高遅延環境における iSCSI シーケンシャルリードアクセス時の TCP 輻輳ウィンドウ制御

豊田 真智子 † 山口 実靖 ‡ 小口 正人 †
† お茶の水女子大学 ‡ 東京大学 生産技術研究所

1. はじめに

大容量のデータを高速に処理するネットワークとして SAN（Storege Area Network）が登場し，その実績は高い評価を得ている．その代表的な技術である FC-SAN はファイバチャネルを用いて構築されるため，導入や管理のコストが問題となり，Ethernet と TCP/IP を用いて構築する IP-SAN に注目が集まっている．IP-SAN の代表的プロトコルである iSCSI は，SCSI コマンドをカプセル化して TCP/IP ネットワーク上に転送する技術であり，サーバ（Initiator）とストレージ（Target）間をシームレスに接続することができる．しかし，iSCSI は複雑なプロトコル構造を持つため，TCP/IP のみで構築されたネットワークより性能が劣化することがわかっており，システム性能は，TCP パラメータである輻輳ウィンドウと関連性があることが確認されている．

そこで本稿では，文献 [1] で提案した輻輳ウィンドウコントロール手法を，高遅延環境における iSCSI シーケンシャルリードアクセスに適用した．輻輳ウィンドウをコントロールしないストレージアクセスでは，輻輳ウィンドウの低下と共にスループットが低下していたが，輻輳ウィンドウをコントロールし，その値が一定値となった後にはスループットも一定値となり，コントロール前より高い性能を示すことが確認された．

2. 輻輳ウィンドウコントロール手法

スループットのばらつきを抑制するために提案した輻輳ウィンドウコントロール手法の概念図を図 1 に示す．本手法は，TCP ソースコードに独自の関数を挿入することで TCP パラメータをモニタすることができる仕組みを Target に実装し，Target からの輻輳ウィンドウ通知を受けて，アプリケーションがストレージアクセスのブロックサイズを調節する仕組みをミドルウェア機能として提供するものである．Linux TCP 実装における輻輳ウィンドウの変化は，一定値となるか，増加後急激に低下するという変化を繰り返すかのどちらかである．また輻輳ウィンドウが低下する原因としては，実験環境に依存するエラーを検出した場合と，ネットワーク状態に依存するエラーを検出した場合に分けることができる．本手法は，実験環境に依存するエラーである送信側のデバイスドライバのバッファが溢れることによる CWR エラーを検出して，輻輳ウィンドウが低下した場合に適用するものとする．実装したコントロール手順は以下の記述に基づくものである．

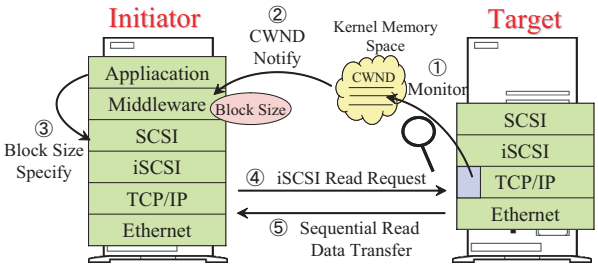

図 1: 輻輳ウィンドウコントロール手法の概念図

1. Target で輻輳ウィンドウをモニタし，変化を観察する
2. 観察中に CWR が検出され，輻輳ウィンドウが低下した場合にはその時の輻輳ウィンドウ値を，輻輳ウィンドウが一定値であると判断した場合には輻輳ウィンドウの限界値を Initiator に通知し，Target でも通知した輻輳ウィンドウ値を記録する
3. 通知を受けた Initiator では，ミドルウェアが輻輳ウィンドウからブロックサイズを決定し，アプリケーションがブロックサイズを再指定する
4. Initaitor から Target にシーケンシャルリードコマンドを送信し，ストレージアクセスを行う
5. Target が Initiator に向けて要求されたブロックサイズのデータ転送を実行する
6. CWR を検出するか，一定値であると判断する度にこの処理を繰り返す

本手法適用後，輻輳ウィンドウは限界値で一定に保たれ，その時のブロックサイズが最適値となる．なお，本手法においてミドルウェアが指定するブロックサイズは以下の式を用いて計算した．

転送ブロックサイズ $[byte]$ = 輻輳ウィンドウ値×最大転送単位（MTU）

本実験時の MTU（Maximum Transmission Unit）は Ethernet の最大セグメント長（1500KB）から TCP/IP ヘッダ（オプションを含む）を除いた 1448KB である．

3. 輻輳ウィンドウコントロール手法を用いた性能測定実験

iSCSI ストレージアクセス時に，輻輳ウィンドウコントロール手法を用いてアクセスブロックサイズを調節した場合の性能評価を行う．Initiator から Target にシーケンシャルリードアクセスを行い，その時の輻輳ウィンドウ，ブロックサイズ，スループットを測定した．この時，最初にアクセスするブロックサイズの初期値を 1024KB に設定して実験を行った．

Controlling TCP Congestion Window on iSCSI Sequential Read Access in a Long-Latency Environment
† Machiko Toyoda, Masato Oguchi
‡ Saneyasu Yamaguchi
Ochanomizu University (†)
Institute of Industrial Science, The University of Tokyo (‡)

表 1: 使用計算機

CPU	Intel Xeon 2.4GHz
Main Memory	512MB DDR SDRAM
OS	Initiator, Target ： Linux2.4.18-3 Dummynet ： FreeBSD 4.9 - RELEASE
NIC	Initiaotr, Target : Intel PRO/1000XT Server Adapter Dummynet : Intel PRO/1000MT Server Adapter

3.1 実験環境

本実験は以下の環境で行った．Initiator と Target 間は Gigabit Ethernet で接続し，TCP/IP 接続を確立した．遠隔ストレージアクセスを想定した実験を行うため，Ethernet の接続途中に人工的な遅延装置として FreeBSD Dummynet[2] を挟み，片道遅延時間を“ 16ms ”，往復“ 32ms ”に設定し，高遅延環境を構築した．Initiator，Target，Dummynet はすべて PC 上に構築し，Initiator，Target には Linux を，Dummynet には FreeBSD をインストールした．iSCSI ネットワークにおける性能を調べるため，Target はメモリモードで動作させ，ディスクアクセスを伴わないように設定した．実験で使用した計算機の環境を表 1 に示す．

また，本実験で用いた iSCSI 実装において，Target にはニューハンプシャー大学 InterOperability Lab が提供する UNH IOL reference implementation ver.3 on iSCSI Draft 18[3] を用いた．しかし，UNH 実装では大きなブロックサイズで read コマンドを発行することができないため，性能測定への影響を考慮し，Initiator には UNH 実装の Initiator と同等の機能を持ち，かつ大きなブロックサイズのデータ転送を行う自作 Initiator を用いた．この自作 Initiator は通常のユーザ空間のアプリケーションとして動作し，iSCSI Target と TCP/IP コネクションを確立して iSCSI プロトコルで通信を行うものである．

3.2 実験結果

前節の実験環境を用いて行った実験の結果として，図 2，図 3 を得た．図 2 は輻輳ウィンドウコントロール手法を用いない場合のストレージアクセスの結果，図 3 は輻輳ウィンドウコントロール手法を用いた場合のストレージアクセスの結果である．輻輳ウィンドウをコントロールしない場合は，輻輳ウィンドウの変化に伴い，スループットも増加，低下の変化を繰り返す．一方輻輳ウィンドウをコントロールした場合は，CWR エラーの検出により Initiator のミドルウェアが機能し，アクセスブロックサイズがやや低い値に設定されるが，輻輳ウィンドウが一定値となった時にはスループットも一定値となり，鋸型に変化していたときよりも高い性能を保ったままストレージアクセスを行うことが確認される．

3.3 考察

ブロックサイズは一度に送信するデータ長であるため，この値を増加させることで送信データ量が増加し，スループットは向上する．しかし，あまりブロックサイズを大きくしすぎるとデータ送信側の送信バッファが溢れ，CWR エラーが生じる．この時 TCP 実装により，一度に送信するデータ量が多すぎたと判断され，送信パケットを減らすために輻輳ウィンドウが大幅に減少する．

遠隔ストレージアクセスなどの高遅延環境においては，パケットを送信してから確認応答パケットが返信されるまでの時間が長くなるため，送信側では何もせずに待っ

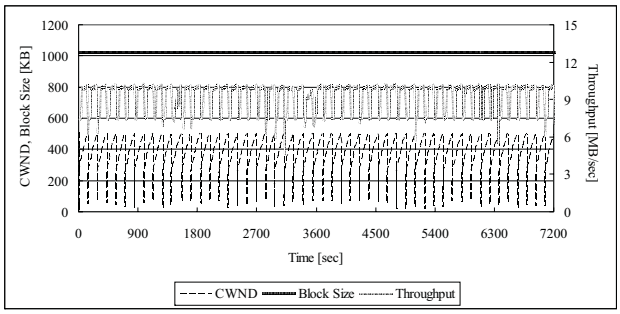

図 2: 輻輳ウィンドウコントロール手法を用いない場合の輻輳ウィンドウ，ブロックサイズ，スループットの時間変化

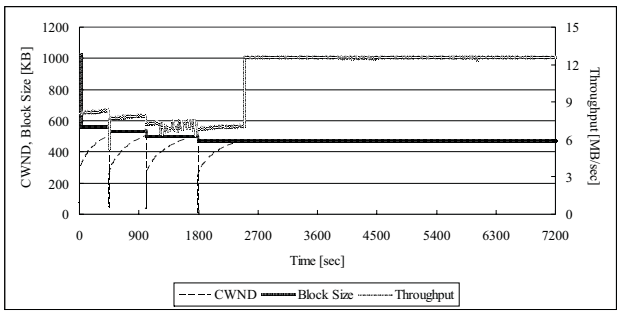

図 3: 輻輳ウィンドウコントロール手法を用いた場合の輻輳ウィンドウ，ブロックサイズ，スループットの時間変化

ているだけの無駄な時間が生じる．そのため，確認応答を待たずに送信できるパケット数を増やすことが待ち時間を短縮し，結果として性能を向上させる．輻輳ウィンドウコントロール手法を用いることで輻輳ウィンドウは最終的に一定となるため，送信パケット数が一定となり，待ち時間を減らすことができる．図 2，図 3 の結果から，このような制御を用いた本手法が有効なものであることが確認された．以上のことから，高遅延環境においてはブロックサイズを大きくするより輻輳ウィンドウを一定値に保った方が性能が向上すると言える．

4. まとめ

iSCSI ストレージアクセスにおいて，輻輳ウィンドウコントロール手法を高遅延環境下に適用し，アクセスブロックサイズを変更させてスループットを向上する手法を提案した．今後は遅延時間が異なる環境の iSCSI ストレージアクセスにおいて本手法を適用し，詳細な性能評価を行いたい．

参考文献

[1] 豊田真智子，山口実靖，小口正人：“ iSCSI アクセス時の TCP 輻輳ウィンドウ制御を用いたシステム性能向上手法の一検討 ”，電子情報通信学会技術研究報告，CPSY2004-50，pp.1〜6，December 2004.

[2] L.Rizzo: “ dummynet ”, http://info.iet.unipi.it/~luigi/ip_dummynet/

[3] InterOperability Lab: Univ, of New Hampshire, http://www.iol.unh.edu/consortiums/iscsi/