6 iSCSIとFCoEによるストレージ構築 ～ストレージネットワークの進化～

小口正人◉お茶の水女子大学

ストレージはクラウドを支える重要な存在である．ストレージシステムの進化に伴い，クラウド内も含め，サーバとストレージの間をネットワークで接続する方式がいくつかの形態で登場し，利用されてきている．本稿ではイーサネットをベースにサーバとストレージが接続されるiSCSI（Internet Small Computer System Interface）とFCoE（Fibre Channel over Ethernet）の2方式を概観する．

iSCSI

＊iSCSI誕生の背景

ストレージシステムは従来，図-1の左側のようにサーバとストレージシステムが密に接続され，SCSI（Small Computer System Interface）などのプロトコルでアクセスが行われていた．この形態をDAS（Direct Attached Storage）と呼ぶ．しかしサーバが特定のストレージにしかアクセスできないと，偏りが生じストレージ全体を効率よく使うことができない．また管理の手間がかかるため，数が増えると管理コストが高くなる．そこで図-1の右側のようにストレージへのアクセス部分をネットワークとし，この上をSCSIプロトコルが流れるような仕組みが作られた．これがSAN（Storage Area Network）である．すなわちサーバ間通信（フロントエンド）はLANで接続され，ストレージアクセス（バックエンド）はSANで接続される形態が，現在のサーバとストレージシステムの標準的な形となった．

SANとしてはFC（Fibre Channel）が広く用いられている．FCは文字通り，光ファイバをベースとしたネットワークである．しかしFCはスイッチもケーブルも高価なものであり，また光ファイバであるため銅線で接続されたイーサネットなどのLANと比べると扱いにくい．さらに接続距離に制限があるなどといった問題もあり，機器の種類が豊富でコストパフォーマンスがよい，汎用のネットワーク技術をストレージ接続に活かすことが望まれている．

そのような要望に応えてiSCSIは登場し，2003年2

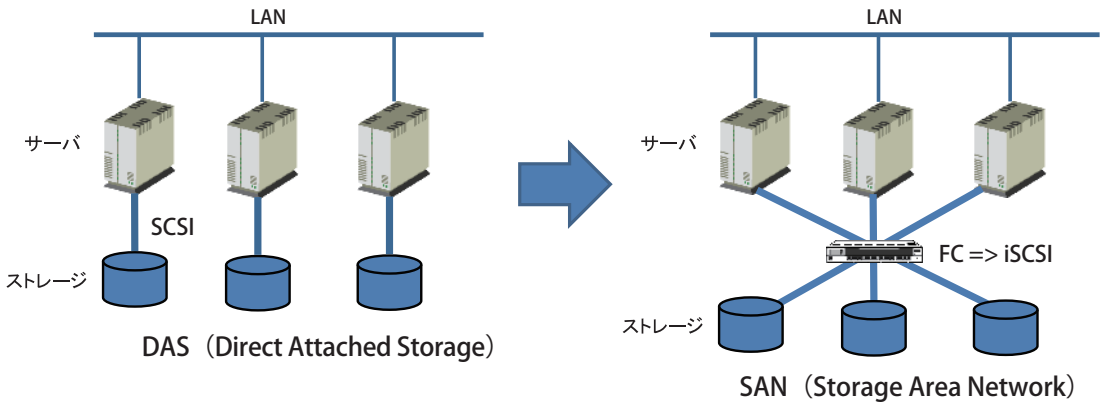

図-1　DASからSANへの進化

月にIETFにより規格が承認された[1)]．iSCSIはSCSIプロトコルをTCP/IPネットワーク上に載せたものであり，汎用ネットワーク環境であるイーサネットとその上のTCP/IPが動いていれば利用することができる．

iSCSIアーキテクチャ

iSCSIはSCSIコマンドや読み書きするデータを，TCP/IPパケットにカプセル化してネットワーク越しに転送する規格である．iSCSIのプロトコルスタックを図-2に示す．ストレージアクセスを行うサーバ側（イニシエータ）では，アプリケーションがローカルストレージへアクセスするかのようにSCSIコマンドを発行するが，iSCSI層がこれを受け取ってTCP/IPのパケットに詰め込み，イーサネット等のネットワークを介してストレージ側（ターゲット）へ送る．ターゲットではTCP/IPパケットからSCSIコマンドを取り出してストレージアクセスを行い，結果を逆の経路でイニシエータへ返す．

以下ではこのカプセル化が具体的にどのように実現されているか，少々細かく説明する．SCSIコマンドをTCP/IPパケットに詰め込むために作られるiSCSIのブロックをiSCSI PDU（Protocol Data Unit）と呼ぶ．この構成を図-3に示す．この図では1つのiSCSI PDUが1つのTCP/IPパケットにカプセル化されているが，大きさがTCPのMSS（Maximum Segment Size）を超える場合には分割されて，各々のセグメントにTCP/IPのヘッダが付く．図-3にはReadのSCSIコマンドメッセージがカプセル化された場合の例が示されている．

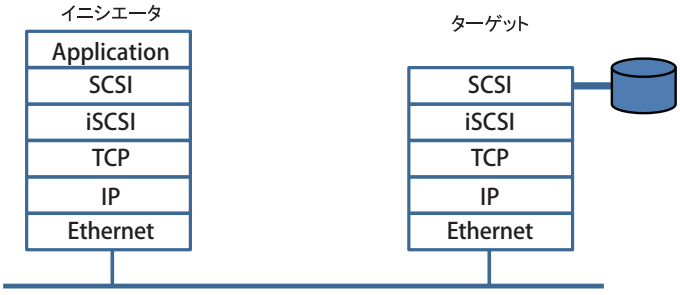

図-2　iSCSIのプロトコルスタック

iSCSI PDUは48バイトのBHS（Basic Header Segment）といくつかのオプションフィールドからなる．SCSIプロトコルによりストレージとの間でRead/Writeのデータが転送される場合には，Data Segmentフィールドが使われる．

SCSIコマンドメッセージをカプセル化する場合，BHSの後ろ16バイトのフィールドには，SCSI CDB

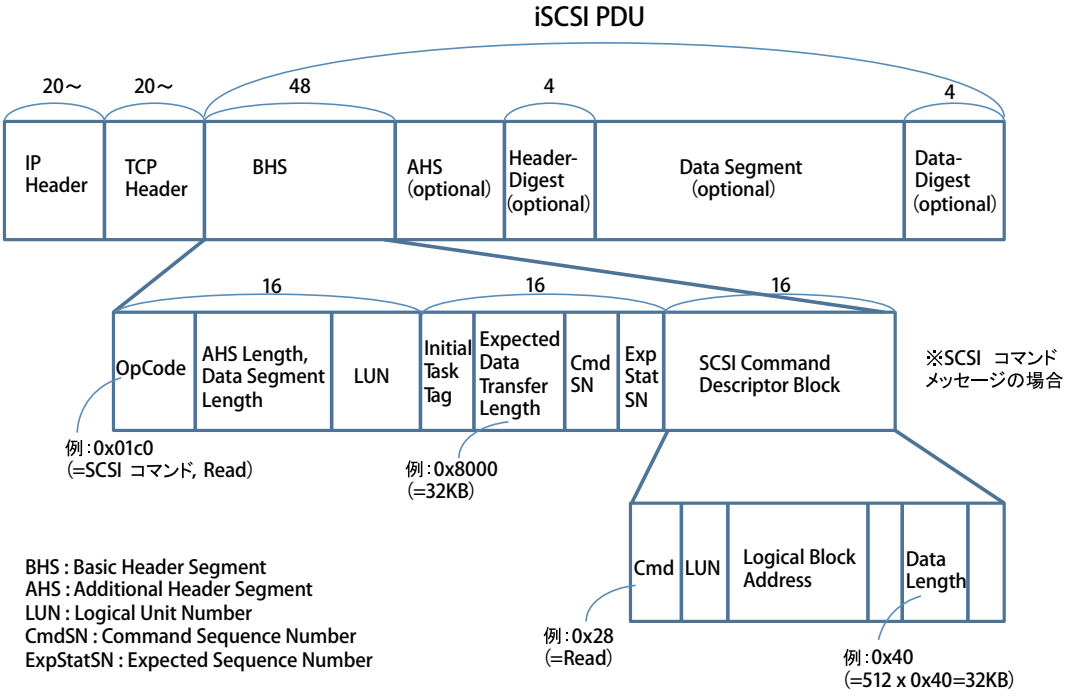

図-3　iSCSI PDU（Protocol Data Unit）

(Command Descriptor Block)，すなわち SCSI コマンドがそのまま入る．ただしその中のいくつかの情報は読み出されて，BHS の前の方のフィールドに埋められる．たとえば BHS の先頭は OpCode というフィールドで，ここには SCSI CDB の中身は何であるか(たとえば SCSI Read コマンドなど）が読み出されて書き込まれる．Expected Data Transfer Length フィールドにも，SCSI CDB 中で指定している Read/Write のデータの大きさが読み出されてバイト単位で記録される(SCSI CDB 中では 512 バイト単位)．

代表的な iSCSI 実装の紹介

前述のように iSCSI は，イーサネットとその上の TCP/IP という汎用のネットワーク環境があれば利用することができる．最新の Linux や Windows においては，iSCSI は容易に利用できるようになっている．

Linux では当初，UNH（University of New Hampshire）により作られた参照実装が多く使われていたが，その後いくつかの実装が現れて改良が進められた．イニシエータドライバでは，Linux-iSCSI（sfnet）と Open-iSCSI がよく使われるようになってきたが，この両プロジェクトは合併し，Linux2.6.16 以降で動作する Open-iSCSI ドライバとして利用されている[2]．またターゲットドライバとしては，iSCSI Enterprise Tareget が広く用いられている[3]．一方 Windows 7 や Windows Server 2008 では，標準で Microsoft iSCSI イニシエータドライバが搭載されている．ターゲットドライバは Windows Storage Server 2008 に搭載されている．

このような導入の敷居の低さにより，iSCSI の利用は拡大してきている．一方，性能面を考慮して iSCSI 用のハードウェアも現れて利用されている．ターゲットとしては iSCSI インタフェースを持ったストレージシステムが，比較的高額なものから，量販店で売られるような割合安価なモデルまで幅広く販売されている．またイニシエータは，TCP/IP の処理を NIC にハードウェア実装する TOE（TCP Offload Engine）や，iSCSI の処理までハードウェア実装した iSCSI HBA（Host Bus Adaptor）が存在し，目的に応じて使い分けられている．

現在の状況を見ると，iSCSI の技術はすでに発展のフェーズに乗ったと言えよう．今後はクラウドを始めとしたサーバシステムにおける利用から，エンドユーザの個人利用まで，幅広く用いられていくものと考えられる．

FCoE

FCoE 誕生の背景

「**iSCSI 誕生の背景**」で述べたように，現在のサーバシステムはフロントエンドを LAN で接続され，バックエンドを SAN で接続される形が一般的になった．LAN のアーキテクチャはイーサネットであり，ギガビットイーサネットから 10Gbps イーサネットへと移行しつつある．一方 SAN として主流なのは，やはり FC である．

FCoE は，INCITS（International Committee for Information Technology Standards)の FC を担当する T11 技術委員会により，FC-BB-5 の一部として標準化が行われた[4), 5)]．FCoE は FC のフレームをカプセル化し，仕様に従って拡張された 10Gbps イーサネット上で運ぶというものである．

FCoE の目的を一言でいうと FC とイーサネットの統合であり，FCoE によって FC のフレームとイーサネットのフレームを，同じインタフェースを通し同じネットワークで運ぶことを目指した．これはユニファイド I/O と呼ばれる．

現在のサーバは，インタフェースをいくつも持っている．たとえば図 -4 の左側のように，制御用のイーサネット，データ転送用のイーサネット，ストレージに繋がる FC があり，それぞれが二重化されているとすると，インタフェースは全部で 6 枚となり，6 本のケーブルがサーバから出てネットワークに繋がる形となる．これに対し，これらの FC とイーサネットをすべて FCoE で統合すると，図 -4 の右側のように，二重化も含めて 2 枚のインタフェース CNA（Converged Network Adaptor）と 2 本のケーブルで済む．

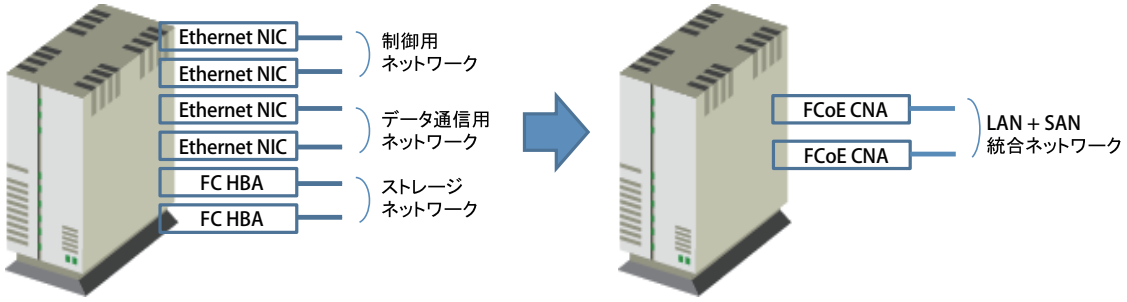

図 -4　FCoE による FC とイーサネットの統合

FCoE が最も期待されている活躍の場はデータセンタである．データセンタでは，FC とイーサネット，さらにはインフィニバンドなど複数の独立したネットワークが運用されており，サーバは複数のインタフェースを持ち，複数のネットワークと接続する形態となっている．しかしながら複数のネットワークが別々に存在するのは，導入コストの面でも管理コストの面でも望ましくない．さらに，ラックマウントやブレードタイプのサーバの場合，物理的な制約から複数のインタフェースを持つことができない場合もあり，その結果，共有ストレージに直接接続されたサーバは，全体のうち一部のみに限られることとなっている．すなわち FCoE を用いれば，統合によりコストが下げられるだけでなく，機器の可用性を上げることができる．さらには FCoE を用いた統合でインタフェースやスイッチの数を減らすことによる省電力効果も期待されている．

FCoE アーキテクチャ

イーサネットと FC，そしてこれを統合した FCoE の階層構造を図 -5 に示す．FCoE は iSCSI と異なり，TCP/IP の上に FC を載せるわけではない．トランスポート層（TCP，UDP）もインターネット層（IP）も用いず，第 1 層と第 2 層にあたるイーサネット部分のみを用い，その上に FC の FC-2 層以上を載せる．ただしこの 1，2 層も既存のイーサネットそのままではなく，拡張された仕様に基づく CEE（Converged Enhanced Ethernet）が用いられる．この CEE については次節で述べる．

FCoE において，FC-2 層以上は変更を加えることなく，既存の FC を用いることができる．したがって FC のネームサービスやゾーニングなどの管理機能や SAN のアプリケーションをそのまま利用できる．これは FC フレームをイーサネットフレームにカプセル化することにより実現されている．

以下ではこのカプセル化がどのように実現されているか，少々細かく説明する．イーサネット，FC および FCoE のフレームを図 -6 に示す．

VLAN 対応のイーサネットフレームは，宛先と送信元 MAC アドレス各 6 バイト，VLAN タグ 4

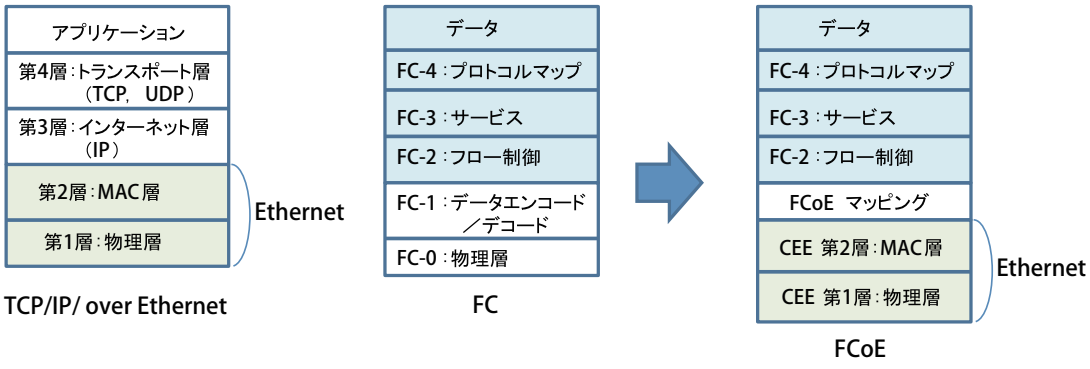

図 -5　イーサネットと FC を統合した FCoE のプロトコルスタック

バイト，タイプ／フレーム長フィールド 2 バイトによりヘッダが構成されている．データの後ろにはトレーラとして FCS（Frame Check Sequence）が 4 バイト続く．そして FCoE ではこのデータ部分に FC のデータが埋め込まれる．タイプ／フレーム長フィールドは FCoE を表す 0x8906 となっており，これにより FCoE フレームであることが識別できる．データの最大長は 2,112 バイトとなっており，ヘッダ等も含めた FCoE フレームは最長 2,148 バイトとなる．したがってジャンボフレームなどに対応するイーサネット環境が必要になる．

CEE について

FCoE は FC フレームがイーサネットへカプセル化されているため，FC から見ると下がイーサネットになっていることは認識できない．しかしながら FC はパケットが喪失せず宛先に届くことを前提として設計されており（ロスレス），パケットロスがあることを前提にトランスポート層で到達保証を行う TCP/IP/ over イーサネット環境とは異なる．したがって FC をイーサネット上に載せようとした場合，ロスレスのイーサネットが必要となってくる．そのための規格が CEE である．CEE は IEEE において，DCB（Data Center Bridging）という名前で標準化が進められている [6]．DCB の用途は FCoE だけではないが，FCoE はその上位に載る有力なアプリケーションといえる．

DCB の主要な規格は以下の 3 つである．

- 802.1Qau：Congestion Notification（CN）：輻輳通知
- 802.1Qaz：Enhanced Transmission Selection（ETS）：拡張送信選択
- 802.1Qbb：Priority-based Flow Control(PFC)：優先度ベースフロー制御

CN はフレームの転送経路において輻輳が発生した際に，これを送信元へ知らせて送信レートの制御を行う．輻輳を検出したスイッチは輻輳通知メッセージを送信元へ送り，これを受け取った送信元は，フレーム送出を抑制することにより輻輳を緩和させる．

ETS ではトラフィックの優先度クラスを設け，各クラスの保証帯域を規定してフレームの配信を行う枠組みである．優先度をプライオリティグループに割り当て，LAN や SAN，プロセス間通信など，異なる種類のトラフィックをグループごとに区別し

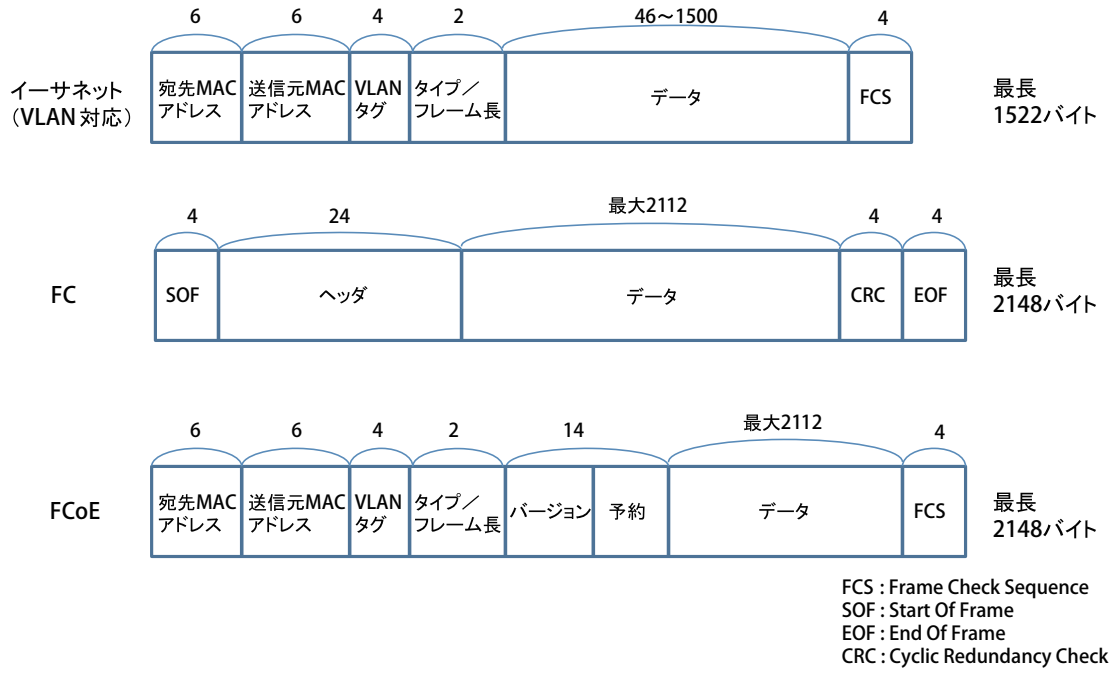

図 -6　イーサネット，FC，FCoE のフレーム

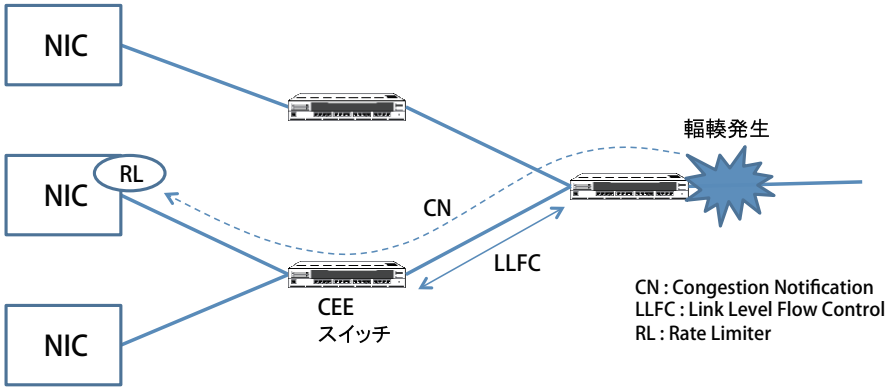

図 -7　CN と PFC の動作の概念図

て扱うことができる．

PFC はスイッチにおいて輻輳が発生した際に，その 1 つ手前のスイッチまたは端末に対して PAUSE フレームを送り，転送を一時的に停止させる．元々 802.3x PAUSE という隣接ノード間でフレーム転送を一時的に停止させるフロー制御の規格があるのに対し，これに優先度クラスを対応させて，優先度別にフレーム転送を一時的に停止できるようにしたものである．

フロー制御が行われている様子を図 -7 に示す．LLFC（Link Level Flow Control）は隣接ノード間のフロー制御で，PFC により実現される．しかし深刻な輻輳が発生した場合，PFC だけでは一時的な対処しかできないため，CN で元栓を絞ることによって，ロスレスのイーサネットを実現している．

FCoE の発展方向

FCoE の導入は，まず最初に FCoE の CNA をインタフェースに持つサーバが導入され，FCoE 対応のスイッチにより既存の LAN/SAN の枠組みに接続されるところから進んでいくであろう．FCoE に対応したストレージが導入されて，エンド・ツー・エンドが FCoE で統合されるようになるまでには，まだ少し時間がかかりそうである．

FCoE による統合のメリットは前述の通りであるが，果たして統合は進むのであろうか．統合を阻害する要因もいくつか考えられる．まずメリットとして考えられているコスト削減が思惑通り進むかということである．これは技術が成熟して安定した安価な製品が出てくるかどうかにかかっているであろう．また性能面で考えると，現在 8Gbps が主流となっている FC に対し，10Gbps イーサネットを用いれば物理的な転送速度は向上するが，FC も 16Gbps へと移行し始めたところである．たとえば SSD のような高速ストレージが用いられる場合には，接続ネットワークの速度が性能に直結するため，少しでも速いことが求められる可能性がある．さらに，LAN と SAN は異なる組織が管理しているようなケースが多いため，これらを統合するには組織を変える必要も出てくるかもしれない．これらの阻害要因が今後どのように変わっていくかが，FCoE 普及の鍵といえるであろう．

参考文献
1) RFC 3720 : Internet Small Computer Systems Interface (iSCSI) (http://tools.ietf.org/html/rfc3720).
2) Open-iSCSI (http://www.open-iscsi.org/).
3) iSCSI Enterprise Target (http://iscsitarget.sourceforge.net/).
4) INCITS Working Draft Proposed American National Standard for Information Technology, Fibre Channel Backbone-5 (FC-BB-5) Rev. 2.00 (http://www.t11.org/ftp/t11/pub/fc/bb-5/09-056v5.pdf).
5) INCITS 462-2010, American National Standard, Fibre Channel Backbone-5 (FC-BB-5) (http://www.techstreet.com/standards/INCITS/462_2010?product_id=1724386).
6) Data Center Bridging, IEEE 802 Tutorial (http://www.ieee802.org/802_tutorials/07-November/Data-Center-Bridging-Tutorial-Nov-2007-v2.pdf).

(平成 23 年 1 月 22 日受付)

小口正人 (正会員) oguchi@computer.org

平成 2 年慶大・理工・電気卒．平成 7 年東大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，東大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．