社団法人　情報処理学会　研究報告
IPSJ SIG Technical Report

2005－DPS－122（21）
2005－CSEC－28（21）
2005／3／22

43Gbps回線を利用したiSCSIの性能測定

寺岡文男[†]　澤木敏郎[†]　金森勇壮[†]
広瀬健志郎[††]　西村純一[††]　名倉正剛[††]

本稿では広域分散 IP ストレージ構築のための基礎実験として、慶應義塾大学の矢上キャンパスと湘南藤沢キャンパス間に敷設された 43Gbps 実験回線を用いて iSCSI と NFS の性能評価を行った。さらに比較のために LAN 環境においても RTT やパケット損失率を変化させて性能評価を行った。その結果、1Gbps 程度の回線の場合は RTT が 20ms 以下では NFS の方が RTT に影響されにくいことがわかった。またパケットロス率が 10^{-4} より高い場合は iSCSI の方が NFS よりも高いスループットを示し、輻輳しやすいネットワークには NFS よりも iSCSI の方が適していることが分かった。

Performance Evaluation of iSCSI on 43Gbps Line

Fumio Teraoka,[†] Toshiro Sawaki,[†] Yuso Kanamori,[†]
Kenshiro Hirose,[††] Jun'ichi Nishimura[††]
and Masataka Nagura[††]

This paper evaluated performance of iSCSI and NFS over 43Gbps experimental line built between Yagami Campus and Shonan Fujisawa Campus of Keio University as a fundamental experiment to construct a wide area distributed IP storage. The results showed that NFS is more tolerant to RTT than iSCSI if RTT is less than 20ms. The results also showed that iSCSI had higher throughput than NFS if the packet error rate is more than 10^{-4}, and that iSCSI is suitable to networks which tends to be congested.

1. はじめに

慶應義塾インフォメーションテクノロジーセンター (ITC) はキャンパスネットワークやワークステーション室などの情報基盤の管理運用を行う組織である。三田キャンパスに ITC 本部が置かれ、各キャンパス (三田、日吉、信濃町、矢上、湘南藤沢) にはその地区の ITC が置かれている。ITC では 2000 年から各キャンパスを 10Gbps の高速回線で接続する実験を開始し、2004 年 10 月からは慶應義塾大学、NTT、NTT 東日本の 3 者による共同実験として、慶應義塾大学矢上キャンパス (横浜市港北区) と湘南藤沢キャンパス (藤沢市) に 43Gbps の超高速実験を開始した。実験期間は 2005 年 3 月末までである。

今回の共同実験の主な目的は超高速回線を使いこなすための通信プロトコルおよび次世代アプリケーションの創出、および超高速ネットワーク運用性の検証である。技術的課題としては大きく次の 3 つが考えられる。1 つ目はシステム技術である。このカテゴリにはたとえば広域分散 IP ストレージ技術の創出などがあげられる。2 つ目は超高速ネットワークに対する制御・管理技術の確立である。このカテゴリにはたとえば超高速ネットワークにおける QoS 制御方式の確立、高性能トランスポートプロトコルの創出などがあげられる。3 つ目は次世代アプリケーションの創出である。このカテゴリにはたとえばデジタルシネマなどの大容量コンテンツの配信や e-Learning、e-Governance などの実証実験などがあげられる。

本稿はこのうちのシステム技術に焦点を当て、広域分散 IP ストレージ構築のための基礎実験として 43Gbps 回線を利用した iSCSI および NFS の性能評価を行い、基礎データを収集した。

2. iSCSI の概要

iSCSI は TCP/IP の上で SCSI コマンドを転送するプロトコルであり、2004 年 4 月に RFC 3720[1] として標準化された。iSCSI によってディスクにアクセス

† 慶應義塾インフォメーションテクノロジーセンター
Information Technology Center, Keio University.
†† 慶應義塾大学大学院 理工学研究科
Graduate School of Science and Technology, Keio University.

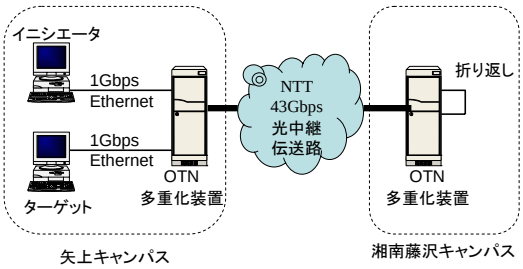

図 1 43Gbps 回線を介した実験環境

するマシンをイニシエータと呼び、実際に SCSI ディスクを持ちイニシエータにディスクを提供するマシンをターゲットと呼ぶ。iSCSI により、イニシエータはターゲットマシンに接続された SCSI ディスクをインターネットを介してローカルな SCSI ディスクと同様にアクセスすることができる。

iSCSI プロトコルそのものの性能測定については多くの研究が行われている。文献[2] では iSCSI によるストレージアクセスの性能劣化原因を解析している。その結果、ブロックサイズの細分化を回避しローカルマシンでのドライバにおける輻輳を回避することにより、TCP/IP のスループットと同程度のスループットが得られることが示されている。

文献[3] では NFS と iSCSI の性能を比較している。解析の結果、データのアクセスに負荷が集中する場合には両者は同等の性能を示し、メタデータのアクセスに負荷が集中する場合は iSCSI の方が NFS の 2 倍程度のスループットが得られることが示されている。

文献[4] ではターゲット側に新しいキャッシング技術を導入することにより、iSCSI を利用したリモートミラーリングのスループットを向上させている。

3. 43Gbps 実験回線とその応用

矢上キャンパスと湘南藤沢キャンパス間に敷設された 43Gbps 実験回線は以下のような構成になっている (図 1 参照)。矢上キャンパスと湘南藤沢キャンパスには 43Gbps OTN 多重化装置と呼ばれる装置が対向で設置され、その間を 43Gbps の光中継伝送路で結んでいる。OTN 多重化装置のユーザ側には 32 チャンネルの 1Gbps の回線と 1 チャンネルの 10Gbps の回線が提供されている。今回は実験のため湘南藤沢キャンパス側で 1Gbps の回線を折り返し、矢上キャンパスに実験機器を接続して実験した。

高速回線を利用した広域分散 IP ストレージシステムとしては以下のような利用方法が考えられる。

(1) iSCSI によってリモートマシンのディスクをローカルな SCSI ディスクと同様に扱う。

(2) NFS によってリモートマシンのファイルシステムをローカルなファイルシステムと同等に扱う。

(3) iSCSI によってリモートマシンのディスクをローカルな SCSI ディスクと同様に扱って RAID を構成し、ローカルディスクの内容を自動的にリモートマシンのディスクにバックアップする。

(1) はまさに SCSI ケーブルの長さの制限をなくしたようなモデルである。基本的にある 1 台のマシンのディスクを複数のマシンから共有することは前提とされていない。(2) は、通常はローカルな環境で用いられている NFS を広域で利用するものであり、ある 1 台のマシンのディスクを複数のマシンから共有することに向いている。(3) は、たとえばローカルディスクとリモートディスクを使用して RAID を構成し、ミラーリングを行えば自動的にローカルディスクがリモートディスクにバックアップされることになる。本稿ではこのうち (1) と (2) について基礎的なデータを収集した。

4. システム構成と測定方法

今回の測定に使用したコンピュータの仕様を表 1 に示す。今回はイニシエータ側およびターゲット側とも一般に公開されているソフトウェアを使用した。この理由は、たとえばイニシエータの専用ハードウェアは特定のオペレーティングシステムでしか動作しないなど、専用ハードウェアにはまだまだ汎用性に問題があるからである。イニシエータ側のソフトウェアとしては Linux-iSCSI Project[5] から配布されているプログラムを使用し、ターゲット側のソフトウェアとしては ARDIS Technologies 社[6] から配布されているプログラムを使用した。

本稿ではユーザが体感する性能に焦点を当てるため、iSCSI や NFS を介したファイルアクセスに関しての性能評価を行う。また今回使用した iSCSI 用ソフトウェアは write の安定性が低いため、測定では read のみを使用した。

ファイルを読み込む場合、プログラムは通常 open() システムコールによってファイルをオープンし、read() システムコールによってファイルの内容を読み込み、最後に close() システムコールによってファイルをクローズする。open() システムコールでは対象となるファイル名 (パス名) を指定するが、このときオペレーティングシステムの内部では以下のような処理が行われる。まず指定されたパス名をオペレーティングシステム内部でのファイルの識別子である i-node 番号に変換する。その際、指定されたパス名に現れるディレク

表 1 コンピュータの仕様

イニシエータ		ターゲット	
CPU	Pentium4, 3.00GHz	CPU	Pentium4, 3.00GHz
L1 キャッシュ	1MB	L1 キャッシュ	1MB
hyper threading	on	hyper threading	on
メモリ	512MB	メモリ	512MB
HDD	3.5”, 73GB, 15,000rpm	HDD	3.5”, 73GB, 15,000rpm
OS	Linux 2.4.20	OS	Linux 2.4.27
		SCSI	ULTRA320

トリの読み込みが行われる。すなわちファイルのオープンという動作に関してもディスクの read が頻繁に発生する。

read() システムコールではまず i-node の構造を読み出し、データが格納されているディスクブロック番号を調べる。このとき、i-node の情報がメモリ上にないときにはディスクアクセスが発生する。その後、得られたディスクブロック番号に該当するディスクブロックを読み出すことによってプログラムが要求するデータを得ることができる。すなわちファイルの読み込みの際にはデータが格納されているディスクブロックだけではなく、ファイルの管理データの読み込みも必要になる場合がある。

一方、オペレーティングシステム内部ではファイルアクセス高速化のためにさまざまな工夫がなされている。メモリ上でのディスクブロックのキャッシュやディスクの先読みなどである。

以上のように実際のファイルアクセスにはさまざまな要因が関連しており、純粋に iSCSI や NFS の性能のみを測定するのは困難である。本稿ではユーザが体感する性能に焦点を当てるため “open()、read() の繰り返し、close()” の実行時間から read のスループットを測定した。

5. iSCSI によるファイルアクセス

以下の測定では共通して 512MB のファイルの sequential read を行い、スループットを測定した。

5.1 単独マシンでのファイルアクセス

最初に、単独のコンピュータに接続された SCSI ディスクに関し、ブロックサイズを変えてスループットを測定した。測定結果を図 2 に示す。グラフの横軸はブロックサイズ (KB) であり、縦軸はスループット (Mbps) である。測定の結果、ブロックサイズによりスループットにばらつきがあることがわかる。ブロックサイズが 2KB, 8KB, 32KB, 128KB のときにはスループットがよく (910〜960Mbps)、1KB, 4KB, 16KB, 64KB, 256KB のときにはスループットが悪い (750Mbps 前後)。

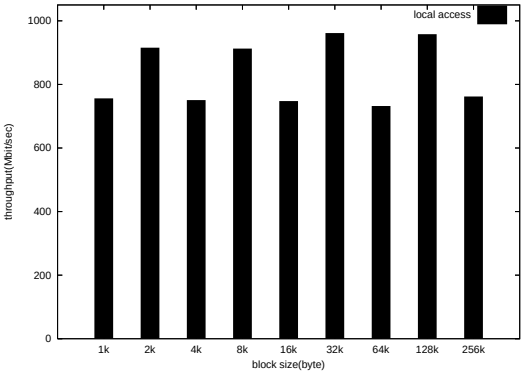

図 2 単独マシンでの SCSI ディスクのスループット

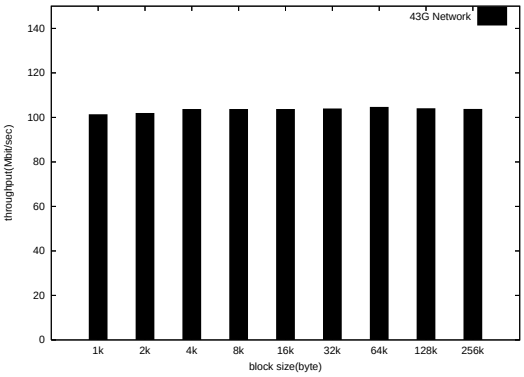

図 3 43Gbps 回線を介した iSCSI によるスループット

5.2 43Gbps 回線を介したファイルアクセス

次にイニシエータマシンとターゲットマシンを 43Gbps 回線で接続した場合の iSCSI によるスループットを測定した。測定環境を図 1 に示す。なお、43Gbps 回線は前述したように湘南藤沢キャンパスで折り返しており、RTT (Round Trip Time) は約 2.8ms である。

測定結果を図 3 に示す。グラフの横軸はブロックサイズ (KB) であり、縦軸はスループット (Mbps) である。測定の結果、ブロックサイズはスループットにあまり影響を及ぼさず、101〜104Mbps 程度であることがわかる。この原因は文献[2] で述べているように、

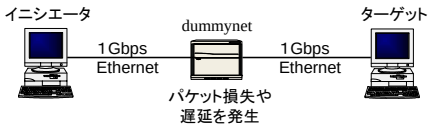

図 4 LAN を用いた測定環境

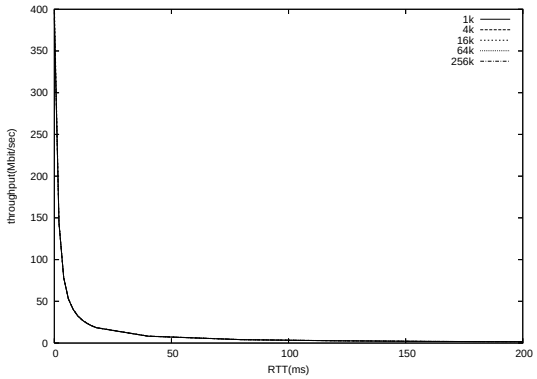

図 5 RTT を変化させた場合の iSCSI によるスループット

アプリケーションが指定したブロックサイズによらず iSCSI ドライバが特定ブロックサイズでの iSCSI Read PDU を送信しているためかもしれないが、今回は確認していない。

図 2 と図 3 を比べると、43Gbps 回線を利用した iSCSI のスループットは単独マシンの SCSI のスループットの約 1/7〜1/9 になっていることが分かる。これは RTT が大きく影響しているものと考えられる。

5.3 RTT を変化させた場合のファイルアクセス

次にイニシエータマシンとターゲットマシンを 1Gbps の Ethernet で接続し、中間に接続したマシンで任意の遅延を発生させた場合のスループットを測定した (図 4 参照)。中間に接続したマシンにおいて dummynet と呼ばれるソフトウェアを実行し、RTT を 0〜200ms で変化させて測定した。測定結果を図 5 に示す。グラフの横軸は RTT (ms) であり縦軸はスループット (Mbps) である。ブロックサイズによる違いはなく、スループットは RTT に反比例して急激に低下することが分かる。

次に RTT が 0〜20ms について詳しく測定した。測定結果を図 6 に示す。43Gbps 回線の RTT は前述の通り約 2.8ms である。当然ではあるが、図 6 における RTT が 2.8ms 付近の値は、図 3 が示す値に近い値になっている。

5.4 エラー率を変化させた場合のファイルアクセス

次に上記と同じ LAN 環境において、dummynet によって回線のパケットエラー率を 1×10^{-6} から 5×10^{-1}

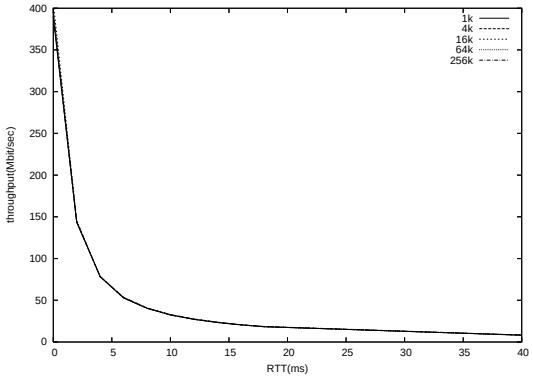

図 6 RTT が 40ms までのときの iSCSI によるスループット

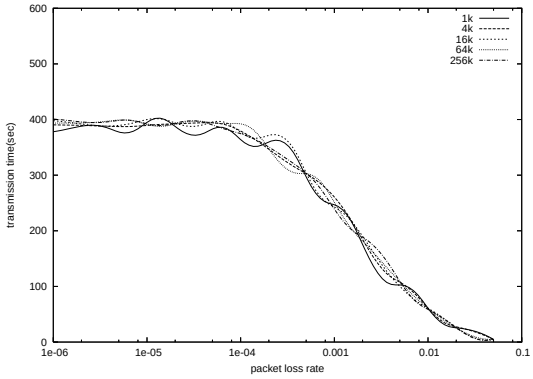

図 7 パケットエラー率を変化させた場合の iSCSI によるスループット

まで変化させた場合のスループットを測定した。測定結果を図 7 に示す。グラフの横軸はパケットエラー率の対数であり縦軸はスループット (Mbps) である。ここでもブロックサイズの違いによるスループットの違いはあまり見られない。

エラー率が 0 から 1×10^{-4} あたりまではスループットの低下はあまり見られない。その後はエラー率の増加とともにスループットも低下している。この結果、輻輳によってパケット損失が頻繁に発生するにつれ iSCSI の性能も急激に悪化することが分かる。

6. NFS によるファイルアクセス

次に NFS によるファイルアクセスの性能評価を行った。文献[7] では NFS と iSCSI の差異を以下のように分析している。プロトコルの観点では、NFS はファイル単位でアクセスするためファイルアクセスプロトコルと呼ぶことができ、iSCSI はブロック単位でアクセスするためブロックアクセスプロトコルと呼ぶことができる。NFS は異なるマシン間でのデータ共有に向

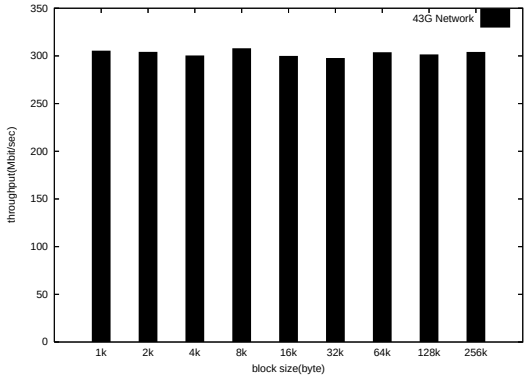

図 8 43Gbps 回線を介した NFS によるスループット

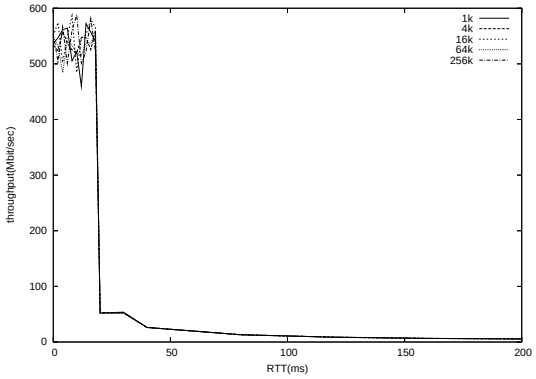

図 9 RTT を変化させた場合の NFS によるスループット

いているが、iSCSI はそうではない。またキャッシングの観点では、NFS はファイルシステムはサーバ側にあるため、サーバ側にファイルシステムのキャッシュが作られる。そのためサーバ側のキャッシュヒットにはネットワーク通信を要することになる。さらに NFS クライアントもデータや管理データをキャッシュする。iSCSI ではファイルシステムはイニシエータ (クライアント) 側にあるため、データや管理データのキャッシュヒットの効率がよい。

本稿では、NFS と iSCSI について第 5 章と同様に 512MB のファイルの read 時間からスループットを計算した。

6.1 43Gbps 回線を介した NFS アクセス

43Gbps 回線を介した NFS アクセスにおいて、ブロックサイズを 1KB, 2KB, 4KB, 8KB, 16KB, 32KB, 64KB, 128KB, 256KB にしたときのスループットを測定した。測定結果を図 8 に示す。グラフの横軸はブロックサイズ (KB) であり、縦軸はスループット (Mbps) である。結果から分かるようにスループットはブロックサイズにほとんど影響されず、298Mbps〜305Mbps 程度である。

第 5.2 節で示したように、同じ環境では iSCSI によるスループットは 101〜104Mbps 程度であり、NFS によるスループットに比べて約 3 倍性能が悪くなっている。この結果から、RTT が 2.8ms 程度では iSCSI よりも NFS の方が効率がよいと言える。次節では遅延の影響をさらに詳しく見ていく。

6.2 RTT を変化させた場合の NFS アクセス

図 4 に示した測定環境において、RTT を 0ms〜200ms に変化させた場合の NFS によるスループットを測定した。結果を図 9 に示す。グラフの横軸は RTT (ms)、縦軸はスループット (Mbps) である。RTT が 20ms あたりまでは RTT の増加に伴う明確なスループット低下は見られないが、20ms あたりで急激にスループットが低下する。

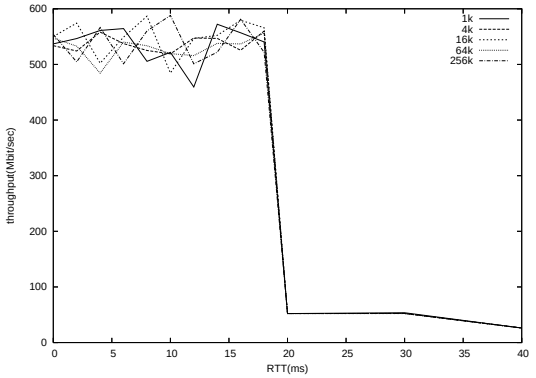

図 10 RTT が 20ms までのときの NFS によるスループット

次に RTT が 0〜20ms について詳しく測定した。測定結果を図 10 に示す。グラフの横軸は RTT (ms)、縦軸はスループット (Mbps) である。図 6 と図 10 を比較すると、RTT が 20ms 以下では NFS の方が iSCSI よりも RTT の影響を受けにくいと言える。たとえば RTT が 16ms のとき、NFS のスループットは約 500Mbps であるが iSCSI のスループットは 40Mbps となっている。これは NFS と iSCSI で使用しているトランスポートプロトコルの違いの影響であると思われる。iSCSI は TCP を使用している。TCP はスライディングウィンドウ方式を採用しているため、広帯域・高遅延の回線 (long-fat pipe) ではなかなかウィンドウが開かず、帯域を使い切れないことが知られている。一方、NFS は UDP を使用しているため TCP のように long-fat pipe の影響が少ないものと思われる。

6.3 エラー率を変化させた場合の NFS アクセス

図 4 に示した測定環境において、パケットエラー率を 1×10^{-6} から 5×10^{-1} まで変化させたときの

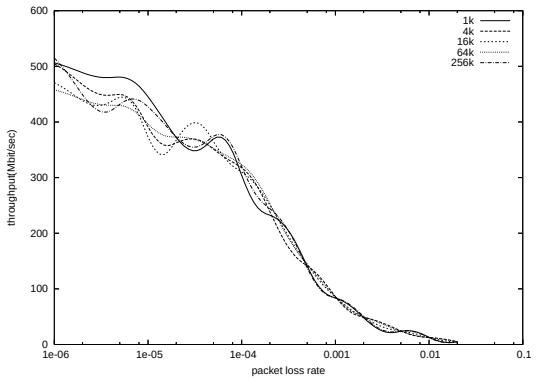

図 11 パケットエラー率を変化させた場合の NFS によるスループット

NFS のスループットを測定した。測定結果を図 11 に示す。グラフの横軸はパケットエラー率の対数、縦軸はスループット (Mbps) である。パケットエラー率の増加に伴いスループットも低下していることが分かる。

図 7 と図 11 を比較すると、パケットエラー率に対しては NFS より iSCSI の方が影響を受けにくいと言える。たとえばパケットエラー率が 0.1%のとき、iSCSI では 240〜260Mbps 程度であるのに対し、NFS では 88〜91Mbps 程度となっている。この理由もやはり両者で使用しているトランスポートプロトコルの差であると思われる。NFS は UDP を使用しているため、パケット損失回復のための再送は NFS 自体が行わなければならない。一方、TCP は 1 ウィンドウ内で 1 パケットが損失する程度のときは高速再送および高速リカバリアルゴリズムによりスループットはさほど低下しない。このため、iSCSI は NFS に比較してスループットがよいと思われる。

7. ま と め

本稿では広域分散 IP ストレージ構築のための基礎実験として 43Gbps 実験回線を利用した iSCSI と NFS の性能評価を行った。同時に比較実験として LAN 環境において遅延やパケットロスレートをエミュレータによって制御して iSCSI および NFS のスループットを測定した。その結果、広帯域・高遅延 (long-fat pipe) においては TCP の性質のため iSCSI のスループットは低下し、1Gbps の帯域では RTT が 20m 以下のときには NFS の方が圧倒的にスループットがよいことが分かった。今回使用した TCP は Linux に標準で搭載されているものなので、パラメータを調整したり TCP window scale option を利用することによりスループットの向上が見込める。一方、パケット損失に関しては iSCSI の方がその影響を受けにくいことが分かった。したがって輻輳しやすいネットワークにおいては NFS より iSCSI の方が適していると言える。今後はさらに iSCSI による RAID を構成し、リモートディスクへの自動的なミラーリングがどの程度のパフォーマンスで実現可能であるかを調査する予定である。

謝　　辞

43Gbps 回線実験の機会を与えていただきました日本電信電話株式会社および東日本電信電話株式会社に深く感謝いたします。

参 考 文 献

1) J.Satran, K.Meth, C.Sapuntzakis, M.Chadalapaka, and E. Zeidner. *Internet Small Computer Systems Interface (iSCSI)*, April 2004. RFC 3720.
2) 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, Vol. J87-D1, No. 2, pp. 216–231, February 2004.
3) Stephen Aiken, Dirk Grunwald, and Andrew R. Pleszkun. A performance analysis of the iscsi protocol. In *Proceedings of the 20th IEEE/11th NASA Goddard Conference on Mass Storage Systems and Technologies (MSS'03)*, pp. 123–134, 2003.
4) Ming Zhang, Yinan Liu, and Qing Yang. Cost-effective remote mirroring using the iscsi protocol. In *Proceedings of 12th NASA Goddard, 21st IEEE Conference on Mass Storage Systems and Technologies (MSST04)*, pp. 101–114, April 2004.
5) Linux-iscsi project. http://linux-iscsi.sourceforge.net/.
6) Linux isci target implementation. http://www.ardistech.com/iscsi/.
7) Peter Radkov, Li Yin, Pawan Goyal, Prasenjit Sarkar, and Prashant Shenoy. A performance comparison of nfs and iscsi for ip-networked storage. In *Proceedings of 3rd USENIX Conference on File and Storage Technologies (FAST'04)*, pp. 101–114, March 2004.