iSCSI遠隔ストレージアクセス時のTCPの振舞に関する一検討

比嘉 玲華† 神坂 紀久子† 山口 実靖‡ 小口 正人†
†お茶の水女子大学 ‡東京工学院大学

1 はじめに

近年、インターネットなどの発達により、企業でも家庭でも、個人の所有する情報量が爆発的に増えてきた。それに伴って問題となってくるのが、データを保全、管理する作業である。データを格納するストレージ装置の増設や万一に備えたデータのバックアップ、データの移管など、面倒な作業が幾つも生じてくる。さらに、データが複数の機器に分散しているという問題もある。

こうした問題に抜本的な解決策を与えるプロトコルとして登場したのが、「iSCSI」である。しかし、iSCSIは、複雑な階層構成で処理されており、バースト的なデータの転送も多いことから、通常の通信と比較して、特に、高遅延環境においては著しく性能が著しく劣化してしまう。また、下位基盤のTCP/IP層が提供できる限界性能を超えることはできず、最大限の性能が発揮できるようTCPパラメータなどを制御することが求められる[2]。そこで、本研究では、遅延装置で高遅延環境を作り、iSCSIのパラメータを最適に設定し、その状況でのTCP輻輳ウィンドウの振舞とスループットを観察して、性能向上の手法を考察、検討する。

2 研究背景

2.1 研究目的

一般に遠隔ストレージへのデータバックアップを考えた場合、データの書き込み量と読み出し量を比べると、書き込み量の方が圧倒的に多いことが容易に想像可能である。また、遠隔ストレージ側には標準的な環境のみを使用し、カスタマイズが不可能である場合も多い。そこで本研究では、iSCSIのInitiator側をカスタマイズして高遅延環境におけるiSCSIのシーケンシャルライトアクセスの性能を高めるための手法を考案、検討する。

2.2 Linux TCP 実装

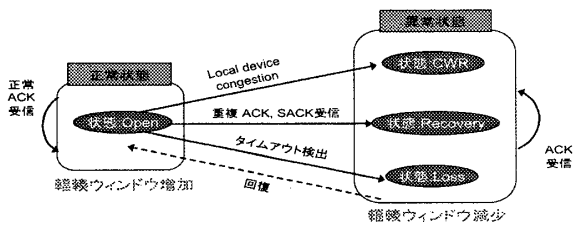

図 1: LinuxTCP の状態遷移

TCPでは，通信能力の制御にウィンドウサイズという概念を用いている．ウィンドウサイズとは，ホストがACKなしに一度に送信できるデータのサイズで，TCPヘッダに含まれる．また，このウィンドウサイズは，データの送信側では輻輳ウィンドウ，受信側では広告ウィンドウと呼ばれ，このどちらか小さい方の値がウィンドウサイズとして用いられる．広告ウィンドウは現在の受信ウィンドウの空き容量を示しており，ACKで送信側に送られる．一方，輻輳ウィンドウは送信側の制御パラメータで，ネットワークの混乱を回避するため送信側がOSのカーネル内において自主的に制限する値である．輻輳制御ではこの輻輳ウィンドウが利用されている．

本実験で用いたLinux OSにおいては，通信時の状態が正常であればACK受信ごとに輻輳ウィンドウは増加するが，エラーが検出されると異常と判断され，輻輳ウィンドウは低下する(図1)．輻輳ウィンドウが低下する原因としては，送信側デバイスドライバのバッファが溢れることによるLocal Congestionエラーを検出した場合（CWR），重複ACK又はSACKを受信した場合（Recovery），タイムアウトを検出した場合（Loss）の3つが挙げられる．また，LinuxのTCP実装では，通信中に一度設定された輻輳ウィンドウは，そのウィンドウの値を使い切らない限りは変化しないという特徴を持ち，この時スループットはほぼ一定の値で安定することが確認されている．

3 研究概要

3.1 実験手順

本研究は、図2の環境で行った。また使用したシステムを表1に示す。InitiatorとTarget間はGigabitEthernetで接続し、TCP/IP接続を確立した。また、Initiator側のTCPソースコードにモニタ用の関数を挿入し、ユーザ空間からもアクセス可能なカーネルメモリ空間に記録する仕組みを作成した。これにより、InitiatorからTargetへのライトアクセス時の輻輳ウィンドウの値が観察可能になる。遅延装置を使い、高遅延環境を作り出し、デフォルトのiSCSIとパラメータ設定を変更したiSCSIを起動して測定を行った。

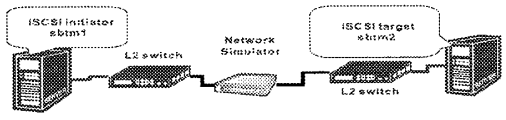

図 2: 実装システム

3.2 iSCSI パラメータ設定

本研究において、iSCSIのパラメータ設定をライトアクセス時における最適な状態になるように調整した。Target側で、Unsolicitedなライト通信を行えるようにし、Unsolicitedなライトの最大長を64KBから256KBへ、

A Study of TCP Behavior on iSCSI Remote Storage Access
† Reika Higa, Kikuko Kamisaka, Masato Oguchi
‡ Saneyasu Yamaguchi
Ochanomizu University (†)
Kogakuin University (‡)

OS	Red Hat Enterprise Linux 2.6.18-8.e.15
CPU	Quad Core Intel Xeon 1.6GHZ
Main Memory	2GB
HDD	73GB SAS×2(RAID0)
RAID Controller	SAS5iR
iSCSI	Initiator : open-iscsi-2.0-865
	Target : iSCSI Enterprise Target(IET)-0.4.15
Network Simulator	ANUE

表 1: 実験環境

Target が受信する PDU の最大セグメント長を 8KB から 128KB へ変更した。

3.3 bonnie++

ハードディスクベンチマークツールとしては bonnie++1.03 を用いた [3]。これはデータベースのような大規模なファイル操作のスループットを測定可能である。また比較的小さなファイルの作成、読込み、削除のスループットも測定可能である。本研究では、Sequential Write(連続書き込み) のスループットを測定した。

4 実験結果

4.1 ローカルディスク、iSCSI アクセスにおけるスループット

遅延装置を使って、片道遅延時間 0,1,2,4,8,16ms の遅延環境を作り、デフォルトの iSCSI と最適パラメータ設定の iSCSI のスループットを測定した。また比較として、ローカルディスク (SAS) アクセスとの性能も測定した。この結果を図 3 に示す。(横軸 ms、縦軸 MB/s) ローカルディスクに高速な SAS ディスクをハードウェア RAID0 構成で用いているため、ローカルアクセスが極めて性能が良いことが確認できた。これと比較すると iSCSI の性能は低くなる。ただし、iSCSI を用いた場合も低遅延環境においては比較的良好なスループットが出ているが、高遅延環境においては、遅延時間と反比例するようにスループットが低下していた。また、最適パラメータ設定の iSCSI のスループットはデフォルトの iSCSI のスループットよりも高い値になっているが、高遅延環境になるにつれて差がなくなっている。

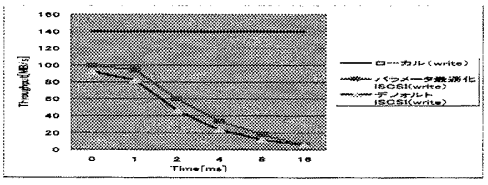

図 3: ストレージアクセスのスループット

4.2 輻輳ウィンドウモニタ

高遅延環境においてスループットが急激に低下する理由として、以下のようなことが考えられる。iSCSI は SCSI コマンドを、TCP/IP パケット内にカプセル化しており、SCSI over iSCSI over TCP/IP over Ethernet という複雑な構成となっている。iSCSI を用いる通信は下位レイヤである TCP/IP の提供するスループットを超えることは不可能であり、TCP の設定や振舞が性能に大きな影響を与えると考えられる。

そこで、輻輳ウィンドウの値をモニタし、振舞を調べた。write システムコール 12 より Direct I/O を行うプロセスを 20 並列で起動し、ターゲットへのライトアクセスを実行した (図 4、図 5)。また、片道遅延時間 8ms の遅延環境において測定した。パラメータを最適化した iSCSI では、デフォルトの iSCSI に比べて、輻輳ウィンドウの成長は比較的早く、またその最大値は約 300 から約 400 へと大きくなることが確認できた。

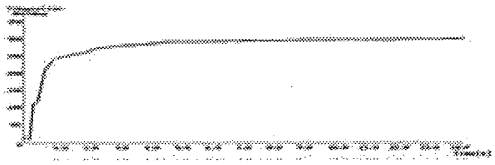

図 4: デフォルト iSCSI 輻輳ウィンドウ

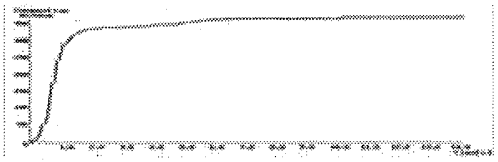

図 5: パラメータ最適化 iSCSI 輻輳ウィンドウ

5 まとめと今後の課題

iSCSI のパラメータを最適な値に設定することで、シーケンシャルライトアクセスにおけるスループットが向上すること、輻輳ウィンドウの値が大きくなることが本研究で確認できた。しかし、高遅延環境になるにつれてスループットが大幅に低下し、差もあまり見られなくなってしまう。その原因としては、大きなブロックサイズで write システムコールを発行しても、SCSI 層において、小さなブロックサイズに分割されてしまうことによるものと考えられる。今後の課題としては、ブロックサイズが分割されないような手法を考案し、さらにその条件下において輻輳ウィンドウを制御するなどして、高遅延環境におけるスループットを高めていく。

6 謝辞

本研究は一部、独立行政法人科学技術復興機構産学共同シーズイノベーション化事業によるものである。

参考文献

[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

[2] 豊田真智子, 山口実靖, 小口正人："高遅延ネットワーク環境における iSCSI リードアクセス時の TCP 輻輳ウィンドウ制御手法の性能評価" SACSIS2005, pp.443-450, 2005 年 5 月

[3] Bonnie++ http://www.textuality.com/bonnie/intro.html_5.c