社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

次世代広域イーサネットにおける大容量データ伝送に向けた複数経路探索手法の一検討

丸川　純平† 　山田　翔太† 　寺澤　　緑† 　清水　　翔† 　石井　大介†

岡本　　聡† 　山中　直明†

† 慶應義塾大学理工学部情報工学科
〒 223–8522 横浜市港北区日吉 3–14–1
E-mail: †marukawa@yamanaka.ics.keio.ac.jp

あらまし　現在の広域イーサネットにおいては，VLAN パスを確立する際に確保可能な帯域がリンク (波長) の帯域に制限されてしまう．この制限を打破するためには，End-to-End での複数パス経路のアグリゲーションが必要となる．本研究では，複数パス経路のアグリゲーションにより仮想的な大容量データ伝送を実現する次世代広域イーサネットにおいて，経路間での遅延差を考慮した複数パス経路の同時発見を実現するために，拡張イーサネットフレームフラッディング手法に基づいた経路発見手法を提案する．また，本アルゴリズムをソフトウェアスイッチに実装し，発見した複数経路における遅延時間を評価した結果を報告する．

キーワード　広域イーサネット，SCTP，並列伝送，遅延測定

Study of Multiple Path Discovery Approach for Mass Data Transmission in Next Generation Wide Area Layer2 Network

Jumpei MARUKAWA†, Shota YAMADA†, Midori TERASAWA†, Sho SHIMIZU†, Daisuke ISHII†, Satoru OKAMOTO†, and Naoaki YAMANAKA†

† Dept. of Information and Computer Science, Keio University
3-14-1 Hiyoshi, Kohoku, Yokohama, 223-8522 Japan
E-mail: †marukawa@yamanaka.ics.keio.ac.jp

Abstract　Currently, bandwidth allocation of VLAN path in Wide Area Ethernet depends on link bandwidth. Aggregation of multiple end-to-end paths is effective to avoid this limit. This paper proposes multiple independent paths discovering method using extended Ethernet frame flooding technique in next generation Wide Area Ethernet which virtually enables high capacity transmission by aggregating multiple paths. Implementing the algorithm to software switches and evaluates the delay of discovered multiple paths.

Key words　Wide Area Ethernet, SCTP, Parallel Transfer, Delay measurement

1. ま え が き

近年，企業の拠点やデータセンタ間を LAN(Local Area Network) 技術であるイーサネットにより通信接続する広域イーサネットサービスが注目されている．図 1 に，広域イーサネットにおける拠点間通信を示す．広域イーサネットサービスでは，ギガビットイーサネット (GE: Gigabit Ethernet) や 10 ギガビットイーサネット (10GE : 10 Gigabit Ethernet) リンクを複数の企業で共有し，各企業に割り当てられた固有の VLAN（Virtual LAN） ID で伝送路を論理的に分離することにより，専用線よりも安価に導入することが可能である．

しかし，今後増加すると見込まれる企業やデータセンタの加入数に伴い，伝送路の要求帯域が広域イーサネットのリンク帯域を上回ってしまうことが懸念される．40G/100G イーサネットの標準化も検討されている [1] が，実用化されるまでに時間を要することが確認されている．そのため，1GE や 10GE の従来網を使用した広帯域伝送技術の早期確立が課題となる．

広域イーサネット網における広帯域経路を実現する手段の 1 つとして，並列伝送技術が挙げられる [2]．ノード間の複数リンクに対して IEEE802.3ad リンクアグリゲーションを行い，並

列伝送を行うことにより，1 本の仮想的な大容量データ伝送路を確立することが可能となる．しかし，リンクアグリゲーションはフロー単位でリンクが決定されるため，各フローの使用可能帯域はリンク帯域に依存してしまうという問題がある．現在はエンドユーザの要求帯域はリンク帯域以下であることが多いが，将来的に転送データ量が増え，エンドユーザの要求帯域がリンク帯域を超過する場合はリンクアグリゲーションを使用不可能である．本問題における解決策として，エンドノード間で複数の VLAN パスを確立した後，フロー内の各フレームに対して VLAN ID をラウンドロビンで付与し，対応する VLAN パスに転送することで，フローの使用可能帯域上限を全リンク帯域にすることが可能である [3]．しかし，転送データをフレーム単位で複数経路に分割して送信する場合，経路間の遅延差により受信側でのフレーム順序逆転が発生するため，TCP/IP 通信におけるフローのスループットは TCP 再送制御により激減する．

パケット順序逆転問題を軽減するプロトコルとして，TCP/IP over SCTP（Stream Control Transmission Protocol）/IP [4] が提案されている．TCP/IP over SCTP は，TCP/IP パケットを SCTP/IP パケットでカプセル化することで，パケット順序逆転により生じる TCP 再送制御が原因となるスループット低下を低減することが可能である．本プロトコルはエッジノードでカプセル化するため，導入時に複雑な機能設定がなく安易であること，エッジノードのみの機能追加で良く安価であることが特長として挙げられる．

しかし，本プロトコルを適用したネットワークにおいて選択した複数経路間での遅延差が大きい場合は，SCTP によるパケット順序逆転の抑制効果を TCP の再送制御によるパケット再送数が上回るため，スループットが低下してしまう．そのため，全経路における遅延測定を行い，遅延差の少ない経路対を選択することが必要となる．

そこで，本論文では複数パス経路のアグリゲーションにより仮想的な大容量データ伝送を実現する次世代広域イーサネットにおいて，経路間での遅延差を考慮した複数パス経路の同時発見を実現するために，拡張イーサネットフレームフラッディング手法に基づいた経路発見手法を提案する．

本稿の構成は以下のとおりである．2 節では，関連研究として，並列伝送および従来の経路探索手法について述べる．3 節では，本論文の対象とする次世代広域イーサネットのアーキテクチャについて述べ，4 節で広帯域伝送路を実現するための複数パス経路発見手法を提案する．5 節では本手法をソフトウェアスイッチに実装したネットワークにおいて，提案手法により算出した遅延時間の評価を行う．最後に 6 節で結論を述べる．

2. 関 連 研 究

2.1 並列伝送技術

データ転送の高速化手法として注目される並列伝送は，広域イーサネット網内のノード間リンク帯域増加や冗長化による信頼性の向上を目的としたリンクアグリゲーションを End-to-End に拡張し，複数経路のパスをアグリゲーションすることにより

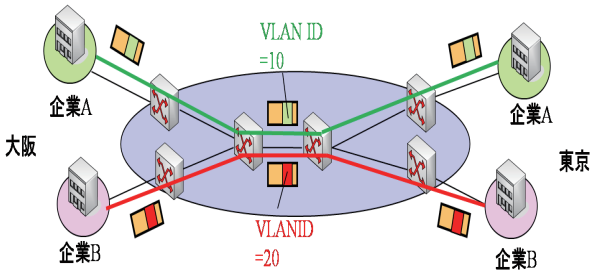

図 1　広域イーサネット

実現される．複数経路上でのデータ転送で使用するパケット負荷分散アルゴリズムとして，フローベースアグリゲーションとラウンドロビンアグリゲーションが存在する．

フローベースアグリゲーションは，エンドノード間のフロー単位で各パスへ割り当てるアルゴリズムである．本アルゴリズムは，各フローに 1 本のパスを割り当てるため，フレーム順序逆転が発生しないというメリットがあるが，各フローの使用可能帯域はリンクの波長帯域に制限されてしまう．

一方，ラウンドロビンアグリゲーションはフレーム単位で各パスへ割り当てるアルゴリズムである．エッジスイッチに入力されたフレームはラウンドロビンで各パスへ出力される．本アルゴリズムは複数パス経路の帯域を最大限使用することができるメリットがあるが，フローごとにフレーム順序逆転が発生する問題がある．フレームの順序逆転は，フレームの経由する経路間の遅延差により発生する．

図 2 に，TCP 通信を 100Mbps リンクのラウンドロビンアグリゲーションにより実現し，1 リンクのみ遅延の値を変化させた場合のスループットを示す [4]．フレームの順序逆転により TCP/IP の再送制御が頻繁に発生し，スループットが減少していることが分かる．さらに，1Gbps，10Gbps の広帯域リンクを使用した場合のスループットは図 2 に示すスループットより低下することが予想される．

スループット低下の原因となる複数経路間の遅延差を低減する手法として，PLB（Packet-based Lane Bundling）[5] が提案されている．フレームに送信時間を挿入することにより，エッジノード間での複数経路におけるフレーム伝送遅延差を測定でき，遅延が最大のフレームに他フレームを受信ノードにおけるバッファリングより合わせることにより，経路間の遅延差を低減することが可能である．しかし，PLB およびバッファは各ノードの NIC（Network Interface Card）にハードウェア実装されるため，導入にコストがかかるという問題がある．

2.2 複数経路発見手法

既存の複数経路発見手法として，パス要求の発生毎にシグナリングを用いた方式が提案されている [6]．あて先までのパスを最短経路ではなく，波長の使用状況やリンク距離を考慮した経路を探索し，確立することにより，パス確立時のブロック率を低減することが可能である．

本方式は，各ノード間のリンク距離情報として VDC（Virtual Distance Cost），帯域使用情報として VLC（Virtual Link Cost）の 2 個のパラメータを設定し，パラメータ情報収集には

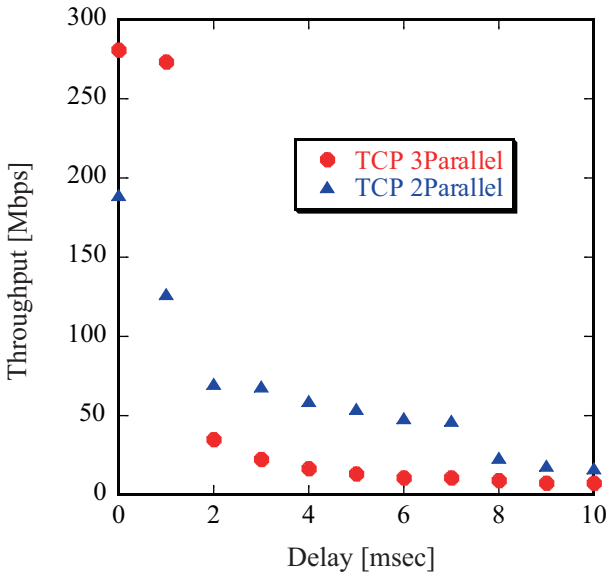

図 2　ラウンドロビンアグリゲーションにおけるスループット

SEARCH 信号を使用する．送信元ノードは SEARCH 信号を隣接ノードに送信し，経由リンクのパラメータを信号に格納することにより，あて先ノードで各経路のパラメータを収集することが可能である．あて先ノードは 1 個目の SEARCH 信号を受信してから一定時間待機し，待機時間内に到着した SEARCH 信号に格納されたパラメータを比較して最適経路を決定する．パスは，最適経路決定後にあて先ノードより RESV 信号を送信することにより，受信ノードでスイッチの設定を行い確立する．

シグナリング方式における SEND 信号は，送信元からあて先までの片方向通信のみのため，ネットワーク内における帯域圧迫は双方向と比較して低減することが可能である．しかし，あて先ノードで経路選択をする本方式は，あて先からの RESV 送信時においてパス確立を失敗した場合，送信元で再度 SEARCH 信号を送信する必要がある．本論文の対象とする複数経路の同時パス確立は，単経路のパス確立と比較してパスのブロック率が高くなることが想定されるため，本方式では SEARCH 信号の再送信率が高くなってしまう．また，片方向通信による遅延測定はノード間の時間同期が必要だが，大規模ネットワークでは全ノードの時間同期をとることは困難である．

そのため，送信元ノードで経路情報を収集可能であり，ノード間の時間同期の必要がない双方向通信による複数経路探索手法が必要である．

3.　次世代広域イーサネット

図 3 に，本論文がの対象とする次世代広域イーサネットアーキテクチャを示す．次世代広域イーサネットでは，ユーザの要求帯域がリンク帯域以上となる場合は，ネットワーク内の複数の Gb イーサネットリンクのアグリゲーションを行い，要求帯域を満たす．GMPLS（Generalized Multi-Protocol Label Switching）で確立した複数経路をアグリゲーションし，エンドユーザ間においてオーバレイネットワーク上で仮想的に大容量のデータ伝送を可能とする．オーバレイネットワークアーキテクチャを用いることにより，エンドユーザはネットワーク内のトポロジやトラヒック，そしてリンク帯域制限を意識せずに，要求 QoS 保証を満たすコネクションを確立することが可能となる．使用帯域に変更が生じた場合は，ユーザが自由に帯域設定変更を行う．ネットワーク内では設定された帯域を満たす複数経路を自動的に再選択し，GMPLS によりパス確立を行う．

例えば，あるユーザが 1Gbps のパスを 3 本アグリゲーションして 3Gbps の仮想伝送路を使用していたとする．その後，トラヒック増加に伴い 5Gbps の伝送路が必要になった場合，所望帯域や要求 QoS を指定することにより，その時点でのネットワーク状況（遅延や帯域使用量）に応じて 5 本の 1Gbps パスを自動的に確立することが可能である．

次世代広域イーサネットを実現するためには，ユーザの要求保証を満たす複数の経路候補を発見する手段が必要不可欠である．

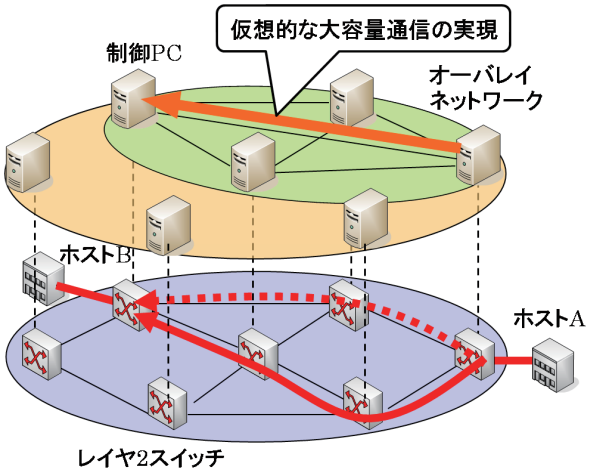

図 3　次世代広域イーサネットアーキテクチャ

図 4 に，GMPLS により制御された広域イーサネット網のネットワークアーキテクチャを示す．ネットワークは，制御プレーンとデータプレーンから構成されており，企業 LAN と接続しているエッジノードおよびフレームを中継するコアスイッチが存在する．

各プレーンの役割を以下に示す．

- 制御プレーン
  - サービスユーザプロファイル機能
  - サービス制御機能
  - ネットワーク管理
  - スイッチ設定
- データプレーン
  - 実データ転送

企業は広域イーサネット網におけるエッジノードへ接続し，広域イーサネット網より提供される仮想伝送路を利用し接続先の企業との通信が可能となる．GMPLS により制御された広域イーサネット網のエッジノード間で確立するパスを L2-LSP（Layer2-Label Switched Path）と呼び，RSVP（Resource reSerVation Protocol）により L2-LSP が確立される．接続要求が発生した

場合，送信エッジノードは OSPF の経路情報を基に宛先エッジノードまでの経路を決定する．次に RSVP により，経路上の資源予約と VLAN の設定が行われ，L2-LSP を確立する．GMPLS により自動的に L2-LSP を確立可能であるため，管理者は従来の VLAN 設定に必要な処理を削減可能となり，企業などのユーザは不必要な L2-LSP を削減し，必要なときに必要な帯域を確立することができる．

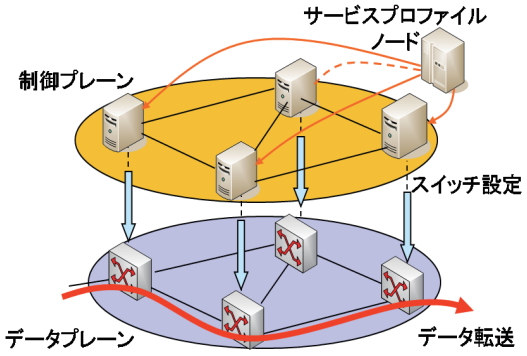

図 4　広域イーサネットアーキテクチャ

4.　複数パス経路の同時発見手法

4.1　概　　要

次世代広域イーサネットにおいて，複数経路の遅延情報を測定するアルゴリズムを提案する．本アルゴリズムを適用したネットワークにおいてエンドユーザによるパス確立要求が発生すると，送信元は遅延測定用のイーサネットフレームをフラッディングし，あて先までの複数経路において同時に RTT(Round Trip Time) を測定する．遅延測定用のイーサネットフレームは，従来のイーサネットフレームを拡張し，あて先を指定したフラッディングを行うことができるようにする．送信元は収集した遅延情報を元に，経路間での遅延差を最小にする経路組を選択することが可能となる．以下に拡張イーサネットフレームフォーマットおよび複数経路発見アルゴリズムを示す．

4.2　拡張イーサネットフレームフォーマット

従来のイーサネットフレームでフラッディングを行う場合，あて先フィールドにブロードキャストアドレスである FF:FF:FF:FF:FF:FF を格納するため，あて先を指定したフラッディングを行うことができないという問題がある．そのため，あて先アドレスを格納できるフレームの定義を行う必要がある．

図 5 に，今回定義した次世代広域イーサネット網のフレームフォーマットを示す．既存のイーサネット技術を利用可能にするため，拡張フレームフォーマットは従来のイーサネットフレームフォーマットからの変更を極力少なくした．ヘッダのフィールド構成は [7], [8] に準拠する．

ヘッダは以下の 6 個のフィールドにより構成される．1)D-MAC(48bit)，2)S-MAC(48bit)，3)VPI(32bit): Type 16bit + VPI 16bit，4)I-TAG(32bit): Type 16bit + Instance ID: 16bit，5) ペイロードタイプ（16bit），6) 拡張ヘッダ（64bit）: TTL(Time to Live) 8bit + D-MAC (48bit) + Sequence Number (7bit) + Flag (1bit).

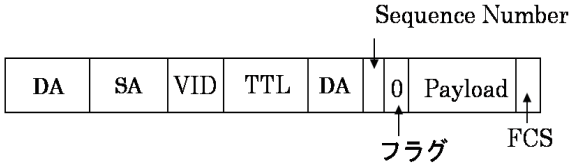

図 5　拡張イーサネットフレームフォーマット

拡張ヘッダにおいて，TTL はイーサネットフレームの網内での永久ループ回避用，Sequence number は送信フレームの特定用，Flag は送信フレームと返信フレームの判別用である．また，D-MAC フィールドにブロードキャストアドレスを格納する場合，VPI フィールドの Type を 0x0800 に設定すると IP，0x0806 にすると ARP，0x8100 にすると 802.1Q VLAN と判別される．提案方式のイーサネットフレームは，IP や ARP と区別するため Type を 0x8100 に設定する．

4.3　フラッディングによる複数経路発見手法

図 6 および図 7 に，フラッディングによる複数経路発見アルゴリズムを示す．

（1） パス確立要求が発生．

（2） 送信元スイッチは，イーサネットフレームヘッダの D-MAC にブロードキャストアドレス，ペイロードタイプに 0x8100，拡張ヘッダ内の D-MAC にあて先アドレス，Flag bit に 0，各出力ポートごとに異なる Sequence Num，ペイロードに送信ポート番号を格納し，フレームのブロードキャストを行う．(図 6（a)）

（3） 拡張イーサネットフレームを受信した中継ノードは拡張ヘッダ内の TTL および D-MAC を確認し，TTL ≠ 0 または D-MAC が自身あてでない場合は受信ポート以外の全ポートにイーサネットフレームの TTL を減算して複製し，転送する．転送する際，ペイロードの中身に受信ポートおよび送信ポート番号を格納する．(図 6（b)）

（4） あて先ノードは，返信フレームの D-MAC に送信元 MAC アドレス，拡張ヘッダ内の Flag bit に 1，ペイロードにポート情報を格納し，ブロードキャストフレームを受信したポートへ送信する．(図 7（c)）

（5） 中継ノードは，Flag bit=1 であることを確認すると，ペイロード内の出力ポート情報に沿ってフレームを転送する．

（6） 送信元ノードは受信した返信フレームの Sequence Number およびペイロード内のポート情報を確認することにより，経路情報とその RTT を算出することが可能となる．(図 7（d)）

図 6（a）において送信元ノード 1 からあて先ノード 7 までのデータ送信要求が発生した場合を考える．送信元ノードは拡張イーサネットフレームの D-MAC アドレスにブロードキャストアドレスである FF:FF:FF:FF:FF:FF，S-MAC アドレスにノード 1 の MAC アドレス，type に 0x8100，拡張フィールドの TTL に 256，D-MAC アドレスにノード 7 の MAC アドレ

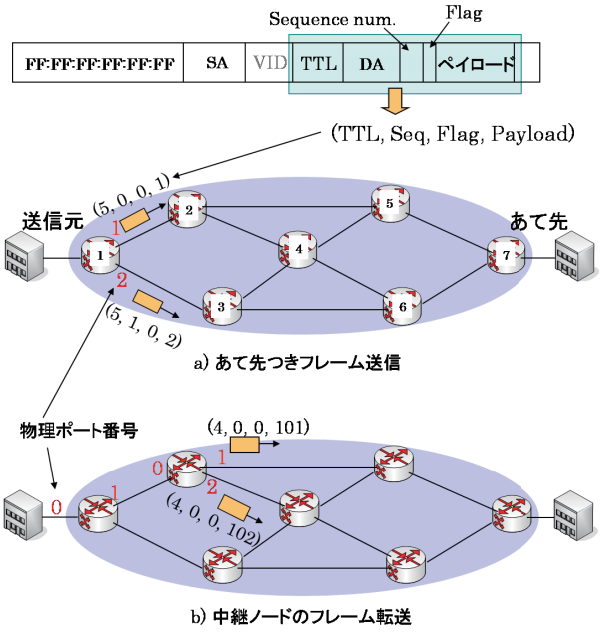

図 6　フラッディングによる複数経路発見アルゴリズム

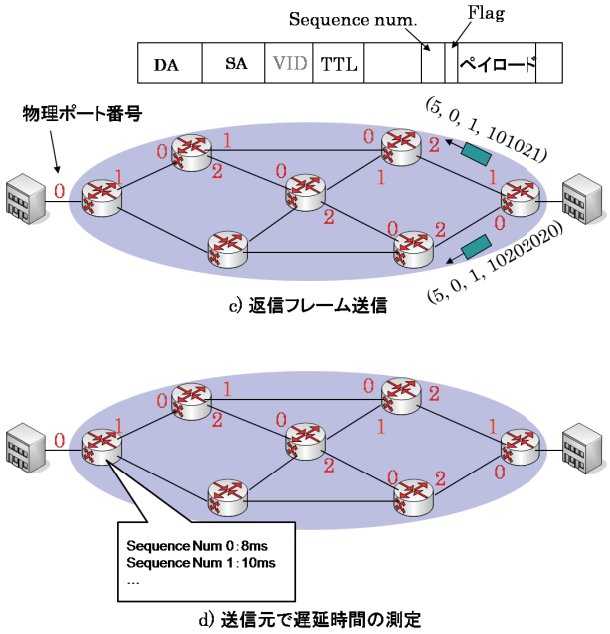

図 7　フラッディングによる複数経路発見アルゴリズム

ス，Flag に 0 を格納する．次に，送信元は全ポートからあて先つきブロードキャストフレームを送信するが，Sequence Num およびペイロードの上位 2Byte に送信ポート番号（1 または 2）を格納してから送信を行う．この時，フレームの送信時間をテーブルに記録する．拡張イーサネットフレームを受信した中継ノード 1 およびノード 2 は，まず TTL=0 ではないこと，そして Flag=0 であることを確認し，ペイロードに受信ポートを追加格納する．その後，ノード 1 および 2 は，受信フレームの TTL を減算，ペイロードに送信ポートを追加格納し，受信ポート以外の全ポートに転送する．あて先ノード 7 までの中継ノードは以上の手順を繰り返す．結果，あて先ノード 7 で受信したフレームのペイロードは，経由したノードが 1，2，5，7 の場合は 101021 となる．この場合，あて先ノード 7 は拡張イーサネットフレームの D-MAC アドレスにノード 1 の MAC アドレス，S-MAC アドレスにノード 7 の MAC アドレス，拡張フィールドの TTL に 256，Flag に 1，ペイロードに 101021 を格納する．中継ノードは（256 − 受信フレームの TTL 値）$*2+1$ を計算することにより，フレームの転送ポート番号をペイロード内より探索する．TTL=254，ペイロード=101021 の場合，転送ポートは 0 となる．Flag=1 のイーサネットフレームを受信した送信元ノード 1 は，送信時間とフレーム受信時間の差分により RTT を測定する．

ペイロードに格納可能なポート数は，イーサネットフレームのペイロードが 1500Byte，スイッチの最大ポート数が 4096 個であることから 1 ポートあたり 2Byte を使用するため，750 ポート，325 ノード分の経路情報を格納することが可能である．本アルゴリズムを適用することにより，送信元において複数経路と遅延時間の情報を保持することができ，経路間での遅延差を考慮した適切な経路選択が可能となる．また，双方向通信型であるため，パス確立失敗時は送信元において保持する経路情報の中から別経路を選択し，パス予約を行うことが可能である．そのため，ブロードキャストの再送信を行う必要がないというメリットがある．

5. 性 能 評 価

本提案アルゴリズムをソフトウェアスイッチに実装し，実験により検証する．図 8 に実験ネットワークを示す．実験ネットワークにおけるノード数は 7 個，リンク数は 10 本，リンク帯域は 1Gbps であり，送信元ノードおよびあて先ノードは固定とする．

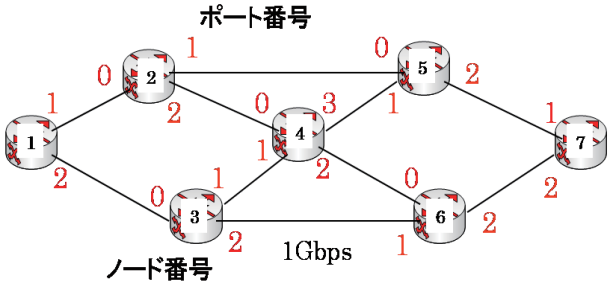

図 8　実験ネットワーク

以上のネットワークにおいて本アルゴリズムを適用し，各経路において得られた遅延情報を表 1 に示す．双方向型通信を用いた各経路の経路情報および実遅延時間の測定を実現し，送信元ノードで経路情報の保持が可能であることが確認された．本アルゴリズムはフラッディングを用いる手法のため，ネットワーク規模が大きくなると，あて先付きブロードキャストフレームの送信によるオーバーヘッドが想定される．また，送信元ノードはフラッディング後に返信フレームの受信まで待機するが，待機する時間が長い場合は他ノードのパス確立によりパス確立をブロックされてしまう可能性がある．そのため，一度に測定するパス数の検討も必要である．

表 1　収集経路情報に対する遅延時間の比較

経路情報	遅延時間 (ms)
1, 2, 5, 7	3.8
1, 3, 6, 7	3.5
1, 2, 4, 5, 7	4.1
1, 3, 4, 6, 7	5.8

6. ま と め

広域イーサネットにおいて，リンクの波長帯域に依存しない広帯域な伝送路の確立を実現するために並列伝送技術を用いた．しかし，並列伝送で使用パス間での遅延差が大きいとスループットが低下してしまうため，各経路の実遅延時間を測定する必要があった．そこで本論文では，拡張イーサネットフレームにあて先ノードの MAC アドレスを格納してフラッディングを行い，あて先までの複数経路におけるラウンドトリップタイムを計測する手法を提案した．本システムをソフトウェアスイッチに実装し，各経路における遅延時間の測定を行った．本システムにより，双方向通信型の遅延測定および経路情報収集を実現した．今後の課題としては，遅延以外の要求 QoS にも対応した経路発見手法の提案を行うことである．

謝　　辞

本研究の一部は，独立行政法人情報通信研究機構 (NICT) の委託研究「λアクセス技術の研究開発」の成果である．関係者各位に深謝する．

文　　献

[1] IEEE P802.3ba 40Gb/s and 100Gb/s Ethernet Task Force, http://www.ieee802.org/3/ba/

[2] IEEE Std 802.3ad-2000, “Amendment to carrier sense multipleaccess with collision detection (CSMA/CD) access-method and physical layer specifications aggregation of multiplelink segments”

[3] 宮城 洋之, 岡崎 裕介, 碓井 亮太, 荒川 豊, 岡本 聡, 山中 直明, “ パラレル転送を適用したグリッドコンピューティング特性の評価,’ ’ 電子情報通信学会技術研究報告, Vol. PN2007-93, pp. 111-116, March 2008.

[4] 山田 翔太, 寺澤 緑, 清水 翔, 石井 大介, 岡本 聡, 山中 直明, “ 大容量データ転送アプリケーションの実現に向けた TCP over SCTP パラレルネットワーキングおよび並列経路選択手法の検討,’ ’ 電子情報通信学会技術研究報告, Vol. PN2009-14, pp. 19-24, August 2009.

[5] S. Kobayashi, Y. Yamada, K. Hisadome, O. Kamatani, O. Ishida, ”Scalable Parallel Interface for Terabit LAN,” IEICE Trans. Commun., vol.E92-B, no.10, pp.3015-3021, Oct. 2009.

[6] 本間 潤一郎, 伊藤 隆範, 石井 大介, 山中 直明, “ WDM ネットワークにおけるシグナリングを用いた経路/波長探索方式,’ ’ 電子情報通信学会技術研究報告, Vol. NS2004-309, pp. 321-324, March 2005.

[7] 岡本 聡, 菊田 洸, 西田 昌弘, 石井 大介, 荒川 豊, 山中 直明, “ 次世代広域レイヤ 2 網実現に向けたプロトコル実装と実証実験,’ ’ 電子情報通信学会技術研究報告, Vol. OCS2008-108, pp. 7-12, January 2008.

[8] S. Yamada, Y. Okazaki, M. Terasawa, S. Shimizu, D. Ishii, S. Okamoto, and N. Yamanaka, “A Study of TCP over SCTP Parallel Networking and Parallel Route Selection Approach for Mass Data Transfer Applications” ONDM2010 Feb.2008.