5A-5

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

藤原啓成† 若宮直紀‡ 志賀賢太†

(株)日立製作所システム開発研究所† 大阪大学大学院情報科学研究科‡

1. はじめに

近年の iSCSI プロトコル[1]の実装の進展と広域 IP 網サービスの低価格化により，企業の PC・サーバ等が遠隔地にあるデータセンタのストレージ装置に直接アクセスできる環境が整いつつある．広域 IP 網サービスは専用線サービスよりも安価であるが，遅延やパケット損失率が大きいため，iSCSI 通信を適用した場合のスループット低下が指摘されている．

本稿では，広域 IP 網を介した長距離アクセス向けに，iSCSI および関連プロトコルレイヤ(TCP，SCSI およびバックアップアプリ：図 1参照)のプロトコルチューニングを検討した．iSCSI のソフトウェア実装[2]をチューニング対象とし，WAN エミュレータによる広域 IP 網の模擬環境により，チューニングの効果を確認したので報告する．

2. 広域 IP 網を介した iSCSI 通信の課題

広域 IP 網を介した iSCSI 通信のスループット低下の原因は，広域 IP 網の遅延の大きさによる帯域遅延積の増大と，パケット損失時の TCP レイヤの通信スループットの低下である．本章では，原因ごとに課題を説明する．

2.1. 高遅延に伴う帯域遅延積の増大

広域 IP 網を介した iSCSI 通信では，長距離アクセスであることにより，エンドポイント間の伝播遅延が大きく，帯域を使い切るために広域 IP 網上に送り出すべきデータ量(帯域遅延積)が増大する．帯域遅延積が増大して TCP のウィンドウサイズおよび送受信バッファサイズを上回ると，帯域を使い切れなくなる．これは，図 1に示す TCP の各上位レイヤのデータ送出量が不十分な場合も同様である．この結果，広域 IP 網の利用率が低下して iSCSI 通信のスループットが低下する．

2.2. パケット損失に伴う TCP 通信のスループット低下

広域 IP 網では，網内の輻輳により IP パケットの損失が発生する．パケット損失が発生すると，TCP の輻輳制御[3]によって，TCP 通信のスループットが 1/2 以下に低下し，その後，1RTT 毎に，TCP 通信のスループットが向上する．しかしながら，長距離アクセスでは RTT が大きい為，パケット損失前のスループットまで回復するには LAN 環境に比べて長い時間がかかる．

このため，TCP を下位プロトコルとする iSCSI 通信もパケット損失が発生時にスループットが低下する．

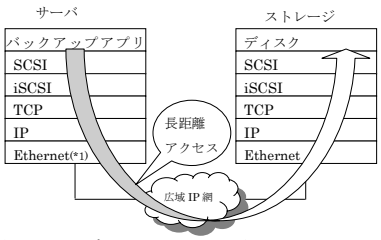

図1 広域 IP 網を介した iSCSI 通信のプロトコルレイヤ

A study on protocol tuning for iSCSI transmission over a wide-area IP network

Keisei Fujiwara†, Naoki Wakamiya‡, Kenta Shiga†

†Hitachi, Ltd., System Development Laboratory

‡Graduate School of Information Science and Technology, Osaka University

表1 プロトコルチューニング一覧

プロトコルレイヤ	パラメータ	プロトコルチューニングの種類			
		チューニング前	チューニング後		
			(1)帯域遅延積の増大対策チューニング	(2)TCP通信のスループット低下対策チューニング	
				(a)	(b)
バックアップアプリ	書込み多重度	1	6	6	6
SCSI	書込み多重度	1	2	8	16
iSCSI	InitialR2T	Yes	No	No	No
	FirstBurstLength	64KB	256KB	256KB	256KB
	MaxConnections	1	1	8	16
TCP	Window Scaling	無効	有効	有効	有効
	送信バッファサイズ	128KB	10MB	10MB	5MB
	受信バッファサイズ	170KB	40MB	40MB	20MB
	sack	無効	無効	有効	有効
	timestamps	無効	無効	有効	有効

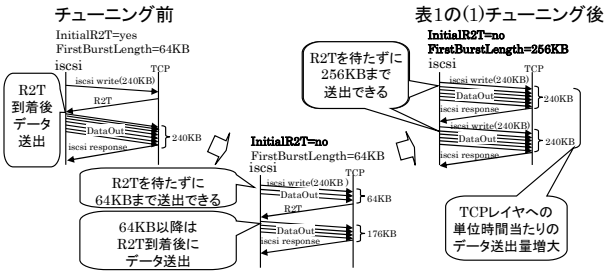

図 2 iSCSI レイヤチューニングの効果

3. プロトコルチューニングの検討

各プロトコルレイヤのチューニングによる前章の課題解決を検討した．高遅延環境下における iSCSI 通信については[4]などで Read の性能向上の検討がなされているが．本稿では，バックアップアプリのシーケンシャル Write の性能向上を検討する．また，企業のエンドユーザがチューニングすることを考慮して，iSCSI のソフトウェア実装[2]で設定可能な iSCSI パラメータをチューニング対象とする．なお，帯域遅延積の大きいネットワークにおける TCP 通信の性能向上としては HighSpeed TCP などの検討もなされているが，本稿では，TOE (TCP/IP オフロードエンジン)などの実システム環境を考慮して，RFC1323 などに示されている一般的に利用可能な TCP パラメータの範囲内でチューニングを検討する．

表 1に検討したプロトコルチューニングの一覧を示す．以下，課題別の検討内容を説明する．

3.1. 帯域遅延積の増大対策チューニング

帯域遅延積の増大による広域 IP 網の利用率の低下を防ぐため，表 1の(1)のチューニングを検討した．

TCP レイヤでは，WindowScaling を有効にし，かつ送受信バッファを拡大する．WindowScaling により TCP のウィンドウサイズは約 1GB まで拡張可能となるが，表 1(1)では送受信バッファサイズを帯域遅延積(表 2の広域 IP 網の模擬値から，帯域遅延積は 100ms×100Mbps＝1.25MB)よりも十分大きいものとした．

iSCSI レイヤでは，InitialR2T を no かつ FirstBurstLength を 256KB とすることで，単位時間当たりの TCP レイヤへのデータ送出量を増大させる(図 2)．

バックアップアプリおよび SCSI レイヤでは，書込み多重度を増やし，iSCSI レイヤへのデータ送出量を増大させる．

以上のチューニングにより，広域 IP 網に対して帯域遅延積を上回るデータ送出量を確保する効果が期待できる．

(*1)Ethernet は富士ゼロックス株式会社の登録商標です。

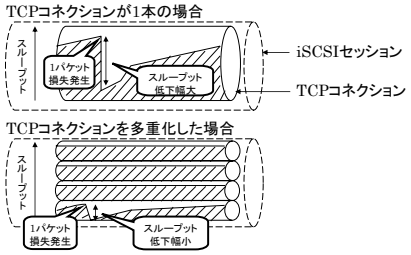

図 3 TCP コネクション多重化による
iSCSI 通信のスループット低下の軽減効果

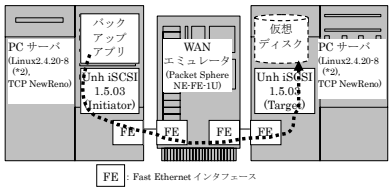

図 4 実験環境

表 2 WAN エミュレータの設定値

設定パラメータ	WANエミュレータの設定値
往復伝播遅延	100ms
ランダムパケット損失率	0.001%, 0.01%, 0.1%, 1%
帯域	100Mbps

3.2. TCP 通信のスループット低下対策チューニング

パケット損失時の，TCP の輻輳制御による TCP 通信のスループット低下の影響を軽減するため，表 1(2)の(a)，(b)のチューニングを検討した．

まず，表 1(2)の(a)について説明する．

TCP レイヤでは，sack オプションの有効化により，パケット再送時間を短縮する．また，timestamps オプションの有効化により，RTT の計測精度を向上させる．これにより，RTT を元に算出される再送タイムアウト時間(RTO)の精度を向上し，不適切な RTO によるパケットの誤再送を防止する．

iSCSI レイヤでは，MaxConnections を 8 とする．これにより，1 つの iSCSI セッション内で 8 本の TCP コネクションを確立でき，各コネクションが広域 IP 網の帯域を分割使用する．この TCP コネクションの多重化には，IP パケット損失時の iSCSI 通信のスループット低下を軽減する効果が期待できる(図 3)．なお，TCP のウィンドウサイズの合計値および送受信バッファサイズの合計値も増大するため，帯域遅延積の増大対策としても有望である．

SCSI レイヤでは，書込み多重度を 8 に変更し，8 本の TCP コネクションに 8 つの書込みが並行して行えるようにする．なお，iSCSI レイヤへのデータ送信量も増大するため，帯域遅延積の増大対策としても有望である．

次に，表 1(2)の(b)について説明する．表 1(2)の(b)では，上記の(a)のチューニングに加えて，iSCSI レイヤの MaxConnections および SCSI レイヤの書込み多重度を 16 に変更する．これにより，(a)の効果の増大を見込める．ただし，実験環境のメモリ資源の制約のため，送受信バッファを縮小した．

以上のチューニングにより，TCP の輻輳制御に伴う iSCSI 通信のスループット低下を軽減する効果が期待できる．

4. 実験

4.1. 実験環境

実験環境は，iSCSI のソフトウェア実装[2]を導入した PC サーバ 2 台および広域 IP 網を模擬する WAN エミュレータで構成した(図 4)．表 2に WAN エミュレータの設定値を示す．

実験にあたり，企業のバックアップサーバが広域 IP 網を介して遠隔地にあるデータセンタのストレージへデータのバックアップを行うことを想定した．

(*2)Linux は登録商標です。

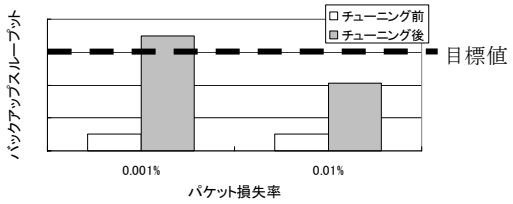

図5 帯域遅延積の増大対策チューニングの効果

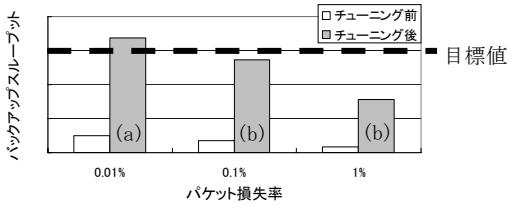

図6 TCP 通信のスループット低下対策チューニングの効果

バックアップスループットの目標値は，あるサーバが行うバックアップのスループットと同等以上であることとし図 5，図 6 中に破線で示している．

4.2. 実験結果

測定したバックアップスループットのグラフを，プロトコルチューニングの種類別に図 5と図 6に示した．なお，結果は 1 回の測定値を示す．ただし，チューニング前の測定値およびチューニング後の目標値付近の測定値については，再現確認のために 2 回の追加測定を実施し，計 3 回の最小値を示した．

(1) 帯域遅延積の増大対策チューニングの効果

図 5は，表 1の(1)のチューニングを適用する前後において，バックアップスループットを測定した結果である．図より，本チューニングが広域 IP 網を介したバックアップスループットの向上に有効であることが確認された．特に，パケット損失率 0.001%時には目標スループットを達成できている．

(2) TCP 通信のスループット低下対策チューニングの効果

図 6は，表 1の(2)のチューニングを適用する前後において，バックアップスループットを測定した結果である．さらなるチューニングによって，図 5と比較してバックアップスループットをより向上でき，パケット損失率が 0.01%の場合にも目標スループットが達成できることが確認された．

5. おわりに

本稿では，広域 IP 網を介した長距離アクセス向けの iSCSI および関連レイヤ(TCP，SCSI およびバックアップアプリ)のプロトコルチューニングを検討し，その有効性を確認する実験を行った．実験の結果，数倍のスループット改善が得られ，検討したプロトコルチューニングの有効性が確認された．ただし，パケット損失率 0.1%および 1%時には，スループット改善の効果はあるが，バックアップスループットとしては不十分であることが確認された．

参考文献

[1] J. Satran 他，“rfc3720: Internet Small Computer Systems Interface (iSCSI)，” IETF, 2004.4.
[2] http://sourceforge.net/projects/unh-iscsi/
[3] M.Allman 他，“rfc2581:TCP Congestion Control，” IETF, 1999.4.
[4] 山口 実靖 他，“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，” 電子情報通信学会和文論文誌 データ工学特集号(和文論文誌 D-I) pp.216-231, 2004.2.