D-002

中規模 IP-SAN での高多重度 I/O 処理の解析と性能向上手法

Analyses and Performance Improvement of IP-SAN with Highly Multiplexed I/Os

山口 実靖† 小口 正人‡ 喜連川 優§
Saneyasu Yamaguchi Masato Oguchi Masaru Kitsuregawa

1．はじめに

iSCSI を用いる IP-SAN は接続距離に限界がない，導入コストが低い，管理技術者が多いなどの利点を持つ反面，性能が FC-SAN より劣り性能向上が重要な課題となっている．性能向上が困難である理由として IP-SAN のプロトコルスタックの複雑さがあげられる．

この複雑さに対し我々は，多段構成である IP-SAN の全層を網羅的に解析でき，少数の I/O 要求の動作を詳細に追跡できる IP-SAN トレースシステムを提案した[1]．これにより個々の I/O 要求が各層を通過する時刻などを観察することが可能となり，I/O 要求が処理されずに待たされている層の発見などが可能となった．しかし，当該システムは少数の I/O 要求が処理される小規模 IP-SAN システムにおいて各 I/O 要求の動作の詳細を観察するのには適しているが，多数(100 以上など)の I/O 要求が処理される IP-SAN システムにおいては，多数の I/O 要求の動作を詳細に考察することが必要となり，解析が困難であるという問題があった．

本稿では，複数のイニシータにより構成され，多数の I/O 要求が並列に発行される非常に複雑な動作をする IP-SAN システムに対して，トレース結果を集計し階層別の並列動作数を算出し，ボトルネック部を抽出する手法を提案する．

2．IP-SAN 解析システム

iSCSI IP-SAN における I/O 処理は，ファイルシステム/RAW デバイス over SCSI over iSCSI over TCP/IP over Ethernet という多段階層の処理により構成される．我々は，オープンソース OS 実装(Linux 2.4.18)と iSCSI 実装(UNH-iSCSI 1.5.02)を用い IP-SAN システムを構築し，全ての層に動作履歴保存機能を追加した．履歴保存機能では，確保したメモリにカーネル内部で発生した I/O イベントとその時刻を保存する．イベントは上位層から下位層への I/O 要求の転送などである．

本機能を用い各層における発行済み I/O 要求数，終了(応答受信)済み I/O 要求数，処理中 I/O 要求数を求めることにより，各層における並列動作 I/O 要求数や並列度を制限している層を抽出することが可能となる．適用例は第 4 章に示す．

3．高負荷中規模 IP-SAN 性能評価

1 台の iSCSI ターゲットと 4 台のイニシータにより構成される中規模 IP-SAN において，各イニシータから複数の I/O 要求(システムコール“read()”)を並列に発行し，その性能を測定した．イニシータとターゲットは PC(CPU Intel Pentium4 1.5GHz，メモリ 384MB)と前章の実装で構築し，ネットワークには Gigabit Ethernet を用い NIC は Intel PRO/1000 XT Server Adapter を用いた．イニシータとターゲットの間に 16[ms]の遅延を人工的に追加し中規模 IP-SAN を模擬した．システムコールは RAW デバイスに対して発行した．測定結果を図 1 の“4init (default)”に示す．横軸は各イニシエータで並列に発行されるシステムコールの数であり，縦軸は 1 秒あたりに全イニシエータで処理されるシステムコール“read()”の合計である．図より 96 まではイニシータあたりの並列発行数を増やすとシステム全体の性能が増加し，96 以上では増加しないことが分かる．

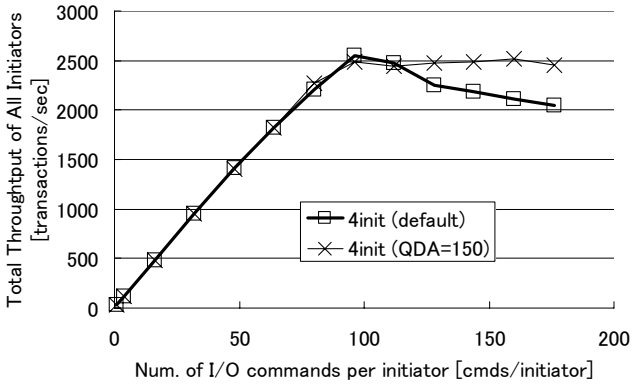

図 1 高負荷中規模 IP-SAN の性能

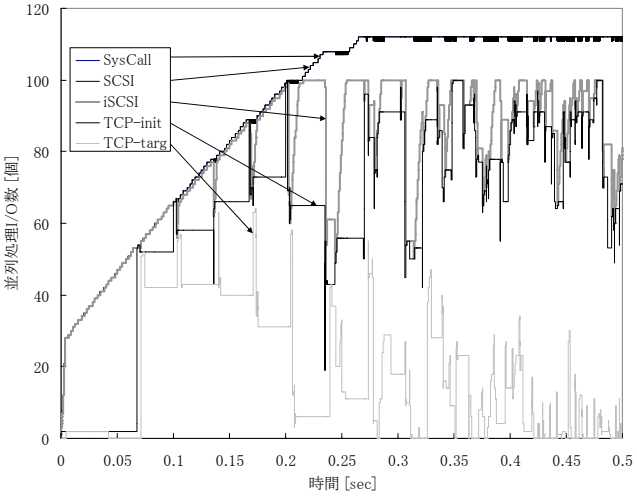

図 2 IP-SAN 各層での並列処理 I/O 数の推移

4．解析とボトルネックの抽出

図 1 の“4init (default)”の結果を第 2 章で提案した手法により解析し，動作の把握とボトルネック部の抽出が可能であることを示す．同結果では，並列度 100 程度において並列度の増加による性能の上昇がとまり，プロトコルスタッ

クのいずれかの層において並列度が制限されていると考えられる．並列発行システムコール数 112 にける，1 組の iSCSI 接続(1 組のイニシエータとターゲット)の各層で処理中 I/O 要求数の推移を算出し，図 2 を得た．処理中 I/O 要求数とは，既に発行されたがまだ終了していない I/O 要求の数である．これは，発行済み I/O 要求の総数と，終了済み I/O 要求数の総数の差により求めることができる．同図において，"SysCall"はシステムコール層，"SCSI"は SCSI 層，"TCP-init"はイニシエータ機の TCP 層，"TCP-targ"はターゲット機の TCP 層において集計した処理中 I/O 要求数を表している．図より，上位のシステムコール層，SCSI 層では並列度が 112 であるが，iSCSI 層で並列度が最大 100 に制限されており，それより下の層では並列度がさらに減少していることが分かる．よって，並列度の制限は iSCSI 層に存在することが分かる．

次に，並列度が制限されている iSCSI 層をさらに細分化し，iSCSI 層内の副層における並列度の推移を図 3 に示す．同図の“iSCSI(Queued)”は iSCSI のキューイング副層における並列度，すなわち iSCSI キューに入れられたが終了していない I/O 処理の数を表している．同様に“iSCSI(Thread)”は iSCSI 層において送信スレッドが起動されたが終了していない I/O 処理の数を表し，“iSCSI(Sock)”は iSCSI 層内においてソケットレベルでの送信処理(TCP 層に処理が渡される)が行われたが終了していない I/O 処理の数を表している．同図より，iSCSI 層における I/O キューイング副層における並列度は 112 まで上昇しているが，送信スレッド起動の副層における並列度が 100 に制限されていることが分かる．よって，並列度の制限はキューイング副層とスレッド起動副層の間に存在することが分かる．

両副層間に存在する送出数の制限は最大命令シーケンス番号(max_cmd_sn)の確認であり，これが十分で無かったために並列度が制限されていたと特定することができる．

本実装においてこの数は QUEUE_DEPTH_ALLOWED として定められており，初期値は 100 である．これを 150 として性能を計測し図 1 の“4init (QDA=150)”を得た．同グラフから，特定された並列度制限を解決することにより性能が向上されたこと(176 並列発行の例において 20%の向上)が確認され，本手法の有効性が確かめられた．

また，QUEUE_DEPTH_ALLOWED を 150 とした環境における各層の処理中 I/O 要求数の推移を図 4 に示す．より規模の大きい IP-SAN 環境を想定しイニシエータとターゲット間の遅延時間を 32[ms]とした．並列発行システムコール数は 160 である．図より，システムコール層および SCSI 層において並列に処理されている I/O 要求の数は 160 にまで至っていることが分かる．これに対し，iSCSI のキューイング副層における並列度は最大で 144 となっており，それ以下の層における並列度は 144 以下となっている．同様にして，ボトルネック部を抽出可能であることが確認できる．

5. おわりに

本稿では，IP-SAN のトレース解析システムのトレース結果を集計し，IP-SAN 各層において並列に処理される I/O 要求数の推移を求め，性能を制限しているボトルネック部を抽出する手法を提案した．また有効性の検証のために，4 台のイニシエータ機と 1 台のターゲット機から構成され，

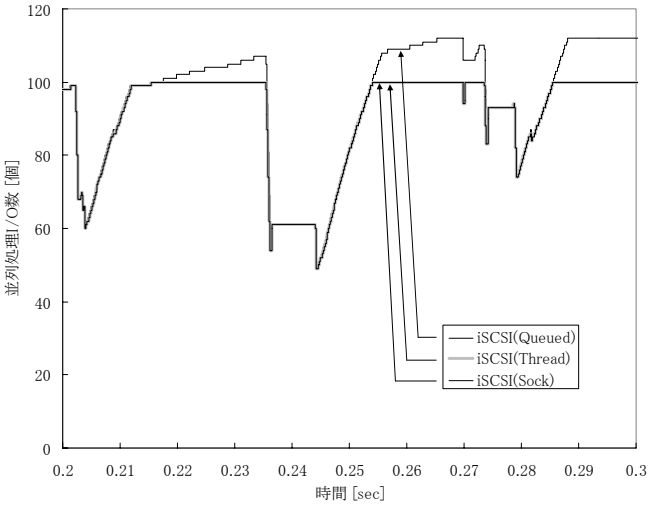

図 3 IP-SAN の iSCSI 層での並列処理 I/O 数の推移

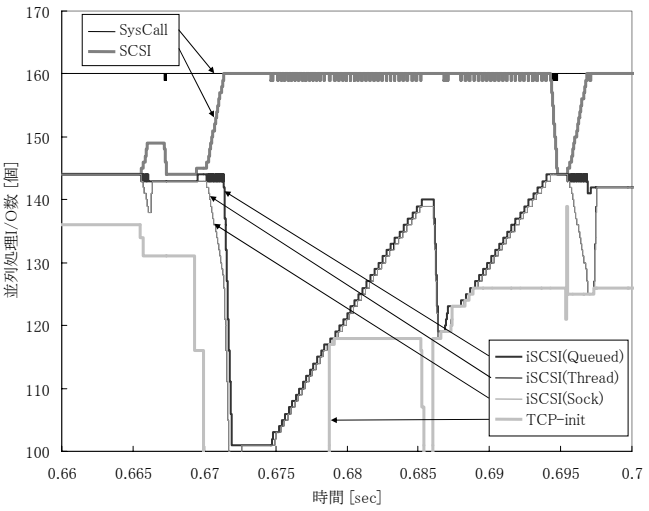

図 4 IP-SAN の各層での並列処理 I/O 数の推移(QDA=150)

数百個の I/O 要求が並列に処理される IP-SAN システムに対して提案システムを適用した．検証の結果，提案手法により当該システムのボトルネック部が抽出され，性能を制限している問題を解決することにより性能が改善されることが確認され，提案手法の有効性が確認された．

今後は，QUEUE_DEPTH_ALLOWED 増加後の性能改善手法の考察や，他の実装を用いての検証などを行っていく予定である．

参考文献

[1] 山口 実靖 小口 正人 喜連川 優, “IP ネットワークストレージシステムのトレース解析”, FIT 2004 第 3 回情報科学技術フォーラム 一般講演論文集 第 2 分冊　社団法人電子情報通信学会, pp. 41-42, 2004.8.20