Vol.2009-DPS-139 No.11
2009/6/19

複数の Target を遠隔地に配置した iSCSI-RAID の性能検証

周 萱[†]　　渡邉 貴之[†]

将来予想される大地震・火山噴火などの大規模な自然災害や人的災害から電子データを保護するためのディザスタリカバリの方法としては，遠隔地に複数配置したストレージ間で同期ミラーリングを行う方法が考えられる．本報告では，iSCSI プロトコルによる IP-SAN 環境を用いて，遠隔地に複数配置したストレージ間で RAID1 システムを構成し，広域ネットワークを想定した環境下でのネットワークの遅延やパケット損失が，シーケンシャル write 性能にどのような影響を与えるのかを検証した．

Performance Evaluation of iSCSI-RAID Consisting of Remotely-Disposed Targets

Xuan Zhou[†] and Takayuki Watanabe[†]

As a disaster recovery method to protect electronic data by large-scale disasters such as the earthquake or the volcanic eruption expected in the future, a method to perform mirroring between the multiple storages places in remote areas is an effective. By this report, an iSCSI-RAID1 system consisting of remotely-disposed targets is configured and evaluated about its sequential write performance.

1. はじめに

現在，様々な分野における情報の電子化やストレージの大容量化に伴い，膨大な電子データがストレージ内に蓄積され，個別拠点内でのデータ処理だけでなく，広域ネットワークを介した拠点間での情報共有や電子商取引のために活用されている．そのため，電子データを保持するストレージに障害が発生すると，データの喪失により多大な損失を被る可能性がある．

一般に，ファイルサーバの可用性を高める方法としては，ストレージを冗長化する RAID (Redundant Array of Independent Disks) 方式がある．また，複数のサーバが同一のストレージを共有することによって，サーバ自体の故障によるサービスの停止を避ける方式もある．特にデータセンタ等の大規模システムでは，複数のサーバ・ストレージを専用ネットワークである SAN (Storage Area Network) によって接続する例が一般的である．

一方，将来予想される大地震・火山噴火などの大規模な自然災害や人的災害では，サーバ・ストレージを設置する拠点全体がダウンする可能性がある．このような災害からデータを保護するためのディザスタリカバリの方法としては，サーバ・ストレージを遠隔地に複数配置し，これらを接続した広域 SAN を構築することにより，遠隔ストレージ間で RAID1 システムを構成し，同期ミラーリングを行う方法が考えられる．ただし，ストレージ間の接続距離は，FC (Fibre Channel) の規格上 10km に制限されるため，大規模災害を想定した 100km オーダのミラーリングは不可能であった．近年，FCIP (Fibre Channel over IP), iFCP (Internet Fibre Channel Protocol)や iSCSI (SCSI over IP) [1]といった既存のストレージプロトコルを IP でカプセル化した IP ベースの SAN 技術（IP-SAN）が標準化され，距離の制約なしに，安価な広域 IP 網を介した遠隔 SAN の構築が可能となった．

特に，iSCSI は FC を介さずに，SCSI コマンドを直接 IP にカプセル化しやり取りを行うため，一般に高価な FC 対応機器を利用する必要がない．また，iSCSI のソフトウェアスタックが主要な OS 上で提供されており，安価かつ容易に遠隔 SAN を構築することができる．しかし，TCP/IP へのカプセル化に要する処理のオーバヘッドや，ネットワーク遅延及びパケット損失の影響により，ストレージアクセスの性能劣化が問題視されている. 前者については, ハードウェアによる TOE (TCP Offload Engine) や CPU の高速化により解決可能であるが，後者については様々な検討が行われている．

本報告では，iSCSI プロトコルによる IP-SAN 環境を用いて，遠隔地に複数配置したストレージ間で RAID1 システムを構成し, 広域ネットワークを想定した環境下でのネットワークの遅延やパケット損失が，シーケンシャル write 性能にどのような影響を

[†] 静岡県立大学大学院経営情報学研究科
Graduate School of Administration & Informatics, University of Shizuoka

与えるのかを検証する．本研究では，ランダムアクセス性能が求められる DBMS などのアプリケーションではなく，大容量データのバックアップや，科学技術計算や CAE（Computer-Aided Engineering）シミュレーションの結果保存などのアプリケーションを想定している．遠隔ミラーリングの性能向上を目的として，複数のプロトコル階層にわたるチューニングを行い，その効果を評価する．更に，サーバ・ストレージ間のネットワーク遅延の不均一性が，遠隔ミラーリングの性能に与える影響について考察する．

2. 関連研究

これまでに，iSCSI の性能検証とその向上に関する様々な先行研究が行われている．例えば，文献[2]では，広域広帯域ネットワークを想定した遅延・輻輳を考慮した環境での iSCSI の性能について論じている．また，文献[3]では，Linux における Target 実装のバリエーションと性能評価について述べている．文献[4]では，iSCSI シーケンシャル read の性能低下が iSCSI PDU のデータサイズに依存することを指摘し，独自の Initiator を用いて，PDU データサイズの増加が性能向上に寄与することを実証している．文献[5]では，広域 IP 網を想定し，高遅延かつパケット損失率の大きい環境において，TCP や iSCSI Target 及び Initiator の設定項目をチューニングすることで，iSCSI の性能が向上することを述べている．また，文献[6]では，高遅延環境において，TCP や iSCSI PDU の最適化を行っても尚，iSCSI の性能が劣化することを述べ，その要因について調査を行っている．

以上の先行研究では，サーバ（Initiator）とストレージ（Target）が 1 対 1 の構成を想定していた．一方，iSCSI を用いて複数の Target 間で RAID を構成した場合の性能評価に関する研究も行われている．文献[7]では，iSCSI を用いた RAID である iRAID を構成し，RAID0 及び RAID5 についての性能を遅延が無視できる環境で評価している．また，文献[8]では，用途として RAID1 によるディザスタリカバリを想定し，1 台の Target を遠隔地に配置し，もう 1 台をローカルサイトに配置した場合のミラーリング性能について論じている．しかし，パケット損失の影響や，複数の Target を遠隔地に配置した場合については述べられていない．

本報告では，広域ネットワークを介して複数台の Target を遠隔地に配置した場合のミラーリング性能について，ネットワーク遅延とパケット損失の影響を加味して評価を行う．

3. iSCSI の概要

3.1 iSCSI システムのプロトコル階層

iSCSI プロトコルによる IP-SAN では，SCSI コマンド（iSCSI リクエスト）を発行する側のデバイスを Initiator と呼び，SCSI コマンドを受け取ってストレージに対して read/write を行い，レスポンスや状態を Initiator へ返送する側を Target と呼ぶ．SCSI コマンドやレスポンスは, iSCSI の PDU (Protocol Data Unit) にカプセル化され, TCP/IP を用いて Initiator/Target 間でやり取りされる．ソフトウェア RAID を用いた場合の iSCSI RAID システムのプロトコル階層を図 1 に示す．図から，SCSI コマンドの伝送のために，iSCSI 以下の複数のプロトコルが介在しており，これらのオーバヘッドがストレージの read/write 性能に少なからぬ影響を及ぼすことが 2 章で述べた関連研究からも報告されている．

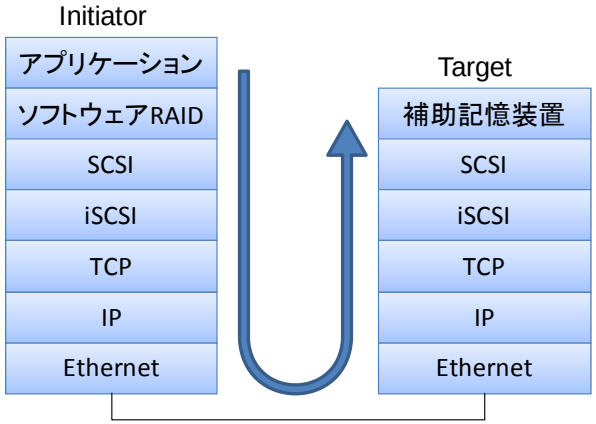


図 1　iSCSI によるソフトウェア RAID システムのプロトコル階層

3.2 シーケンシャル write 時のフロー

iSCSI では，SCSI コマンド（Command Descriptor Blocks: CDB）を iSCSI PDU にカプセル化してやり取りを行う. PDU の先頭にはヘッダー情報として 48Byte の BHS（Basic Header Segment）が格納されており，続いて SCSI コマンドの実行に必要なパラメータである DataSegment が格納されている．また，BHS の先頭 1Byte が PDU の種別を示す Opecode である．主な PDU の種別と Opecode の例を表 1 に示す．

次に，シーケンシャル write 時における Initiator と Target 間での基本的な iSCSI PDU のやり取りを図 2 に示す．まず，Initiator 側のアプリケーションによって書き込み要求が発行されると，書き込みは SCSI レイヤにおいて複数の SCSI Write コマンドに分割される．iSCSI レイヤでは，Initiator は Target に対して SCSI Write CDB を内包した SCSI Command PDU を送信する．この時，総書き込み Byte 数は，BHS の Expected Data Transfer Length（EDTL）フィールドに設定され，Target に通知される．

表 1　主な PDU の種別

発行元	PDU の種別	Opecode
Initiator	SCSI Command	0x01
	Login Request	0x03
	SCSI Data-Out	0x05
	Logout Request	0x06
Target	SCSI Response	0x21
	Login Response	0x23
	SCSI Data-In	0x25
	Logout Response	0x26
	Ready To Transfer	0x31

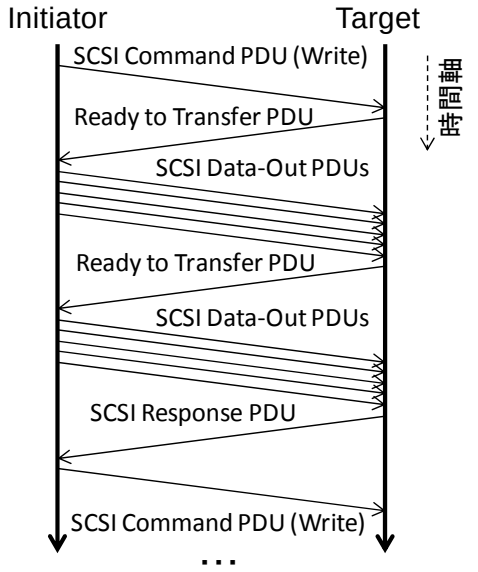


図 2　シーケンシャル write 時の基本的な iSCSI PDU フロー

次に, Target はデータの受取が可能であれば, Ready to Transfer (R2T) PDU を Initiator へ返す．その後，Initiator は，SCSI Data-Out PDU の DataSegment に書き込みデータを格納し，Target へ送信する．もしも，PDU のサイズが TCP の最大セグメント長（MSS）を超える場合，PDU は TCP レイヤで MSS 毎に分割されて送信される．Target は，データの受取が終了すると，Initiator へ SCSI Response PDU を返す．以上の処理を SCSI Write コマンド毎に繰り返す.

一方，iSCSI PDU 内の DataSegment の最大サイズは，Initiator 及び Target 内の MaxRecvDataSegmentLength（MRDSL）パラメータで制限されている．このパラメータは，iSCSI セッションの開始時に，Initiator と Target が自身の受信可能な最大サイズを互いに通知し合うことで決定される．従って，Initiator は Target から通知された MRDSL 以上の DataSegment を持つ PDU を Target に送信できない．ただし，Initiator は Target に対して，複数の連続した Data-Out PDU をバースト的に送信することができる．連続的に送信可能な Data-Out PDU の最大データサイズは，Target から通知される R2T PDU 内の Desired Data Transfer Length（DDTL）フィールドに制限される. また，DDTL の最大値は Target 内の MaxBurstLength パラメータで制御可能である．R2T PDU は，一連の連続的な Data-Out PDU を受信するたびに Initiator へ通知される.

例えば，MRDSL が 8192Byte で，SCSI Write CDB の書き込み Byte 数が 262144Byte の場合，R2T の DDTL が 16384Byte と通知されたとすると，Data-Out PDU の DataSegment は 8192Byte に制限されるため，2 個の Data-Out PDU が連続的に Target に送信される．Target は，2 個の Data-Out PDU を受信後，再度 R2T PDU を Initiator に通知する.

4. iSCSI シーケンシャル write の最適化

4.1 TCP レイヤでの最適化

iSCSI プロトコルは, 下位レイヤのプロトコルとして TCP を利用している. 従って, TCP レイヤでの最適化が iSCSI の性能向上に有効である[5]．一般に，遠隔地間を想定した広域ネットワークでは，帯域幅遅延積（Bandwidth Delay Product: BDP）が増大する．従って，TCP セグメントの送信・受信バッファは，想定される BDP から勘案して十分大きいサイズを確保しておくことが求められる[9][10].

一方で，パケット損失が発生すると，TCP の輻輳ウインドウが減少し，通信効率が低下する．輻輳ウインドウの制御アルゴリズムには，古典的な Reno アルゴリズムを始めとして複数の方法が提案されている[6].

4.2 iSCSI レイヤでの最適化

iSCSI によるシーケンシャル write では，SCSI Data-Out PDU を効率的に伝送することが，書き込み性能向上の鍵となる．従って，より大きな DataSegment を持つ PDU を連続的に送信できるように，Target の MRDSL や MaxBurstLength を設定することが有効である. MRDSL の設定可能値は 512～16777215 であり, 標準は 8192 であるが, Target

の実装により上限値は異なっており，本研究で使用した Linux 用の iSCSI Enterprise Target では 262144（=256KB）が上限となっている[11].

次に，iSCSI では，InitialR2T=No と設定することで，SCSI Command PDU (Write) の発行による書き込み時に，Target からの初回の R2T PDU による許可を省略して，Data-Out PDU を送信することができる. このようにして Target に送信されるデータを，Unsolicited Data と呼ぶ[1]. また, ImmediateData=Yes と設定することで, SCSI Command PDU (Write) 自体に書き込みデータを含めて送信することができる.

また，1 つの iSCSI セッション内に複数の TCP コネクションを確立して，同時並行的に iSCSI プロトコルのやり取りを行うことも仕様上は可能であり，MaxConnections パラメータによって設定が可能である[12]. これによって, パケット損失発生時の TCP スループットの低下を分散させることができる．ただし，MaxConnections の上限値は実装により異なり，iSCSI Enterprise Target では 1 に固定されている.

以上の最適化は，Target と Initiator が 1 対 1 の環境を想定したチューニングであるが，本研究において想定している RAID1 によるミラーリング環境においても，同様のチューニングを行い，その効果を評価する.

5. 実験環境と実験結果

5.1 実験環境と条件

本研究では, 図 3 に示すような環境を構築し, ベンチマークを行った. Initiator, Target, Dummynet のすべての NIC Port は Gigabit Ethernet に対応しているが，回線の帯域幅としては安価な広域 IP 網を想定して 100Mbps とした．また，WAN 回線による遅延やパケット損失を模擬するために，FreeBSD Dummynet を用いている．実験に用いたハードウェア／ソフトウェア環境の一覧を表 2 に示す.

ベンチマークプログラムとしては，C 言語を用いたオリジナルのプログラムを作成し, 10MB のランダムデータを write() 関数の同期書き込みによって処理が完了する時間を計測した．計測は，以下の条件の組み合わせごとに各 20 回行い，スループットの平均値を算出した.

まず，TCP 及び iSCSI の設定項目を表 3 に示した「標準」「TCP 最適化」「TCP+iSCSI 最適化」と変化させ，チューニングの効果を測定した．また，ネットワーク遅延の影響を評価するため，Dummynet の片道遅延の設定を 0ms から 50ms まで 5ms 間隔で変化させ，測定を行った．一方，パケット損失の影響については，Dummynet において損失率 0.0%の場合と，損失率 0.01%での結果を測定した.

すべての実験は，1 台の Target のみに書き込みを行う single 構成と，2 台の Target に Software RAID を介してミラーリング書き込みを行う raid 構成の両者に対して実施した.

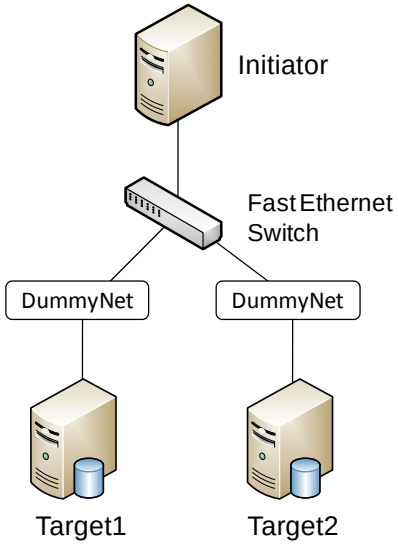


図 3　実験環境の模式図

表 2　ハードウェア／ソフトウェア環境

機器	種別	バージョン等
Initiator	Hardware Spec	Intel Core2Duo 2GHz, 2GB RAM, Marvell 88E8056 PCI-E Gigabit Ethernet Controller
	OS	FedoraCore Linux 8 (64bit kernel 2.6.26.8-57)
	Software RAID	mdadm v2.6.7
	Software Initiator	Open-iSCSI 2.0-865
Target	Hardware Spec	AMD Athlon 3500+ 2.2GHz, 512MB RAM, Broadcom NetXtreme BCM5721 Gigabit Ethernet PCI Express, 80GB SATA HDD
	OS	FedoraCore Linux 8 (64bit kernel 2.6.24.3-12)
	Software Target	iSCSI Enterprise Target 0.4.15
Dummynet	OS	FreeBSD 7.0 - RELEASE

表 3　設定値一覧

	標準	TCP 最適化	TCP+iSCSI 最適化
TCP バッファサイズ	128KB	16MB	16MB
iSCSI MRDSL	64KB	64KB	256KB
iSCSI FastBurstLength	64KB	64KB	256KB
iSCSI MAxBurstLength	64KB	64KB	256KB
iSCSI InitialR2T	Yes	Yes	No
iSCSI ImmediateData	No	No	Yes

5.2 実験結果

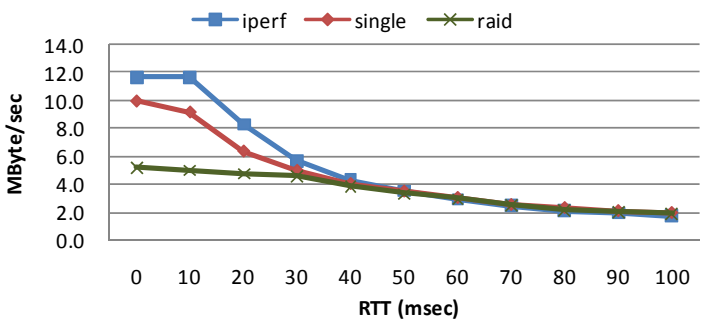
(1) パケット損失率 0.0%におけるスループット

Dummynet において，パケット損失率を 0.0%とした場合のスループットの結果を図 4 に示す．raid 環境における測定では，Initiator と Target1 間（以降，I→T1，Initiator と Target2 間（以降，I→T2）の往復遅延時間（RTT）が等しくなるように設定した．また, 比較として, iperf (version 2.0.2)を用いた 300 秒間の平均スループットを計測し，図 4 に併記した．

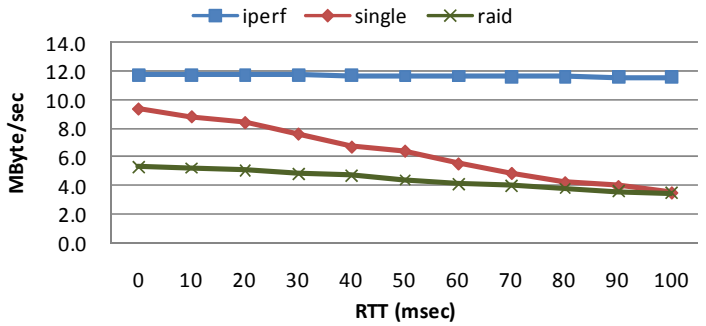
図 4 (a)から，最適化を行わない場合には，RTT が大きくなると BDP が TCP のバッファサイズを超えるためボトルネックとなり，iSCSI の性能が頭打ちなることがわかる．実際に，BDP の定義から，100Mbps の帯域幅で RTT が 10ms において BDP は約 122KB となり，RTT が 10ms を超えるとバッファを使い果たし，以降スループットは RTT に反比例して減少する．一方, single 構成と raid 構成とを比較すると，RTT が 0ms ではスループットに約 2 倍の差があるが, RTT の増加とともに両者の差は縮まり, 30ms 以降は差が無くなっていることがわかる．

次に，図 4 (b)から，TCP のバッファサイズを拡大することによって，iperf の結果が RTT によらず常に帯域の上限である 11.9MB/s（≒100Mbps）に近接した値となっており，TCP レイヤにおいて最適化が図られていることがわかる．一方，iSCSI のスループットは，RTT の増加に比例して右肩下がりで減少している．この時，raid 構成の傾きに比較して，single 構成の傾きが急角度となっており，両者のスループットは RTT が 100ms でほぼ一致している．

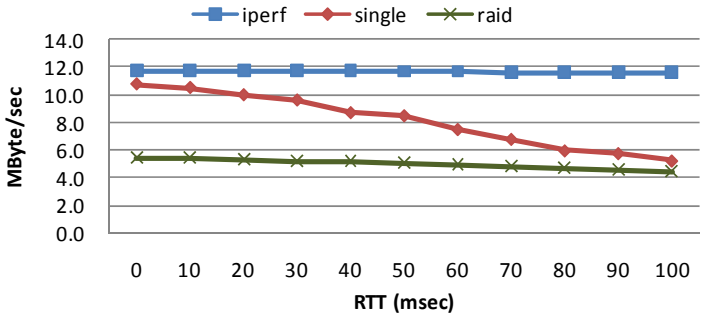
次に, TCP レイヤでの最適化に合わせて iSCSI レイヤにおける最適化を行った場合，図 4 (c)に示すように，single，raid 構成ともに，全体的にスループットが底上げされていることがわかる．しかし，TCP のみの最適化と同様に，スループットは RTT に比例しながら負の傾きで減少している．特に，single 構成における減少の傾きは，iSCSI レイヤにおける最適化の前後でほぼ等しくなっており，iSCSI レイヤにおける最適化によってスループット全体が約 1.8MB/s ほど上方に平行移動していることがわかる．



(a) 標準



(b) TCP 最適化



(c) TCP + iSCSI 最適化

図 4 パケット損失率 0.0%の場合のスループット

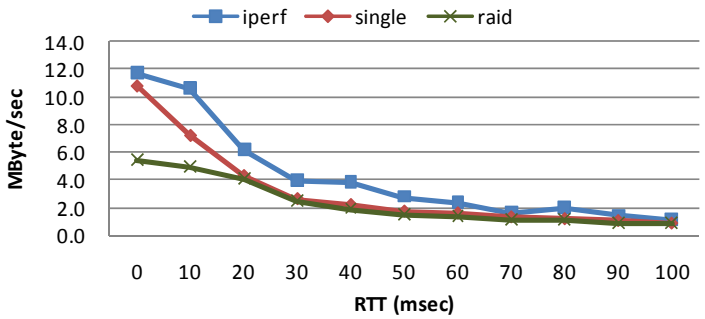


図 5 パケット損失率 0.01%におけるスループット

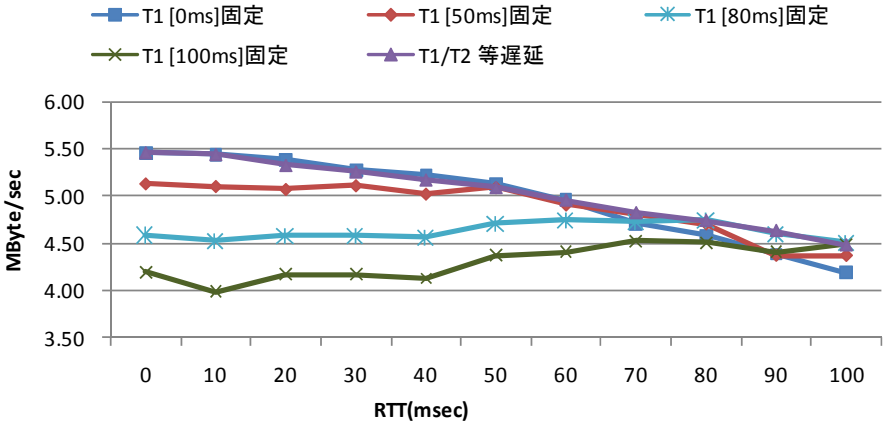


図 6　遅延が不均一の場合におけるスループット

(2) パケット損失率 0.01%におけるスループット

続いて，Dummynet において，パケット損失率を 0.01%に設定した場合のスループットの結果を図 5 に示す．ここで，表 3 の条件としては，TCP + iSCSI 最適化を設定し，TCP の輻輳制御アルゴリズムとしては Reno アルゴリズムを用いている．

結果から，パケット損失の影響によって，TCP レイヤでのスループットの低下がボトルネックとなり，iSCSI の性能が著しく低下していることがわかる．また，single 構成と raid 構成とを比較すると，RTT が 0ms ではスループットに約 2 倍の差があるが，RTT の増加とともに両者の差は縮まり，20ms 以降は差が無くなっていることがわかる．

(3) 遅延時間が不均一な場合のおける raid 環境のスループット

最後に，Dummynet において，パケット損失率を 0.0%に設定し，尚且つ，I→T1 と I→T2 の RTT を不均一に設定した場合のスループットの結果を図 6 に示す．ここで，表 3 の条件としては，TCP + iSCSI 最適化を設定している．

結果より，I→T1 および I→T2 の遅延を等しくした場合（均一条件）に比較して，I→T1 の遅延を 0ms, 50ms, 80ms, 100ms に固定し，I→T2 の遅延を変動させた場合（不均一条件）の何れもが，スループットの面で均一条件を同等か下回っていることがわかる．例えば，図 6 から，I→T1 = 50ms，I→T2 = 0ms でのスループットに比較して，I→T1 = 50ms，I→T2 = 50ms でのスループットは同等であることがわかる．従ってスループットの上限は，より遅延の大きい Target によって決定されているといえる．

また，Target 間の遅延差が大きい場合，不均一条件よりも均一条件のスループットが改善していることがわかる．例えば，I→T1 = 100ms，I→T2 = 0ms に比較して，I→T1 = 100ms, I→T2 = 100ms のスループットが向上していることがわかる．この傾向は，遅延差が大きくなるほど顕著となっている．

6. まとめ

本報告では，複数の Target を遠隔地に配置した iSCSI による RAID1 ミラーリングシステムについて，そのシーケンシャル write 性能の検証を行った．結果から，TCP レイヤ，iSCSI レイヤにおける最適化は，RAID1 によるミラーリングシステムに対しても有効であり，RTT の増加によって，Target1 台への書き込み性能と RAID1 への書き込み性能が漸近すること，またパケット損失の発生によりその傾向が顕著となるとの結果が得られた．

また，RAID1 システムにおいて，シーケンシャル write 性能はより遠方に配置した Target の RTT に支配され，他方の Target は Initiator に近接して配置しても性能は向上しないことがわかった．また，Target 毎の RTT の不均一性が大きい場合，等速条件よりもスループットが低下する結果が得られた．

今後の予定としては，本研究で得られた結果について理論的なモデル化を行い，より詳細な検討を行う．また，パケット損失を考慮した場合について，複数の輻輳制御アルゴリズムを切り替えて，パフォーマンスを比較することを検討している．

謝辞　本報告の基礎となる卒業研究を行った，平成 19 年度静岡県立大学卒業生の伊藤翔君に，謹んで感謝の意を表する．

参考文献

1) J. Satran, K. Meth, C. Sapuntzakis, M. Chadalapaka, and E. Zeidner, “Internet Small Computer Systems Interface,” RFC3720, April 2004.

2) W. T. Ng, B. Hillyer, E. Shriver, E. Gabber, and B. Özden, “Obtaining High Performance for Storage Outsourcing,” in the USENIX Conference on File and Storage Technologies, pp.145-158, Jan. 2002.
3) Fujita Tomonori and Ogawara Masanori, “Analysis of iSCSI Target Software,” Proceedings of the international workshop on Storage network architecture and parallel I/Os, pp. 25-32, 2004.
4) 山口実靖, 小口正人, 喜連川優, “iSCSI 解析システムの構築と高遅延環境におけるシーケンシャルアクセスの性能向上に関する考察,” 電子情報通信学会論文誌 D-I, J87(2), pp.216-231, Feb. 2004.
5) 藤原啓成, 若宮直紀, 志賀賢太, “広域 IP 網を介した iSCSI 通信におけるプロトコルチューニングの一検討,” 情報処理学会第 68 回全国大会, 5A-5, pp.1-55 - 1-56, March 2006.
6) 比嘉玲華, 松原幸助, 岡廻隆生, 山口実靖, 小口正人, “輻輳ウィンドウ及びパケット解析を用いた iSCSI 遠隔ストレージアクセスの評価,” 電子情報通信学会技術研究報告 コンピュータシステム研究会 CYPS2008, pp.37-42, Dec. 2008.
7) Xubin.He, Praveen Beedanagari, and Dan Zhou, “Performance evaluation of distributed iSCSI RAID,” Proceedings of the international workshop on Storage network architecture and parallel I/Os, pp.11-18, 2003.
8) 小山芳樹, 山口実靖, 淺谷耕一, “iSCSI を用いる遠隔ミラーの動作解析,” 電子情報通信学会第 19 回データ工学ワークショップ (DEWS2008) 論文集, D4-6, March 2008.
9) 鶴正人, 熊副和美, 尾家祐二, “長距離高速通信のための TCP 性能改善技術の動向,” 情報処理学会誌, Vol.44, No.9, pp.951-957, Sep. 2003.
10) Lawrence Berkeley National Laboratory, TCP Tuning Guide: http://www-didc.lbl.gov/TCP-tuning/（2009 年 5 月 15 日確認）.
11) The iSCSI Enterprise Target Project: http://iscsitarget.sourceforge.net/（2009 年 5 月 15 日確認）.
12) 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,” 電子情報通信学会技術研究報告 情報ネットワーク研究会, pp.1-6, June 2008
13) iperf: http://sourceforge.net/projects/iperf（2009 年 5 月 15 日確認）.