D_016

複数イニシエータ接続 iSCSI ストレージの性能に関する考察

Performance Analysis of iSCSI Storage with Multiple Initiators

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

1. はじめに

近年，計算機システムが扱うデータ容量の増大にともないストレージ管理コストが増大し，これが計算機システムの大きな問題となっている．ストレージ管理コスト削減の方法として SAN(Storage Area Network)を用いてストレージを集約する方法があり，SAN の一つに iSCSI を用いる IP-SAN がある．

本稿では，IP-SAN を構成する各層にモニタ機能を追加し中規模の IP-SAN システムの動作解析する手法を提案し，これを用いて多数の I/O 要求が発生する高負荷状態の IP-SAN システムの性能に関する考察を行う．

2. 複数イニシエータ接続 IP-SAN 解析システム

提案システムの実装として 1 台の iSCSI ターゲットと最大 5 台の iSCSI イニシエータ群からなる IP-SAN システムを構築した．イニシエータ群とターゲットの間には FreeBSD Dummynet を用いて人工遅延装置を配置し，広域 IP-SAN システムを模擬した．イニシエータ群とターゲットは PC を用いて構築しその仕様は CPU Intel Pentium 4 1.5GHz，メインメモリ 375MB，NIC Intel PRO/1000 Server Adaptor，OS Linux 2.4.18 となっている．また，iSCSI 実装はニューハンプシャー大学 IoL が開発する iSCSI ドライバ 1.5.02 を用いた．

iSCSI を用いる IP-SAN では，アプリケーションから I/O 要求が発行されるとイニシエータ機上で(1)カーネルのシステムコール層，(2)ファイルシステム層とブロックデバイス層またはローデバイス層，(3)SCSI 層，(3)iSCSI 層，(4)TCP/IP 層，(5)Ethernet 層を経由し，ネットワークを越え，ターゲット機上で(6)Ethernet 層，(7)TCP/IP 層，(8)iSCSI 層，(9)SCSI 層を経由しストレージにアクセスする．また，その応答は上記の層を逆向きに経由しアプリケーションに伝えられる．我々はカーネルやデバイスドライバに対して各層を I/O 要求が通過するイベントを記録し，IP-SAN の振る舞いを追跡できる解析システムを構築した[1].

3. 高負荷時 IP-SAN 性能評価

前章の IP-SAN システムにおいて 1 台のターゲットに対して複数台のイニシエータから同時に多数の微少な I/O 要求を発行しその性能を計測した．具体的には各イニシエータ上で iSCSI 接続の raw デバイスに対して 512 バイトのシステムコール read()を連続して発行するベンチマークプロセスを多数並列に実行させた．ターゲットはファイルモードで動作させ要求データが常にメインメモリキャッシュ上にある状態で計測を行い，イニシエータ台数を 1 台から 5 台に，各イニシータ上のベンチマークプロセス数を 1 個から 512 個に変化させて計測を行った．また，イニシータ群とターゲットの間の遅延時間は 4 ミリ秒とした．

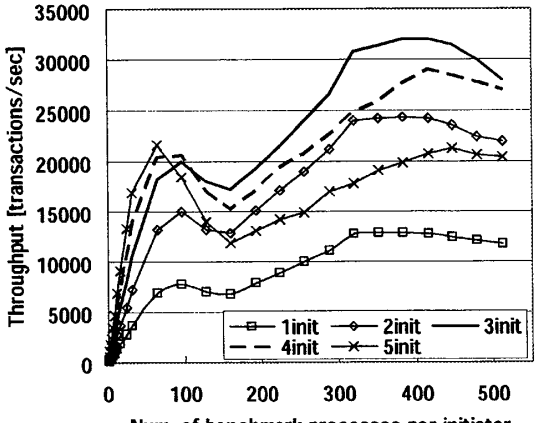

図 1：高負荷 IP-SAN システムの性能

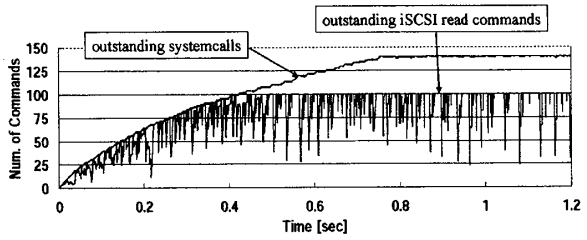

図 2：140 プロセス時の並列処理要求数

上記実験を行い，図 1 の結果を得た．同図の横軸が各イニシータ上で動作させたベンチマーク数であり，縦軸は全イニシエータの全プロセスの合計速度である (速度は単位時間あたりのシステムコール処理数)．同図より少数ベンチマーク時(1 個から 100 個程度)はプロセス数の増加に伴い全プロセス合計性能が増加するが，それを超える負荷(100 個から 140 個程度)を与えると合計性能が逆に減少することが確認された．また，イニシータ数 3 まではイニシータ数の増加に伴い全イニシエータの合計性能が増加するがこれを超えると逆に減少することも確認された．

4. IP-SAN システムの動作解析

提案解析システムを用い前章の実験結果のイニシエータ数 4，プロセス数 140 時の IP-SAN システムの振る舞いを観察し，図 2 を得た．同図は，処理中のシステムコールの数と，処理中の iSCSI 命令の数の時間変化を表している．前者は発行されたが終了していないシステムコールの数，後者は iSCSI 層により下位層(TCP/IP 層)に発行されたがまだ応答を受信していない iSCSI 命令の数を表している．140 プロセス並列に実行した場合，140 個のシステムコー

† 工学院大学
‡ お茶の水女子大学
§ 東京大学生産技術研究所

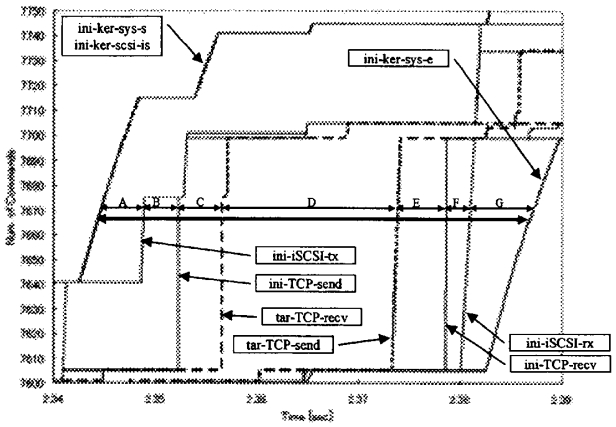

図3：4イニシエータ100プロセス時動作解析

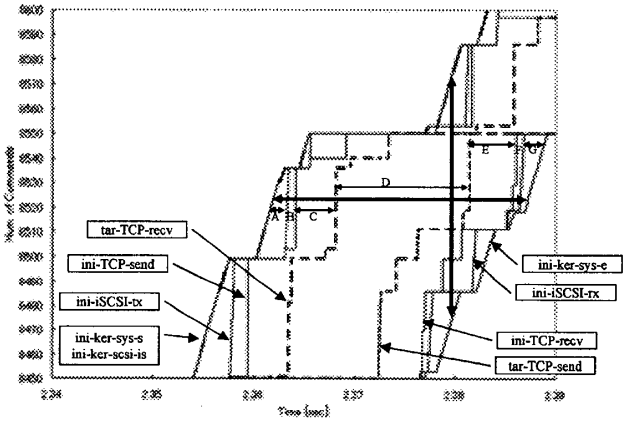

図4：4イニシエータ140プロセス時動作解析

ルが並列に発行されその応答を待っている状態になるがiSCSI層は最大で100個のiSCSI命令しか発行せず，これらの間で並列性が100に制限されていることが確認できる．

次に，イニシエータ数4，プロセス数100，140時における IP-SAN の振る舞いを比較しプロセス数の増加に伴い合計性能が低下する原因を考察する．IP-SAN の各層を通過する I/O 要求の中から特定のイニシエータに関するもののみを抽出し，各層を通過した I/O 要求の累積数の時間変化を図3，図4に示す．図3が100プロセス時，図4が140プロセス時の累積通過要求数である．図内の ini-ker-sys-s がイニシエータカーネルにおけるシステムコールの発行，ini-ker-scsi-is がイニシエータカーネルにおける SCSI 要求の発行，init-iSCSI-tx が iSCSI ドライバにおける iSCSI 命令 PDU の送出，ini-TCP-send がイニシエータ TCP における iSCSI 命令 PDU パケットの送出，tar-TCP-recv がターゲット TCP における iSCSI 命令 PDU の受信，tar-TCP-send がターゲット TCP における iSCSI 応答 PDU の送信，ini-TCP-recv がイニシエータ TCP における iSCSI 応答 PDU の受信，ini-iSCSI-rx が iSCSI ドライバにおける iSCSI 応答 PDU の受信，ini-ker-sys-e がイニシエータカーネルにおけるシステムコールの終了を表している．ただし，システムコールの発行と SCSI 命令の発行はほぼ同時であり図内では重なって表現されている．横軸が時刻であり，縦軸は各層を通過した I/O 要求と応答の累積である．

図内に太線で記した水平の両方向矢印の区間があるシステムコールの発行時刻から終了時刻までを表しており，区間 A が SCSI 層による SCSI 命令の発行と iSCSI 層による iSCSI 命令 PDU の送出の間に要した時間であり，以下同様に区間 B が iSCSI PDU の発行と TCP 層のパケットの送出の間の時間，C がパケットのネットワーク転送時間，D がターゲットの処理時間，E がパケットのネットワーク転送時間，F が TCP 層のパケット受信と iSCSI 層の PDU 受信の間の時間，G が iSCSI 層の応答 PDU の受信とシステムコールの終了の間の時間となる．

また，図3に太線で記した垂直の両方向矢印の区間は発行されたシステムコール数と終了したシステムコール数の差であるため，現在処理中のシステムコールの数を表しており，以下同様に各階層間の差がその階層間で処理中の I/O 要求数を表している．例えば線 ini-TCP-send と線 tar-TCP-recv の差は，ネットワーク転送中の I/O 要求の数を表現している．

図3，図4を比較し，各システムコールの所要時間が大きく増加していることが確認できる(両図の水平矢印部の例では25ミリ秒と42ミリ秒)．各層ごとの処理時間では，ネットワーク転送時間(両図の C 部と E 部)はほぼ同程度であり，ターゲット機の処理時間(両図の D 部)は増加しているがシステムコール所要時間の増加と比べると小さいことが分かる．これに対して両図の A 部や G 部が大幅に増加しており，主な性能劣化要因は負荷の増大による計算機の I/O 処理時間の増大であるこが確認でき，特に iSCSI イニシエータ層における処理時間の増大の影響が大きいことが分かる．また，本実験は片道4ミリ秒の擬似的な中規模ネットワーク環境で行ったが，ネットワーク遅延時間が全処理時間内でしめる割合が小さいことも確認できる．この様に提案解析システムを用いることで複数の計算機で構成され多数の I/O 要求が発行される複雑な IP-SAN システムの振る舞いを視覚的に観察可能となり，性能の考察を行うのに有用であることが確認された．

5. おわりに

本稿ではストレージ機器が数台のサーバ計算機に接続されている中規模の IP-SAN とその解析システムを構築し，高負荷時の IP-SAN システムの性能について考察した．IP-SAN システムを提案解析システムにより観察した結果，IP-SAN を構成する各層の処理時間や各層内で処理中の I/O 要求の量などが視覚的に観察可能となり，IP-SAN の振る舞いを把握するのに有用であることが確認された．本実験環境の例においては高負荷時の IP-SAN システムの主な性能劣化原因は，負荷が集中するストレージ機器の性能の低下やサーバ計算機とストレージ機器を結ぶネットワークの転送速度の低下ではなく，イニシエータ計算機の処理時間の増大であることなどが確認された．

今後は，各層間の処理時間や各層内で処理中である要求数の平均や変動などの統計情報を求め，さらなる考察を行う予定である．

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,”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，pp. 41-42，September 2004.