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

Performance Evaluation and Improving of Sequential Storage Access using the iSCSI Protocol in Long-delayed High throughput Network

山口 実靖♥　小口 正人♦
喜連川 優♠

Saneyasu YAMAGUCHI　Masato OGUCHI
Masaru KITSUREGAWA

現在，安価でかつ規模拡張性の高い SAN 構築方法や，データセンターなどのストレージアウトソーシング手法として iSCSI が注目されている．本稿では高遅延広帯域環境における iSCSI プロトコルを用いたシーケンシャルリードの性能について述べる．まず，高遅延広帯域環境における iSCSI シーケンシャルリードの性能評価を行う．結果，既存の OS システムなどでは遅延の増加に伴いその性能が大きく劣化することが分かった．次に，iSCSI プロトコルの振る舞いを説明し，性能劣化原因がブロックサイズの小ささにあることを述べる．最後にこの問題を回避することによりその性能が大きく改善されることを示す．

The iSCSI protocol draws attention as a method of configuring SAN with low cost as well as a method of storage outsourcing such as data center. Performance of sequential access via the iSCSI protocol in long-delayed high throughput network is discussed in this paper. At first, performance of sequential access with iSCSI is measured in long-delayed high throughput network. The result shows that the performance severely decreases as network delay increases. We describe behavior of sequential access with iSCSI and show that accessing with small block size causes deceasing of performance. Performance of accessing with large size block is described. Accessing with large size block can increase the performance significantly.

1. はじめに

超大容量のデータを高速に処理するためのシステムとして，SAN(Storage Area Network)[1]が注目を集めている．SANを導入しストレージを集約することによりその管理コストが大きく削減できると言われており，その実績は高い評価を得ている．しかし，現世代のFC(Fibre Channel)を用いて構築するSAN(以下“FC-SAN”と呼ぶ)は，最大接続距離の短さや，ハードウェアコストの高さなど，問題点も明らかになってきており，TCP/IP と Ethernet で構築するIP-SAN への期待が高まっている．IP-SAN用データ転送プロトコルとしても iSCSI プロトコルが2003年2月にIETF[2]により承認されその期待はますます高まっている．本稿では，高遅延環境におけるiSCSIプロトコルによるシーケンシャルリードの性能について述べる．iSCSIの登場は広域SANやデータセンターによるストレージアウトソーシングなどを可能とし，今後は高遅延環境における性能は重要となってくると考えられる．シーケンシャルリードは一般的ストレージアクセス方法の一つで大規模データマイニング，マルチメディアDB，デジタルライブラリやこれらのためのデータのバックアップで用いられその性能向上はとても重要と思われる．本稿では第0章においてiSCSIプロトコルを用いたネットワーク越しのストレージアクセスの性能評価を行い，その性能を示す．これによりiSCSIの性能はネットワーク遅延時間の増加に伴い激しく劣化してしまうこと，ネットワークが提供できる性能と比べてiSCSI層で得られる性能は大きく劣ることが確認された．次に第0章において，iSCSIプロトコルの振る舞いを元にその性能低下がブロックサイズの小ささに起因していることを示す．そして，第5章においてこの問題を回避することにより性能劣化を大幅に抑えられることを示す．

2. 研究背景

2.1 iSCSI

iSCSI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ネットワーク(インターネットやEthernet LANなど)越しにSCSIアクセスを実現するSCSI over TCP/IPのためのプロトコルである．データセンターなどのストレージアウトソーシング，データバックアップ(特に災害復旧目的)や次世代SAN(IP-SAN)などの応用に有効であると期待されており，2003年2月にIETFにより承認された．IP-SANはEthernetとTCP/IPを用いて構築するSANであり，FCを用いて構築する現世代のSANの，FC接続距離の限界，相互接続性，FCハードウェアのコストの高さ，FC技術者の少なさなどの問題を解決するSANとして注目されている．TCP/IPの接続距離の長さ，技術者の多さ，導入コストの低さは実証済みであるがその性能の問題が残されており本稿ではiSCSI使用時の性能について述べる．

2.2 本研究の位置付け

本稿では，iSCSI の最大の特徴であるSCSIプロトコルのTCP/IPネットワーク上での転送に注目し考察をする．これらの影響と他の要素の影響と分離するため，本稿ではストレージデバイスが十分に高速と見なせる環境において考察を行う．以下の議論はiSCSI性能の上限を示し，ストレージデバイスが十分に高速であってもネットワークにより性能が激しく劣化してしまうこと，およびその解決策を示すものである．実環境の性能の考察にはさらにストレージデバイスの振る舞いを考慮する必要があるが，高遅延環境においてはネットワークの振る舞いが性能に対し支配的になり実性能はこれ近いと予測する．

♥ 正会員　東京大学生産技術研究所
sane@tkl.iis.u-tokyo.ac.jp
♦ 正会員　お茶の水女子大学　oguchi@computer.org
♠ 正会員　東京大学生産技術研究所
kitsure@tkl.iis.u-tokyo.ac.jp

2.3 関連研究

文献[3]において，Ng らは独自の SCSI over IP 実装を用いて 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が遅延時間にほぼ反比例することを指摘している．また，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． しかし，一貫性の問題やアプリケーション変更のない手法については言及していない．文献[4]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，ネットワークの影響が大きい高遅延環境についての考察はなされていない．

3. iSCSI シーケンシャルリードの性能評価

本章では，ネットワーク遅延時間と iSCSI を用いたシーケンシャルリードの性能の関係について述べる．

3.1 実験環境

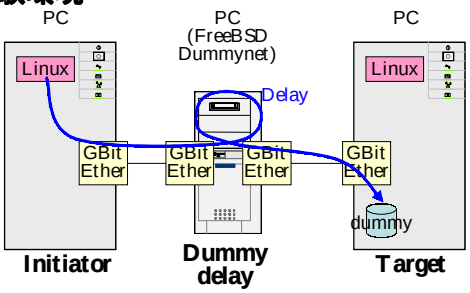

図 1 実験環境

図 1 の環境において，シーケンシャルリード性能評価実験を行った．Target(ストレージ)と Initiator(サーバ計算機)の間に人工遅延装置 Dummynet[5]をはさみ，Target - Initiator 間で TCP コネクションを確立し iSCSI 接続を行う．“Initiator-Dummynet”間および“Dummynet-Target”間は Gigabit Ethernet クロスケーブルで接続を行う．iSCSI 実装はニューハンプシャー大学 InterOperability Lab [6]が提供する reference 実装(iSCSI draft 18[7]準拠 ver.3)を用いた(以下この実装を UNH と呼ぶ)．Target は前述の理由によりメモリモード（これは無限に高速なストレージとみなせる）で動作させたため実ストレージアクセスはともなわない．Initiator，Target，Dummynet は全て PC 上に構築した．PC は CPU が 1.5GHz Pentium 4，メインメモリ 128MB，Initiator と Target の OS が Linux 2.4.18，Dummynet の OS が FreeBSD 4.5-RELEASE，NIC が Intel PRO/1000 XT Server Adapter (Initiator と Target は 1 枚，Dummynet は 2 枚の NIC を搭載)である．性能測定は Initiator 計算機の OS 上からシングルスレッドアプリケーションで iSCSI 接続の SCSI デバイスの raw デバイスに対して **read()**システムコールを繰り返し呼び出すことによりシーケンシャルリードを行い，そのスループットを計測した．システムコール時のブロックサイズ(以下これを“SC ブロックサイズ”と呼ぶ)は 500KB である．TCP 受信 Window サイズは 1MB である．

3.2 実験結果

上記の測定実験を行い図 2 の結果を得た 横軸(1 Way Delay)は，Dummynet により人工的に生成した Initiator と Target の間の片道遅延時間である．“0ms”は Dummynet を経由するが人工的には遅延を作成しなかった場合であり，実際は片道 140μs 程度である．図中の“iSCSI(DEF)”が，上記の実験条件における iSCSI シーケンシャルリードのスループットである(“DEF”は“DEFAULT”の略でチューンを行っていない場合の性能を意味する)．図中の“Socket”は，“iSCSI(DEF)”と同条件(片道遅延，TCP Window Size が等しい)における単純なソケット通信のスループットであり，これが実験環境(iSCSI にとっての下位層)が提供できる限界の通信速度と見なせる．“Socket”の上限が約 40[MB/s]であるのは Dummynet の限界である．測定結果より**(1)iSCSI スループットは遅延の増加に伴い大きく劣化する**ことおよび，**(2)下位層(“Socket”参照)は高いスループットを提供できるのにもかかわらず iSCSI 層では低い性能しか得られない**ことが確認された．

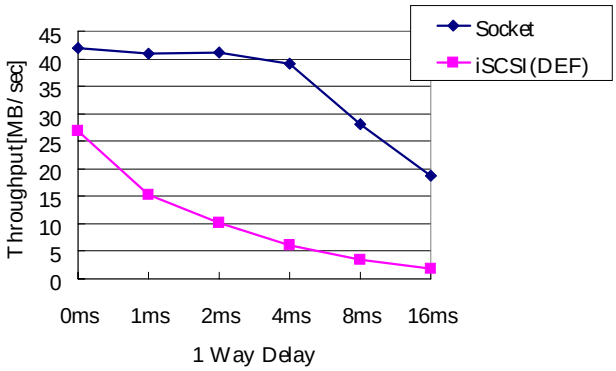

図 2 iSCSI シーケンシャルリードスループット

4. 性能低下要因の考察

一般に高遅延環境におけるレスポンスタイムの悪化は不可避であるが，iSCSI ではスループットも大きく劣化してしまうことが前章の実験により確認された．また，iSCSI スループットは基本的に下位層のスループット(“Socket”)を超えることができないため iSCSI 層による性能劣化を最小限に抑えることが重要であるが iSCSI 層がこれを大きく劣化させていることも確認された．本章では，iSCSI シーケンシャルアクセスの振る舞いを説明し，性能低下の原因およびその回避手法について述べる．

4.1 iSCSI シーケンシャルリードの振る舞い

ブロックサイズが小さいときの iSCSI シーケンシャルリードは図 3 の振る舞いを繰り返すことにより行われる．

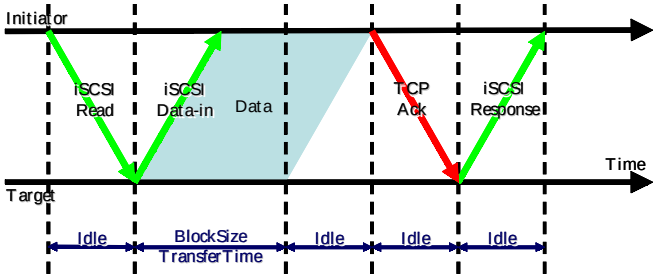

図 3 iSCSI シーケンシャルリードの振る舞い

すなわち，(1)Initiator が iSCSI Read PDU(Protocol Data Unit)を Target に送信し(図 3 中の“**iSCSI Read**”)，(2)Target がデータを返し(“**iSCSI Data-in**”および

“Data”), (3)InitiatorがTCP Ackを送信し(“TCP Ack”), (4)Targetが iSCSI Responseを送信する(“iSCSI Response”), を繰り返す（以下この iSCSI Readサイクルを“iRdサイクル”と記す）．実例として前章の実験環境において片道遅延時間 8msの場合のパケット遷移図を図 4に示す．“iRdサイクル”内にはネットワークを使用せずに相手の送信データの到着を待つ時間（図 3中の“Idle”）が 4個含まれている．以上より“iRdサイクル”の時間，スループットは以下の通りモデル化される．

$$\text{サイクル時間} = 4\times\text{片道遅延時間} + \text{データ転送時間}$$

$$\text{データ転送時間} = \frac{\text{ブロックサイズ}}{\text{下位層スループット}}$$

$$\textit{iSCSI}\text{スループット} = \frac{\text{ブロックサイズ}}{4\times\text{片道遅延} + \frac{\text{ブロックサイズ}}{\text{下位層スループット}}}$$

ブロックサイズとは，iSCSI Read PDU内に記述されているバイト数である(“SCブロックサイズ”と区別するためこれを“PDUブロックサイズ”と呼ぶ). モデルより iSCSIスループットは“PDUブロックサイズ”と下位層スループットに対し単調増加，遅延時間に対して単調減少であることが分かる．モデルと第 0節の実測値の差は，遅延 1msで 9%，2msで 8%，4msで 4%，8msで 1%，16msで 4%である．ただし，第 0節で後述する理由により“PDUブロックサイズ”は 125KBとしてモデルスループットを計算した．

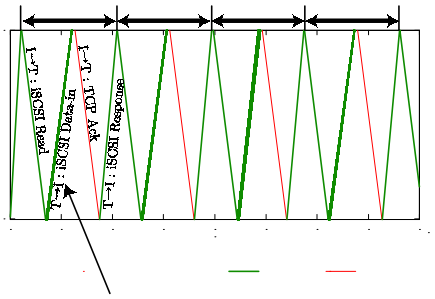

図 4 パケット遷移図

4.2 性能の低下の理由

次に，実環境[1]において iSCSIの性能が著しく低下する理由を説明する．第 0章の測定は Linux OS上から iSCSI接続のリモートストレージの rawデバイスに対して**read()**システムコールを発行することにより測定した. OSやドライバ等の実装に強く依存することであるが，一般にシステムコールが発行されるとキャラクタデバイス，SCSIドライバ，iSCSI Initiatorドライバ等を経由し TCP/IP実装に送信データ(これは iSCSI PDUである)が渡される．これらを経由した結果発行したシステムコールがそのまま TCP/IPに伝えられるとは限らない．我々が TCP/IP層でパケットを観察したところ第 0節の実験環境の例では，大きな“SCブロックサイズ”の**read()**が発行されると，127.5KBの“PDUブロックサイズ”の iSCSI Read PDUに分割され，1 PDUずつ順に送信とデータの受信を行う．本実験の例では，500KBの“SCブロックサイズ”は，“PDUブロックサイズ”が 127.5KBの PDU 3個と 117.5KBの PDU 1個(合計 4 PDU)に分割される．第 0節で述べたように“iRdサイクル”にはアイドル時間が含まれるため大きな“SCブロックサイズ”が小さな“PDUブロックサイズ”群に分割されることは性能を大きく劣化させる．本測定の例では“iRdサイクル”内におけるネットワークを使用しないアイドル時間の比率は片道遅延時間 1msにおいて 57%，2msで 73%，4msで 84%，8msで 88%，16msで 91%と計算され，このアイドル時間がスループット低下の主たる要因であることが分かる．一般に小さなブロックによるアクセスは性能を低下させるがネットワークを経由する iSCSIではこの影響が非常に大きくなると言える．

[1] ただし，実ストレージアクセスは考慮していない

5. 性能向上手法

5.1 遅延隠蔽手法

モデル化を行うことにより性能向上手法の発見は容易となる．第 0節のモデルより iSCSIのスループットは，(1)ブロックサイズ(“PDUブロックサイズ”)を大きくする，(2)下位層(TCP/IP層)のスループットを向上させる，(3)片道遅延時間を短縮させる，により向上できると考えられる．まず手法(1)であるが，前述のように性能劣化の最大の要因が小さな“PDUブロック”によるアイドル時間であるため“PDUブロックサイズ”を大きくしアイドル時間比率を相対的に下げることは大きな効果があると期待できる．手法(2)は 10Gigabit Ethernetの使用や TCP offloadエンジンなどにより実現されるが，第 0節で示したようにブロックサイズが小さいままでは“iRdサイクル”内における実通信時間の割合は少ないためこれにより短縮される時間の割合も少ない．手法(3)は TCP offloadエンジンやマルチスレッド化により実現されるが，TCP offloadエンジンにより短縮される時間はネットワークの遅延時間と比べて小さく，効果も小さいと考えられる．アプリケーションをマルチスレッド化しコネクションを複数確立することは仮想的な遅延時間の短縮に効果があると期待される．しかし，マルチスレッド化および複数コネクションの管理がアプリケーション開発者に与える実装負荷は少なくないため本稿ではこれに言及しない．以上の理由により**(1)“PDUブロックサイズ”の拡大**が最も効果的であると考える. またこれまでに言及されてないが TCP/IPおよび iSCSIにはそれぞれ受信 Windowサイズ，MaxBurstLength等のパラメータが定められておりこれらも同様に十分に大きな値とする必要がある．

5.2 評価実験

前章の手法(“PDUブロックサイズ”の拡大)による高遅延環境における iSCSIスループットの向上を評価する実験を行った．第 0節で述べた理由によりアプリケーションが大きな“PDUブロックサイズ”の iSCSI Read PDUを発行できないため，簡易 iSCSI Initiatorを試作し測定を行った．試作 Initiatorはカーネル空間で動作するデバイスドライバではなくユーザ空間で動作するアプリケーションである．試作 Initiatorは Target計算機と直接 TCP/IPコネクションを確立し iSCSIプロトコルに基づいてシーケンシャルリードを行う．iSCSI Login処理，iSCSI Read処理のみが実装されており ErrorHandlingなどの処理は実装されていない．UNH実装同様 iSCSI draft18に準拠している．第 0節の環境における試作 Initiatorを用いた iSCSIシーケンシャルリードの性能は図 5の“iSCSI(KI)”の様になった．比較のために図 2の“Socket”および“iSCSI(DEF)”も併せて記す．“PDUブロックサイズ”はそれぞれ，125KB，1MB，2MB，4MBである．図中の“iSCSI(DEF)”は UNH Initiatorと UNH Targetを用いており，“iSCSI(KI)”は試作 Initiatorと UNH Targetを

用いている．“iSCSI(DEF)”と“iSCSI(KI)”を比較し“PDU ブロックサイズ”を大きくすることにより遅延増加に伴うスループットの劣化およびソケット通信に対する iSCSI プロトコルによる劣化が大幅に抑えられ，iSCSI 性能が大きく向上されることが確認された．特に，高遅延環境において大きなスループットの向上が確認された．“iSCSI(DEF)”と“iSCSI(KI)”の違いとして“PDU ブロックサイズ”の他に SCSI ドライバ等のドライバを経由するか否かの違いが存在するが“iSCSI(DEF)”(これは“SC ブロックサイズ”が 500KB であり，平均“PDU ブロックサイズ”が 125KB である)と“iSCSI(KI) 125K”を比較することにより高遅延環境においてその差は十分に小さいことが分かり，本実験における性能向上は“PDU ブロックサイズ”によるものであることが確認できる．片道遅延 16ms におけるブロックサイズと iSCSI スループットの関係を図 6 に示す．同図からもブロックサイズを大きくすることによりスループットが向上することが確認できる．ただし，片道遅延 16ms において 10.4[MB/sec] のスループットは下位層により与えられるスループットの半分強に減少していることを意味し，さらなる工夫も必要と言える．

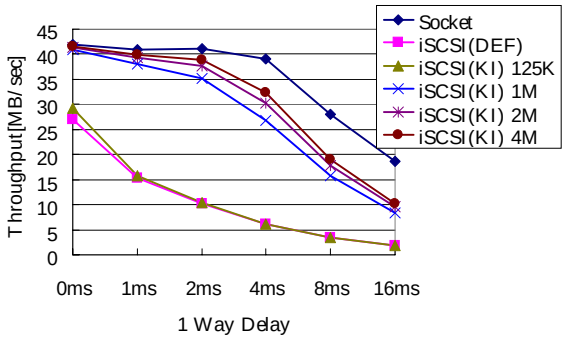

図 5　試作 Initiator iSCSI スループット

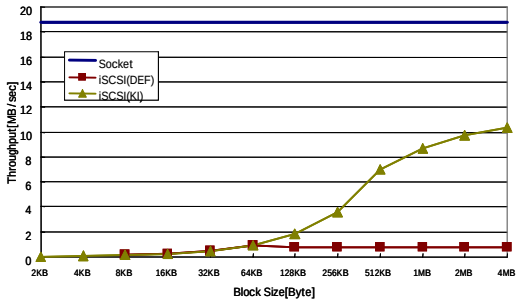

図 6　ブロックサイズと iSCSI スループット

6. おわりに

本稿では高遅延広帯域ネットワーク環境下で iSCSI プロトコルを用いてシーケンシャルアクセスを行うときの性能について述べた．まず，実験により一般的な OS システム上で iSCSI プロトコルを用いるとネットワーク遅延の増加に伴いシーケンシャルアクセスのスループットが著しく低下することを示した．つぎに，iSCSI プロトコルの振る舞いを説明しスループット低下の原因が Read ブロックサイズの小ささにあることを述べた．そして，試作 iSCSI Initiator を用いて大きなブロック単位でのシーケンシャルアクセスの性能を評価し，スループットの劣化を大幅に抑えられることを示した．遅延の増加によるスループットの著しい低下は数 ms 程度から観察されており，これの回避は iSCSI を実用する上で重要になると考える．特に，広域 SAN などではこれが重要である．本稿で示した実験例の範囲でも 片道遅延 4ms 程度までは下位層の提供する速度とほぼ等しい速度が得られており，片道遅延 16ms の状況において 10[MB/sec] 程度のスループットは確認されている．各ネットワークインフラの状況に依存するが提供されているネットワークの性能が十分速いとき，国内程度の距離であれば iSCSI プロトコルを用いて十分なスループットが確保できると言える．今後は，高遅延環境(8ms, 16ms やそれ以上)におけるさらなる性能の向上，実ストレージを用いての評価，シーケンシャルリードリードライトアクセスやランダムアクセスの評価やその性能向上，などを行っていく．

[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 : http://www.ietf.org/
[3] Wee Teck Ng, Bruce Hilly Elizabeth Shriver, Eran Gabber, Banu Ozden, ``Obtaining High Performance for Storage Outsourcing", ¥textit{Proc. FAST 2002, USENIX Conference on File and Storage Technologies}, January 28-29, 2002, pp. 145-158
[4] Prasenjit Sarkar and Kaladhar Voruganti, ``IP Storage: The Challenge Ahead", ¥textit{Proc. of Tenth NASA Goddard Conference on Mass Storage Systems and Technologies}, April 2002
[5] L. Rizzo, "dummynet", http://info.iet.unipi.it/~luigi/ip¥_dummynet/
[6] The University of New Hampshire's InterOperability Lab
[7] 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

山口 実靖　Saneyasu YAMAGUCHI
東京大学生産技術研究所 学術研究支援員．2002 年 東京大学大学院博士課程修了，工学博士．iSCSI を用いたネットワークストレージシステムの性能向上の研究に従事．情報処理学会正会員．

小口 正人　Masato OGUCHI
お茶の水女子大学理学部情報科学科助教授．1995 年 東京大学大学院工学系研究科博士課程修了，工学博士．並列・分散処理，計算機ネットワークに関する研究に従事．IEEE, ACM, 電子情報通信学会，情報処理学会各会員．

喜連川 優　Masaru KITSUREGAWA
1978 年東京大学工学部電子工学科卒．1983 年同大大学院工学系研究科情報工学博士課程了．工学博士．同年同大生産技術研究所講師．現在，同教授．平成 15 年 4 月より、同所 戦略情報融合国際研究センター長．データベース工学，並列処理，Ｗｅｂマイニングに関する研究に従事． 情報処理学会理事，SNIA-Japan 顧問，ACM SIGMOD Japan Chapter Chair．平成 9，10 年本学会データ工学研究専門委員会委員長． VLDB Trustee，IEEE ICDE，PAKDD，WAIM ステアリングコミティメンバ