チャンクサイズ固定時のスループット
Fig. 11　Throughput when chunk size is fixed.

（縦軸）への影響が大きいことが分かる．

iSCSI-APT の適用基準として本節の冒頭で設定した 60 s に対して，N の変更回数を仮に 12 回とすれば，1 回の計測時間を 5 s 以内とする必要がある．この条件に適合するチャンクサイズは，図 10 から iSCSI PDU 単位にして 100 から 200 個（図中の x100 もしくは x200）程度に制約される．

一方，**図 11** はチャンクサイズを固定値とした場合の，スループット G のシミュレーション結果を示す．図 10 からは，TCP コネクション数 N（横軸）が大きい場合，計測時間（縦軸）が短くなり，図 11 からは，チャンクサイズ（図中の x100 などの x に続く数字）が小さいほど，スループット（縦軸）計測の誤差（グラフ間の乖離）が拡大することが読み取れる．この結果，たとえばチャンクサイズの上限を，200 個以下とした場合のスループットのピーク値が，実際のスループットのピーク値と一致しない可能性が生じる．

予備シミュレーションの結果，チャンクサイズ固定による G の計測方法では，iSCSI-APT 処理時間（TCP コネクション数の収束時間）とスループット G の計測精度の間にトレードオフの関係があり，基準とする収束時間 60 s という条件に対しては，適合することが困難であることが分かった．

ただし，チャンクサイズ固定によるスループット G の計測方法は，iSCSI-APT 動作時間を短くするという観点では適さないものの，データ転送時間が長く，iSCSI-APT 動作によるオーバヘッドが比較的小さくなる状況（たとえば，1 時間の連続転送など）では適用も可能である．

以上の予備シミュレーションの結果，本論文ではスループット G の算出手法として，チャンクサイズを固定として計測する方式を採用せず，計測時間を固定してスループット G を算出する方式のみで実施し，解析する．

なお，固定値の計測時間 T_m 中にカウントされる iSCSI PDU 数の合計を不定サイズのチャンクと定義してスループット G を算出することから，T_m と G の関係は以下となる．

$$G = \text{通過した iSCSI PDU 数}/T_m$$

TCP コネクションには，すでにデータ転送中のアクティブなコネクションと，コネクションが確立されていても，iSCSI READ が流れないコネクションがあり，iSCSI-APT 制御は両者間のバランスを調整する処理である．すなわち，iSCSI-APT 制御は，データ転送実行中にアクティブな TCP コネクション数 N を変更する．本シミュレーションの処理も同様である．

本シミュレーションでは，比較評価に際して，iSCSI-APT の収束した TCP コネクション数によって，当初より固定運用される iSCSI データ転送を採用する．当該方式は，予約帯域に対してスループットを最大化する TCP コネクション数に一致するとは限らないが，まったく異なる数値にもならないと考えられることから，基準として用いる．

3.4　iSCSI-APT による TCP コネクション数 N 最適値の探索手法

図 12 に，L1-BoD サービス利用時に，iSCSI-APT の起動を想定する 3 つの状況（Case）を示す．Case 1 は，回線開通時の起動であり，与えられた帯域に対してスループットが最

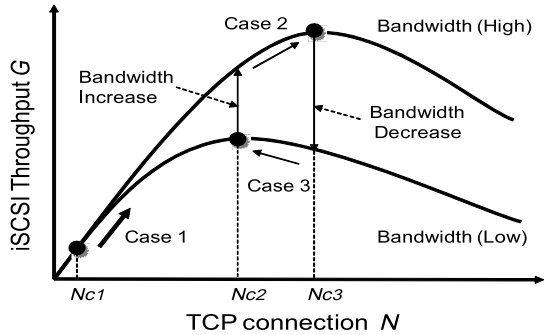

図 12　TCP コネクション数 N 最適値の探索
Fig. 12　Search of the optimum number of TCP connections N.

大値となる TCP コネクション数を探索する処理を行う．次の Case 2 は，Case1 により処理が完了し，安定状態となった以降に，L1-BoD サービスのパス変更により，帯域が増加した場合の再起動処理である．最後の Case 3 は，Case 2 の反対に，帯域が減少した場合の再起動処理である．なお，帯域変動に際して，中継するノード特性には変動がないものと仮定する．

異なる 2 つの帯域の下で N と G の関係は，理想的には 2 つの「上に凸」の形状を形成する[5]．実際には，帯域変更直後，過渡的に iSCSI スループットは安定しない．このため，計測に際して，iSCSI-APT の適切な作動タイミングは 1 つの課題である．本論文のシミュレーションでは，作動タイミングと，帯域変動にともなう iSCSI-APT 処理の再起動の要不要は議論の範疇外として，帯域変更時から 1 回分の計測間隔（T_m）の時間後に iSCSI-APT を起動させることとした．

まず，Case 1 には文献 9)–11) で提示した手法が利用可能である．一方で，Case 2 と Case 3 に際して Case 1 同様の手法による係数積を使用した N の増減を行うと収束時間が大きくなるため，N の微調整が可能な加減算による調整方法を適用する．さらに，Case 1 の増加方向探索（N が増加する方向への探索）手法に，Case 3 の減少方向探索も可能とするためのアルゴリズム修正を行う．すなわち，帯域の増加（図中の Low → High）に際して，（図中の $N_{c2} \rightarrow N_{c3}$）探索，同様に，帯域の減少（High → Low）に際して，（$N_{c3} \rightarrow N_{c2}$）探索の処理を追加する．

Case 1 から 3 を統合した処理の流れを以下に整理する．N を最適化する iSCSI-APT 処理は，2 つのステージ（ステージ 1，ステージ 2）で構成される．

ステージ 1 では N をアルゴリズム動作時の初期値 N_i から増加方向探索，もしくは減少方向探索することでスループットの最大値を N の最適値 N_{opt} の含まれる N の幅（以降，ブラケットと呼ぶ）を導出する．初期値 N_i には，図 12 中の TCP コネクション数 N_{c1}，N_{c2}，N_{c3} を適用する．

$$N_i \leftarrow \begin{cases} N_c1 & \text{(Case 1)} \\ N_c2 & \text{(Case 2)} \\ N_c3 & \text{(Case 3)} \end{cases} \tag{1}$$

そのうえで，

$N \leftarrow N_i$

として G を計測する．次に，式 (2) に基づき

$$N \leftarrow \begin{cases} k \times N & \text{(Case 1)} \quad (k > 1) \\ N + L & \text{(Case 2)} \\ N - M & \text{(Case 3)} \end{cases} \tag{2}$$

（ただし，L，M は正の整数）

N を増減させたうえで G を計測する．計測した G が前回の算出値との間で以下の式 (3) の関係

$$G(N) < G(N_{-1}) \tag{3}$$

となる（前回値よりスループットが低下する）まで，式 (2) による N の再設定と G の計測を繰り返す．式 (2) が満たされた場合，過去 3 回分の履歴にある 3 つの N を組として，

$(N_{-2},\ N_{-1},\ N) = (l, m, r)$

をブラケットに定め，ステージ 2 に移行する．

ステージ 2 では，ステージ 1 で算出した N の最適値 N_{opt}（測定スループットの唯一の極大点を示す N）の含まれる，最初のブラケットの幅から，黄金分割探索法を適用して N の最適値 N_{opt} を探索する．

具体的には，まず，式 (4) と黄金比 ν を用いて，ステージ 1 で決定したブラケット（l, m，r）から，新たな N を算出する．

$$N \leftarrow \begin{cases} l + (m-l)\nu & \text{if } |m-l| > |r-m| \\ m + (r-m)\nu & \text{otherwise} \end{cases} \quad (4)$$

$$\nu = \frac{3-\sqrt{5}}{2} \approx 0.382$$

再び G を計測し，式 (4) の関係が成り立つ（TRUE）場合，式 (6) を用いて新たなブラケットを導出する．式 (4) の関係が成り立たない（FALSE）場合は，式 (7) を用いて新たなブラケットを導出する．そのうえで，新たに導出したブラケットを用いて，式 (4) から N を再設定する．

$$\text{TRUE or FALSE} \leftarrow \{G(N) > G(m)\} \quad (5)$$

$$(l, m, r) \leftarrow \begin{cases} (m, N, r) & \text{if } m < N \\ (l, N, m) & \text{otherwise} \end{cases} \quad (6)$$

$$(l, m, r) \leftarrow \begin{cases} (l, m, N) & \text{if } m < N \\ (N, m, r) & \text{otherwise} \end{cases} \quad (7)$$

以上のステージ 2 内の過程を繰り返し，(l，m，r）が連続する整数となった場合

$N_{opt} \leftarrow m$

とする．

上記の修正アルゴリズムにより，L1-BoD リンクの利用中に増減が発生する場合や，同一の L1-BoD リンク内で複数の iSCSI セッションを動作させる場合でも，iSCSI-APT の TCP コネクション数 N 調整機能を，帯域に追随して適切に動作させることで N を帯域の変動ごとに再設定した iSCSI プロトコルの利用が可能となる．

4. シミュレーション結果

4.1 回線開通時のシミュレーション結果

4.1 節のシミュレーションでは，**表 2** のパラメータを使用する．遠隔ストレージとのデータ転送には，書き込みと読み出しの 2 つが存在するが，読み出しの評価を行う．そのため，遠隔ストレージにおいて iSCSI Read 到着後，iSCSI Data-in が発行されるまでの処理遅延の影響に留意する必要がある．Storage I/O delay（以降，$T_{\mathrm{I/O}}$）のモデル化に際して，各 iSCSI コネクションに公平なアクセス遅延を付加することとし，所定の遅延幅に一様分布するランダムな読み出し遅延を適用する．当該モデルにより，ストレージ装置内のデータ処理構造が，システム評価全体に影響しないよう配慮した．なお，読み出し遅延時間の決定に用いる一様分布の初期値には，複数回実施したシミュレーションごとに異なる値を付与する．その結果，統計的には以降の結果に影響しないことを確認済みである．

表 2　ネットワークパラメータ（回線開通時）
Table 2　Network parameters (when link start up).

ネットワーク帯域	1	[Gbit/s]
RTT	10	[ms]
iSCSI Max Burst Length	1	[Mbyte]
iSCSI PDU Length	1	[Mbyte]
Storage I/O delay（$T_{\mathrm{I/O}}$）	3-4 [uniform]	[ms]
TCP コネクション数 N の初期値	4	
ネットワークリンクのキュー長	500	[packet]
ネットワークリンクのキュー管理方式	DropTail（FIFO）	
TCP ソケットバッファサイズ	64	[kbyte]

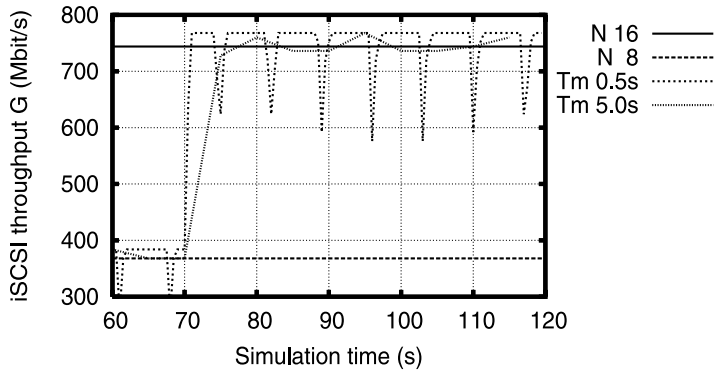

図 13　$N = 8$ から $N = 16$ へ変更時のスループット
Fig. 13　Throughput when $N = 8$ changes to $N = 16$.

最初の**図 13** は，計測時間 T_m の長短が，計測するスループットに及ぼす影響を示す．ここでは，計測時間 T_m が 0.5 ms と 5 ms の 2 つの場合について，TCP コネクション数 N をシミュレーション時刻 70 s に 8 から 16 に変更したときのスループット G の変動を示している．この図より，スループットを誤差なく計測するためには，計測時間 T_m に数秒程度

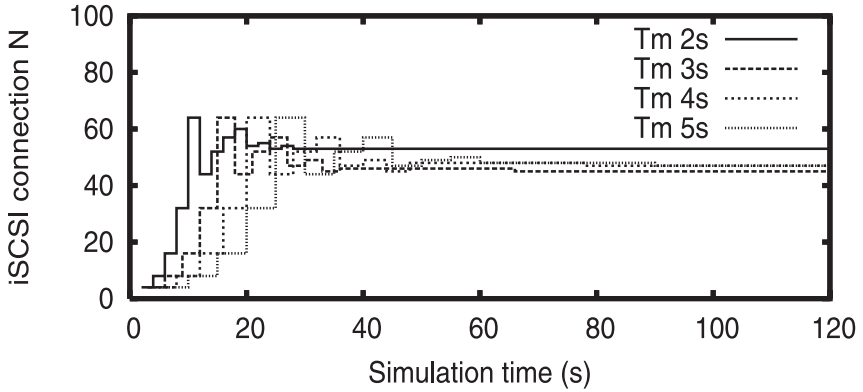

図 14　iSCSI-APT 動作時の iSCSI コネクション多重度 N の収束
Fig. 14　Convergence of iSCSI connection multiplicity N using iSCSI-APT.

が必要であり，小さすぎると実際のスループットとの誤差が大きくなり，iSCSI-APT を動作させることが困難であることが分かる．

以上の理由に基づき，秒単位の計測時間 T_m に対して，iSCSI-APT を動作させた場合の，TCP コネクション数の収束（最適化）の時間推移を示したのが，**図 14** である．

図 14 は，計測時間 T_m を 2 s から 5 s まで 1 s 単位に調整し，シミュレーションを行っている．T_m の選択により，TCP コネクション数 N の収束値に数本の違いが発生している．この理由として，2 つが推定できる．1 つは，iSCSI スループットのピーク値が先鋭でなく，台形状（複数のピーク）を形成すること，もう 1 つは図 13 で示した T_m によるスループット G の変動である．たとえば，$T_m = 5$ s の条件におけるシミュレーションでは，60 s 以内に，$N = 47$ への収束が確認できた．

iSCSI-APT 動作時の各コネクションの振舞いを確認するため，図 14 の，2 本の TCP コネクションのスループット，輻輳ウィンドウ（Cwnd）値を **図 15**，**図 16**，および **図 17** に示す．

いずれも，$T_m = 5$s の条件における計測値である．2 本の TCP コネクションには，それぞれ No.1，No.40 の記号が付与されているが，これは iSCSI-APT の処理過程で，当該コネクションが利用された以前に使用された TCP コネクション数に 1 を加算した番号である．すなわち，No.1 は iSCSI-APT 起動時より常時利用しているコネクションであり，No.40 は，iSCSI-APT 起動後，20 s ほど経た段階で初めて利用された 40 番目のコネクションで

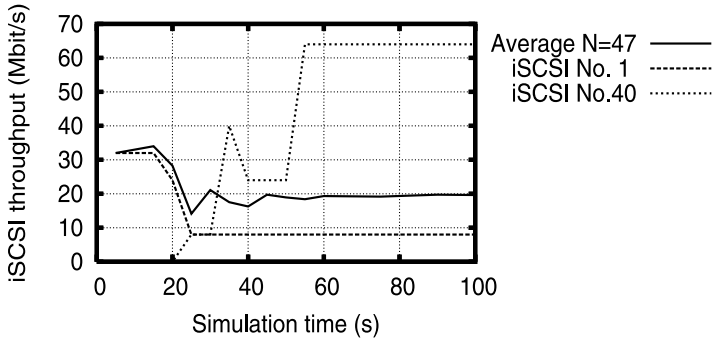

図 15　iSCSI コネクションのスループット比較例
Fig. 15　Throughput comparison example of iSCSI connection.

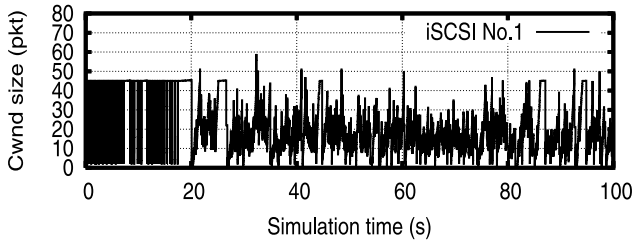

図 16　図 15 iSCSI No.1 の輻輳ウィンドウ（Cwnd）
Fig. 16　Congestion window (Cwnd) of iSCSI No.1 in Fig.15.

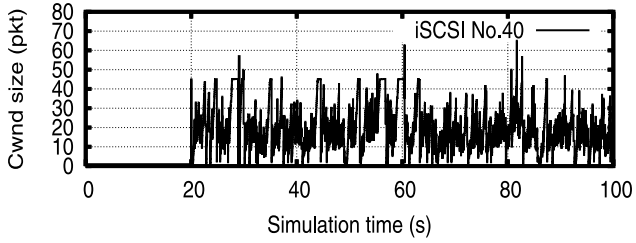

図 17　図 15 iSCSI No.40 の輻輳ウィンドウ（Cwnd）
Fig. 17　Congestion window (Cwnd) of iSCSI No.40 in Fig.15.

ある．図 15 には，両コネクションのスループットに加えて，iSCSI-APT による収束値である 47 本のコネクションの平均スループットも示している（図中の Average $N = 47$）．

この結果，No.1，No.40 双方のコネクションの総スループットへの寄与は大きく異なり，No.1 が時間とともに 8 Mbit/s まで総スループットへの寄与を低下させている一方で，No.40 が 64 Mbit/s と，8 倍も寄与していることが読み取れる．図 16 は，上記，No.1 のコネクションの輻輳ウインドウサイズを示し．図 17 は，No.40 のコネクションの輻輳ウインドウを示す．双方の輻輳ウインドウを比較すると，やや No.40 の平均が高く見えるが，この解析だけでは両者の違いがはっきりと視認できるレベルにはない．

次に，iSCSI-APT を使用しない場合と使用する場合の，iSCSI PDU の読み出しに要する時間の分布 A を計測し，**図 18** と**図 19** に示す．このとき，図 18 の計測条件として，iSCSI-APT の収束値である $N = 47$ と $T_m = 5$ s を用いる．図中のドット 1 つが，iSCSI PDU 受信時刻（横軸）を意味し，一方，縦軸が iSCSI Read 発行から，iSCSI Data-in 受信完了までに要した時間を示している．両者を比較し，明確な相違として指摘できる点は，次の 2 点である．

iSCSI-APT を利用した場合，

(1)　iSCSI Read 完了時間が 1.5 s を上回る頻度が高まる．

(2)　iSCSI Read 完了時間が 0.2 s 程度の頻度が高まる．

なお，本節では，$T_{\mathrm{I/O}}$ を考慮したシミュレーションを行ったが，$T_{\mathrm{I/O}}$ の有無によるシミュレーション効果の違いを確認するため，**図 20** を用いる．

図 20 は，図 19 と同一の表 2 の条件の下，$T_{\mathrm{I/O}} = 0$ ms に変更を行い，シミュレーションを実施している．図 19 と図 20 の違いから，$T_{\mathrm{I/O}}$ の増大，すなわち，ストレージ品質の低下が，iSCSI 読み出し時間分布の広がりに直結することが視認できる．

上記の分散傾向は，iSCSI PDU Length の増加によって顕著になる．**図 21** と**図 22** は iSCSI PDU Length を 1 Mbyte から 3 Mbyte に拡大したとき（その他のパラメータは同一のまま）の分布 B を示す．

この結果，iSCSI-APT のような，すでに複数の TCP コネクションを用いたデータ転送が行われている状態で，コネクション数の増減を行うと，データ転送効率の高いコネクションと効率の悪いコネクションの 2 極に分離していく傾向が明らかとなった．ただし，効率の良いコネクションは必ずしも，先に利用を開始したコネクションと限らず，後から利用を開始したコネクションとなることも図 15 から明らかとなっているが，どのような条件のコネクションが高速データ転送可能で，どのような条件はそうでないかについては，まだ明らか

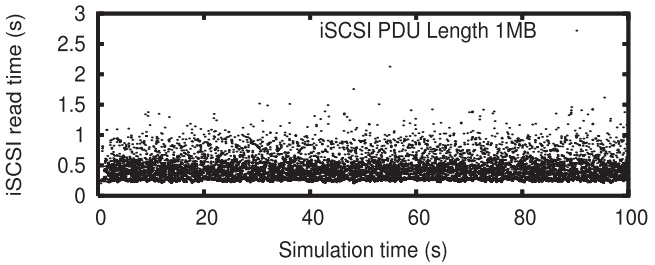

図 18　iSCSI 読み出し時間分布 A（N 固定）
Fig. 18　Distribution of iSCSI read cycle A (N fixed).

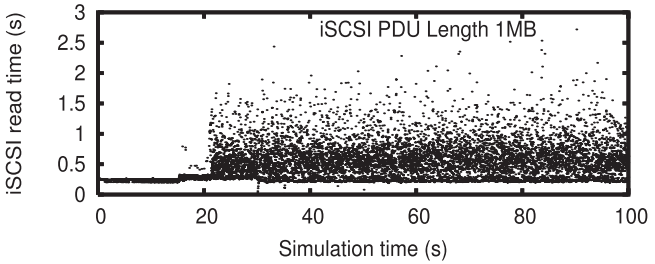

図 19　iSCSI 読み出し時間分布 A（iSCSI-APT）
Fig. 19　Distribution of iSCSI read cycle A (iSCSI-APT).

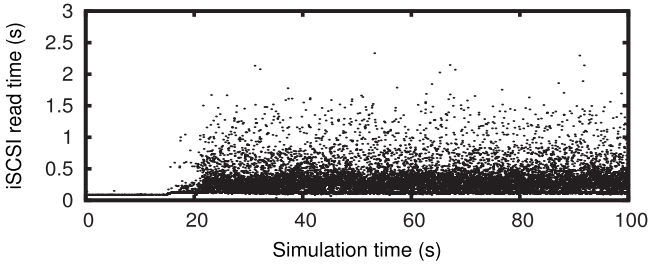

図 20　iSCSI 読み出し時間分布 A2（iSCSI-APT: $T_{\mathrm{I/O}} = 0$ms）
Fig. 20　Distribution of iSCSI read cycle A2 (iSCSI-APT: $T_{\mathrm{I/O}} = 0$ms).

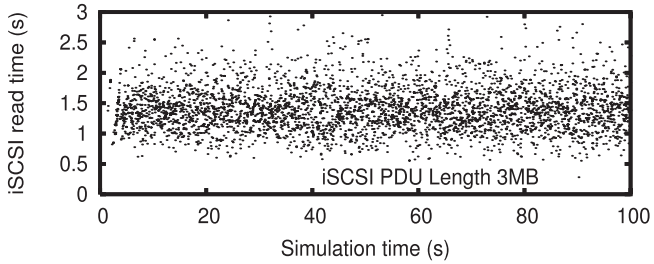

図 21　iSCSI 読み出し時間分布 B（N 固定）
Fig. 21　Distribution of iSCSI read cycle B (N fixed).

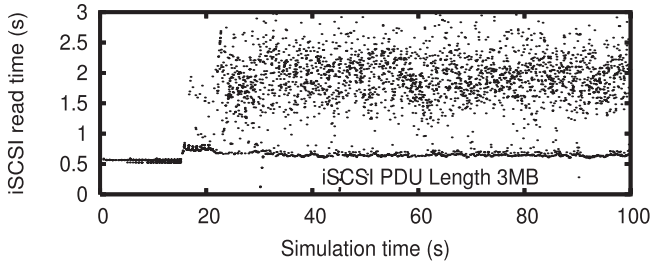

図 22　iSCSI 読み出し時間分布 B（iSCSI-APT）
Fig. 22　Distribution of iSCSI read cycle B (N fixed).

ではない．

4.2　帯域変動時のシミュレーション結果

L1-BoD サービスでは，固定の帯域を一定時間利用できるだけでなく，連続的なパス設定・解放を使用することで，帯域利用の中途で，帯域の増減が可能な仕様となっている[21)]．また，文献 21) によれば，パス設定を指定しない場合，設定されるパスの通過する L1 スイッチの数などにより，パス設定および解放時間が分単位で遅延する可能性がある．このため，連続する帯域変動を想定する場合，分単位での帯域変動もありうることを念頭に置く必要がある．このため，回線開通時だけでなく，いったん，最適化を完了した TCP コネクション数 N を，帯域増減後に，iSCSI-APT の再駆動によって再調整する方式についてシミュレーションを行う．

文献 21) の記載に基づき，変動帯域の単位を 150 Mbps として，帯域変動シナリオ**図 23** を設定し，iSCSI-APT の動作シミュレーションを行った．実際の L1-BoD サービスでは，パス変更の最小時間は 15 分であるが，連続するパス設定・解放の重畳時間による帯域変化も視野に入れ，本シナリオのような 50 s ごとの帯域変更を設定した．評価に際しては，N 固定の方式との比較を行う．

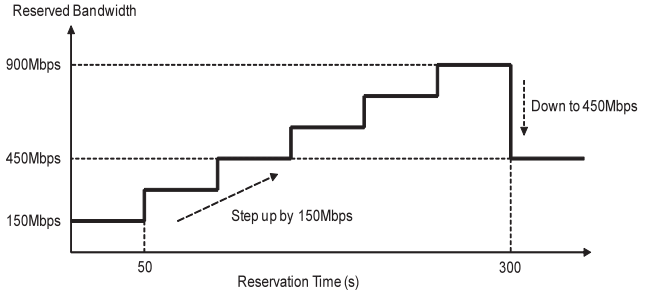

図 23　L1-BoD の帯域変動シナリオ
Fig. 23　Scenario of L1-BoD path changes.

表 3　ネットワークパラメータ（帯域変動時）
Table 3　Network parameters (when path changes).

RTT	34	[ms]
Storage I/O delay（$T_{\mathrm{I/O}}$）	0	[ms]
iSCSI-APT 初期係数 k	2	
帯域増加時増数 L	2	
帯域減少時減数 M	2	
TCP ソケットバッファサイズ	128	[kbyte]

なお，実装に際しては，バックアップサーバが予約結果の情報を持たない前提のため，帯域変動を自動検知する機能が別途必要であるが，本シミュレーションでは帯域変動後，一定の時間を経て，iSCSI-APT の再駆動を行う．

シミュレーションで使用したパラメータは**表 3** のとおりである．記載のないパラメータは，表 2 を使用する．表 2 との主な違いはアルゴリズムの違いにより，帯域増加時の増数と帯域減少時の減数が追加されていることと，RTT = 34 ms を用いることである．RTT = 34 ms は，実測に基づく北海道大学ノードと九州大学ノード間のネットワーク遅延から採用している．なお，ストレージ装置の読み出し時間は無視した．**図 24** 以降，TCP コネクション数 N を固定したデータ転送方式（以降，単に MC/S 方式と呼ぶ）によるスループットの時間

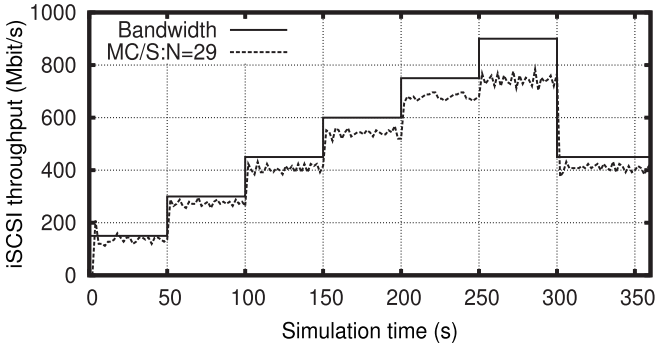

図 24　シナリオに基づく MC/S 方式のスループット（$N = 29$ に固定時）
Fig. 24　MC/S throughput on the scenario ($N = 29$ fixed).

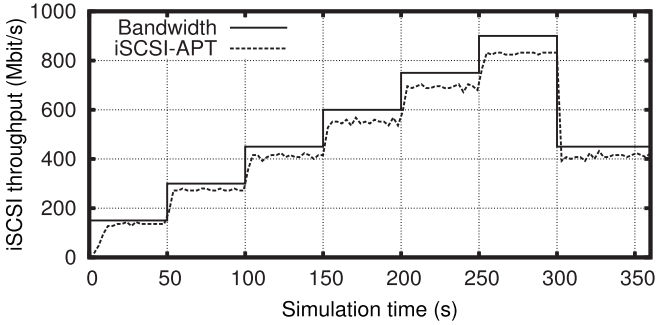

図 25　シナリオに基づく iSCSI-APT 利用時の区間スループット
Fig. 25　iSCSI-APT throughput on the scenario.

推移を示す．図 24 では，MC/S 方式で $N = 29$ に固定する場合，200 s までは予約帯域に追従するものの，250 から 300 s の区間での追従が十分でないことが分かる．

一方，**図 25** の iSCSI-APT を利用した場合は，MC/S 方式と比べて，250 から 300 s の区間においても追従性が向上するだけでなく，帯域減少を含む，すべての区間において，帯域追従後のスループットの安定性が高いことが分かる．ここでいう安定性とは，帯域変動後のスループットのバースト性が抑えられることをいう．

図 26 は，iSCSI-APT 利用時の N の時間推移である．回線開通時の N 収束に要する時間は 30 s 程度だが，以降は，本論文で新たに追加した Case 2，3 の N 調整効果により，20 s

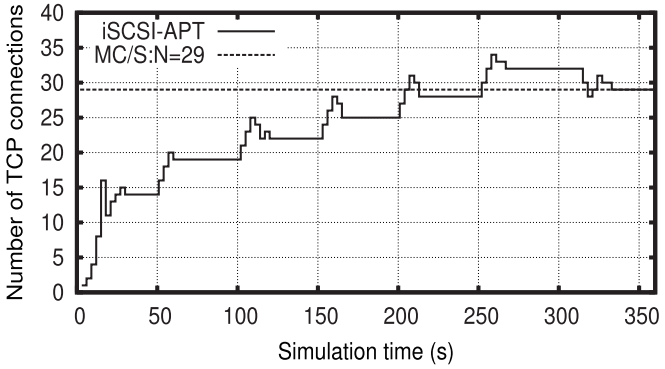

図 26　MC/S 方式と iSCSI-APT 利用時の TCP コネクション数
Fig. 26　Number of TCP connections of MC/S and iSCSI-APT.

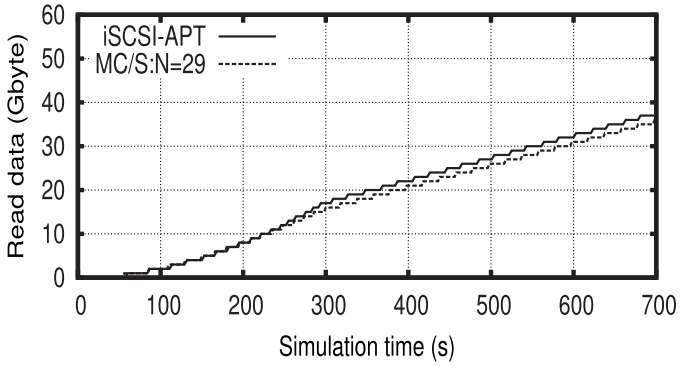

図 27　MC/S 方式と iSCSI-APT 利用時の読み出しデータ量
Fig. 27　Received data of MC/S and iSCSI-APT.

程度での収束を繰り返し，最終的に $N = 29$ に収束していることが分かる．最終的な評価指標の 1 つとして iSCSI プロトコルによるデータ転送量（本シミュレーションでは読み出し）への効果について述べる．**図 27** は，両方式による読み出しデータ量の時間推移を示したものである．本来，未知の帯域に対して，TCP コネクションをどのように設定するかは，事前調整などに基づき設定されることが一般的と考えられるが，ここでは iSCSI-APT 方式の収束値 $N = 29$ による MC/S 方式との比較を行う．図 27 からは，700 s 経過の段階で 2 Gbyte

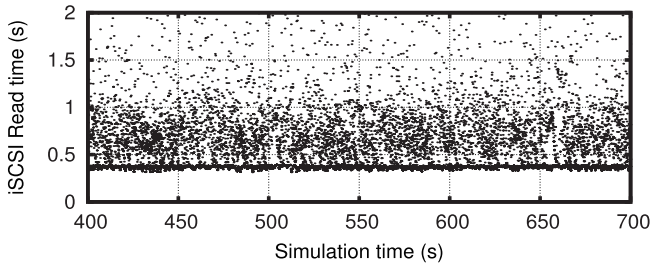

図 28　iSCSI 読み出し時間分布 C（N 固定）
Fig. 28　Distribution of iSCSI read cycle C (N Fixed).

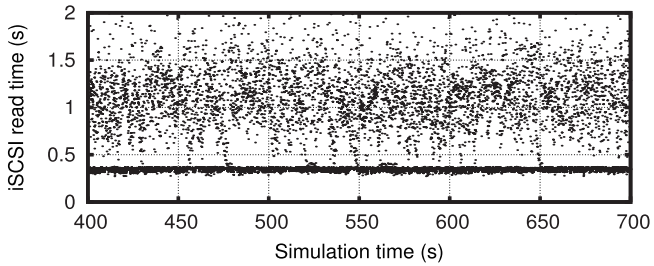

図 29　iSCSI 読み出し時間分布 C（iSCSI-APT）
Fig. 29　Distribution of iSCSI read cycle (iSCSI-APT).

のデータ転送量の差が生じている．当該シナリオによる両方式のデータ転送能力の違いは，データ転送量の 5% ほどであり，提案方式が少し上回っている．ただし，MC/S 方式については，事前のパラメータ調整などで，さらに改善の余地があることから，iSCSI-APT 方式は MC/S 方式に比べて少し上回るか同品質程度と考えることができる．この結果，あらかじめ帯域と遅延量が分からない帯域予約型サービスのような状況下でも，提案方式は，従来方式と同等以上のデータ転送能力を実現でき，帯域予約サービスの帯域予約結果にかかわらず，高い品質のデータ転送を行えることが確認できた．

最後に，300 s 経過以降，ともに $N = 29$ である両方式の特性の違いを明らかにするため，iSCSI Read 読み出し時間分布 C を測定した（**図 28**，**図 29**）．

図 29 の分布は，図 28 に比べて，2 極化傾向が顕著である．iSCSI Read 時間が伸張する iSCSI PDU の増加にもかかわらず，スループットが向上する理由の説明のため **表 4** を示す．

表 4　iSCSI 読み出し時間分布 C
Table 4　Distribution of iSCSI read cycle.

iSCSI read time (s)	≤0.35	≤0.70	≤1.40	≤2.00	≥2.00	PDU 数
図 28	8.6%	67.7%	20.5%	2.1%	1.1%	15,383
図 29	**34.0**%	33.4%	24.7%	**5.4**%	**2.4**%	15,660

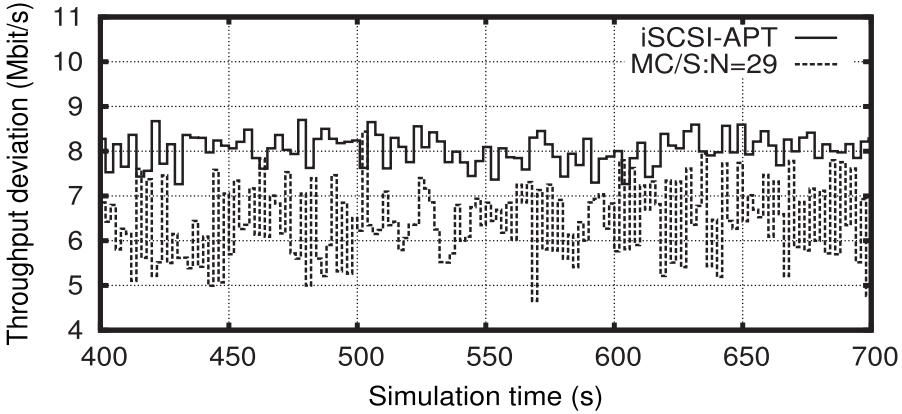

図 30　MC/S 方式と iSCSI-APT 利用時のスループット偏差
Fig. 30　Throughput deviation: MC/S vs iSCSI-APT.

表 4 は，図 28 と図 29 でプロットした 400 s から 700 s シミュレーション時間内の各 iSCSI PDU を iSCSI Read に要した時間結果（縦軸）によって分類している．この統計結果から，iSCSI-APT 適用時の iSCSI Read 時間の短縮と伸張の 2 つの傾向（2 極化）は，前者が多数（0.35 s 未満の比率が，8.6% に対して 34.0%）を占めることが分かる．このため，iSCSI Read 時間の伸張する後者が増加（1.40 s 以上の比率が，3.2% に対して 7.8%）したとしても，結果的に iSCSI Read を完了する PDU 数の増加（15,383 に対して 15,660），すなわち，スループットの向上につながることが分かる．

上記，表 4 に示される傾向は，400 s から 700 s 区間のマクロ的な特性を示す．そこでさらに，iSCSI-APT の計測間隔 T_m 区間での特性を解析するため，TCP コネクション別スループットに対する偏差，スループット偏差（Throughput deviation）を算出する（**図 30**）．スループット偏差は，表 4 が示す iSCSI 読み出し時間分布を，実際に利用した TCP コネクションごとに系列化し，TCP コネクション間のスループットのばらつきを算出することに

なる．このため，スループット偏差が大きいほど，iSCSI 読み出し時間の大きい PDU が特定の TCP コネクションに集中しているものと推定できる．また，スループット偏差の時間変動からは，TCP コネクション間の相互干渉により生じる各 TCP コネクションのスループット変動を検知でき，時間変動が小さいほど各 TCP コネクションが同じ状態を継続しているものと推定できる．

具体的なスループット偏差の評価は，これまでと同様，iSCSI-APT の収束が完了（400 s 以降，700 s まで）し，$N = 29$ で安定的に動作している状態と，同じく $N = 29$ を固定で利用する MC/S 方式との比較により示す．計測間隔は $T_m = 5\,\mathrm{s}$ とする．スループット偏差は，TCP No.i のスループットを G_i として，以下により算出する．

$$Throughput\ \ deviation: \ \sqrt{\frac{\sum_{i=1}^{N}(G_i - G)^2}{N}} \tag{8}$$

算出した図 30 では，iSCSI-APT を用いる方式は，MC/S 方式（$N = 29$）と比較して，TCP コネクション間のスループット偏差が，より大きく，時間変動による最大値と最小値の幅が，より小さい．このため，iSCSI-APT により形成される並列 TCP コネクションによるデータ転送は，各 TCP コネクションのデータ転送能力が時間とともに変動せず，能力の高い TCP コネクションは高いままに，低い TCP コネクションは低いままに推移するものと推定できる．

上記の傾向は，文献 31) で報告された傾向に類似している．スループット偏差を抑制することを目的とした研究に文献 27) があるが，当該研究は TCP コネクション数の固定を前提に，最大のスループットを得ることを目的としており，本研究のような TCP コネクション数の動的制御は対象としていない．また，iSCSI イニシエータ側で制御を行うことなく，たとえば，中継するルータにおいてアクティブキュー管理機構を導入することによっても，スループット偏差を低下させることは可能と考えられる．しかし，iSCSI-APT は固定数の TCP コネクションを用いて最大のスループットを得ることが目的ではなく，最大のスループットを得る TCP コネクション数を探索するアルゴリズムである．このため，必ずしもスループット偏差を低下させなければならない必然性はないが，iSCSI 層上位からアプリケーション層において，読み出し遅延を許容しない実装が，仮に行われた場合などは，再送要求などによって，最終的なスループットが低下する可能性も考えられるため，スループット偏差の抑制は，今後の改良に向けた選択肢として留意したい．

5. まとめと今後の課題

本論文では，筆者らの提案する iSCSI プロトコルの TCP コネクション数自動調整機構 iSCSI-APT を，国立情報学研究所（NII）の学術情報ネットワーク SINET3 の L1-BoD サービスに代表される DCN に適用することを想定したシミュレーションモデルの構成とその性能評価を行った．

その結果，DCN で利用可能な高速・広帯域ネットワークを用いて，十数分以上の連続データ転送を行う場合は，iSCSI-APT 機能を適用する iSCSI データ転送のスループットが，同数の TCP コネクション数を利用した，コネクション数固定の iSCSI データ転送方式に比べて転送効率が高いことを示した．

さらに，あらかじめ利用回線の帯域幅やネットワーク遅延が不明な場合，あるいは，回線の利用中に帯域幅が変動する予約型のネットワークの利用に際しても，iSCSI-APT はスループットを最大化する TCP コネクション数を自動探索可能であることを示した．

iSCSI-APT を用いた高速データ転送機構は，従来より HPC（High Performance Computing）分野で行われている，特殊な実装や専用装置などを用いることなく，あるいは，あらかじめ特定の回線を対象とした予備実験，チューニング過程などを必要とせずに，高効率の帯域利用を可能とする技術であり，経済的に普及させることが可能である．特に，すでに普及した iSCSI ストレージ装置をそのまま利用することが可能である点を特徴とする．

一方で，iSCSI-APT は iSCSI プロトコルによる連続データ転送用途に特化した付加機能であるため，アクセスするストレージとの間でランダムアクセスや小さなデータの読み書きをともなうアクセスに対してはオーバヘッドとなり，汎用機能として導入することができない．また，同一帯域を複数のアプリケーションで共用する環境では，正確なスループット G を計測できないため，適用できないなどの導入条件がある．その他，本論文の対象とする帯域予約機能を前提としない環境であれば，iSCSI-APT の導入により運用の利便性向上は期待できるが，回線利用効率の改善に関して大きな効果を保証することはできない点などに留意する必要がある．

本論文で用いたシミュレーション手法は，複数の TCP コネクションを用いて並列にデータ転送を行う iSCSI 以外のプロトコルにも容易に転用可能であり，DCN 上での利用に対して，同様の考察を行うことが可能である．

今後は，Linux 実装ではすでに利用可能な種々のアルゴリズムである High-Speed TCP [37]，Scalable TCP [38]，FAST TCP [39] などを適用可能なシミュレーション環境の整備と L1-

BoD サービスに対応した帯域変動検知手法の研究を進める予定である．また，並行して iSCSI-APT の Linux 実装を進め，L1-BoD を利用したシステム構築を進める予定である．

謝辞　本研究を実施するにあたり，高品質遠隔バックアップ試験へのご支援をいただく，国立情報学研究所漆谷重雄教授，北海道大学情報基盤センター高井昌彰教授，九州大学情報基盤研究開発センター岡村耕二准教授に感謝する．

参 考 文 献

1) Satran, J., Meth, K., Sapuntzakis, C., Chadalapaka, M. and Zeidner, E.: Internet small computer systems interface (iSCSI), RFC 3720 (Proposed Standard) (2004). Updated by RFCs 3980, 4850, 5048.
2) Ng, W.T., et al.: Obtaining high performance for storage outsourcing, *Proc. 1st USENIX Conference on File and Storage Technologies* (*FAST 2002*), pp.145–158 (2002).
3) Kancherla, B.K., Narayan, G.M. and Gopinath, K.: Performance evaluation of multiple TCP connections in iSCSI, *Proc. 24th IEEE Conference on Mass Storage Systems and Technologies* (*MSST 2007*), IEEE Computer Society, pp.239–244 (2007).
4) Qiu, L., Zhang, Y. and Keshav, S.: On individual and aggregate TCP performance, *Proc. Internetl Conference on Network Protocols*, pp.203–212 (1999).
5) Ito, T., Ohsaki, H. and Imase, M.: On parameter tuning of data transfer protocol GridFTP in wide-area Grid computing, *Proc. 2nd International Workshop on Networks for Grid Applications* (*GridNets 2005*), pp.415–421 (2005).
6) Ito, T., Ohsaki, H. and Imase, M.: GridFTP-APT: Automatic parallelism tuning mechanism for data transfer protocol GridFTP, *Proc. 6th IEEE International Symposium on Cluster Computing and the Grid* (*CCGrid 2006*), pp.454–461 (2006).
7) Nishijima, T., Inoue, F., Ohsaki, H., Nomoto, Y. and Imase, M.: On maximizing IP-SAN throughput over TCP connections with automatic parallelism tuning for long-fat networks, *Proc. 3rd Workshop on Middleware Architecture in the Internet* (*MidArc 2009*), pp.251–254 (2009).
8) Nishijima, T., Inoue, F., Ohsaki, H., Nomoto, Y. and Imase, M.: Performance evaluation of block device layer with automatic parallelism tuning with heterogeneous IP-SAN protocols, *Proc. 1st Workshop on High Speed Network and Computing Environments for Scientific Applicationst* (*HSNCE 2010*) (2010).
9) 井上史斗，大崎博之，野本義弘，今瀬　真：iSCSI 複数コネクションの多重度制御によるスループット最大化手法，データ工学ワークショップ（DEWS 2008）(2008).
10) 野本義弘，大崎博之，井上史斗，今瀬　真：TCP コネクション多重度制御が iSCSI スループットに与える影響，信学技報，pp.1–6 (2008).
11) Inoue, F., Ohsaki, H., Nomoto, Y. and Imase, M.: On maximizing iSCSI throughput using multiple connections with Automatic Parallelism Tuning, *Proc. 5th IEEE International Workshop on Storage Network Architecture and Parallel I/Os* (*SNAPI 2008*), pp.11–16 (2008).
12) 野本義弘，大崎博之，井上史斗，今瀬　真：レイヤ 1 帯域予約モデル上での TCP 多重度数制御手法の特性解析，信学技報，pp.357–362 (2009).
13) Kittananun, J., Inoue, F., Ohsaki, H., Nomoto, Y. and Imase, M.: Performance evaluation of iSCSI-APT on SINET3 with layer-1 on-demand bandwidth allocation, 信学技報，pp.23–28 (2009).
14) Ohsaki, H., Kittananun, J., Inoue, F., Nomoto, Y. and Imase, M.: Performance evaluation of iSCSI-APT (iSCSI with Automatic Parallelism Tuning) on SINET3 with layer-1 bandwidth on demand service, *Proc. 8th Asia-Pacific Symposium on Information and Telecommunication Technologies* (*APSITT 2010*) (2010).
15) Summerhill, R.: The new Internet2 network, *Proc. 6th Global Lambda Integrated Facility* (*GLIF*) *Meeting* (2006).
16) Campanella, M.: Development in GEANT2: End-to-end services, *Proc. 6th Global Lambda Integrated Facility* (*GLIF*) *Meeting* (2006).
17) Urushidani, S., Matsukata, J., Fukuda, K., Abe, S., Ji, Y., Koibuchi, M., Yamada, S., Shimizu, K., Takeda, T., Inoue, I. and Shiomoto, K.: Layer-1 bandwidth on demand services in SINET3, *Proc. IEEE Global Communications Conference, Exhibition and Industry Forum* (*Globecom 2007*), pp.2286–2291 (2007).
18) 漆谷重雄：[招待講演] SINET3 のネットワーク設計と運用，信学技報，pp.1–6 (2008).
19) Urushidani, S., Abe, S., Ji, Y., Fukuda, K., Koibuchi, M., Nakamura, M., Yamada, S., Shimizu, K., Hayashi, R., Inoue, I. and Shiomoto, K.: Design of versatile academic infrastructure for multilayer network services, *IEEE Journal on Selected Areas in Communications*, pp.253–267 (2009).
20) Urushidani, S., Fukuda, K., Ji, Y., Koibuchi, M., Abe, S., Nakamura, M., Yamada, S., Shimizu, K., Hayashi, R., Inoue, I. and Shiomoto, K.: Implementation and evaluation of layer-1 bandwidth-on-demand capabilities in SINET3, *Proc. IEEE International Conference on Communications* (*ICC 2009*), pp.1–6 (2009).
21) 漆谷重雄，清水香里，福田健介，林　理恵，鯉渕道紘，田沼浩行，計　宇生，今井邦宏，阿部俊二，井上一郎，中村素典，塩本公平，山田茂樹：レイヤ 1 帯域オンデマンドサービスシステムの開発，信学会論文誌，Vol.J92-B, No.7, pp.1039–1049 (2009).
22) Tseng, J., Gibbons, K., Travostino, F., Laney, C.D. and Souza, J.: Internet Storage Name Service (iSNS), RFC 4171 (Proposed Standard) (2005).
23) Lu, Y. and Du, D.H.C.: Performance study of iSCSI-based storage subsystems, *IEEE Communications Magazine*, pp.76–82 (2003).
24) Lu, Y., Farrukh, N. and Du, D.H.C.: Simulation study of iSCSI-based storage sys-

tem, *Proc. 12th NASA Goddard & 21st IEEE Conference of Mass Storage Systems and Technologies* (*MSST 2004*), pp.101–110 (2004).
25) Gauger, C.M., Kohn, M., Gunreben, S., Sass, D. and Perez, S.G.: Modeling and performance evaluation of iSCSI storage area networks over TCP/IP-based MAN and WAN networks, *Proc. 2nd International Conference on Broadband Networks* (*BROADNETS 2005*), pp.915–923 (2005).
26) Motwani, G. and Gopinath, K.: Evaluation of advanced TCP stacks in the iSCSI environment using simulation model, *Proc. 22nd IEEE/13th NASA Goddard Conference on Mass Storage Systems and Technologies* (*MSST 2005*), pp.210–217 (2005).
27) 亀澤寛之，中村　誠，稲葉真理，平木　敬ほか：長距離・高バンド幅通信における並列 TCP ストリーム間の調停の実現，*Proc. Symposium on Advanced Computing Systems and Infrastructures* (*SACSIS2004*), pp.425–432 (2004).
28) Sivakumar, H., Bailey, S. and Grossman, R.L.: PSockets: The case for application-level network striping for data intensive applications using high speed wide area networks, *Proc. 2000 ACM/IEEE Conference on Supercomputing* (2000).
29) Hacker, T.J., Athey, B.D. and Noble, B.: The End-to-End Performance Effects of Parallel TCP Sockets on a Lossy Wide-Area Network, *Proc. 16th IEEE-CS/ACM International Parallel and Distributed Processing Symposium* (*IPDPS*), pp.434–443 (2002).
30) Lu, D., Quao, Y., Dinda, P. and Bustamante, F.: Modeling and Taming Parallel TCP on the Wide Area Network, *Proc. 19th IEEE International Parallel and Distributed Processing Symposium* (2005).
31) Nakamura, M., Inaba, M. and Hiraki, K.: Fast Ethernet is sometimes faster than Gigabit Ethernet on LFN – Observation of congestion control of TCP streams, *Proc. International Conference on Parallel and Distributed Computing And Systems* (*PDCS2003*), pp.854–859 (2003).
32) Fall, K. and Varadhan, K.: *The ns manual* (*Formerly ns Notes and Documentation*) (2007), available at ⟨http://www.isi.edu/nsnam/ns/doc/ns_doc.pdf⟩.
33) Mesnier, M.: *NS modeling of iSCSI*, Carnegie Mellon University (2003).
34) Vishwakarma, S. and Bagaria, S.: iSCSI simulation study of storage system, *Proc. 10th International Conference on Computer Modeling and Simulation* (*UKSim 2008*), pp.703–707 (2008).
35) Open Grid Forum: GridFTP v2 protocol description, available at ⟨http://www.ggf.org/documents/GFD.47.pdf/⟩.
36) Verhelst, W.: Network block device (TCP version), available at ⟨http://nbd.sourceforge.net/⟩.
37) Floyd, S.: HighSpeed TCP for large congestion windows, RFC 3649 (Experimental) (2003).
38) Kelly, T.: Scalable TCP: Improving performance in high speed wide area networks, *ACM SIGCOMM Computer Communication Review*, pp.83–91 (2003).
39) Wei, D., Chen, J., Low, S. and Hegde, S.: FAST TCP: Motivation, architecture, algorithms, performance, *IEEE/ACM Trans. Networking* (*TON*), pp.1246–1259 (2006).

(平成 22 年 5 月 28 日受付)
(平成 22 年 11 月 5 日採録)

野本　義弘（正会員）

1983 年大阪府立大学工学部電気工学科卒業．1985 年同大学大学院博士前期課程修了．同年日本電信電話（株）入社．現在，同社サービスインテグレーション基盤研究所主任研究員．IMS/SDP アプリケーション，クラウド基盤利用の研究開発に従事．名古屋工業大学大学院博士後期課程在学中．電子情報通信学会，IEEE 各会員．

大崎　博之（正会員）

1995 年大阪大学大学院基礎工学研究科修士課程修了．1997 年同大学院基礎工学研究科博士課程修了．同年同大学院基礎工学研究科情報数理系助手．1999 年同大学情報処理教育センター助手．2000 年同大学サイバーメディアセンター助手．2002 年同大学大学院情報科学研究科情報ネットワーク学専攻助教授．2007 年同大学院情報科学研究科情報ネットワーク学専攻准教授．情報ネットワークに関する研究に従事．電子情報通信学会，IEEE 各会員．博士（工学）（大阪大学，1997 年）．

井上　史斗

2007 年大阪大学工学部情報工学科卒業．2009 年同大学大学院博士前期課程修了．同年 KDDI（株）入社．在学中，ネットワークストレージ用通信プロトコルの高速化に従事．電子情報通信学会会員．

今瀬　　真（正会員）

1977 年大阪大学大学院基礎工学研究科修士課程修了．同年日本電信電話公社入社．武蔵野電気通信研究所，基礎研究所，ソフトウェア研究所等において情報通信工学の基礎研究に従事．1983 年より 1 年間米国ワシントン大学計算機・通信研究センター客員研究員．帰国後，NTT マルチメディアネットワーク研究所，情報流通プラットフォーム研究所にて実用化研究を指導．2002 年大阪大学大学院情報科学研究科教授，2007 年 8 月同大学院情報科学研究科研究科長．工学博士（大阪大学，1984 年）．