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.

データセンターにおける輻輳回避のための ルーティング用論理トポロジ構築手法

下間　雄太†　　大下　裕一†　　村田　正幸†

† 大阪大学 大学院情報科学研究科　〒 565–0871 大阪府吹田市山田丘 1–5
E-mail: †{y-shimotsuma,y-ohsita,murata}@ist.osaka-u.ac.jp

あらまし　データセンターでは、サーバー間が連携して多量のデータを処理することにより、様々なオンラインサービスを提供している。サーバー間のトラヒックパターンは 1 秒以下といった短い間隔で著しく変化する。そのため、このような急激なトラヒックパターンの変化に対応しつつ、サーバー間に十分な帯域を確保可能な手法が必要とされている。トラヒックパターンが変化する状況下で、サーバー間に十分な帯域を確保するためには、トラヒック状況に応じた経路を設定することが必要となる。しかしながら、短い間隔で発生するトラヒック変動に対応した適切な経路を集中制御で設定することは、トラヒック情報の収集・経路計算の負荷の問題があり、困難である。この問題に対して、本稿では、ネットワーク内の各スイッチがデータセンターネットワークのサブセットとなる複数の論理トポロジの情報を保持し、論理トポロジの情報と自身の出力リンクの輻輳状況をもとに、トラヒックの転送に用いる論理トポロジを自律的に決めることにより、トラヒックパターン変動後も十分な帯域を確保することが可能な経路制御手法を提案する。また、本稿では、この経路制御手法に用いる論理トポロジを特定のリンクに負荷が集中しないように構築する手法も提案する。そして、シミュレーション評価により、提案手法が、トラヒックが存在する全サーバーペア間に十分な帯域を確保することができることを示す。
キーワード　データセンター、論理トポロジ、自律的ルーティング

Configuration of Robust Logical Topologies for Avoiding Congestions in Data Centers

Yuta SHIMOTSUMA†, Yuichi OHSITA†, and Masayuki MURATA†

† Graduate School of Information Science and Technology, Osaka University
1–5 Yamadaoka, Suita, Osaka 565–0871, Japan
E-mail: †{y-shimotsuma,y-ohsita,murata}@ist.osaka-u.ac.jp

Abstract　Servers in a data center cooperate with each other to handle a large data, and the traffic pattern between servers changes significantly within a second. To keep the performance of the data center, sufficiently large bandwidth should be provided between any server pairs even in such frequent and significant traffic changes. Because of too frequent traffic changes, the central control cannot be applied to handle such traffic changes. In this paper, we propose a method to provide a sufficiently large bandwidth without central control. In this method, we preconfigure the multiple logical topologies which are the subset of the physical network. Then, each node selects the logical topologies used to send the traffic based on the traffic information monitored directly by it. In this paper, we also propose a method to configure the logical topologies for our routing method. In this paper, we demonstrate that our method can provide a larger bandwidth than ECMP.
Key words　data center, logical topologies, autonomous routing

1. はじめに

近年、クラウドコンピューティングなど、データセンターを介して提供されるサービスが増加し、データセンター内で処理されるデータも著しく増加している。データセンターでは、多くのサーバーが連携することにより多量のデータの処理を行っ

ている。サーバー間を流れるトラヒックは、サーバー間での連携により処理されるデータによって異なり、1 秒以下といった短い時間で著しく変化する [1-3]。そのため、そのような著しいトラヒックパターンの変化が起きた場合であっても、ネットワークの一部に負荷が集中することなく、通信を行っているサーバー間に十分な帯域を確保する経路を設定することが必要となる。

このようなトラヒック状況に応じて、動的にネットワーク内の経路を変更することは、経路の計算を行うサーバーを配置することにより、集中的に行う手法の検討が進められてきた [3, 4]。これらの手法では、データセンター内の各サーバー間あるいはサーバーラック間に流れるトラヒックの観測データを経路計算サーバーが定期的に収集し、そのトラヒック情報をもとに、適切な経路を計算し、ネットワーク内の各機器の設定を変更する。上述のようにデータセンター内では 1 秒といった短い間隔でトラヒック変動が発生するため、トラヒック変動に追随した経路制御を行うには、制御間隔も 1 秒以下と短くすることが必要となる。しかしながら、短い間隔での集中制御による経路変更は、短い間隔でのトラヒック情報収集がネットワークにかける負荷や、適切な経路を計算するのにかかる計算時間の問題があり、困難である。

そのため、データセンターにおけるトラヒック変動に追随した経路制御は、各機器が自身が把握可能な情報をもとに自律的に行うことが望ましいと考えられる。文献 [5] では、FatTree 型のデータセンターネットワークにおいて、各機器がより根に近いスイッチにトラヒックを転送する際に、ランダムにトラヒックの転送先のスイッチを選択することにより、負荷分散を行う手法が提案されている。しかしながら、この手法は、FatTree 型のネットワークにしか適用できないという問題や、根から葉に転送される際には、輻輳箇所を迂回する経路を確保できないといった問題がある。

本稿では、任意のネットワーク構造をもつデータセンターネットワークにおいて適用可能な、各機器が自律的に輻輳箇所を迂回し、サーバー間に十分な通信帯域を確保することが可能な経路制御手法を提案する。本手法では、ISP を対象とし、故障箇所を各ノードの判断によって迂回する経路を確保する手法として提案されたマルチトポロジルーティング [6, 7] を輻輳箇所の解消に用いる。マルチトポロジルーティングでは、ネットワークのサブセットとなる複数の論理トポロジをあらかじめ構成する。そして、故障の発生を検出したノードは、故障箇所が含まれない論理トポロジを選択し、パケットに選択した論理トポロジの情報を書き込んだ上で転送することにより、故障箇所の迂回を行う。

本稿では、マルチトポロジルーティングを輻輳の解消に用いるために、(1) 各機器が、転送に用いることができる論理トポロジにおいて、もっとも宛先までのホップ数が短いものを候補論理トポロジとし、候補論理トポロジのうち、論理トポロジに含まれる自身が接続しているリンクの負荷が最も低いものを自律的に選択することによる負荷分散、(2) 各機器が把握した十分な帯域が確保できず、転送に用いることができない論理トポロジの情報をパケットに埋め込んで転送することによる、高負荷リンクの迂回という拡張を行った。

また、本稿では、上記のルーティング手法に適した論理トポロジの設計方法も提案する。提案手法では、故障・輻輳箇所を迂回した際にも、特定のリンクにトラヒックが集中しないように、論理トポロジ群の設計を行う。

シミュレーション評価により、輻輳発生時にも、局所的な情報で当該箇所を迂回可能であり、さらに、特定のリンクにトラヒックが集中することなく、全サーバー間で高い通信帯域を確保できることを明らかにする。

本稿の構成は次のとおりである。2. 章では、データセンターにおける論理トポロジを用いた自律的な経路制御手法を提案し、3. 章では論理トポロジの構築手法を提案する。4. 章で提案手法の性能評価を行い、最後に 5. 章で本稿のまとめについて述べる。

2. データセンターにおける論理トポロジを用いた自律的ルーティング

本稿では、あらかじめ複数の論理トポロジを構成する。各論理トポロジは、データセンターネットワーク内の全ノードと、ネットワーク内のリンクのサブセットが含まれる。そのため、全リンクが使用可能であれば、いずれの論理トポロジ上の経路にトラヒックを流しても宛先まで到着可能であり、高負荷で利用できないリンクが存在した場合は、当該リンクを含まない論理トポロジを用いることにより、高負荷リンクを迂回し、宛先までトラヒックを送信可能な経路を見つけることができる。本稿では、このような複数の論理トポロジの情報を各スイッチが共有し、自律的に転送に用いる論理トポロジを選択することにより、輻輳箇所を迂回しつつ負荷分散を実現し、全サーバー間に十分な帯域を提供する経路制御手法を提案する。

2.1 各スイッチにおける制御

本制御では、以下の手順により、各スイッチは転送先の論理トポロジを選択する。

（1） 利用可能な論理トポロジの集合を得る。

（2） 利用可能な論理トポロジのうち、当該論理トポロジを用いた際の宛先ノードまでのホップ数が最短の論理トポロジを候補論理トポロジとする。

（3） 候補論理トポロジのうち、当該論理トポロジに含まれる自身が接続するリンクの負荷がもっとも小さい論理トポロジを選択する。

以下に各手順の詳細について述べる。

2.1.1 利用可能な論理トポロジの集合の取得

上記の手順のうち、利用可能な論理トポロジの集合は、全論理トポロジから、(1) 到着したパケットのヘッダに埋め込まれた情報から取得した、他の機器で検出された利用不可能な論理トポロジ、(2) 自分自身が接続しているリンクのうち、高負荷や故障のため利用不可なリンクを検出した場合は、宛先までの転送先経路に当該リンクを含む論理トポロジを除外したものとする。また、自身が高負荷や故障のため利用不可なリンクを検出した場合は、利用不可となった論理トポロジの情報をパケッ

トのヘッダ部に書き入れ、転送先で当該論理トポロジが用いられることを防ぐ。これにより、高負荷なリンクや故障箇所を迂回した経路を自律的に設定することが可能となる。

2.1.2 候補論理トポロジの選択

本手法では、宛先までのホップ数が最も少ない論理トポロジを候補論理トポロジとする。宛先ごとに各論理トポロジを用いた場合の最短ホップ経路を計算し、論理トポロジとホップ数の対応表をあらかじめ作成しておくことにより、瞬時にホップ数がもっとも少ない論理トポロジの集合を取得することが可能である。ホップ数の最も少ない論理トポロジのみを候補論理トポロジとすることにより、ホップ数の増大により、より多くの機器の帯域を消費してしまうといった問題を解消することができる。

2.1.3 論理トポロジの選択

本手法では、候補論理トポロジのうち、当該論理トポロジに含まれる自身が接続するリンクの負荷がもっとも小さい論理トポロジを選択する。リンクの負荷は、ポートごとに観測されたトラヒック統計情報や、待ち行列の長さを取得することにより得る。リンクの負荷が低い論理トポロジを選択することにより、トラヒックの集中を避け、負荷を分散させることができる。

2.2 論理トポロジの利用可否情報のパケットの埋め込み方法

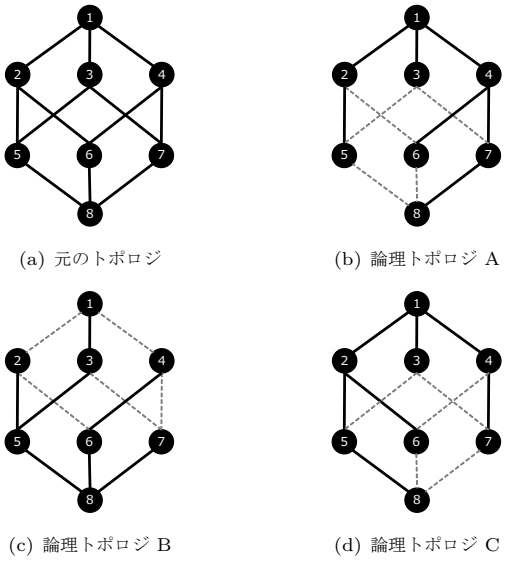
論理トポロジ数が n の場合、n ビットの情報で全論理トポロジの利用の可否を表すことができる。この n ビットの情報をパケットに書き込むことにより、他のスイッチにも自身が検出した利用不可な論理トポロジの情報を伝達することができる。

n ビットの論理トポロジ利用可否情報の埋め込み先のひとつの例として、IPv6 ヘッダ内の宛先アドレスフィールドが考えられる。データセンター内は、単一の管理者が管理しているため、内部の機器の IP アドレスは任意のポリシーで設定することができる。また、IPv6 のアドレス長は 128bit であり、そのうち n ビット分を論理トポロジの利用の可否を表す情報として用いる。4. 章の評価によると、論理トポロジ数が少なくても高負荷リンクを迂回し、全サーバー間に十分な帯域を確保することができるため、n は十分小さく、このような論理トポロジの利用可否情報の埋め込みを行った場合であっても、残りの $128 - n$ ビットでデータセンター内の機器を識別することが可能だと考えられる。

2.3 動 作 例

図 1 は、8 ノード、12 リンクのトポロジおよびそのトポロジから構築された論理トポロジの例を、図 2 に、当該論理トポロジを用いたトラヒック転送の例を示す。

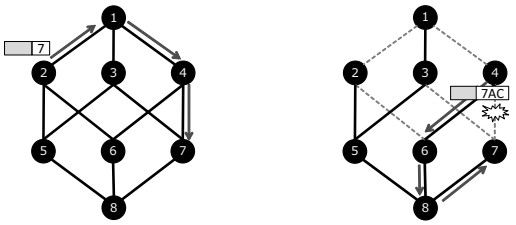
図 2(a) の場合、いずれのリンクの負荷も低く、利用可能であるため、ノード 2 からノード 7 へは、最短ホップ経路である $2 \rightarrow 1 \rightarrow 4 \rightarrow 7$ の順でパケットを転送する。それに対して、ノード 4–7 間のリンクが高負荷で利用できない場合は、図 2(b) に示すように、ノード 4 は使用できない論理トポロジ A、C の情報をパケットに付け、論理トポロジ B を使用し、ノード 6 にパケットを転送を行う。そして、その後、論理トポロジ A、C を使わずにパケットの転送を行うことで高負荷リンクの回避を行っている。



(a) 元のトポロジ　(b) 論理トポロジ A

(c) 論理トポロジ B　(d) 論理トポロジ C

図 1　元のトポロジおよび論理トポロジの例



(a) 全リンクの負荷が低い場合　(b) ノード 4–7 間のリンクが高負荷の場合

図 2　リンク故障が発生した場合の論理トポロジを用いた自律的ルーティングの例

3. 論理トポロジの設計

本稿の手法では、論理トポロジの組み合わせにより、故障・輻輳発生時に確保可能な迂回経路の数や迂回先でトラヒックが集中するかどうかが決まる。そのため、可能な限り多くの迂回経路を確保できるようにしつつ、迂回先でトラヒックが集中しないように論理トポロジの設計を行う必要がある。しかしながら、論理トポロジのすべての組み合わせを調べ、最適な組み合わせを選択することは、候補となる組み合わせが膨大となるため困難である。そこで、本節では、発見的手法により論理トポロジの組み合わせを求める。本手法では、論理トポロジを 1 つずつ設計・追加を行う。その際に、候補となる論理トポロジに対し、その論理トポロジの適切さを示す指標を計算し、最もよい論理トポロジを選択する。以降、候補となる論理トポロジの構成、論理トポロジの適切さを示す指標について述べる。

3.1 候補となるトポロジ

各論理トポロジは、以下の性質を持つことが望まれる。(1) 全スイッチ間を接続したものであること。(2) 高負荷リンクが発生した際に当該論理トポロジが利用不可となる可能性が低いこと。本稿の手法では、高負荷リンクが発生した場合、宛先までの経路上に当該リンクが含まれる論理トポロジを利用不可と

することにより、高負荷なリンクを迂回した経路を確保する。しかしながら、その結果、多くの論理トポロジが利用不可となることがあれば、確保できる迂回経路数が著しく減ってしまい、その後に発生した別の高負荷なリンクを迂回した経路を確保できない。

本稿では、上記の性質を持つように、各候補トポロジは全スイッチを接続するように構成した木構造として構築する。木構造として構築することにより、全スイッチを接続しつつも、各論理トポロジに含まれるリンクの数を最小とすることができる。リンクの数を最小とすることにより、高負荷なリンクが発生した際に、そのリンクが各論理トポロジに含まれる確率を低く抑えることができる。

以下に、N ノードのデータセンターネットワークトポロジが与えられた際に、候補となる論理トポロジの集合 C を構築する手順を示す。

初 期 状 態

データセンターネットワーク内の各ノードを木の根ノードとして指定した論理トポロジを構築する。その結果、根ノードのみを含んだ論理トポロジが N 個、候補論理トポロジ C 内に存在する状態となる。

論理トポロジの成長

C に含まれる各論理トポロジ G に、以下の条件を満たすノード n を追加することにより、論理トポロジを成長させる。

- n は論理トポロジ G に含まれていない
- 論理トポロジ G 内のいずれかのノードと直接接続している

n が直接接続している G 内のノードが複数存在する場合は、そのいずれと接続した場合も、候補論理トポロジとし、候補論理トポロジの集合 C に追加される。その結果、n が直接接続している G 内のノード数分ほど、C に含まれる候補論理トポロジは増える。論理トポロジの成長は、C に含まれるすべての論理トポロジが、データセンターネットワーク内の全ノードを含むまで繰り返し行われる。

以上の方法により、データセンターネットワークのサブセットで構成した木構造はすべて網羅することができる。

3.2 指　　標

本論理トポロジ設計手法において、追加する論理トポロジを選択する際には、以下の点について比較し、適切な論理トポロジを選択する必要がある。(1) 当該論理トポロジを追加することにより確保可能な迂回経路数。迂回経路をより多く確保することにより、高負荷のリンクが生じた場合にも、当該リンクを迂回し、広帯域の通信経路を確保できる可能性を高めることができる。(2) 各ノードが利用不可となった場合の影響の大きさ。特定のノードがほとんどの論理トポロジで多数のサーバー間トラヒックの中継地点となっている場合、当該ノードの故障や当該ノードのリンクが高負荷となると、利用可能な論理トポロジが著しく少なくなり、迂回経路を確保できなくなる。そのため、特定のノードが利用不可となった場合の影響が大きくなることは避ける必要がある。 (3) 迂回時のトラヒックの分散。迂回先でトラヒックが集中し、新たな輻輳が発生することを防止できることが求められる。

以降、上記の要件を満たすものを選択するための指標について述べる。

3.2.1 迂回経路数

これまでに構築した論理トポロジ上、追加する候補の論理トポロジ上で、各スイッチ間の経路を求める。そして、その経路の集合のうち、リンクを共有しない、互いに疎な経路の本数を数える。全スイッチ間の経路集合に対して計算された互いに疎な経路の本数の最小値を当該論理トポロジが追加された際に確保可能な迂回経路数とする。このように計算された、迂回経路数が多くなるような論理トポロジを追加することにより、いずれのリンクが使用不可となった場合にも、迂回経路を用い、十分な帯域の通信路を見つけることが可能となる。

3.2.2 各ノードの故障の影響

トラヒックを中継する可能性の高いノードほど、故障や当該ノードに関するリンクが高負荷となった影響が大きい。この影響を調べる指標として、本稿では、ノード平均媒介中心性を定義する。ノード n のノード平均媒介中心性 B_n^{node} は以下のように定義される。

$$B_n^{\mathrm{node}} = \frac{1}{|G|}\sum_{g\in G}\left(\sum_{s,d\in S}\frac{R_g^{\mathrm{node}}(s,n,d)}{R_g(s,d)}\right)$$

ただし、G は論理トポロジの集合、S はノードの集合、$R_g(s,d)$ は論理トポロジ g 上の s–d 間の最短ホップ経路の本数、$R_g^{\mathrm{node}}(s,n,d)$ は論理トポロジ g 上の最短ホップ経路のうち、n を通るものの本数である。

ノード平均媒介中心性は、全ノード間にトラヒック量 1 のトラヒックが流れ、そのトラヒックを流す論理トポロジをランダムに選択した場合に、当該ノードを経由するトラヒック量の期待値を意味する。ノード平均媒介中心性が大きいほど、当該ノードを経由するトラヒックが多く、当該ノードを用いた転送が故障や高負荷により困難となった場合に、転送先の候補数が減るトラヒックが多いことを示す。そのため、特定のノードの故障や過負荷がトラヒックの収容に与える影響を少なくするために、ノード平均媒介中心性が大きなノードを生じないように論理トポロジを構成することが望ましい。そこで、本稿では、各論理トポロジの候補に対して、当該論理トポロジを加えた際の最大のノード平均媒介中心性を計算し、その値が小さくなる論理トポロジを選択する。

3.2.3 迂回時のトラヒックの分散

高負荷のリンクを迂回するように論理トポロジが選択された際に、トラヒックが集中するかどうかについて判断する指標として、本稿では、リンク平均媒介中心性を定義する。リンク l のリンク平均媒介中心性 B_l^{link} は以下で定義される。

$$B_l^{\mathrm{link}} = \frac{1}{|G|}\sum_{g\in G}\left(\sum_{s,d\in S}\frac{R_g^{\mathrm{link}}(s,l,d)}{R_g(s,d)}\right)$$

ただし、G は論理トポロジの集合、S はノードの集合、$R_g(s,d)$ は論理トポロジ g 上の s–d 間の最短ホップ経路の本数、$R_g^{\mathrm{link}}(s,l,d)$

は論理トポロジ g 上の最短ホップ経路のうち、リンク l を通るものの本数である。

リンク平均媒介中心性は、全ノード間にトラヒック量 1 のトラヒックが流れ、そのトラヒックを流す論理トポロジをランダムに選択した場合に、当該リンクを経由するトラヒック量の期待値を意味する。リンク平均媒介中心性が高いリンクほど、高負荷のリンクが生じ、当該リンクを迂回するような論理トポロジが選択された際に、トラヒックが集中する確率が高いリンクと考えられる。そのため、本稿では、各論理トポロジの候補に対して、当該論理トポロジを加えた際の最大のリンク平均媒介中心性を計算し、その値が小さくなる論理トポロジを選択する。

3.2.4 指標間の優先順位

本稿の論理トポロジ設計手法では、候補となる論理トポロジの中から、追加する論理トポロジを選択する際に、上記の指標を計算することにより、適切な論理トポロジを選択する。その際には、より多くのリンクが利用不可となった場合でも、適切な経路を見つけることができることが、トラヒック変動へ追随した経路制御では最も重要となるため、迂回経路数を最優先とし、迂回経路数が同じ候補論理トポロジが複数存在した場合は、各ノードの故障の影響を比較する。さらに、各ノードの故障の影響も同じ候補論理トポロジが複数存在した場合は、迂回時のトラヒックの分散を比較する。

4. 評　　価

本評価では、提案手法によって設計された論理トポロジを用いた自律的ルーティングをデータセンター内で一般的に用いられている ECMP [8] と比較する。

4.1 シミュレーション環境

本評価では、評価対象の各トポロジ上で各経路制御手法を動作させ、経路制御手法で定められた経路にトラヒックの収容を行う。トラヒックは各サーバーラック間に発生するものとし、発生したトラヒックのうち、一定量のトラヒックずつ、各経路制御手法によって負荷が小さく、追加のトラヒックを収容する経路として適切だと判断された経路に収容を行う。そして、空帯域が存在する経路が経路制御手法により見つからず、さらなるトラヒックの収容ができなくなった時点で、トラヒックの収容を終了する。

上記の手順で、空帯域が存在する経路がなくなった時点で各サーバーラック間を流れているトラヒック量が、各ネットワーク構成・経路制御手法の組み合わせにおいて、収容可能な最大のトラヒック量であると考えることができる。そこで、本評価は、空帯域が存在する経路がなくなった時点において、収容できたトラヒック量を調べることにより、サーバーラック間に確保できた通信帯域を評価する。

4.1.1 評価対象のトポロジ

本評価では、データセンター内で一般的に用いられているネットワーク構成である、FatTree と Torus の 2 種類のネットワーク構造を用いた。本評価に用いる FatTree は 4 ポートのスイッチを 20 台用いて構築されており、最下位層のスイッチのみをサーバーラック内のサーバーに接続した。また、本評価では Torus は 4 x 4 の構造で 4 ポートのスイッチ 16 台を用いて構築し、全ノードにサーバーラック内のサーバーを接続した。また、本評価では、スイッチの各ポートの通信帯域は 10 Gbps としている。

4.1.2 トラヒックの発生方法

本評価では、FatTree の場合は 2 割のサーバーラック間でトラヒックを発生させ、Torus の場合は 1 割のサーバーラック間でトラヒックを発生させた。トラヒックを発生させるサーバーラックはランダムに選択し、各評価 100 パターンのトラヒック状況で評価を行った。

4.1.3 評 価 指 標

本評価では、通信を行っているサーバーラック間の通信帯域の最小値および通信を行っている全てのサーバーラック間の通信帯域の合計値を評価指標として用いる。

4.2 Torus における評価

図 3 は 100 パターンのトラヒックに対して Torus における通信を行っているサーバーラック間の通信帯域の最小値の累積分布関数 (CDF) を表す。また、図 4 は 100 パターンのトラヒックマトリクスに対して Torus における通信を行っている全てのサーバーラック間の通信帯域の合計値の累積分布関数 (CDF) を表す。

図 3 に示されているように、提案手法は論理トポロジの数に関係なく ECMP より、サーバーラック間の通信帯域の最小値が大きいことがわかる。これは提案手法では、各ノードが自律的に論理トポロジを使用した経路制御を行うことで負荷を分散させることができ、さらに高負荷なリンクが発生した際に、当該リンクに接続しているノードが適切な論理トポロジを選択し迂回経路にトラヒックを転送することができるためである。それに対して、ECMP では、輻輳が発生した際にも最短経路しか選択できず、輻輳箇所を避ける際にホップ数が長くなるような経路にトラヒックを転送することができない。その結果、ECMP では輻輳箇所を経由して通信を行っているサーバーラック間に大きな通信帯域を確保できない。

さらに図 4 より、論理トポロジの数が 10 あるいは 15 の場合は、提案手法の方が ECMP より通信帯域の合計値が大きいことがわかる。このことから、提案手法では輻輳発生時に迂回路にトラヒックを転送することで、当該トラヒックのホップ数が大きくなりネットワークの利用資源の量が増える可能性は考えられるものの、当制御により全体的な通信帯域を向上させることができていることがわかる。

また、論理トポロジ数に注目して比較を行うと、論理トポロジ数が大きくなればなるほど、サーバーラック間に確保可能な通信帯域の最小値も、総通信帯域も向上するものの、論理トポロジ数が 5 の場合であっても、サーバーラック間に最低限確保が必要な通信帯域は ECMP よりも著しく大きい。つまり、少ない論理トポロジ数であっても、高負荷のリンクを迂回した経路を設定することができ、通信帯域を確保するのに十分な制御が可能であるといえる。

4.3 FatTree における評価

図 5 は 100 パターンのトラヒックに対して Fattree における

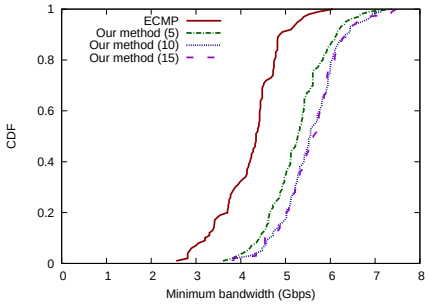


図 3　Torus におけるサーバーラック間の通信帯域の最小値

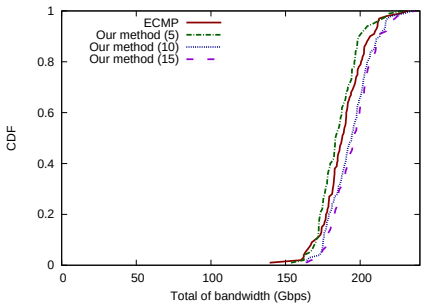


図 4　Torus における全サーバーラック間の通信帯域の合計

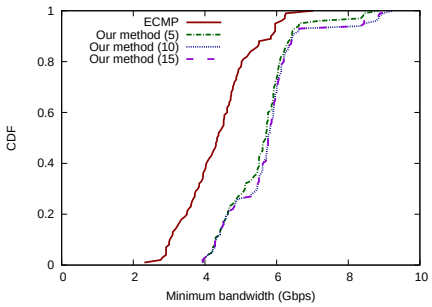


図 5　Fattree におけるサーバーラック間の通信帯域の最小値

通信を行っているサーバーラック間の通信帯域の最小値の累積分布関数 (CDF) を表す。また、図 6 は 100 パターンのトラヒックに対して Fattree における通信を行っている全てのサーバーラック間の通信帯域の合計値の累積分布関数 (CDF) を表す。

Fattree に関しても、Torus と同様に図 5 に示されているように、提案手法は論理トポロジの数に関係なく ECMP より、サーバーラック間の通信帯域の最小値が大きくなっている。さらに図 6 より、通信帯域の合計値も、論理トポロジ数が 5 の場合は ECMP と同程度、論理トポロジの数が 10 あるいは 15 の場合は ECMP よりも大きな通信帯域を確保できている。

以上より、提案手法はトポロジによらず、トラヒック状況に応じて通信帯域を確保できるような経路を選択できていることがわかる。

5. ま と め

本稿では、マルチトポロジルーティングを応用した、負荷分散および環境変動への対応を考慮した論理トポロジを用いた自律的ルーティング手法を提案した。また、本稿では、迂回が発生した場合であっても、トラヒックが特定のリンクに集中することを避けることができるように、論理トポロジ群を設計する手法についても提案した。Torus および Fattre の 2 種類のトポロジに対して、提案手法と ECMP との比較評価を行った結果、輻輳発生時に局所的な情報のみで当該箇所を迂回可能であり、さらに迂回を行った際に、特定のリンクにトラヒックが集中することがなく、全サーバー間に十分な通信帯域を確保できることを明らかにした。

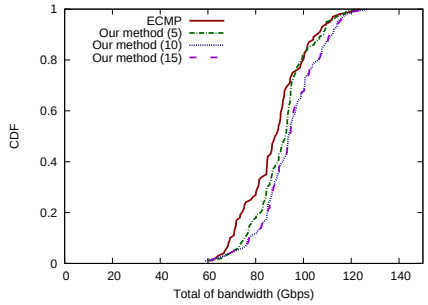


図 6　Fattree における全サーバーラック間の通信帯域の合計

謝　　辞

本研究の一部は、情報通信研究機構（NICT）の委託研究「高機能光電子融合型パケットルータ基盤技術の研究開発」の成果による。ここに記して謝意を表す。

文　　献

[1] T. Benson, A. Anand, A. Akella, and M. Zhang, “The Case for Fine-Grained Traffic Engineering in Data Centers,” in *Proceedings of the 2010 internet network management conference on Research on enterprise networking*, pp. 1–6, Apr. 2010.

[2] S. Kandula, S. Sengupta, A. Greenberg, P. Patel, and R. Chaiken, “The Nature of Datacenter Traffic: Measurement and Analysis,” in *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement*, pp. 202–208, Nov. 2009.

[3] T. Benson, A. Anand, A. Akella, and M. Zhang, “MicroTE: Fine Grained Traffic Engineering for Data,” in *Proceedings of ACM CoNEXT*, pp. 1–12, Dec. 2011.

[4] M. Al-Fares, S. Radhakrishnan, B. Raghavan, N. Huang, and A. Vahdats, “Hedera: Dynamic Flow Scheduling for Data Center Networks,” in *Proceedings of the 7th USENIX conference on Networked systems design and implementation*, pp. 281–295, Apr. 2010.

[5] A. Greenberg, J. Hamilton, N. Jain, S. Kandula, C. Kim, P. Lahiri, D. Maltz, P. Patel, and S. Sengupta, “VL2: A scalable and flexible data center network,” *ACM SIGCOMM Computer Communication Review*, vol. 39, pp. 51–62, Aug. 2009.

[6] A. K. et al, “Resilient Routing Layers for Recovery in Packet Networks,” in *Proceedings of Dependable Systems and Networks*, pp. 238–247, June 2005.

[7] M. C. Scheffel, C. G. Gruber, T. SchwabeCorresponding, and R. G. Prinz, “Optimal multi-topology routing for IP resilience,” *International Journal of Electronics and Communications*, vol. 60, pp. 35–39, Jan. 2006.

[8] C. HOPPS, “Analysis of an Equal-Cost Multi-Path Algorithm,” RFC 2992, Internet Engineering Task Force, Nov. 2000. `http://tools.ietf.org/html/rfc2992`.