社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

バースト性を考慮した高遅延ネットワーク環境下における iSCSI シーケンシャルアクセスの性能向上に関する考察

山口　実靖†　　小口　正人††　　喜連川　優†

† 東京大学 生産技術研究所 概念情報工学研究センター〒 153-8505 目黒区駒場 4-6-1
†† お茶の水女大学理学部情報科学科〒 112-8610 東京都文京区大塚 2-1-1

あらまし　iSCSI プロトコルが 2003 年 2 月に IETF により承認され, 遠隔ストレージへのアクセスプロトコル, 遠隔バックアップ, IP ネットワークを用いた広域 SAN などへの応用が期待されている. しかし, 最適化を行っていない初期環境を用いて高遅延ネットワーク上の iSCSI ストレージアクセスを行うとその性能は遅延の増加に伴い大きく劣化していく. 遠隔環境で iSCSI プロトコルを用いる場合は, 遅延による性能の劣化を最小限に抑えることが重要となる. 本稿では, 高遅延環境における大きなブロックサイズでの iSCSI シーケンシャルリードの性能について述べる. まず, シーケンシャルリード実行時の SCSI, TCP/IP, Ethernet 各層の振る舞いの解析結果を示し, SCSI プロトコルによる発生するトラフィックのバースト性が輻輳を発生させ TCP フロー制御による出力の制限を招いていることが性能低下要因であることを示す. そして, この問題に対処することにより iSCSI 性能が大きく向上されることを示す.

キーワード　SAN，iSCSI プロトコル，解析システム，高遅延環境，シーケンシャルアクセス

Performance Improving of Sequential Storage Access using iSCSI Protocol in Long-delayed Network with Considering Bursty Traffic

Saneyasu YAMAGUCHI†, Masato OGUCHI††, and Masaru KITSUREGAWA†

† Center for Conceptual Processing of Multi-Media Informaition, IIS University of Tokyo, 4-6-1 Komaba Meguro-Ku, Tokyo, 153-8505 Japan
†† Department of Information Sciences Faculty of Science Ochanomizu University, 2-1-1 Otsuka Bunkyo-ku, Tokyo, 112-8610 Japan

Abstract　The iSCSI protocol was approved by IETF in February 2003. Storage outsourcing, remote backup and wide area SAN with iSCSI became possible. To restrain performance reduction by network latency is important to obtain high performance in accessing remote storage with the iSCSI protocol. Performance of iSCSI access is influenced on by SCSI protocol, iSCSI protocol, network device and network thus an analysis system which can totally analyze these layers to obtain high performance. In this paper, we introduce our iSCSI analyze system, and we also describe the analysis of iSCSI access in long-delayed network and point out the cause of performance reduction with our analysis. By solving it, the performance of iSCSI access can significantly increase and obtain nearly equal the highest performance of the network environment.

Key words　SAN, iSCSI protocol, Analysis System, Long delayed Network, Sequential Access

1. はじめに

近年, 計算機システムが処理するデータの量は飛躍的に増大し, それに伴うストレージシステムの管理コストの増大が計算機システムの大きな問題の一つとなっている. 超大容量ストレージの管理コストの削減方法として SAN(Storage Area Network) [1] の導入や, データセンター等の遠隔ストレージの利用によるストレージ管理のアウトソーシングなどが登場してきた. SAN はストレージ専用の高速ネットワークであり, SAN の導入により従来サーバ計算機の周辺機器としてサーバ単位で管理されていたストレージを集約して管理することが可能となる. ストレージの集約によりその管理コストを大幅に削減することが可能であり, 既に多くの企業により導入されている. しかし FC(Fibre Channel) を用いる現世代の SAN(FC-SAN) はその実用に伴い, ①FC の接続距離に限界がある, ②FC 管理技術者が少ない, ③FC 導入コストが高い, などの問題点も明らか

となり, これらの問題を解決する IP-SAN が次世代 SAN として期待されるようになってきている. IP-SAN は Ethernet と TCP/IP を用いて構築する SAN であり, ①接続距離に制限が無く広域な SAN が構築可能である, ②管理技術者が多い, ③導入コストが低い, などの利点がある. IP-SAN 用のストレージアクセスプロトコルとしては iSCSI プロトコル [2] が 2003 年 2 月に IETF [3] により承認され iSCSI を用いた広域 SAN, IP-SAN, iSCSI によるデータセンター等の遠隔ストレージへの SCSI プロトコルによるシームレスなアクセスの実現が期待されている.

本稿では, iSCSI ストレージアクセスの解析システムおよびそれによる性能向上について述べる. iSCSI プロトコルは SCSI プロトコルを TCP/IP プロトコルの中にカプセル化し TCP/IP ネットワーク上で転送するプロトコルであり, プロトコルスタックは SCSI over iSCSI over TCP/IP となり, 多くの場合 SCSI over iSCSI over TCP/IP over Ethernet となる. iSCSI ストレージアクセスは上記のプロトコル各層を経由し行われるため, これら全階層がその性能に影響を与える. よって, iSCSI の性能を考察するためにはこれら各層を網羅的に解析しその性能を考察する必要がある. 我々はこれら各層を観察し iSCSI 性能の劣化原因を発見できる “iSCSI 解析システム” を構築した. 本解析システムを遅延ネットワーク上における iSCSI ストレージアクセスに適用した結果, バースト性の高い SCSI プロトコルを TCP プロトコルの上位に配置することにより輻輳が発生しそれが性能を低下させていることを確認することができ, 検出された問題点に対処することによりスループットを大きく向上させることが可能であった.

本稿は以下のように構成される. まず, 第 2. 章において研究背景を紹介する. 第 3. 章において, iSCSI における各層の振る舞いとその解析システムについて述べる. 第 4. 章において, 高遅延ネットワーク環境における iSCSI ストレージアクセスに対する解析とそれによる性能向上について述べる. 第 5. 章において関連研究を紹介する. 最後に, 第 6. 章において本稿のまとめを述べる.

2. 研究背景

本章において, 研究背景として iSCSI 解析システム, 本稿で解析を行う高遅延ネットワーク環境における iSCSI ストレージアクセスの重要性について述べる.

2.1 iSCSI ストレージアクセス解析システム

SCSI プロトコルを TCP/IP プロトコルにカプセル化して遠隔ストレージに対する SCSI アクセスを実現する iSCSI では, 上位層から依頼された SCSI アクセスが SCSI 層, iSCSI 層, TCP/IP 層, Ethernet 層を経由してネットワークで転送される. よって, これらのうちの一層における問題により iSCSI のエンドツーエンドの性能は低下することとなり, 性能の考察にはこれら全層の網羅的な解析が必要となる.

2.2 高遅延ネットワーク環境における iSCSI ストレージアクセス

iSCSI などによる IP-SAN の登場により, FC-SAN では実現できなかった, 広域 SAN が実現できるようになる. また, データセンターなどのストレージサービスプロバイダーの利用における遠隔ストレージアクセスプロトコルとして汎用の TCP/IP ネットワークと SCSI プロトコルを用いることが可能である iSCSI が有用であると考えられる. このように, 高遅延環境上で iSCSI プロトコルを用いる需要は非常に高いと言える.

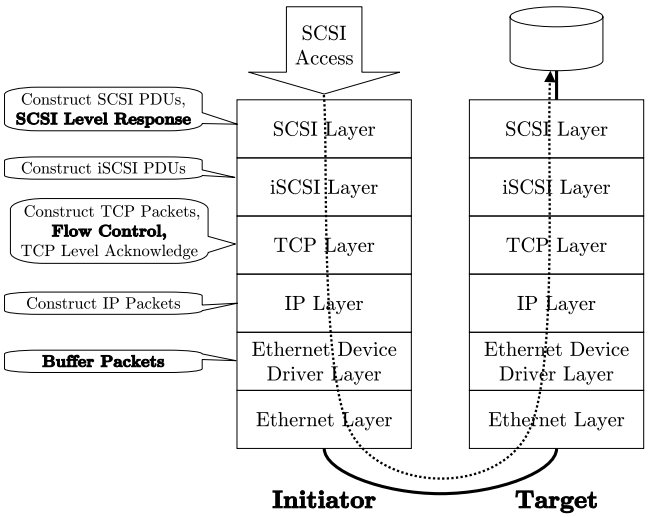

図 1 iSCSI プロトコルスタックの機能
Fig. 1 Functions in iSCSI protocol stack

我々は文献 [4] において, 最適化を行っていない初期環境で iSCSI を用いると高遅延環境における iSCSI シーケンシャルリードのスループットがネットワーク遅延の増加に伴い大きく減少すること, および iSCSI シーケンシャルリードスループットがネットワーク環境が提供できる限界スループット (ソケット通信により単純なデータ転送を行い得られるスループット. 以下これを “素のソケット通信” と呼ぶ) より大きく劣り iSCSI ではネットワークを十分に利用できていないことを述べた. そして, その原因が iSCSI 層におけるブロックの小ささにあることと, それを拡大することにより iSCSI スループットを大きく向上させることが可能であることを述べた. 上記の iSCSI 層のブロックサイズ (SCSI 層のブロックサイズはこれに等しい) の小ささは実際にネットワークを流れるパケットを監視することにより確認することが可能である. しかし, 文献 [4] の解析から得られた考察では遅延の大きな場合 iSCSI により得られるスループットはネットワークの提供できる限界性能と比べ依然大きく劣っていた. そこで, カーネル内部で処理される TCP/IP の振る舞いや, ネットワークカードドライバなども含め, プロトコルスタックを網羅的に観察および解析できる iSCSI 解析システムを開発した.

3. iSCSI ストレージアクセスの振る舞い

本章において, 多段のプロトコル構成で行われる iSCSI ストレージアクセスの各層の振る舞いと, その解析について述べる.

3.1 iSCSI ストレージアクセスの構成

iSCSI プロトコルスタックは図 1 の様な構造になっている. iSCSI アクセスにおいては上位層から SCSI アクセスを受け,

これが SCSI, iSCSI, TCP/IP, Ethernet の各層を経てネットワークを超えターゲットに転送さる．また，各層の機能も図 1 の様になる．SCSI 層は上位層からデータ要求 (Read の場合) 等を依頼され，SCSI プロトコルに基づきターゲットデバイスからそれを獲得する．SCSI 層において確認応答 (SCSI Response) を行うため，確認応答待ち状態において処理が停止し性能に影響を与える可能性がある [4]．iSCSI 層は SCSI 層から依頼された SCSI アクセスを iSCSI PDU(Protocol Data Unit) にカプセル化し TCP 層に転送する．TCP/IP 層は，iSCSI 層から依頼された iSCSI PDU を MSS(Maximum Segment Size) 毎に分割し下位層の Ethernet に転送する．TCP 層においても確認応答 (TCP Ack) が行われ，広告 Window，輻輳 Window の 2 種の Window サイズが Ack を受信せずに送信できるデータ量の上限値となる．よって，SCSI 層同様に送信が停止され性能に影響を与える可能性がある．Ethernet デバイスドライバは，高速な計算機内部のデータ転送とそれと比較し低速なネットワークにおけるデータ転送のバッファを行う．多くの環境において，NIC デバイスドライバ以下 (計算機のバス，NIC，ネットワーク) の転送速度が最も低速であるため送信要求量に時間的偏りがある場合は，ドライバ内におけるその緩衝が重要となる．Ethernet 層およびそれに接続されたネットワークでは実際のパケットの転送を行う．一般に許容されるパケットサイズの最大値はこの Ethernet 層が最小であり，MTU(Maximum Transfer Unit) は Ethernet 層により決定される．これにより送受信処理回数が決定され性能に影響を与える可能性がある．また，ネットワーク内で発生したパケットの損失は TCP 層において検出され TCP のフロー制御に影響を与え，結果 iSCSI 性能に影響を与える可能性がある．

3.2 iSCSI の解析

iSCSI ストレージアクセスは前節の各層を全て経由し，各層がその性能に影響を与えるために，iSCSI 性能を考察するためにはこれら各層を網羅的に解析する必要がある．我々はこれらを網羅的に解析できる iSCSI 解析システムを構築した．本解析システムは主に①通信内容のプロトコル翻訳 (SCSI, iSCSI, TCP/IP 各層に対応)，②パケット転送の時間軸上における可視化，③TCP フロー制御の監視，④簡易版イニシエータを用いた iSCSI ストレージアクセスの生成とその解析，などの機能を持つ．

①“プロトコル翻訳機能” では，実際にネットワークで転送されているパケットの内容を IP, TCP, iSCSI, SCSI の各プロトコルに基づき翻訳し，その意味を確認することが可能である．iSCSI では，ネットワークで転送されるパケットには IP ヘッダ，TCP ヘッダ，iSCSI ヘッダ が記載されその後に SCSI PDU が包含されることとなる．本翻訳機能では，上記の SCSI, iSCSI, TCP, IP の全プロトコルに対応し，プロトコル内容を確認することができる．

②“パケット転送の時間軸上における可視化” 機能では，図 2 の様に実際にネットワーク上を移動するパケットを時間軸上に描画可能であり，ネットワークの使用率等を視覚的に確認することが可能である．

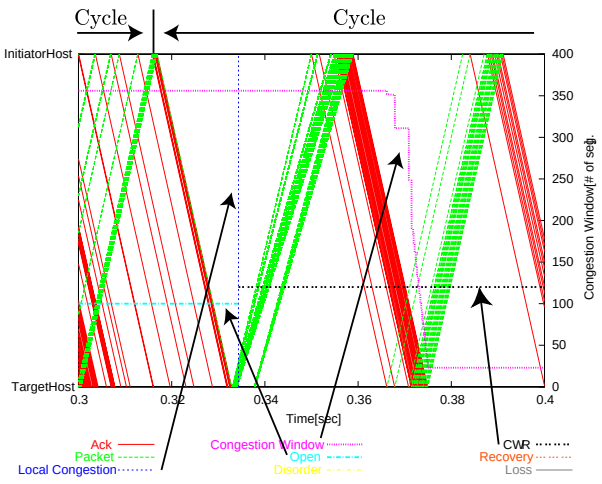

図 2　パケット転送

Fig. 2　Packet Transfer

簡易版 iSCSI イニシエータと UNH iSCSI ターゲット使用．片道遅延時間 16m 秒，iSCSI ブロックサイズ 4MB．

また，後述する TCP フロー制御モニタ機能による解析結果を併せて描画することによりパケットの送受信の振る舞いと，TCP フロー制御の振る舞いの関係を確認することが可能である．図 2 の例では，時刻 0.33 秒近くにおいてターゲット計算機が過度に多いパケットの送信を試みその結果 “Local Congestion” を招いているのが確認できる (詳細は第 4.3 節において後述する)．

③“TCP フロー制御の監視機能” では，TCP 実装のフロー制御の状態をカーネル外部から監査することを可能とする．一般に TCP 実装はネットワークの輻輳崩壊を回避するためにフロー制御を行い，上位層からデータの送信を依頼されても必ずしもその全てを即時に送信せず輻輳 Window を送信の上限値として送信を制限する．本解析システムの “TCP フロー制御の監視” 機能では，イニシエータ計算機およびターゲット計算機の TCP 実装が行っているフロー制御の状態を観察することを可能とする．多くの実装では TCP 実装はカーネル空間で実行されカーネル空間内で輻輳 Window の値の制御を行うため本解析システムにおける機能のうち “TCP フロー制御の監視” 機能のみ実装依存となっており，本解析システムでは Linux TCP 実装を用いて構築されている．Linux TCP 実装では Ack 受信毎に輻輳 Window が単調に増加し続け，Sack 受信，ローカルデバイス輻輳，確認応答のタイムアウトなどのイベントにより TCP 実装が輻輳を検出すると輻輳 Window が縮小し出力スループットが制限されることとなる．本解析システムの “TCP フロー制御監視” では，輻輳 Window の推移，TCP 実装が輻輳と判断し輻輳 Window を縮小させる原因となるこれらのイベントの発生をカーネル外部から確認することができる．

④“iSCSI ストレージアクセスの生成機能” では，OS のドライバ等の制限を受けずに純粋な iSCSI の性能を評価するための手法として，簡易イニシエータ実装を用いて iSCSI PDU を自由に生成し発行する機能を提供している．簡易版イニシエータは iSCSI 仕様に基づいて iSCSI PDU を生成しソケット API

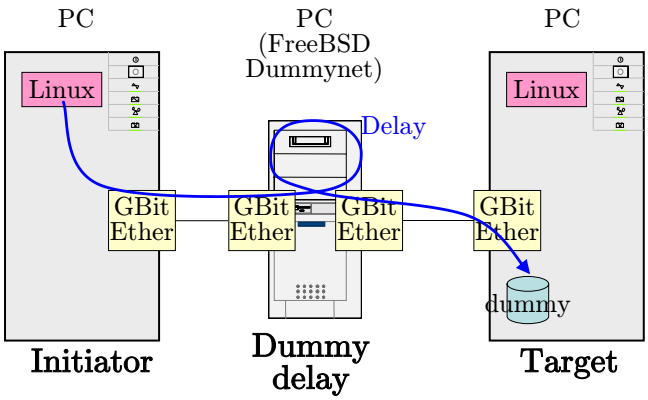

図 3　実 験 環 境
Fig. 3　Experimental Environment

を用いて直接ターゲット計算機に対して PDU を送信するため, OS 付随の SCSI ドライバ等の制限を受けずに意図した iSCSI PDU を送信することが可能となる. iSCSI 仕様は後述の UNH 同様に draft 18 [2] 準拠となっている. また, 簡易版イニシエータ実装となっており計測に必要な login 処理, Read 処理などが実装されておりエラー処理等の機能は実装されていない. 簡易版イニシエータは, カーネル空間で動作するデバイスドライではなくユーザ空間で動作し iSCSI PDU をソケット API で送受信する.

4. 高遅延環境における iSCSI ストレージアクセスの解析と性能向上

本章では, 第 3. 章で述べた解析システムを高遅延環境における iSCSI シーケンシャルリードアクセスに対し適用し, その性能劣化原因の検出および検出点の改善により性能の向上が可能であることを示す.

4.1　高遅延環境における iSCSI シーケンシャルリード

我々は, 文献 [4] において最適化を行っていない初期環境を用いて iSCSI アクセスを行うと iSCSI ブロックサイズが小さくなり, 高遅延環境において iSCSI シーケンシャルリードを行うとネットワーク使用率が非常に低くなりそのスループットが大きく劣化してしまうことを述べた. そして, iSCSI ブロックサイズを大きくすることによりそのスループットを大きく向上させることが可能であること, ブロックサイズを拡大しても iSCSI スループットは素のソケット通信より大きく劣ることを示した. 以下では, ブロックサイズの問題が解決された状況において iSCSI 性能がソケット通信の性能より大きく劣る原因の解析を行う.

4.2　実 験 環 境

以降の実験は, 本節の実験環境において行った. 図 3 のように, iSCSI イニシエータ (サーバ) と iSCSI ターゲット (ストレージ) を Gigabit Ethernet で接続して TCP/IP 接続を確立する. Ethernet の接続は, 途中に人工的な遅延装置として FreeBSD Dummynet [5] を挟みクロスケーブルで接続した.

iSCSI イニシエータは開発した簡易版イニシエータを, ターゲットの実装はニューハンプシャー大学 InterOperability Lab [6] が提供する reference implementation を用いた (以下, この iSCSI の実装を UNH と呼ぶ). これらは全て iSCSI draft 18 準拠となっている. iSCSI の性能評価実験環境を表 1 に記す. イニシエータ, ターゲット, 遅延装置はすべて PC 上に構築し, イニシエータとターゲットには Linux を, 遅延装置 には FreeBSD をインストールした. イニシエータ, ターゲット PC の詳細は表 2 の通りであり, 遅延装置の PC の詳細は表 3 の通りである.

表 1　性能評価実験環境 1
Table 1　Environment for Performance Evaluation 1

iSCSI Initiator, Target	UNH IOL Draft 18 reference implementation ver. 3
iSCSI MaxRecvDataSegmentLength	16777215 Byte
iSCSI MaxBurstLength	16777215 Byte
iSCSI FirstBurstLength	16777215 Byte
ベンチマーク	Single Thread

表 2　性能評価実験環境 2 : 使用計算機
Table 2　Environment for Performance Evaluation 2 : PC Specification

CPU	Pentium 4 2.80GHz
Main Memory	1GB
OS	Linux 2.4.18-3
Network Interface	Gigabit Ethernet Card Intel PRO/1000 XT Server Adapter

表 3　性能評価実験環境 3 : 使用計算機
Table 3　Environment for Performance Evaluation 3 : PC Specification

CPU	Pentium4 1.5GHz
Main Memory	128MB
OS	FreeBSD 4.5-RELEASE
Network Interface	Gigabit Ethernet Card Intel PRO/1000 XT Server Adapter × 2

実験は, イニシエータ計算機上の簡易版イニシエータ (これを “iSCSI(KI)” と記す) から iSCSI プロトコルに基づき iSCSI Read PDU を発行しターゲットからデータを受信し, そのスループットを計測した. 純粋な iSCSI プロトコルの影響を考察するために iSCSI ターゲットはメモリモードで動作をさせており, ディスクへのアクセスは伴っていない. これは無限に高速なストレージデバイスと見なせる. また, TCP 広告 Window サイズは 2MB である.

4.3　ブロック細分化回避後の解析

本節において, 大きなブロックサイズ (4MB) による iSCSI アクセスの解析結果について述べる. 4MB iSCSI ブロックサイズによる iSCSI シーケンシャルリードにおけるパケット転送時間軸表示は前述の図 2 の様になった. 解析結果の図 2 より時刻 0.33 秒程度においてターゲット計算機に “iSCSI Read(4MB)” が到着し, ターゲット計算機が多数のパケットの転送を試み,

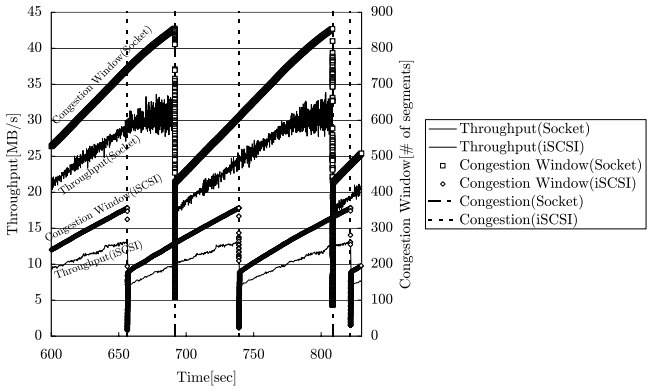

図 4　スループット, 輻輳 Window, TCP 実装のイベント検出
Fig. 4　Throughput, Congestion Window, TCP Event
iSCSI ブロックサイズ 4MB, 広告 Window 2MB, 片道遅延 16ms, 簡易版 iSCSI イニシエータおよび UNH iSCSI ターゲットを使用

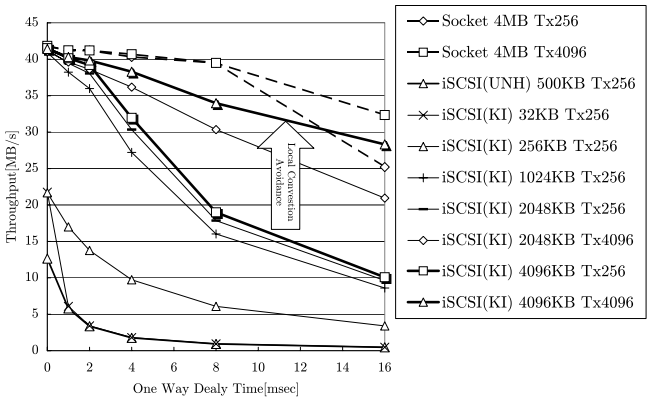

図 5　実験結果：ローカルデバイス輻輳回避
Fig. 5　Experimental Result : Local Congestion Avoidance

ローカルデバイス輻輳が発生していることが確認できる. そして, その結果 TCP 実装は正常な状態 “TCP_CA_Open” から輻輳 Window 縮小の状態 “TCP_CA_CWR” に遷移していること, および状態遷移後に輻輳 Window が減少していることが解析結果より確認できる.

また, 素のソケット通信および iSCSI ブロックサイズを 4MB とした iSCSI(KI) のスループットの時間推移および TCP 実装の輻輳 Window の時間推移は解析の結果, 図 4 となった. また TCP フロー制御監視機能で監視した TCP 実装内のローカルデバイス輻輳の検出および TCP 実装が状態 TCP_CA_CWR に遷移するイベントは同図内に “Local Congestion” と表示されている(注1). 解析システムにより得られた図 4 より, 素のソケット通信では輻輳 Window は十分に大きな値 (約 850) まで上昇しループットも十分に大きな値となっているのに対して, iSCSI(KI) では輻輳 Window の値が 356 に上昇した時点で必ず TCP 実装がローカルデバイス輻輳を検出し輻輳 Window を縮小し, その結果 TCP 実装が出力を制限し, iSCSI のスループットもこれに同期して低下していることが確認された. 素のソケット通信は, TCP が持つセルフクロッキングによりそのバースト性が時間経過とともに減衰しバーストの無い送信を行っているのに対し, iSCSI(KI) においては TCP の上位層である SCSI 層が TCP 層とは独立に確認応答 (SCSI Response) を行い SCSI Read コマンド 1 回毎に同期を取るため TCP の持つセルフクロッキングによるバースト性の減衰は機能せず時間経過後もバーストが残る. この結果, 輻輳 Window の上昇がローカルデバイス輻輳の発生につながりやすくスループットの低下を招いている. これは SCSI プロトコルを TCP プロトコルの上で転送する iSCSI プロトコルの本質的な問題である. ローカルデバイス輻輳は, TCP 実装が上位層から転送を依頼されたデータを下位層に転送要求を出しこれが失敗することにより検出される. 本実験環境の例では TCP 実装が下位層である Ethernet カードのデバイスドライバに対し過度な量のデータのエンキューを依頼することにより発生し, デバイスドライバ内におけるパケットの識別子の枯渇が起こっている.

(注1)：イベントは縦線で表現されているが, イベントは “発生した時刻” のみが意味を持つ

4.4　解析による検出された問題点の回避と性能向上

前節において, iSCSI ブロックサイズ 4MB の iSCSI シーケンシャルリードの性能低下要因がローカルデバイス輻輳であることを本解析システムにより確認した. 本節において NIC デバイスドライバのバッファ数を向上させ TCP 層からの多量のエンキューに対する耐久性を向上させローカルデバイス輻輳を回避することによる性能の向上を評価する. 本実験で用いたネットワークカードはパケット識別子数を 80 以上 4096 以下の範囲で指定することが可能であり, 前節の実験は初期値の 256 個を用いた. これを 4096 とし性能を評価した結果を図 5 に示す.

図 5 に横軸はイニシエータとターゲットの間に配置した遅延装置により人工的に作成した片道の遅延時間であり, 縦軸は各測定結果のスループットである. 各測定のおける “Socket”, “iSCSI(KI)”, “iSCSI(UNH)” は順に, 素のソケット通信, iSCSI(簡易版イニシエータと UNH ターゲット), iSCSI(UNH イニシエータと UNH ターゲット) のスループットである. 本稿ではブロックサイズの小ささの問題を解決する前の性能 [4] については言及されていないが参考のために iSCSI(UNH) としてそれを併記した. 各測定におけるバイト単位 (KB,MB) の数値は各測定におけるブロックサイズである. 素のソケット通信においては, ソケット API への 1 回の送受信要求サイズであり, iSCSI(KI) においては簡易版イニシエータが発行する iSCSI Read PDU 内に記載されている SCSI コマンド Read のブロックサイズであり, iSCSI(UNH) においては read() システムコールを行う際のブロックサイズである. iSCSI(UNH) の測定では実際は 500KB の要求が 32KB ごとに分割され iSCSI Read PDU が発行されることとなる(注2). ブロックサイズ 4MB 以下の測定結果も同様に参考ために併記した. Tx とは NIC ドライバにおける識別子数のことである. “Tx 256” は初期状態, “Tx 4096” はそれを最大値まで拡大した状態であり, デバイスドライバ内において各個数までの

(注2)：分割サイズはシステムの実装に依存するが, 本稿で用いた Linux カーネル 2.4 の例においてはカーネル再構築によりこれを 128KB まで拡大可能である

パケットをバッファすることが可能となる．図 5 よりローカルデバイス輻輳の回避による性能の向上が確認され，ブロックサイズ 4MB の場合，ローカルデバイス輻輳回避以前 (iSCSI(KI) 4096KB Tx 256) と比べ，片道遅延時間 8ms において 1.79 倍，16ms において 2.81 倍の性能向上が確認された．iSCSI(UNH) に対しては 片道遅延時間 16ms の例において 60.5 倍の性能向上がなされた．またシステムの提供できる限界スループット（素のソケット通信）に対する iSCSI 層が原因となるスループットの劣化は，識別子数増加前 (iSCSI(KI) 4096KB Tx256) は片道遅延時間 4ms において 31% ，8ms において 52% ，16ms において 60% であったのに対し識別子数の増加 (iSCSI(KI) 4096KB Tx4096) により 4ms において 6% ，8ms において 14% ，16ms において 12% となり，iSCSI プロトコルを使用することによるスループットの劣化の大幅な削減がなされた．

上記のように解析システムによりカーネル内の振る舞いも含め総合的に解析することが可能となり，解析により適切な改善点を考察することにより iSCSI 性能は大きく向上できることが確認された．

5. 関 連 研 究

文献 [7] において，Ng らは独自の SCSI over IP 実装を用いて IP ストレージの性能に関する詳細な測定と考察を行っている．同文献では 8KB のブロックサイズにおけるシーケンシャルアクセスの性能を測定し，そのスループットが遅延時間にほぼ反比例することが示されている．また SCSI over IP を用いるにあたってネットワークの手前におけるキャッシュの適用や，アプリケーションによるプリフェッチが効果的であると指摘している．アプリケーションレベルの応用性能測定も行っており，性能に関してファイルシステムも考慮した考察がなされている．本論分では，アプリケーション層ではなく，iSCSI 層ならびにその下位層を詳細に解析することにより，iSCSI の適用が強く期待されている遠隔バックアップ等に必要なシーケンシャルアクセスの高速化を目指しており，研究の方向性が異なる．

文献 [8] において，Sarkar らは低遅延環境におけるブロックサイズと iSCSI スループットの関係を紹介している．低遅延環境においては CPU による処理がスループットを制限するため，さらなる高性能を得るためにはハードウェアによる TCP/IP 処理と iSCSI 処理が重要であると主張している．しかし iSCSI を用いた遠隔バックアップで直面する高遅延環境についての考察はなされていない．

6. おわりに

本稿では，我々の実装した iSCSI ストレージアクセスの解析システムを紹介し，高遅延環境におけるブロックサイズの大きな iSCSI シーケンシャルリードアクセスに対して解析を行い性能低下要因が SCSI プロトコルのバーストにより発生する輻輳およびそれによる TCP のフロー制御による出力の制限であることを確認した．そして，その回避により iSCSI 性能を大きく向上させることが可能であることを示した．今後は本解析システムを適用して iSCSI ストレージアクセスのさらなる解析を進め，スループット向上のための下位層も含めた最適化，異なるアクセスパターンにおける iSCSI 内部の性能劣化の回避などを検討する予定である．

文　　献

[1] 喜連川優, “ストレージネットワーキング”, オーム社出版局, 2002

[2] IETF IPS, `http://www.ietf.org/internet-drafts/draft-ietf-ips-iscsi-20.txt` (`draft 18` は現在公開されていない) `http://www.ietf.org/html.charters/ips-charter.html`

[3] IETF : `http://www.ietf.org/`

[4] 山口実靖 小口正人 喜連川優, “高遅延広帯域ネットワーク環境下における iSCSI プロトコルを用いたシーケンシャルストレージアクセスの性能評価ならびにその性能向上手法に関する考察”, 電子情報通信学会 第 14 回データ工学ワークショップ, 2003 年 3 月

[5] L. Rizzo, ”dummynet”, `http://info.iet.unipi.it/ luigi/ip_dummynet/`

[6] The University of New Hampshire’s InterOperability Lab `http://www.iol.unh.edu/consortiums/iscsi/iscsi_linux.html`

[7] Wee Teck Ng, Bruce Hilly Elizabeth Shriver, Eran Gabber, Banu Ozden, “Obtaining High Performance for Storage Outsourcing”, *Proc. FAST 2002, USENIX Conference on File and Storage Technologies*, January 28-29, 2002, pp. 145-158

[8] Prasenjit Sarkar and Kaladhar Voruganti, “IP Storage: The Challenge Ahead”, *Proc. of Tenth NASA Goddard Conference on Mass Storage Systems and Technologies*, April 2002