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.

パスネットワーク・パケットネットワーク併用型 データセンターチップの評価

大下　裕一† 　村田　正幸†

† 大阪大学 大学院情報科学研究科
E-mail: †{y-ohsita,murata}@ist.osaka-u.ac.jp

あらまし　データセンター内の処理を低消費電力で実行するために、多数のコアを収容したチップを構成するデータセンターチップというアプローチが提唱されている。このアプローチでは、多数のコアを収容したチップ上で、データセンターで行われている並列処理を実行することにより、電力効率よく、必要な処理を行うことが期待できる。データセンターチップにおいては、コア間が通信を行い、連携することで、多量のデータの処理を行う。そのため、データセンターチップにおいても、ネットワークは重要な役割を担う。本稿では、チップ内の通信を収容するのにかかる消費電力と、チップ内の通信の遅延を考慮した上で、データセンターチップのネットワーク構成の評価を行う。本評価を行うにあたり、本稿では、まず、低消費電力で目標とする遅延以内に始点・終点間の通信を収容することができるような、仮想ネットワークの構築手法と、仮想ネットワーク上での経路制御手法を提案する。そして、シミュレーションにより、提案した経路制御手法を用いたデータセンターチップの構成の評価を行い、パケットネットワークの上に、パスネットワークを積層し、パスネットワークの設定により仮想ネットワークを構築できるようにした構成が、低消費電力で多くのトラヒックを目標遅延以内に収容できることを明らかにする。
キーワード　ネットワークオンチップ, データセンター, path network, packet network

Evaluation of Data center chip using path and packet networks

Yuichi OHSITA† and Masayuki MURATA†

† Graduate School of Information Science and Technology, Osaka University
E-mail: †{y-ohsita,murata}@ist.osaka-u.ac.jp

Abstract　One approach to reducing the energy consumption is to use the on-chip data centers, which are integrated circuit chips that performs the tasks in a data center. In an on-chip data center, cores communicates with each other to perform tasks. Therefore, a network within an on-chip data center plays an important role. In this paper, we evaluate the network structures in an on-chip data center, considering the energy consumption and the delay between cores. To evaluate network structures, we propose a method to configure the structure of the virtual network and the routing within an on-chip data center. Then, we evaluate network structures by simulation using our method, and demonstrate that the network structure that enables to configure the virtual network accommodate more traffic with a small energy consumption without causing a large delay.
Key words　Network on Chip, Data center, path network, packet network

1. はじめに

近年、大量のデータをデータセンターにおいて処理するクラウドサービスの普及が急速に進んでいる。今後、さらに多種多様のデバイスがインターネットに接続される IoT 時代を迎えるようになるにつれて新たなサービスが登場し、さらにクラウド普及が促進されると予測される。実際、多種多様のデバイスからの大量のデータを処理するため、データを生成するデバイスとデータセンターの間のネットワークエッジに処理装置（いわゆるマイクロ〜ピコデータセンター）を配するフォグコンピューティングやエッジコンピューティングと呼ばれる仕組みも提唱されている [1]。また、移動体通信網においてもモバイルエッジコンピューティングが提唱されるなど、Google や Facebook などに代表される従来の大規模データセンターとは異なる小型のデータセンターが必要になっている。一方で、クラウドサービスが普及するにつれ、そのデータ処理にかかる消費電力が大きな課題となってきている。現在、クラウドサービスはデータセンターを経由して提供されており、クラウドサービスの隆盛とともに多数のデータセンターが構築されるようになった。データセンターの消費電力は、2010 年には全世界の電力消費の 2% 近くを占め、さらにその消費電力は増加し続けている [2]。例えば、2013 年においては、米国のデータセンターだけで 910 億キロワットというニューヨーク市の全世帯の消費電力の 2 倍以上の消費電力を消費しており、2020 年には 1,400 億キロワットに達するという予測もされている。従って、今後、多量のデータを処理するクラウドサービスのさらなる持続的成長のためには、サービス提供に必要なデータ処理を低消費電力で実現することが急務となっている。

クラウドサービスを低消費電力で提供するために、データセンターの消費電力削減に向けた取り組みは最近急速に進められてきた。たとえば、各サーバーイメージを仮想化し、任意のサーバーコンピュータで仮想サーバーを動作させることができる環境を準備した上で、各仮想サーバーの負荷に応じて動的に移動することにより、より多くのサーバーコンピュータをスリープさせ、消費電力を削減する手法の検討 [3, 4] や、時間ごとに必要な通信帯域に応じてデータセンターネットワークの経路変更やスケジューリングにより、不要なネットワーク機器をスリープさせることによる低消費電力化の検討も行われている [5, 6]。しかし、これらの手法では、データセンターが高負荷状態で動作している時間帯は短いという前提に基づいており、データセンターの消費電力を根本的に削減することはできない。今後、クラウドサービスにおける処理がさらに増大することを考慮すれば、前述のような低消費電力化手法にとどまらず、データセンターで行われる処理自身を低消費電力化する革新的技術が必要となる。

データセンターにおいて、多量のデータを処理するタスクは、サブタスクに分割した上で、各サブタスクをサーバーに割り当て、必要に応じて他のサーバーから情報を取得しながら、並列的に処理される [7]。そのため、このような並列処理を低消費電力で、かつ十分な速度で実行できれば、クラウドサービスを低消費電力で提供できるようになる。そこで、データセンター内の処理を低消費電力で実行するために、多数のコアを収容したチップを構成するデータセンターチップというアプローチが提唱されている [8]。このアプローチでは、多数のコアを収容したチップ上で、データセンターで行われている並列処理を実行することにより、電力効率よく、必要な処理を行うことが期待できる。

データセンターチップにおいては、コア間が通信を行い、連携することで、多量のデータの処理を行う。そのため、データセンターチップにおいても、ネットワークは重要な役割を担う。我々は、データセンターチップ内のネットワーク構成として、これまでに、チップ上に 3 次元にネットワークを積層する技術を用い、パケットネットワークの上にパスネットワークを積層した構成を提案してきた [9]。本構成では、パスネットワークを経由してパケットスイッチ間にパスを構築することにより、仮想ネットワークを構築する。そして、チップ内の通信は構築された仮想ネットワーク上に収容される。本構成では、通信状況に応じて、仮想ネットワークの構成を変更することにより、少ない消費電力でチップ内の通信を収容することができる。しかしながら、これまでの検討では、仮想ネットワークを制御する際には、消費電力のみを考慮し、チップ内の通信にかかる遅延については考慮してこなかった。

そこで、本研究では、チップ内の通信を収容するのにかかる消費電力に加え、チップ内の通信の遅延を考慮した上で、データセンターチップ構成の評価を行う。本評価を行うにあたり、本稿では、まず、目標とする遅延以内に始点・終点間の通信を収容することができるような、仮想ネットワークの構築手法と、仮想ネットワーク上での経路制御手法を提案する。そして、シミュレーションにより、提案した経路制御手法を用いたデータセンターチップの構成の評価を行い、パケットネットワークの上に、パスネットワークを積層し、パスネットワークの設定により仮想ネットワークを構築できるようにした構成が、低消費電力で多くのトラヒックを目標遅延以内に収容できることを明らかにする。

2. データセンターチップ

データセンターチップは、サーバーに該当する処理を行うコアと、各コア間を結ぶネットワーク、及びメモリからなる。データセンター内の処理はサーバー間連携により行われることから、データセンターチップでは、各コアのみならず、チップ内ネットワークが重要な役割を担う。

図 1 にデータセンターチップの構成を示す。本ネットワークでは、パケットスイッチで構成されるパケット通信層とパススイッチで構成されるパス通信層、コア層からなる。各コアは、パケットスイッチに接続することにより、チップ内ネットワークに接続する。各パケットスイッチは、隣接パケットスイッチ、コアのほかにパス通信層へのリンクを持つ。これにより、パス通信層の設定により、任意のパケットスイッチ間にパスを構築することが可能となる。パスを構築することにより、パケットスイッチ間の接続構成を変更することが可能となり、コア間の

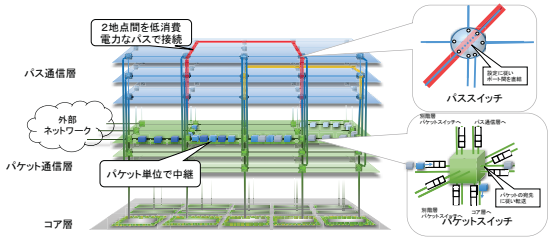

図1　パス・パケット併用型データセンターチップ

通信状況に合わせたネットワーク構成（以降、仮想ネットワークと呼ぶ）を構築可能である。

3. データセンターチップ内ネットワーク制御

3.1 概　　要

本検討では、データセンターチップ内のネットワーク制御は、チップ内に配置された集中制御コントローラにより行われるものとする。集中制御コントローラは各コアから通信需要の統計情報を定期的に収容する。そして、収集した統計情報に合わせて、現在の状況に合わせて、パスネットワークを設定し、仮想ネットワークを構築、仮想ネットワーク上の経路を設定する。

データセンターチップでは、集中制御コントローラが各タイムスロットごとに以下の手順を繰り返すことにより、チップ内ネットワークを制御する。

（1） チップ内のコアから通信需要の統計情報を得る

（2） 現在の通信需要に適した仮想ネットワーク、仮想ネットワーク上の経路を計算する

（3） 現在の仮想ネットワークから、計算された仮想ネットワークに含まれない仮想リンクを削除した場合の経路を計算し、経路設定を投入する

（4） 新しい仮想ネットワークの設定を投入する

（5） 仮想ネットワークの設定が完了後、2で計算された経路設定を投入する。

本制御では、通信需要に適した仮想ネットワーク、仮想ネットワーク上の経路を計算する手法が重要となる。以降、仮想ネットワーク上の経路、過疎ネットワークの構造の計算方法について述べる。

3.2 制 御 目 標

本稿では、仮想ネットワーク、仮想ネットワーク上の経路は、以下の目標に合わせて制御するものとする。

- 全通信について、要求される遅延性能を満たすこと。
- 1の条件を満たす経路・トポロジのうち、消費電力を最小化すること。

以下、上記の各指標について以下に述べる。

3.2.1 遅　　延

データセンターチップでは、コア間の通信に発生する遅延は、コア間の通信が収容された経路上のパケットスイッチで発生する遅延の合計となる。

ここで、コア間の通信 f において発生する遅延、D_f は、

$$D_f = \sum_{l \in L_f} d_l$$

とあらわされる。ここで、d_l はリンク l で発生する遅延、L_f はコア間の通信 f が経由するリンクの集合を示す。

本稿では、各パケットスイッチは、各宛先ポートに対してバッファを持ち、1クロックにつき各ポートから1パケットを隣接するパケットスイッチに送出する。そのため、リンク l において発生する遅延は、リンク l の送信元のパケットスイッチ n_l において、リンク l 宛のポートのバッファ内待ちパケット数に1を加えた値となる。

ここで、バッファ内待ちパケット数が k である確率を $P_l(k)$ とすると、d_l は以下のように定まる。

$$d_l = \sum_k kP_l(k) + 1$$

平衡状態においては、バッファ内待ちパケット数の確立については、以下の条件が成り立つ。

$$P_l(k) = \sum_i p_l(i) P_l(k+1-i)$$

ただし、$p_l(i)$ は、1クロックあたりにパケットスイッチ n_l に到着するリンク l 宛のパケット数を示し、経路を定めることにより定まる。

3.2.2 消 費 電 力

パスネットワークの消費電力は、パケットネットワークと比べ、十分に小さいと考えられる。そこで、本稿では、パケットネットワークの消費電力に焦点を当てる。パケットネットワークの消費電力は、パケットスイッチの消費電力の合計である。パケットスイッチの消費電力は、当該パケットスイッチを経由するパケット数に比例する。そのため、パケットスイッチを経由するパケット数の合計を減らすことにより、データセンターチップの消費電力を抑えることができる。

3.3 経路計算手法

本稿では、仮想ネットワーク上の経路は、各通信の遅延を目標値以下に抑えつつ、パケットスイッチを経由するパケットの総数を最小化するように経路を定める。全通信の経路を一度に決める最適化問題は、長い計算時間を要する。そこで、本稿では、収容する必要のある通信1本ずつ経路を決定する。その際に、各通信の収容先の経路は、最短ホップ経路を優先することにより、低消費電力な経路に収容する。本稿では、トラヒック量が多い通信から順に経路を決める。これは、トラヒック量の多い通信をホップ数の大きな経路に収容すると、大きな消費電力がかかるためである。そのため、トラヒック量の多い通信の経路を優先的に決めることにより、トラヒック量の多い通信をホップ数が短い経路に収容する。

各通信の収容先の経路は、以下の手順で求める。

（1） 仮想ネットワークのトポロジを G とする

（2） トポロジ G 上の最短ホップ経路を求める

（3） 2で求まった経路にトラヒックを収容した場合の遅延を計算する。

（4） 3で計算された遅延が制約を満たしている場合は、2

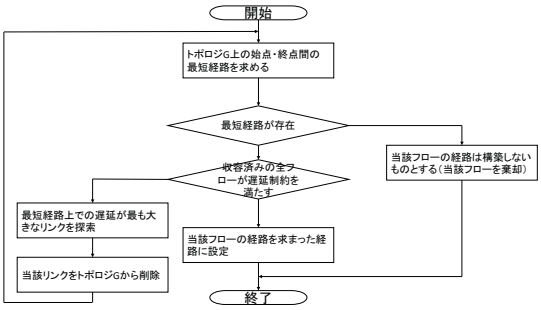

図 2　経路制御手順

で求まった経路に通信を収容する。満たしていない場合は、遅延が最も大きいリンクを G から削除し、手順 2 に戻る。
図 2 に、本手順をフローチャートで示す。

3.4　仮想ネットワーク計算方法

構築可能な仮想ネットワークの構造は、パケットスイッチ数を N、パケットスイッチあたりのポート数を P とすると、${}_{N^2}C_{NP}$ 種類存在し、その中から適切なネットワーク構造を選択するには、長時間の計算を要する。そこで本稿では、以下の手順で、仮想リンクを 1 本ずつ追加することにより、適切な仮想ネットワークを構築する。

（1） 解候補のトポロジの集合 C、パケットネットワークのトポロジを入れる

（2） 解候補 C に含まれる各トポロジ c について、仮想リンクを追加したトポロジを C に入れる

（3） 解候補のトポロジをランク付けし、上位 N 個の仮想トポロジのみを解候補に残し、それ以外を C から削除する。

（4） 手順 2 で新たなトポロジが生成されない場合は、ランク 1 位の仮想トポロジを構築する。それ以外は、手順 2 に戻る。

上記の手順でのランク付けは、当該仮想ネットワーク上での経路を計算した上で、以下の指標をもとに行う。

- 遅延の制約を満たして収容可能な通信の本数が多い方を優先する
- 遅延の制約を満たして収容可能な通信の本数が同じものについては、パケットスイッチが経由するパケットの総量が小さいものを優先する。

4.　評　　価

本評価では、提案した仮想ネットワーク制御手法を用いて、パスネットワーク・パケットネットワーク併用型のデータセンターチップの有効性を確認する。

4.1　比較するデータセンターチップ構成

4.1.1　パスネットワーク・パケットネットワーク併用型（仮想ネットワーク構築可）

本評価では、8 × 8 の格子上にパケットスイッチを配置したパケットネットワーク層を構築する。同様にコア層も 8 × 8 の格子上にコアを配置し、各コアは 1 つのパケットスイッチに接続する。パケットネットワーク層のパケットスイッチは、同じ階層の隣接するパケットスイッチへのリンク、コアからのリンクに加え、パスネットワーク層へのリンクを持つ。パスネットワーク層では、任意に 2 点間を接続するようなパスを構築することができるように構成する。つまり、パケットスイッチのポートがある限り仮想リンクを構築することが可能とする。

4.1.2　パスネットワーク・パケットネットワーク併用（直接パスのみ）

本構成では、前節の構成と同様に、8 × 8 の格子上にパケットスイッチを配置したパケットネットワーク層を構築する。同様にコア層も 8 × 8 の格子上にコアを配置し、各コアは 1 つのパケットスイッチに接続する。ただし、パスネットワーク層は、パケットネットワーク層とは接続せずに、コアと接続する。これにより、各コアから、任意のコア 1 つに対して、パスネットワーク層を経由せずに直接通信することができる。前節の構成と本構成を比べることにより、仮想ネットワークを構築できるように、パスネットワークとパケットスイッチを接続することの効果を示すことができる。

4.1.3　パスネットワークなし

本構成では、8 × 8 の格子上にパケットスイッチを配置したパケットネットワーク層を構築する。同様にコア層も 8 × 8 の格子上にコアを配置し、各コアは 1 つのパケットスイッチに接続する。ただし、パスネットワークは配置しない。本構成と比較することにより、パスネットワーク層を導入する効果を示すことができる。

4.2　評 価 環 境

本評価を行うにあたり以下の環境を用いた。

4.2.1　容　　量

本評価では、パケットスイッチは、各ポートあたり、1 Unit/Clock のトラヒック量の転送が可能であるもののとした。

4.2.2　通信レート

本評価では、ランダムに選択したコア間で通信を発生させた。選択されたコア間で発生する通信レートは均等な値とした。各コアから流入するトラヒック総量を特に明記しない限り、0.5 Unit/Clock とした。また、トラヒック総量は 0.1 から 0.5 Unit/Clock と変化させ、トラヒック量の影響を調べた。

4.3　評 価 指 標

本評価では、以下の 2 つの指標で評価を行う。

4.3.1　棄　却　率

本評価では、通信の収容先の経路を決める際に、目標遅延を満たす通信は収容し、目標遅延を満たすことができない通信は棄却することとする。そこで、本評価では、収容を試みた通信の本数に対して、収容できなかった通信の本数を棄却率と定義する。しすて、棄却率を調べることにより、どれだけ多くの通信を遅延性能を満たしつつ収容できたかという観点での評価を行う。

4.3.2　パケットスイッチを経由するトラヒック量の総和

本研究では、パス通信層の消費電力は、パケット通信層と比べて極めて少ないと想定し、パケットスイッチによる消費電力で評価を行う。各パケットスイッチが消費する電力は、経由するトラヒック量に比例する。そこで、パケットスイッチを経由するトラヒック量の総和を求めることにより、データセンター

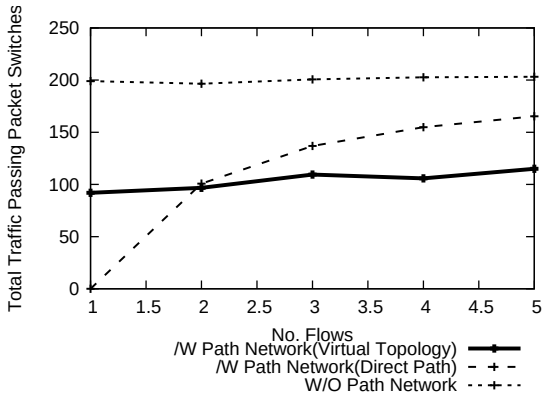

図 3　通信の本数のパケットスイッチを経由するトラヒック量への影響

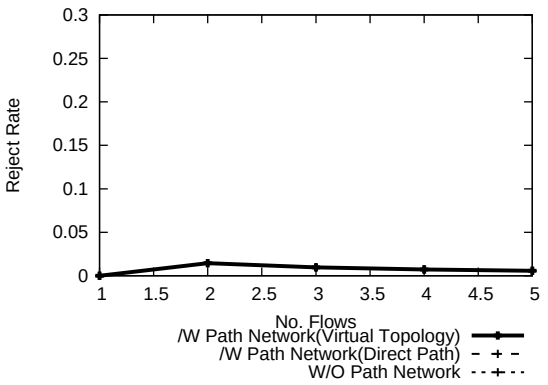

図 4　通信の本数の棄却率への影響

チップ内のネットワークが消費する電力を求めることができる。

4.4　通信本数の影響

まず、本評価では、通信本数の影響を調べる。本評価では、全トラヒックを 100 クロック以内に宛先まで到達できるように収容した。図 3 に通信の本数のパケットスイッチを経由するトラヒック量への影響を示す。図では、横軸は各コアあたりの通信相手の数、縦軸はパケットスイッチを経由するトラヒック総量を示す。図より、パスネットワーク・パケットネットワーク併用（直接パスのみ）は、各コアあたりの通信相手の数が１つのみであれば、パケットスイッチネットワークにトラヒックを転送する必要はなく、パケットスイッチを経由するトラヒック総量は０となる。しかしながら、各コアあたりの通信の本数が２以上になると、パケットスイッチを経由するトラヒック総量は増加する。それに対して、パスネットワーク・パケットネットワーク併用型（仮想ネットワーク構築可）は、各コアあたりの通信の本数が２以上であっても、パケットスイッチを経由するトラヒック総量を少なく抑えることができる。その結果、通信相手の数が３つ以上の場合は、パスネットワーク・パケットネットワーク併用（直接パスのみ）よりも、パケットスイッチを経由するトラヒック総量を少なくなる。これは、この構成では、現在の需要に合わせて、仮想ネットワークを構築することにより、複数本の通信が経由するパケットスイッチを減らすことができるような仮想リンクを構築することができるためである。

図 4 は、通信の本数の棄却率への影響である。図の横軸は各コアあたりの通信相手の数、縦軸は棄却率を示す。図より、本評価では、いずれの構成においても、棄却はほとんど起きていないことが分かる。

4.5　遅延制約の影響

次に、遅延制約を変えた影響を調べる。

図 5 に遅延制約のパケットスイッチを経由するトラヒック量への影響、図 6 に遅延制約の棄却率への影響を示す。

図 5 より、パスネットワーク・パケットネットワーク併用型（仮想ネットワーク構築可）が遅延制約によらず、パケットスイッチを経由するトラヒック量がもっとも小さい。

また、図 6 より、遅延制約を厳しくした場合は、パスネットワークがない場合や、パスネットワーク・パケットネットワーク併用（直接パスのみ）では棄却率が高くなる。しかしながら、パスネットワーク・パケットネットワーク併用型（仮想ネットワーク構築可）では、全通信を２０クロック以内に宛先まで届けるという条件下においても棄却は生じていない。

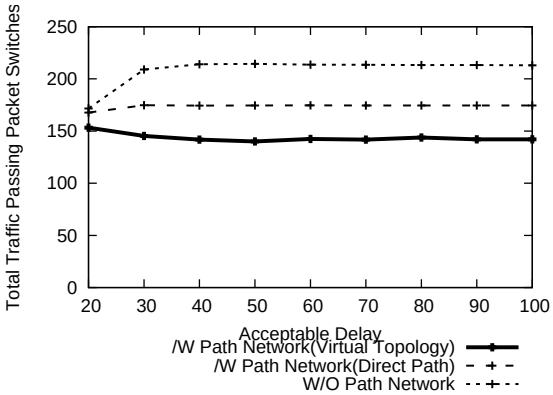

図 5　遅延制約のパケットスイッチを経由するトラヒック量への影響

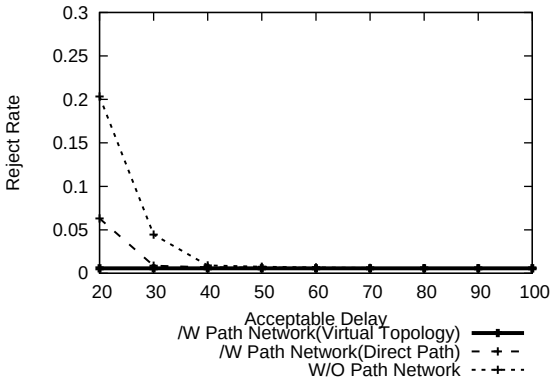

図 6　遅延制約の棄却率への影響

この結果より、現在の需要に合わせて仮想ネットワークを構築することができるようにしたネットワーク構成は、低消費電力での通信の収容のみならず、通信を低遅延で収容することにも寄与できることが分かる。

4.6　トラヒック量の影響

最後に総トラヒック量の影響を調べた。本評価では、各コアが１クロック当たりに送出するトラヒック量を 0.1 から 0.5 まで変化させた。また、本評価では、遅延の制約を２０クロック

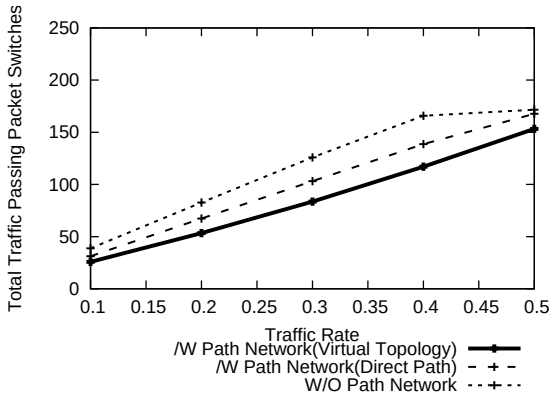

図 7　トラヒック量のパケットスイッチを経由するトラヒック量への影響

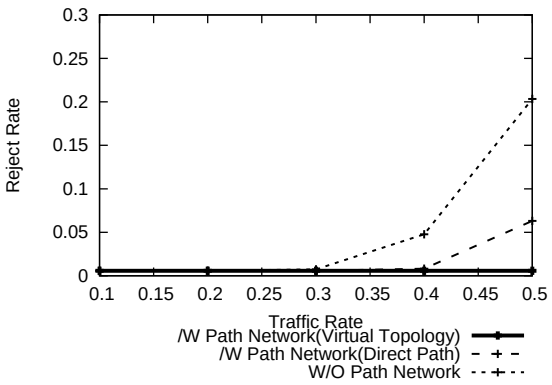

図 8　トラヒック量の棄却率への影響

とした。

図 7 に発生したトラヒック量のパケットスイッチを経由するトラヒック量への影響、図 8 に棄却率への影響への影響を示す。

図 7 より、発生したトラヒック量が増えると、パケットスイッチを経由するトラヒック量も増えることが分かる。また、発生したトラヒック量によらず、パスネットワーク・パケットネットワーク併用型（仮想ネットワーク構築可）が、パケットスイッチを経由するトラヒック量がもっとも小さい。

また、図 8 より、トラヒック量が増加すると、パスネットワークがない場合や、パスネットワーク・パケットネットワーク併用（直接パスのみ）では棄却率が高くなる。それに対して、パスネットワーク・パケットネットワーク併用型（仮想ネットワーク構築可）では、トラヒックレートが 0.5 の場合でも棄却を生じない。

つまり、多くのトラヒックが存在する場合にも、パスネットワーク・パケットネットワーク併用型（仮想ネットワーク構築可）が、低消費電力で多くのトラヒックを収容するのに有効であることが分かる。

5. ま と め

本稿では、チップ内の通信を収容するのにかかる消費電力に加え、チップ内の通信の遅延を考慮した上で、データセンターチップ構成の評価を行った。本評価を行うにあたり、本稿では、まず、目標とする遅延以内に始点・終点間の通信を収容することができるような、仮想ネットワークの構築手法と、仮想ネットワーク上での経路制御手法を提案した。そして、シミュレーションにより、提案した経路制御手法を用いたデータセンターチップの構成の評価を行い、パケットネットワークの上に、パスネットワークを積層し、パスネットワークの設定により仮想ネットワークを構築できるようにした構成が、低消費電力で多くのトラヒックを目標遅延以内に収容できることを明らかにした。

謝　　辞

本研究は戦略的情報通信研究開発推進事業 (SCOPE) の委託により行われた。

文　　献

[1] F. Bonomi, R. Milito, P. Natarajan, and J. Zhu, “Fog computing: A platform for internet of things and analytics,” *Big Data and Internet of Things: A Roadmap for Smart Environments*, pp. 169–186, Mar. 2014.

[2] NRDC, “America’s data centers consuming and wasting growing amounts of energy.” `http://www.nrdc.org/energy/data-center-efficiency-assessment.asp`, Feb. 2015.

[3] L. Liu, H. Wang, X. Liu, X. Jin, W. B. He, Q. B. Wang, and Y. Chen, “Greencloud: a new architecture for green data center,” in *Proceedings of the 6th international conference industry session on Autonomic computing and communications industry session*, pp. 29–38, ACM, 2009.

[4] A. Bergen, R. Desmarais, S. Ganti, and U. Stege, “Towards software-adaptive green computing based on server power consumption,” in *Proceedings of the 3rd International Workshop on Green and Sustainable Software*, pp. 9–16, ACM, 2014.

[5] B. Heller, S. Seetharaman, P. Mahadevan, Y. Yiakoumis, P. Sharma, S. Banerjee, and N. McKeown, “Elastictree: Saving energy in data center networks.,” in *NSDI*, vol. 10, pp. 249–264, 2010.

[6] D.-J. Li, Y.-T. Yu, W. He, K. Zheng, and B. He, “Willow: Saving data center network energy for network-limited flows,” *IEEE Transactions on Parallel and Distributed Systems*, 2014.

[7] J. Dean and S. Ghemawat, “Mapreduce: simplified data processing on large clusters,” *Communications of the ACM*, vol. 51, no. 1, pp. 107–113, 2008.

[8] M. Kas, “Toward on-chip datacenters: a perspective on general trends and on-chip particulars,” *The Journal of Supercomputing*, vol. 62, no. 1, pp. 214–226, 2012.

[9] T. Ikeda, Y. Ohsita, and M. Murata, “3d network structures using circuit switches and packet switches for on-chip data centers,” *International Journal On Advances in Networks and Services*, vol. 7, no. 1 & 2, pp. 1942–2644, 2012.