iSCSI 遠隔ストレージアクセスにおける 各プロトコルレイヤの最適化による性能向上の実現

比嘉 玲華　　(指導教員:小口 正人)

1　はじめに

近年，ストレージ管理コスト低減などの目的で SAN(Storage Area Network) の導入が進んでおり，その中でも，より経済的でコストパフォーマンスの高い IP-SAN のプロトコルである iSCSI は大いに期待されている．また，テロや自然災害を考えた場合，データをできるだけ遠くに保存するというリモートバックアップは非常に重要であり，リモートバックアップのプロトコルの研究は学術的興味のみならず，産業界からも強く望まれている．よりコストパフォーマンスの高いシステムが利用される昨今の時勢を考えた場合，リモートバックアップのプロトコルとして iSCSI の適用が望まれる．しかし iSCSI をリモートバックアッププロトコルとして実用化するには様々な問題がある．最も大きな問題として，iSCSI は高遅延環境になるほど性能が劣化してしまうというデメリットが一般的に知られている．遅延時間 (RTT) に対するスループットを測定した図 4 の”default”のグラフからわかるように，本実験環境においても高遅延環境下における性能劣化が確認されている．

そこで本研究では，リモートバックアッププロトコルとして iSCSI を使用することを目的として，高遅延環境における iSCSI シーケンシャルライトアクセスの性能向上を実現するための手法の提案と実装を行い，その実行性能を解析および評価した．

2　iSCSI リモートストレージアクセス性能向上システムツール

2.1　実験システム

本研究において, Initiator と Target 間は GigabitEthernet で接続し，TCP/IP コネクションを確立した．Target のストレージには SAS ディスクを用い RAID コントローラによる RAID0 構成で接続した．実験環境を表 1 に示す．

表 1: 実験環境

OS	Red Hat Enterprise Linux 2.618-8.e.15
CPU	Quad Core Intel Xeon 1.6GHz
Main Memory	2GB
NIC	Intel PRO/1000PT Server Adaptor on PCI Express
HDD	73GB SAS×2(RAID0)
RAID Controller	SAS5/iR
iSCSI	Initiator : open-iscsi-2.0-870
	Target : iSCSI Enterprise Target(IET)-0.4.15
Network Analyzer	ClearSight Network Recorder
Network Simulator	ANUE

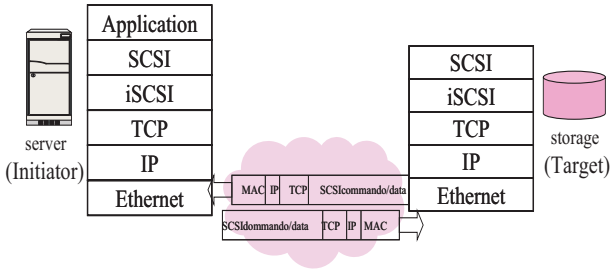

図 1: Configuration of iSCSI

2.2　解析システムツール提案

iSCSI ストレージアクセスは図 1 のように多段のプロトコルで構成されるため，全層を経由して通信処理が行われることから全層が性能劣化原因となる可能性があり，性能向上について考察するにはこれら全層を網羅的に解析する必要がある．そこで図 2 のようなシステムツールを構築した [1]．このシステムツールを使用することで，ネットワーク上を飛び交うパケット解析と Initiator におけるカーネル解析を行うことが可能となる．

システムツール構成要素の一つに我々のオリジナルのツールである「カーネルモニタ」がある．これは TCP カーネルの振舞をモニタするツールである．カーネル内部の TCP ソースにモニタ関数を挿入しカーネルを再コンパイルすることで，一般にはユーザ空間からは見ることができない TCP のパラメータ情報などを可視化できるというツールである．これによりモニタできるようになった値には，輻輳ウィンドウやソケットバッファキュー長といった TCP パラメータ情報や log trace timestamp がある．その他のシステムツール構成要素としては，tcpdump コマンド，ネットワークアナライザがある．これらによって，ネットワーク上を飛び交うパケットを解析し，パケット情報を得ることが可能になる．このシステムツールを効果的に使用しすべての層の解析，最適化を行うことで，iSCSI 遠隔ストレージアクセスの性能低下の原因がどこにあるのか詳しく解析していく．

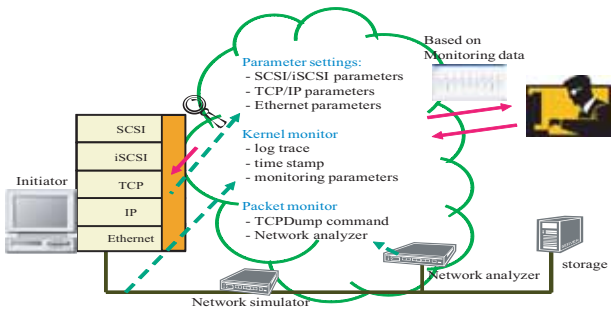

図 2: 解析システムツール概要

3　複数の層にまたがる解析と最適化

複数の層にまたがる最適化を行った．その際アクセスブロックサイズを 4MB，広告ウィンドウを通信の妨げにならない程度の値に設定した．アクセスブロックサイズを 4MB に設定したときのスループットの理論値を図 4 の”theoretical value”に示す．

3.1　Ethernet 層における最適化

NIC パラメータの最適化を行った．その結果 RTT20ms においてデフォルト時と比較して約 5% の性能向上が確認された．

3.2　iSCSI 層における最適化

本実験においては iSCSI のパラメータをライトアクセス時において最も高い性能が出るように最適化した．その結果，RTT20ms においてデフォルト時と比較して約 5 倍の性能向上が確認された．NIC パラメータと iSCSI パラメータ最適化を合わせた結果を図 4 の”optimized iSCSI parameter”に示す．しかし，高遅延環境での性能劣化という問題は解消されておらず，この値は理論値には尚も及ばない．

3.3　ネットワーク上のパケット解析

高遅延環境での性能劣化問題を解析すべくネットワーク上を飛び交うパケットを詳細に解析した．RTT20ms における解析結果を図 3 に示す．ブロックサイズ 4MB で送信したパケットが約 1MB ずつに断続して送信されていることがわかった．上位層には送るべきデータが存在するにも関わらず，送信パケットの断続が生じていることが性能低下の大きな要因だと考えられる．また，TCP ACK がきっかけでパケット送信再開が始まっていることから，パケット送信断続の原因は TCP 層のどこかにあるということが確認された．

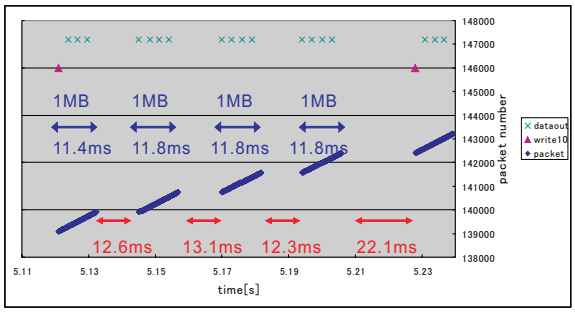

図 3: パケット解析

3.4　TCP 層における解析と最適化

トランスポート層に性能劣化の要因が考えられる場合，一般的には，広告ウィンドウ，輻輳ウィンドウ，ソケットバッファのいずれかである．

3.4.1　ウィンドウ解析

実際の iSCSI 通信時における広告ウィンドウ，輻輳ウィンドウの値を解析した [2]．その結果，広告ウィンドウ，輻輳ウィンドウともにパケット送信の断続を引き起こす値ではないことが分かった．

3.4.2　ソケットバッファ解析と最適化

ソケット通信時と iSCSI 通信時におけるソケットバッファのキュー長を解析した結果，明らかなる差異が確認された．iSCSI 通信時には，不必要なタイムアウト待ちが頻繁に生じており，送信されるべきデータが上位層に存在していても，キューが割当られることなく，パケット送信の断続が引き起こされていることがわかった．そこで，カーネルを解析し条件分岐点となるコードに対して対処して，ソケットバッファ最適化を行った結果，タイムアウト待ちは格段に減り，キュー長も増大し，図 4 の”optimized socket buff”に示す通り，RTT20ms においてデフォルト時と比較して約 7 倍の性能向上が達成された．またその値は理論値と匹敵する性能であることから，iSCSI リモートストレージアクセスの高遅延環境における性能劣化問題は解決されたということができる [1][3]．

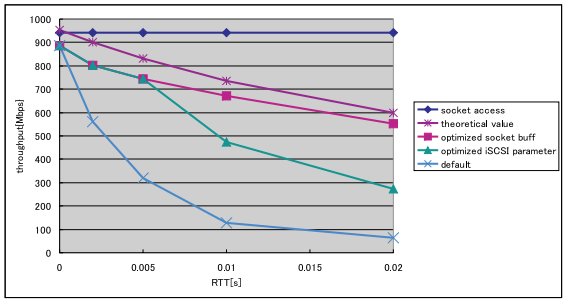

図 4: 各最適化後のスループット比較

4　まとめと今後の課題

本研究では iSCSI リモートストレージアクセスの性能向上を実現するためのシステムツールの提案と実装を行った．複数の層にまたがる解析と最適化の結果，TCP 層におけるソケットバッファ最適化がもっとも有効であった．iSCSI はローカル内での使用を想定したものであり，高遅延環境においてバーストアクセスが行われることは想定外であったことが原因として考えられる．

今後はその他の iSCSI を使用した解析をすることで汎用性の証明をするとともに，より現実的な環境での解析を行うことで iSCSI リモートストレージアクセス性能向上を実現するための有効な手法を提案していきたい．

参考文献

[1] Reika Higa et al. :”Analytical System Tools for iSCSI Remote Storage Access and Performance Improvement by Optimization with the Tools,” ANTS2009, New Delhi, India, December 2009

[2] Reika Higa et al. :”Optimization of iSCSI Remote Storage Access through Multiple Layers,” TeNAS’2009 in conjunction with AINA-09, pp.612-617, May 2009

[3] 比嘉玲華，松原幸助，岡廻隆生，山口実靖，小口正人：「iSCSI リモートストレージアクセスの性能向上を実現する手法の提案と実装」コンピュータシステム研究会（CPSY），信学技報，Vol.109, No.296, CPSY2009-38, pp.19-24，京都，2009 年 11 月