情報処理学会論文誌 Vol. 49 No. 10 3645–3656 (Oct. 2008)

推薦論文

VPN 複数経路接続時における iSCSI ストレージアクセスの性能評価

千島 望[†1] 山口実靖[†2] 小口正人[†1]

近年，ストレージの管理コスト低減などの目的で SAN の導入が進んでおり，IP ネットワークを利用した IP-SAN として iSCSI が期待されている．遠隔バックアップなどを実現するために iSCSI を広域ネットワークに適用することを考えた場合，VPN を利用する場合が多い．そこで本研究では，ネットワークストレージの性能と信頼性を向上させることを目的として，iSCSI 複数コネクション設定と VPN マルチルーティング機能を用いて VPN 広域ネットワーク内を複数経路で接続し，iSCSI ストレージアクセスの特性を解析して性能評価を行った．各経路の遅延時間を変化させた際のスループットや TCP 輻輳ウィンドウの変化などを測定した結果，VPN 複数経路上の各 iSCSI コネクションがどのように振る舞い，それが性能にどのような影響を与えているのか明らかにすることができた．

Performance Evaluation of iSCSI Storage Access through Multi-routing VPN

Nozomi Chishima,[†1] Saneyasu Yamaguchi[†2] and Masato Oguchi[†1]

Recently, SAN is introduced for the purpose of the storage management cost reduction, and iSCSI is expected as IP-SAN that uses IP network. SAN is expected for the realization of remote backup of data, and VPN can adapt iSCSI to a wide area network. In order to improve performance and reliability of network storage, multi-routing iSCSI access is realized using a multi-routing function of a VPN router. We have observed the feature of iSCSI storage access and evaluated its performance. According to the evaluation of throughput and TCP congestion window when a delay time is changed at each VPN connection, the behavior of each iSCSI connection and its impact on performance are clarified.

1. はじめに

近年，インターネット技術の進展やマルチメディアアプリケーションの普及などにより，ユーザが蓄積し利用するデータ容量が爆発的に増加している．これにともないストレージの増設，管理コストの増大が問題となっている．そこで SAN（Storage Area Network）が登場し，広く用いられるようになった．SAN はサーバとストレージを物理的に切り離し，各ストレージとサーバ間を相互接続してネットワーク化したもので，これにより各サーバにばらばらに分散していたデータの集中管理が実現された．

一般に SAN としてはファイバチャネルを用いる FC-SAN（Fibre Channel - SAN）が広く利用されている．しかし FC-SAN はファイバチャネルを用いているため高価となり，また接続距離に制約がある．一方，SAN に IP ネットワークを利用した IP-SAN として iSCSI が期待されている[1),2)]．iSCSI は，これまで DAS（Direct Attached Storage）で使われてきた SCSI コマンドを TCP/IP パケット内にカプセル化することにより，サーバ（Initiator）とストレージ（Target）間でデータの転送を行う．今後インターネット技術の発展により，ギガビットクラスの回線の実現が期待され，iSCSI の有用性もさらに高まると考えられる．

現状において，SAN は主にサーバサイト内のみでしか使用されていない．しかし遠隔バックアップなどを目的として，離れたサイトのサーバとストレージを SAN で接続することが強く期待されている．そこで本稿では，VPN（Virtual Private Network）を利用することにより，ローカル環境で使用されている iSCSI を用いて広域ネットワーク上でリモートアクセスを行うことを検討した．さらに，性能と信頼性がより高い通信を実現するため，VPN 広域ネットワーク内に複数経路を構築した．

iSCSI は複雑な階層構成のプロトコルスタックで処理されており，バースト的なデータ転送も多いことから，通常のソケット通信と比較して，特に高遅延環境においては性能の劣化が著しく，下位基盤の TCP/IP 層が提供できる限界性能を超えることはできない[3)]．さらに広域環境で iSCSI 複数経路アクセスを行う場合，経路により遅延時間やネットワーク性能

†1 お茶の水女子大学
Ochanomizu University
†2 工学院大学
Kogakuin University
本論文の内容は 2007 年 7 月のマルチメディア，分散，協調とモバイル（DICOMO2007）シンポジウムにて報告され，DSM 研究会主査により情報処理学会論文誌ジャーナルへの掲載が推薦された論文である．

が異なるため，iSCSI で適応的なパケット処理を行うことが望ましい．そこで本稿では，そのような iSCSI 複数経路アクセスの特性を調べるため，VPN 複数経路接続において異なる遅延時間を持つ経路を構築して実験を行い，iSCSI ストレージアクセスの性能評価を行う．

ここで信頼性に関しては，複数経路を用いて同一データを送ることにより耐障害性向上などが期待できるが，本稿においては，その基盤である VPN 複数経路接続を構築し，データを複数経路に交互に送るデータ通信を行った．

本稿の構成は以下のとおりである．2 章で VPN 複数経路接続 iSCSI ストレージアクセスについて述べ，3 章で VPN のマルチルーティング機能を利用した本実験システムの概要を述べる．4 章で経路ごとの iSCSI アクセスについての性能評価結果を示し，5 章で各経路の遅延時間による影響を比較する．6 章では iSCSI 複数経路アクセスのモデル化と性能評価結果の解析を行い，最後に 7 章でまとめる．

2. VPN 複数経路接続 iSCSI ストレージアクセス

2.1 iSCSI 複数コネクション

IP-SAN の代表的なプロトコルに iSCSI がある．iSCSI は SCSI コマンドを TCP/IP パケットでカプセル化する規格で，iSCSI により SAN を IP 機器だけで構築することが可能となる．一方で**図 1** のように複雑な階層構成をとることになり，下位のプロトコルの限界性能を超えることはできない．また，iSCSI には長距離アクセスの実現が期待されているが，ギガビットクラスの太い回線を用いた場合の遅延帯域積の問題も指摘されている．そこで iSCSI における性能や信頼性を向上させる手法の実現が求められている．

iSCSI は様々なチューニングを行うことが可能である．本実験で用いたニューハンプシャー大学が提供する UNH-iSCSI の実装では，1 つの iSCSI セッション内に複数の TCP コネクションを確立するように設定することができる[4]．さらにこのコネクションをポート番号と対応付けることができる．つまり，**図 2** に示すようにターゲットの 1 つの IP アドレス，1 つの iSCSI ドライブにポート番号の異なる複数のコネクションを接続することが可能である．本研究では iSCSI のこの特性を利用して，iSCSI 複数経路アクセスを実現した．

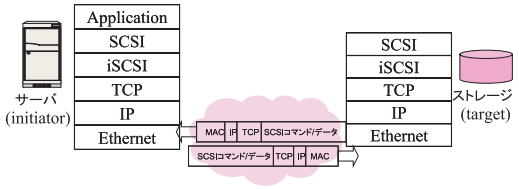


図 1　iSCSI
Fig. 1　iSCSI.

2.2 VPN 複数経路接続

VPN は，インターネットや通信事業者が持つ公衆ネットワークを使って，拠点間を仮想的に閉じたネットワークで接続する技術である．安価であるという公衆網のメリットを活かしつつ，機密性の低さを暗号化などの別の方法で補うことにより，「実質的な専用網」を実現できるということが VPN の利点である．一方，専用網と異なりネットワークの品質は保証されない場合が多い．

本稿では非常災害対策などを目的とした iSCSI による遠隔バックアップなどの評価を行うため，**図 3** に示すように VPN ルータで接続したリモート環境にネットワークストレージを設置し，iSCSI を広域ネットワーク環境に適用することを想定して実験を行った．この場合，広域ネットワーク内の VPN 越しにアクセスを行うため，VPN ルータを通ることによってネットワークの帯域幅が制限され，スループットが著しく低下することが起こりう

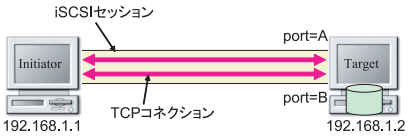


図 2　iSCSI 複数コネクション
Fig. 2　iSCSI multiple connections.

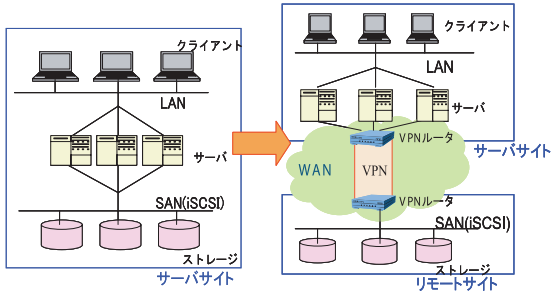


図 3　VPN 利用モデル
Fig. 3　Utilization model of VPN.

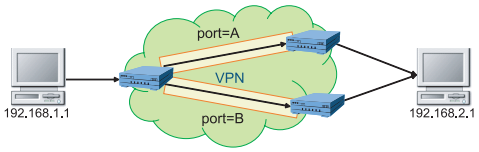


図 4　VPN マルチルーティング機能
Fig. 4　VPN multi-routing funciton.

る[5]．さらに広域ネットワーク内は不安定な通信路であることが想定される．そこで本稿では，VPN 広域ネットワーク内を複数経路で接続することを考えた．これにより，データ転送の性能や信頼性，ネットワークの耐障害性なども向上すると考えられる．

ただし iSCSI 複数経路の構築は，アプリケーションなど上位層に対しては透過的に実現したい．本実験で用いた VPN ルータ Fujitsu Si-R570 はマルチルーティング機能を有している[6]．マルチルーティング機能を使用すると，**図 4** に示すようにポート番号などの情報を利用して同じあて先 IP アドレスを持つネットワークへ複数の経路を用いて送信することが可能となる．それぞれの通信内容に応じて通信経路を分離することができるため，片方の回線をバックアップ用に用いたり，音声データは専用線を用いそのほかの通信は公衆網を用いたりするなどと設定することができる．本研究ではこの機能を利用し，iSCSI 複数コネクション設定と対応付けることにより，コネクションごとに異なる経路を構築することを可能にした．また本研究で用いた VPN ルータは L3 における VPN を利用できる IP アクセスルータで，VPN の暗号化プロトコルには IPsec の DES，3DES，AES を用いており，3DES 時には 500 Mbps の暗号化速度を実現する．しかし，本研究の手法はこの環境に限らず，複数経路接続の設定ができるルータを用いて構築した VPN であれば適用することができる．

iSCSI は通常ギガビットクラス以上の太いネットワーク上で用いられるが，途中に VPN ルータの暗号化処理速度などによりスループットが決まる細い回線が挟まることにより，トラフィックとして大いに性質の異なるものになる[7]．この場合，iSCSI の性能は下位層である TCP の振舞いにより決まってくるため，TCP パラメータなどを観測して iSCSI の性能との関連性を明らかにする必要がある．

2.3　TCP 輻輳ウィンドウ

TCP では，通信能力の制御にウィンドウサイズという概念を用いている．ウィンドウサイズとは，ホストが確認応答パケット（Acknowledgement: ACK）なしに 1 度に送信できるデータの量である．また，データの送信側では輻輳ウィンドウ，受信側では広告ウィンドウという値が決定され，このどちらか小さい方がウィンドウサイズとして用いられる．広告ウィンドウは現在の受信バッファの空き容量を示しており，ACK のヘッダにその情報が含められて送信側に送られる．一方，輻輳ウィンドウは送信側の制御パラメータで，ネットワークの混雑を回避するため送信側が自主的に制限する値である．

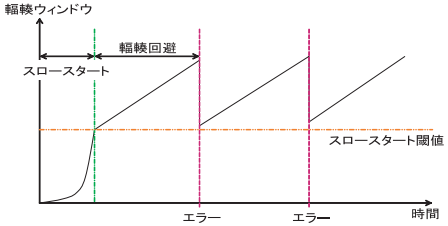


図 5　輻輳ウィンドウの変化
Fig. 5　Behavior of congestion window.

輻輳制御ではこの輻輳ウィンドウが利用されている．輻輳制御はネットワークの混雑解消の方法として TCP が行う機能である．通信開始時にはスロースタートと呼ばれるアルゴリズムに従って指数関数的に輻輳ウィンドウが大きくなる．これによりトラフィックが急激に増加するので，ネットワークが輻輳状態になる可能性がある．これを防ぐため，スロースタート閾値という値を用意し，輻輳ウィンドウがその大きさを超えると輻輳回避と呼ばれるフェーズに入り，一次関数的な増え方となる．そしてエラーが検出されると輻輳ウィンドウは急激に低下し，通常これらを繰り返すことで鋸型のグラフとなる．この様子を**図 5** に示す．

また本実験で用いた LinuxOS における TCP の状態遷移を**図 6** に示す．LinuxTCP においては，通信時の状態が正常であれば ACK の受信ごとに輻輳ウィンドウは増加するが，エラーが検出されると異常と判断され，輻輳ウィンドウは低下する．輻輳ウィンドウが低下する原因としては，送信側デバイスドライバのバッファが溢れることによる Local Congestion エラーを検出した場合（CWR），重複 ACK または SACK を受信した場合（Recovery），タイムアウトを検出した場合（Loss）の 3 つがあげられる．さらに Linux の TCP 実装では，通信中に 1 度設定された輻輳ウィンドウは，そのウィンドウ値を超えるデータ量が送られない限りは変化しないという特徴を持ち，このときスループットはほぼ一定の値で安定することが確認されている．

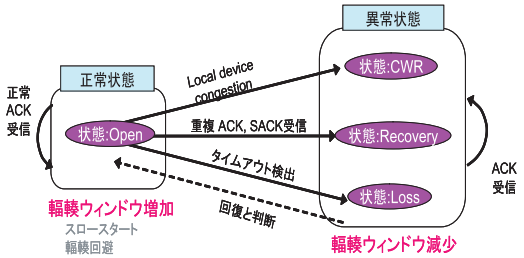


図 6　LinuxTCP の状態遷移
Fig. 6　State transition of LinuxTCP implementation.

2.4　本研究の評価内容と関連研究

本研究では以上に紹介した知識と技術を利用し，VPN 複数経路接続時における iSCSI ストレージアクセスの性能評価を行う．関連する研究としては，以下のようなものがあげられる．

VPN を用いない複数経路通信の研究としては，TCP を複数経路対応に拡張することで，エンド・エンド間において，上位層・下位層に対して透過的に複数経路通信を行う M/TCP がある[8]．M/TCP の場合は TCP 層を変更する必要があるが，本研究における VPN 複数経路接続においては，上位層が複数の TCP コネクションを束ねて利用できるアプリケーションの場合，両端の通信ノードのプロトコルスタックを変えることなく，ポート番号などの値をもとに複数経路を実現することができる．

iSCSI と TCP の関係を評価した研究としては，iSCSI のソフトウェア実装を使用した場合と，TCP の性能を向上させる TOE（TCP Offload Engine）や HBA（Host Bus Adaptor）を使用した場合との比較などが行われている[9]．またファイル操作やベンチマークプログラム実行時における iSCSI と NFS との比較も報告されている[10]．一方 iSCSI 層と TCP 層の間の処理を最適化することにより，特別なハードウェアを使用せず iSCSI プロトコル処理スループットの性能を向上させる手法が提案されている[11]．iSCSI の Initiator や Target の実装について，詳細に解析を行った結果なども報告されている[12),13]．

我々は，これまでに iSCSI を用いたアプリケーション実行性能と TCP パラメータの相関関係の評価を行った[14]．その結果，広告ウィンドウの値を制限することで，輻輳ウィンドウの値も制限でき，それによって実行性能にも影響が出ることが確認された．また，VPN 利用時のネットワークや，iSCSI ストレージアクセスによる性能測定と TCP 輻輳ウィンドウの振舞いを観察した[5),7]．

iSCSI ストレージアクセスにおいて TCP 輻輳ウィンドウを制御する研究としては，輻輳ウィンドウ値を動的にコントロールする手法がある[15]．この手法は，まず Target の OS のカーネルに輻輳ウィンドウモニタ関数を挿入し，これによりモニタした輻輳ウィンドウの変化を観察して，Initiator にその値を通知する．通知を受けた Initiator は輻輳ウィンドウの値に基づきブロックサイズを再指定して，シーケンシャルリードアクセスを行うというものである．この手法を適用し輻輳ウィンドウを限界値で一定に保った場合には，高遅延環境において最大 28%のスループットの向上が確認されている．またトランスポート層を標準的な TCP Reno から他の TCP に置き換えて iSCSI に適用した際のシミュレーション結果なども報告されている[16]．

一方 iSCSI 複数コネクションに関する研究として，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のプロトコルチューニングの検討が行われ，その有効性が確認されている[17]．また TCP 複数コネクションにおいて，各々の TCP 制御情報をコネクション間で共有することにより複数コネクション間の公平性を保つ Fair-TCP を，iSCSI 複数コネクションに適用した結果も報告されており，情報を共有することによって複数コネクションにおける制御がより適切なものとなり，性能向上が見られた[18]．しかし iSCSI 複数コネクションを複数の経路に割り当てて利用する手法に関しては，これまで検討が行われてこなかった．

本研究では，iSCSI 複数コネクションを VPN 複数経路に乗せてアクセスする手法の評価を行った．このような組合せを実現することにより，性能向上だけでなく信頼性の向上も期待できる．また広域環境で iSCSI 複数経路アクセスを行う場合，経路によりネットワーク遅延やネットワーク性能が異なるため，iSCSI で適応的な対処を行うことが望ましい．そこで本研究では，iSCSI 複数経路アクセスの特性を明らかにするために，VPN 複数経路接続において異なる遅延時間を持つ経路を構築し，iSCSI アクセスの性能評価を行った．

3.　実験システム

3.1　TCP 輻輳ウィンドウモニタツール

本実験では，TCP 輻輳ウィンドウをモニタするツールを構築した．**図 7** に示すように，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した．ここでモニタできるようになったものには，輻輳ウィンドウの値のほか，各種エラーイベントの発生（Local device congestion，重複 ACK，SACK 受信，タイムアウト検出）などがある．ま

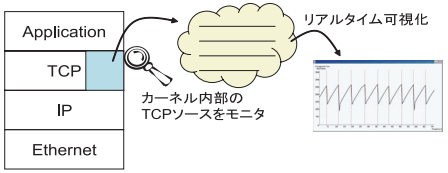


図 7　TCP 輻輳ウィンドウモニタツール
Fig. 7　TCP congestion window monitor tool.

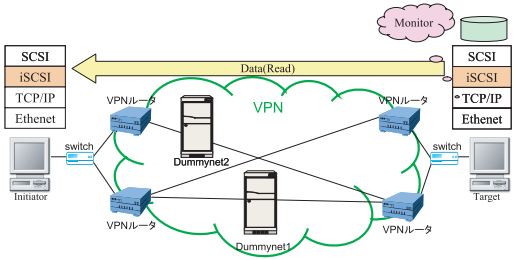


図 8　実験システムの概要
Fig. 8　An overview of experimental system.

た X11 ウィンドウシステムライブラリ関数を用いて，モニタした値をリアルタイムに可視化することもできる．このツールを用い，TCP 輻輳ウィンドウの振舞いなどを解析することにより，iSCSI 複数経路アクセスの評価を行った．

3.2　VPN 複数経路アクセス制御システム

本実験では，VPN ルータを用いて複数経路を構築し，経路ごとに異なる遅延時間を設定し iSCSI ストレージアクセスを実行したときの，性能と輻輳ウィンドウを評価するために**図 8** に示す実験環境を構築した．

iSCSI ストレージアクセスを行う Initiator とストレージを提供する Target の間に VPN ルータを 4 台を挟み，複数経路アクセスが実行できるように構築した．さらにそれぞれの経路に，遠距離アクセスを想定して人工的な遅延装置である FreeBSD Dummynet を挿入した[19)]．Dummynet では，パラメータとして片道遅延時間を設定する．そこで本実験においては，国内程度の距離への遠隔バックアップを想定し，0 msec，1 msec，2 msec，4 msec，8 msec，16 msec，32 msec の 7 つのパラメータを設定して測定した．

この実験システムにおいて iSCSI の複数コネクション設定と VPN ルータのマルチルーティング機能を用いて通信制御を行った．以下では iSCSI リードアクセス，すなわち Target から Initiator へデータが転送される場合について説明しているが，逆も基本的に同じである．まず iSCSI 単数コネクション VPN 単数経路通信の場合は**図 9** のような経路を通る．また iSCSI 複数コネクション VPN 単数経路通信の場合は，**図 10** に示すとおり同一経路上を 2 つのコネクションが張れるように iSCSI を設定した．さらに iSCSI 複数コネクション VPN 複数経路通信の場合は**図 11** に示すとおりコネクションごとに経路が異なるように VPN ルータの設定を行った．このとき Target から送られるパケットは，図 11 のように右

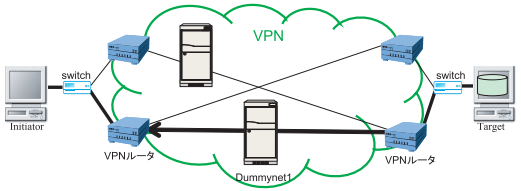


図 9　iSCSI 単数コネクション VPN 単数経路
Fig. 9　iSCSI single connection/VPN single route.

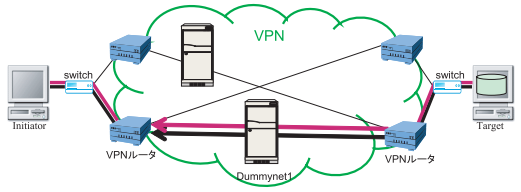


図 10　iSCSI 複数コネクション VPN 単数経路
Fig. 10　iSCSI multiple connections/VPN single route.

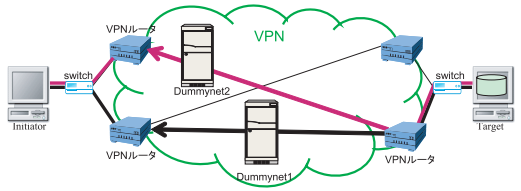


図 11　iSCSI 複数コネクション VPN 複数経路
Fig. 11　iSCSI multiple connections/VPN multi-routing.

下の VPN ルータに送られるように設定する．そして VPN ルータのマルチルーティング機能により，左の 2 つの VPN ルータ宛てにパケットが転送される．このとき，ポート番号の違いにより iSCSI コネクションごとに上下の VPN ルータに分かれるように設定した．

Initiator と Target には，OS は Linux2.4.18-3，CPU は Intel Xeon 2.4 GHz，Main Memory は 512 MB DDR SDRAM，NIC は Intel Pro/1000XT Server Adapter on PCI-X（64 bit，100 MHz），iSCSI は UNH IOL reference implementation ver.3 on iSCSI Draft 18 を用いた[4]．そして Dummynet1 には FreeBSD4.9-RELEASE，Dummynet2 には FreeBSD6.2-RELEASE を用いた．また VPN ルータには Fujitsu Si-R570 を用いた[6]．これは 3DES 暗号化速度最大 500 Mbps を実現する．

この実験環境において，TCP 輻輳ウィンドウモニタツールを起動し，iSCSI シーケンシャルリードアクセス時の性能や TCP パラメータの振舞を観察した．本実験ではストレージアクセスのみの性能を評価するため，Initiator 側では raw デバイスを使用することにより，キャッシュの影響を排除した．また，iSCSI ストレージアクセスにおけるネットワーク性能に焦点を当てて評価を行うため，Target は UNH 実装が提供するメモリモードで動作させ，ディスクアクセスをともなわないようにした．

4. 経路ごとの実行結果の比較

4.1 スループット測定結果

図 12 は iSCSI 単数コネクション VPN 単数経路，iSCSI 複数コネクション VPN 単数経路，iSCSI 複数コネクション VPN 複数経路の各ケースにおいて，片道遅延時間を変化させたときの iSCSI ストレージアクセススループット比較のグラフである．この実験においてブロックサイズは 2 MB に設定した．また，iSCSI 複数コネクション VPN 複数経路において，経路ごとの遅延時間は同じ値を設定した．

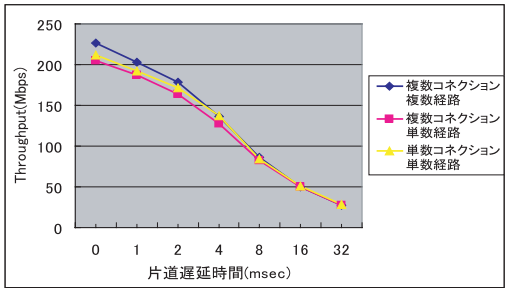


図 12　片道遅延時間とスループット比較
Fig. 12　Comparison of throughput with various delay times.

どの場合も片道遅延時間を長くするとスループットは急激に減少した．遅延時間が短いときには複数コネクション複数経路の場合が一番性能が良く，続いて単数コネクション単数経路，複数コネクション単数経路となっている．

まず複数コネクション複数経路の場合にスループットが単数コネクション単数経路より向上した理由を考察する．Initiator と Target はギガビットイーサネットで接続しているが，VPN ルータの暗号化速度は最大 500 Mbps である．また本実験の VPN 単数経路接続ネットワークのソケット間通信のスループットを測定したところ 330 Mbps 程度の性能であった．一方，VPN ルータをはさまずに Initiator と Target 間を接続した場合，スループットは 400 Mbps 程度であった．したがって VPN ルータの暗号化処理が通信のボトルネックとなっており，複数経路接続にしたことにより，Initiator 側の VPN ルータでの暗号化処理が分散されたため負荷が軽くなったと考えられる．ただし Target 側の VPN ルータは単数のままとなっており性能向上には影響を与えていない．また Initiator 側の VPN ルータの暗号化処理のボトルネックを解消した結果，Target 側の VPN ルータの性能が支配的になっていると考えられる．しかし，VPN ルータは現在急速に性能向上しており，今後本アクセス手法はより有効なものになると期待できる．

これに対し，2 つの回線を用いたにもかかわらず，2 倍近い性能が得られていない理由を調べたところ，iSCSI は各コネクションへのパケット振り分けをラウンドロビンによって行っているため，複数コネクション複数経路の場合でも 2 つの回線を 1 つずつ交互に使っているだけで，2 つの回線を同時には使っていないためであることが分かった．

また図 12 では，遅延時間を長くするとどの場合でも性能に変化がなくなってきている様子が分かる．これは，遅延時間が短いときは複数経路にすることで，ルータでの処理が軽減され性能が向上したと考えられるが，高遅延環境にすると，経路の方がボトルネックとなり性能に差がなくなってきたためであると考えられる．

4.2 輻輳ウィンドウの比較

図 13 は単数コネクション単数経路の場合，**図 14** は複数コネクション単数経路の場合，**図 15** は複数コネクション複数経路の場合に，iSCSI シーケンシャルリードアクセス通信を行ったときの輻輳ウィンドウをモニタした様子である．このときのブロックサイズは 2 MB，片道遅延時間は 2 msec に設定した．またグラフ横軸に時間，縦軸に輻輳ウィンドウとエラーイベント番号を表示した．ここで ErrorNo.2 が Local device congestion（CWR エラー），

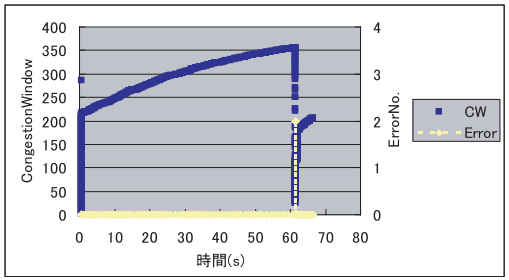


図 13　輻輳ウィンドウ（単数コネクション単数経路）
Fig. 13　Congestion window (single connection/single route).

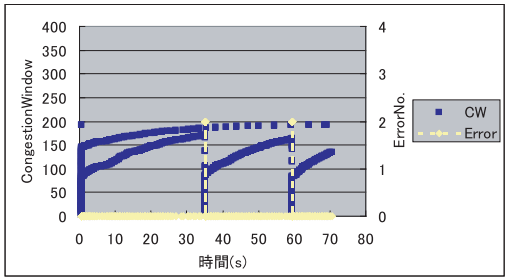


図 14　輻輳ウィンドウ（複数コネクション単数経路）
Fig. 14　Congestion window (multiple connections/single route).

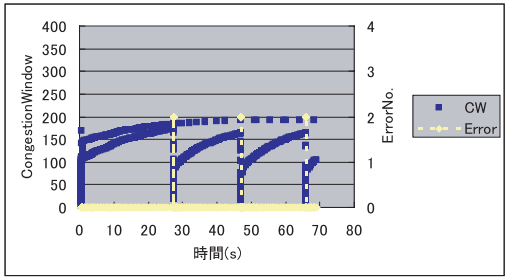


図 15　輻輳ウィンドウ（複数コネクション複数経路）
Fig. 15　Congestion window (multiple connections/multi-routing).

3 が重複 ACK，SACK を受信したこと，4 がタイムアウトを検出したことを示す．また 0 から ErrorNo までを破線で表示し，エラーが起こったタイミングを確認しやすくしているが，破線が他の ErrorNo に関係しているということではない．

図 13 に示された ErrorNo.2 の縦の破線はこの時点で Local device congestion（CWR エラー）が起こったことを表しており，これは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エラーである．図 13 の単数コネクション単数経路の場合，輻輳ウィンドウは約 350 パケットまで増加した後，CWR エラーが検出され急激に減少している．

図 14 の複数コネクション単数経路，図 15 の複数コネクション複数経路の場合は，ほぼ同様な変化をしており，図 13 の単数コネクションの場合と比較すると大きく異なっていることが分かる．ここで輻輳ウィンドウモニタツールは，2 つのコネクションのうち一方でも輻輳ウィンドウの値が変わったら表示されるようになっている．このためそれぞれのコネクションとも独立で値が変わり，ランダムにグラフ表示されることになる．すなわちコネクションごとの輻輳ウィンドウは，一方はなだらかに増加し一定になっており，もう一方は鋸型になっている．

ここで複数コネクションにした場合，なだらかに増加するグラフと鋸型のグラフになった理由を考察する．2 つのコネクションのうちどちらか片方のコネクションが輻輳ウィンドウを使いきったら，ACK が返るまで次の iSCSI アクセス行われなくなる．iSCSI コネクションへのパケット振り分けはラウンドロビンで実行されているため，片方のアクセスが止まってしまうともう片方のアクセスも止まることになる．したがってもう一方のコネクションには輻輳ウィンドウ分を使いきる量のパケットが送られないため，Linux TCP 輻輳ウィンドウの特徴により片方の輻輳ウィンドウは一定となったと考えられる．

次に単数コネクションと複数コネクションの輻輳ウィンドウの違いを比較する．この 2 つのグラフを比べると，検出されるエラーの回数は，複数コネクションの時の方が単数コネクションのときより多くなっている．また，複数コネクションのなだらかに増加するグラフと鋸型のグラフの輻輳ウィンドウを足し合わせると，最大で単数コネクションの時の輻輳ウィンドウの値である 350 パケットに近い値をとっていることが分かる．複数コネクションの場合，なだらかに増加するグラフと鋸型のグラフになるので，その両方の輻輳ウィンドウ分だけパケットは送信されることになる．したがって Target のデバイスドライバのバッファが溢れる頻度が高くなるものと考えられる．

5. 各経路の遅延時間による影響の評価

次に iSCSI 複数コネクション VPN 複数経路接続において，異なる遅延時間を持つ経路を構築して実験を行った．

5.1 スループット測定結果

図 16 は経路ごとに遅延時間を変えていったときのスループット比較のグラフである．横軸は図 11 における Dummynet2 に設定した片道遅延時間，縦軸はスループットをとっている．また，グラフは上から図 11 における Dummynet1 に設定した片道遅延時間が 0，1，2，4，8，16，32（msec）の場合である．

どのグラフも片道遅延時間を大きくするとスループットは減少しているが，Dummynet1，2 の遅延時間が大きくなるにつれてその差は小さくなり，32 msec のときにはほぼ一定の値をとっている．このように片方の遅延時間が大きくなると，もう一方の経路は十分速く通信できる状態でもスループットは低下してしまう．これは iSCSI の振り分けがラウンドロビンであることによる影響であり，パケットを交互に経路に送っているため，遅延時間の大きい経路のほうが通信全体のボトルネックとなっている．これにより経路の遅延時間が実行中に変化した場合も，たとえば Dummynet1 の経路の遅延時間が実行中に大きくなった場合，それにつられて全体の通信性能は低減してしまう．

また，Dummynet1 の遅延時間 2 msec，Dummynet2 の遅延時間 8 msec（これを「遅延 2-8」と表す．以下同様）のときのスループットは 123 Mbps であるが，遅延 8-2 のときは 91 Mbps となっており，Dummynet1 の遅延時間の方が大きく影響している．同様に，遅延 4-16 と遅延 16-4 を比較しても遅延 16-4 のときの方がスループットが低くなっている．このことはすべての場合について同様であり，Dummynet1 の遅延時間の方がスループットに大きく影響しているということできる．これは VPN ルータの 2 経路への振り分けがまったく対等ではなく優先順位などが付いていることによるものと考えられる．これについては次節で議論する．

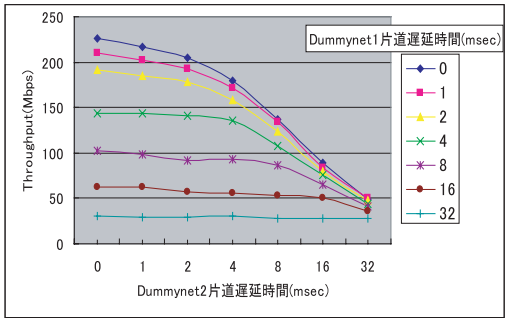


図 16　片道遅延時間とスループット比較
Fig. 16　Comparison of throughput with various delay times.

5.2 輻輳ウィンドウの比較

次に輻輳ウィンドウの違いを比較する．**図 17**，**図 18**，**図 19** はそれぞれ遅延 2-2，遅延 2-16，遅延 16-16 の場合である．

まず Dummynet1 の遅延時間は固定し，Dummynet2 の遅延時間を大きくしていったときの比較をする．図 17 と図 18 を比べると，図 18 の方が明らかに実行時間が長くなっている．また，輻輳ウィンドウの傾きもなだらかになっている．さらにエラーに着目してみると，遅延 2-2（図 17）のときには ErrorNo.2 の Local device congestion による CWR エラーだけなのに対し，遅延 2-16（図 18）のときには ErrorNo.4 のタイムアウトを検出した

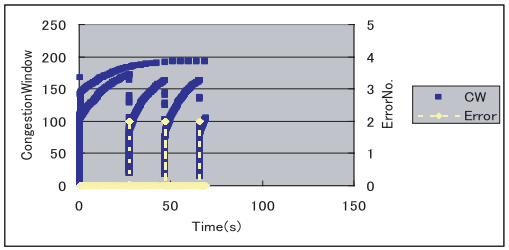


図 17　輻輳ウィンドウ（遅延 2-2）
Fig. 17　Congestion window (delay 2-2).

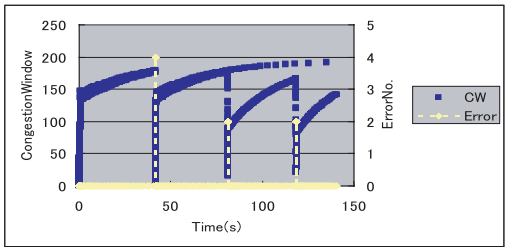


図 18　輻輳ウィンドウ（遅延 2-16）
Fig. 18　Congestion window (delay 2-16).

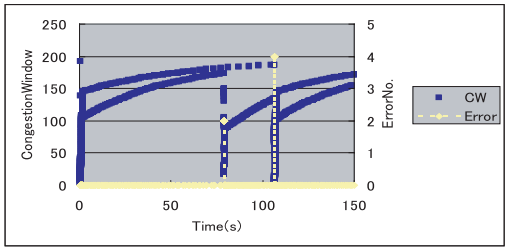


図 19　輻輳ウィンドウ（遅延 16-16）
Fig. 19　Congestion window (delay 16-16).

ことによるエラーが起きている．ここで ErrorNo.4 が検出されると輻輳ウィンドウは片方のグラフのみが減少するのではなく，両方の輻輳ウィンドウが減少しているのが分かる．

次に，Dummynet1 の遅延を大きくしていくと図 19 のようになる．ますます輻輳ウィンドウの傾きはなだらかになり，実行時間は長くなった．また，ErrorNo.4 のタイムアウトもたびたび起こるようになり，スループットも減少した．これらより，片方の遅延時間を長くするとそれにともない両方のコネクションの輻輳ウィンドウが影響を受けており，もう片方には十分な通信帯域が残っているにもかかわらず，それが有効に活用できていない．

また，2 つのコネクションのどちらがどのグラフであるかを確かめたところ，つねに上側にあるグラフが Dummynet1 の方の経路を通るもので，下側にあるグラフが Dummynet2 の方の経路を通るものであることが分かった．これは VPN ルータの設定によるものである．VPN ルータでマルチルーティング機能を利用する場合，1 つの VPN ルータから 2 つのあて先にパケットを振り分けるように設定をする．この設定の記述において，先に記述されたものが優先順位が高くなり，後に設定されたもののほうが優先順位が低くなる．この場合，Dummynet1 の方の経路を先に設定していたので優先順位が高く，そのためこちら側のコネクションはつねに先に輻輳ウィンドウを確保することができる．

6. iSCSI 複数経路アクセスのモデル化と性能評価結果の解析

実測された値を解析するため，iSCSI 複数経路アクセスにおいて経路ごとの遅延時間の変化によるスループットのモデル化を行った．

まず，iSCSI リードアクセスの 1 周期は**図 20** のように表すことができる．このとき 2 つのトラフィックは遅延時間が異なっている場合でも，リードサイクルの周期は同じ長さになっていると考えられる．つまり，その周期は遅延の長い方のみで決まり，もう一方のトラフィックはデータ送信後は待ち状態になる．そこでまず，Dummynet2 を通る経路 2 の遅延時間が Dummynet1 を通る経路 1 のものより長いときを考える．図 20 を参考に，経路 1 の遅延時間を D1，経路 2 の遅延時間を D2 とし，経路 2 のデータ送信にかかる時間を T とすると 1 周期の長さは以下の式で表される．

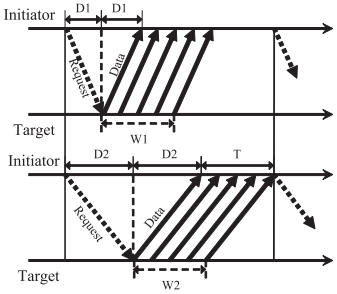


図 20　両経路上の iSCSI リードアクセスの 1 周期
Fig. 20　1 cycle of iSCSI read access on each route.

$$1\text{ 周期の長さ} = 2 * \mathrm{D2} + \mathrm{T}$$

このときの経路 1 の輻輳ウィンドウを W1 とし，経路 2 の輻輳ウィンドウを W2 とする．1 周期の間に送られたデータ量は 1 つのパケットの大きさが約 1.5 KB であるため，およそ以下の式で表される．

$$\text{各周期に送られるデータ量} = 1.5\,\mathrm{KB} * (\mathrm{W1} + \mathrm{W2})$$

したがってスループットは以下の式となる．

$$\text{スループット} = \frac{\text{各周期に送られるデータ量}}{1\text{ 周期の長さ}} \tag{1}$$

$$= \frac{1.5\,\mathrm{KB} * (\mathrm{W1} + \mathrm{W2})}{2 * \mathrm{D2} + \mathrm{T}} \tag{2}$$

前章までに述べた輻輳ウィンドウのモニタ結果より W1 + W2 = 約 350 パケット であり，また遅延時間ごとのスループットの測定結果を代入すると，T の値を計算することができる．その結果この T の値は，D1 を一定としたとき D2 によってあまり変化せず，D1 の関数であることが分かった．すなわち D1 が決まると経路 1 における輻輳ウィンドウ W1 が決まり，その値により W2 が決まるものと考えられる．そこで D1 の値ごとの T の値を実測値から計算して求め，これを式 (2) に代入すると，各 D2 の値からスループットを求める

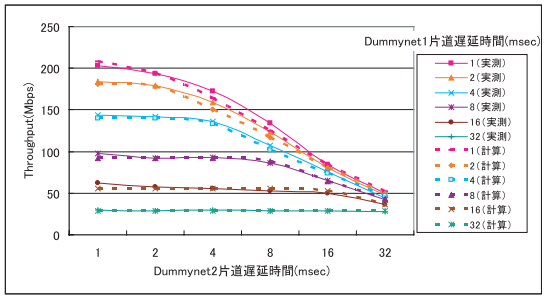


図 21　スループットの測定値と計算値の比較
Fig. 21　Comparison of mesured value and calculate value of throughput.

式が得られる．

次に D1 が D2 より長い場合を考える．このとき式 (2) における D1 と D2 を入れ替え同様に計算すると，再び T は D1 の値により決まる D1 の関数であることが確かめられた．したがってこの場合，式全体が D1 の関数になっており，D2 の変化にはほとんど影響されずスループットは一定値となる．すなわち D1 が D2 より長い場合にも，D1 の値によってまず W1 が決まり，その値から W2 が決定されていると考えられる．D2 が D1 より長い場合と比較して考えると，経路 1 と経路 2 の通信は対等には動作していないことが分かる．

以上の考察をもとに，**図 21** にスループットモデル式の計算値と実測値を比較したグラフを示す．図 21 に示されたように計算値と実測値がほぼ等しいことから，このモデル化は適切なものであるといえる．

またここで，T が D1 のみの関数となったことを考察する．前章で述べたように，本実験で用いた VPN ルータは，経路ごとに優先順位が割り振られる．したがって 2 つの経路は対等ではなく，優先順位の高い経路 1 の輻輳ウィンドウ W1 が全体の性能に大きく影響することになる．つまり，経路 1 の遅延 D1 によって輻輳ウィンドウ W1 が決まり，W1 によって経路 2 の輻輳ウィンドウ W2 が決定されるため，T はいずれの場合においても D1 の関数になっていたと考えられる．

7. ま と め

本研究では，iSCSI ストレージアクセスにおいてコネクションを単数，複数に変化させ，VPN 接続も単数経路，複数経路に変化させたときのスループットの違いと輻輳ウィンドウの振舞いを観察し，比較した．また遅延時間の異なる経路を用いて iSCSI 複数コネクション VPN 複数経路を行ったときのスループットと輻輳ウィンドウの振舞いを評価した．

その結果，iSCSI 単数コネクション VPN 単数経路と iSCSI 複数コネクション VPN 単数経路を比較すると，特に高遅延環境において輻輳ウィンドウの振舞いが大きく変化している様子が分かった．これは iSCSI コネクションが複数になったため，輻輳ウィンドウがコネクションごとに振り分けられ，それぞれ異なる動作をしているためであると考えられる．さらに片方のコネクションにより iSCSI アクセスが止められた場合，もう一方のコネクションでは輻輳ウィンドウを使いきれずほぼ一定の値となっている．また，iSCSI 複数コネクション VPN 単数経路と iSCSI 複数コネクション VPN 複数経路を比較すると，スループットは複数コネクション複数経路の方が多少向上するが，輻輳ウィンドウはほとんど変化が見られない．これは Initiator 側の VPN ルータが複数となったため暗号化処理が軽減されたものと考えられる．

また，高遅延環境においてはスループットはどの場合もほとんど変化がなくなることが分かった．これは経路の遅延が大きくなることで，その影響が支配的になったためであると考えられる．

遅延時間の異なる経路を用いた実験では，片方の遅延時間が大きくなるとそれに引きずられてスループットも低下してしまった．これも iSCSI がパケットの振り分けをラウンドロビンで行っているからであると考えられる．輻輳ウィンドウは遅延時間を大きくしていくと，Local device congestion だけではなくタイムアウトによってもエラーが起きていることが分かった．タイムアウトのエラーが起こった場合には，その影響が片方の経路だけにとどまらずもう一方の経路の輻輳ウィンドウも引き下げられている．

そこで経路により遅延時間が異なる場合の性能を明らかにすべく，iSCSI 複数経路アクセスのモデル化を行い，性能の解析的評価を行った．その結果，VPN 複数経路における各 iSCSI コネクションの振舞いを明らかにし，実験で得られたスループットにきわめてよく適合するモデル化を行うことができた．

最後に，VPN を利用しない複数経路接続を行った場合，暗号化などの VPN ルータにおける処理ではなく通信経路などがボトルネックになることが考えられる．しかしその場合にも，性能向上の度合いは異なると思われるが，複数経路接続することで性能は向上すると予想される．ただし WAN 上で遠隔アクセスを行う場合には VPN を用いることが望ましいため，本稿では VPN 接続環境における単数経路および複数経路を用いた iSCSI 遠隔アクセスを比較評価した．

今後は経路によって遅延時間が異なる場合，iSCSI のパケットの振り分けを適切に制御し性能が向上するような通信を実現させたい．

謝辞　本研究は一部，独立行政法人科学技術振興機構戦略的創造研究推進事業 CREST によるものである．

参 考 文 献

1) Internet Small Computer Systems Interface (iSCSI). http://www.ietf.org/rfc/rfc3720.txt?number=3270
2) SCSI Specification. http://www.danbbs.dk/˜dino/SCSI/
3) 山口実靖，小口正人，喜連川優：高遅延広帯域ネットワーク環境下における iSCSI プロトコルを用いたシーケンシャルストレージアクセスの性能評価ならびにその性能向上手法に関する考察，電子情報通信学会論文誌，Vol.J87-D-I, No.2, pp.216–231 (2004).
4) InterOperability Lab., Univ. of New Hampshire. http://www.iol.unh.edu/consortiums/iscsi/
5) 千島　望，豊田真智子，山口実靖，小口正人：VPN 接続環境における TCP パラメータと通信性能の相関関係評価，FIT2006, L-042 (2006).
6) 富士通 IP アクセスルータ GeoStream Si-R シリーズ GeoStream Si-R570. http://fenics.fujitsu.com/products/sir/sir570/index.html
7) 千島　望，豊田真智子，山口実靖，小口正人：iSCSI アクセス時の VPN 環境における TCP 輻輳ウィンドウ制御手法の検討，DBWS2006, pp.709–712 (2006).
8) 板谷俊輔，相田　仁：複数経路プロトコル M/TCP の FreeBSD への実装の性能向上，電子情報通信学会情報ネットワーク研究会，IN2005-192, pp.213–218 (2006).
9) Sarkar, P., Uttamchandani, S. and Voruganti, K.: Storage over IP: When Does Hardware Support help?, *Proc. 2003 USENIX Conference on File and Storage Technologies* (*FAST2003*), pp.231–244 (2003).
10) Radkov, P., Yin, L., Goyal, P., Sarkar, P. and Shenoy, P.: Performance Comparison of NFS and iSCSI for IP-Networked Storage, *Proc. 2004 USENIX Conference on File and Storage Technologies* (*FAST2004*), pp.101–114 (2004).
11) Joglekar, A., Kounavis, M.E. and Berry, F.L.: A Scalable and High Performance Software iSCSI Implementation, *Proc. 2005 USENIX Conference on File and Storage Technologies* (*FAST2005*), pp.267–280 (2005).
12) 藤田智成，矢田浩二：iSCSI イニシエータの設計についての一考察，情報処理学会研究報告，2005-OS-99(21), pp.135–140 (2005).
13) 藤田智成，小河原成哲：iSCSI ターゲットソフトウェアシステムの解析，情報処理学会論文誌コンピュータシステム，Vol.46, No.SIG3(ACS8), pp.38–50 (2005).
14) 千島　望，豊田真智子，山口実靖，小口正人：iSCSI における TCP パラメータとアプリケーション実行性能の相関関係評価，第 68 回情報処理学会全国大会，pp.131–132 (2006).
15) 豊田真智子，山口実靖，小口正人：iSCSI ストレージアクセスにおける TCP 輻輳ウィンドウコントロール手法の提案と性能評価，電子情報通信学会論文誌，Vol.J90-D, No.2, pp.359–372 (2007).
16) Motwani, G. and Gopinath, K.: Evaluation of Advanced TCP Stacks in the iSCSI Environment using Simulation Model, *Proc. 22nd IEEE / 13th NASA Goddard Conference on Mass Storage Systems and Technologies* (*MSST2005*), pp.210–217 (2005).
17) 藤原啓成，若宮直紀，志賀賢太：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，第 68 回情報処理学会全国大会，pp.155–156 (2006).
18) Kancherla, B.K., Narayan, G.M. and Gopinath, K.: Performance Evaluation of Multiple TCP connections in iSCSI, *Proc. 24th IEEE Conference on Mass Storage Systems and Technologies* (*MSST2007*), pp.239–244 (2007).
19) Rizzo, L.: dummynet. http://info.iet.unipi.it/˜luigi/ip_dummynet/

(平成 19 年 12 月 30 日受付)
(平成 20 年 7 月 1 日採録)

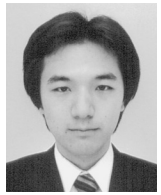
推 薦 文

DSM 関連論文 19 件中で論文評価委員の得点合計で第 1 位になった論文である．内容的には VPN を利用して iSCSI を広域ネットワークに適用するために，高遅延環境における iSCSI の性能を評価しており，その評価結果は今後の iSCSI と VPN を利用した SAN の活用を期待させる優れた論文である．

（分散システム/インターネット運用技術研究会主査　藤村直美）

千島　望

平成 20 年お茶の水女子大学大学院人間文化研究科博士前期課程修了，理学修士．平成 18 年同大学理学部情報科学科卒業．在学中は iSCSI の研究に従事．現在はゴールドマン・サックス・ジャパン・ホールディングスに勤務．



山口　実靖（正会員）

平成 14 年東京大学大学院工学系研究科電子情報工学専攻博士課程修了．博士（工学）．同年より東京大学生産技術研究所学術研究支援員，産学官連携研究員，日本学術振興会特別研究員．平成 18 年工学院大学工学部講師．平成 19 年同大学同学部准教授．iSCSI を用いたネットワークストレージシステムの性能向上の研究に従事．電子情報通信学会，日本データベース学会各会員．

小口　正人（正会員）

平成 2 年慶應義塾大学理工学部電気工学科卒業．平成 7 年東京大学大学院工学系研究科電子工学専攻博士課程修了．博士（工学）．学術情報センター中核的研究機関研究員，東京大学生産技術研究所特別研究員，中央大学研究開発機構助教授，お茶の水女子大学理学部情報科学科助教授を経て，平成 18 年より同教授．ネットワークコンピューティング・ミドルウェアに関する研究に従事．IEEE，ACM，電子情報通信学会各会員．