様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２９　年　　６　月　１５　日現在

機関番号：　３４４１６
研究種目：　挑戦的萌芽研究
研究期間：　2014～2016
課題番号：　２６５４００３９

研究課題名（和文）超高速大容量データセンタネットワークに適した新しいトラヒック制御

研究課題名（英文）New Traffic Engineering Methods for High-Speed Data Center Networks

研究代表者
山本　幹（Yamamoto, Miki）

関西大学・システム理工学部・教授

研究者番号：３０２１０５６１

交付決定額（研究期間全体）：（直接経費）　　2,700,000円

研究成果の概要（和文）：本研究課題は、データセンタネットワーク環境に適した新たなトラヒック制御の確立を目的とし、1) 時間軸方向へのトラヒック分散を図る輻輳制御、2)空間軸方向へのトラヒック分散を図るトラヒックエンジニアリング、3)大規模トラヒック収容を可能とする光ネットワーク設計、の３つのテーマに取り組んだ。具体的には、データセンタ環境に適した新しいエンドツーエンド型ならびにスイッチ介在型輻輳制御の開発、動的なデマンド変化に分散的に対応するトラヒックエンジニアリング手法の開発、ボトルネックリンク発生を抑制する光ネットワーク設計技術の開発を行った。

研究成果の概要（英文）：Datacenter networks have quite different requirements from the current wide area Internet. In this study, we try to develop new traffic engineering methods suitable for data center networks from the following 3 aspects, 1) congestion control enabling traffic distribution in time dimension, 2) traffic engineering distributing traffic in space dimension, and 3) optical network design. As our research results of this study, we develop several end-to-end new congestion control methods realizing fair share of bottleneck links in data center networks, new switch-assisted approach congestion control methods. And we develop a new traffic engineering method of distributed approach which dynamically controls traffic volumes according to dynamic traffic change in data center. We also develop optical path design method for data center networks which minimizes traffic volume on the bottleneck link.

研究分野：　情報通信工学

キーワード：　ネットワーク　データセンタ　トラヒック制御　ネットワーク運用　輻輳制御

１．研究開始当初の背景

クラウドコンピューティングのコア部分に位置するデータセンタにおいては、各サーバに柔軟に設定される仮想マシン構成により、そのトラヒック分布は不均質に変動する。このようなトラヒック変動に対応するには、トラヒック制御、輻輳制御が必須である。データセンタに要求される条件は、従来からの要求以上の高速大容量やロスレス通信という厳しいものであり、データセンタの環境にあわせた新しいトラヒック制御、輻輳制御の開発が求められる。

２．研究の目的

本研究課題は、データセンタネットワーク環境に適した新たなトラヒック制御理論の確立を目的としている。具体的には、ホットスポットに対応する新しいトラヒック制御、輻輳制御の開発を目指し、
(1) 時間軸方向への分散を図る輻輳制御
(2) 空間軸方向で対応するトラヒック制御
(3) 帯域設計により大規模トラヒック収容を可能とする光ネットワーク設計
の３つの観点で研究を進める。

３．研究の方法

上記３つの観点から、具体的に以下の方法で研究を実施した。
(1) 時間軸方向への分散を図る輻輳制御

データセンタ輻輳制御は、ネットワーク内のスイッチが積極的に介在するアプローチと、エンドツーエンドで対応するアプローチの、二つに分類できる。本研究課題は、両者の観点で研究を遂行した。

・スイッチ介在アプローチ

IEEE 802.1Qau にて標準化されているスイッチ介在型輻輳制御 QCN(Quantized Congestion Notification)に対し、データセンタ内の多数の仮想マシンの OS 更新などに利用されるマルチキャスト通信に適応させた新しい方式として、研究代表者は QCN/BS (QCN with Bottleneck Selection)を提案している。この方式には、段数の異なるフロー間で不公平性の問題があることが本研究課題における基礎評価で明らかとなった。この不公平性の生じる原因分析と、これを解決する手法の提案を行った。

・エンドツーエンドアプローチ

データセンタでは、DCTCP(Data Center TCP)が、待ち行列長を安定させ高いスループットを得る優れたエンドツーエンド型方式として知られている。本研究課題では、多段ボトルネック環境での基礎評価により、DCTCP では多段フローのスループットが低く抑えられるという不公平性が生じることを明らかにした。この問題点を解決する手法として、送信ホストが最悪ボトルネックリンクを把握する新しい手法 MB-DCTCP(DCTCP for Multiple Bottleneck)を提案した。

また、1 対多通信であるマルチキャスト通信に対し、エンドツーエンド型の全く新しい輻輳制御 DCMC(Data Center Multicast Congestion control)を提案した。

(2) 空間軸方向で対応するトラヒック制御

本研究課題では、トラヒックエンジニアリング技術によるデータセンタネットワークの最適ルーチングを検討した。通常のトラヒックエンジニアリングは、ある時点における送受信ノード間の通信要求量（デマンド）に応じて、数理計画法等を用いた最適化により集中制御で実現する。しかしこの場合、デマンドの時間的変化が起こるたび、もしくは、ある一定期間をおいて再び最適化を行うような制御となり、ネットワーク全体の構成を同時に変更する必要がある。データセンタでは、ジョブの実行要求に合わせて、仮想マシン集合の参加離脱が頻繁に発生し、その度にデマンドが変動する。よって、管理者が集中制御でトラヒックエンジニアリングを行うことは望ましくなく、動的なデマンドの変化に対応する手法が求められる。この問題に対応するために、マルコフ近似を用いたデータセンタトラヒックエンジニアリング手法の提案を行った。

(3) 光ネットワーク設計

本研究課題では、データセンタネットワークの階層構成の上位に位置するコア部への、光ネットワーク技術の適用を検討した。データセンタのコア部分においては、多量のトラヒックが発生するため、低消費電力で高速大容量伝送が可能な光ネットワークの導入が望ましい。光ネットワークでは波長分割多重を用いて波長単位でトラヒック制御を行うため、通常のパケット交換よりもトラヒックの粒度が大きい。そのため、そのトラヒック特性を考慮したネットワーク設計が必要となる。本研究課題では、ボトルネックリンクの発生を抑制する光データセンタネットワーク設計手法を提案した。

４．研究成果

本研究課題で得られた成果について、2.研究の目的欄に示した３つの観点に関して個々に記述する。

(1) 時間軸方向への分散を図る輻輳制御

・スイッチ介在アプローチ

マルチキャスト通信に対応する新しいデータセンタ輻輳制御方式として、複数に分岐する経路の中から、最悪輻輳状態にあるボトルネックリンクを選択し、これに送信レートを調整する QCN/BS を提案している。この方式は、１対１通信であるユニキャストにおいて、経路上に複数ボトルネックが存在する場合にも適用できることが予想される。今回、3 段の long flow(フロー1)と１段のフロー(フロー2,3,4)が存在し、３段のタンデムリンクのそれぞれをフロー１と他の１段フローの一つずつが共有するモデル（３リンクす

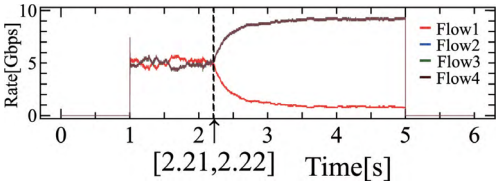


図 1 QCN/BS の不公平性発生事例

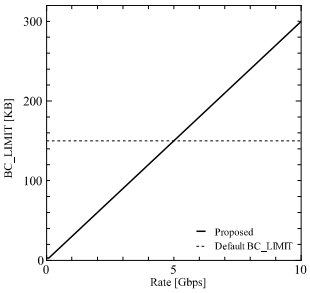
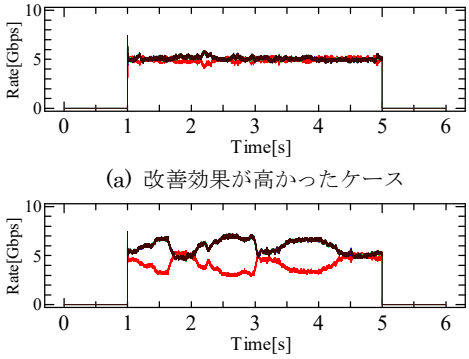


図 2 Adaptive BC_LIMIT の設定

べてがボトルネックとなる）を用いて基本性能評価を行った。図１に示すように、段数の多いフロー１のスループットが劣化する結果が得られた。

この理由を詳細に分析した[雑誌論文[1]]。①１段目のボトルネックリンクでの高い利用率を経た出力過程において、フロー１の２段目以降の到着過程に瞬時的なバーストが生じる。このため、フロー１が２段目以降のボトルネックリンクで瞬時的に待ち行列長を高く見積もり、輻輳フィードバックを多く受け取る。②瞬時的に多く受け取るフィードバックによりフロー１が送信レートを下げたのち、他の１段フローが空いた帯域を埋める形で送信レートを上昇させる。③一旦フロー１の送信レートが低い不公平状態に陥ると、多段ボトルネックからのフィードバックによりレート上昇機会が減るため、long flow であるフロー１はレートが低いまま推移する。という３つのことが相互に関連し、long flow はレートを下げる機会が多くあることに加え、一旦レートが下がると上昇させる機会が減ることで、不公平状態に陥ると回復が望めないことが明らかとなった。

この問題を解決する手法として、すでに研究代表者が QCN に対して提案している、Adaptive BC_LIMIT という手法を、新たに QCN/BS に適用することを提案した。QCN では、送信ノードが自身の送信バイト数をカウントしており、これが BC_LIMIT で設定される値に達する度に、送信レートをレート増加アルゴリズムに従い増加させる。従来は図２に示すように BC_LIMIT として一定値を設定していたが、Adaptive BC_LIMIT においては送信レートに応じて設定値を変化させている。この設定により、送信レートが低い場合にはレート増加タイミングの時間間隔が短くなり、より速くレートを増加させることとなる。逆に送信レートが高い場合には、レート増加



(a) 改善効果が高かったケース

(b) 改善効果が低かったケース

図 3 QCN/BS with Adaptive BC_LIMIT のスループット特性

タイミングはより遅くなる。すなわち、送信レートをより公平なところに安定させるよう、各フローがレート調整することが期待できる。

図３に、QCN/BS に Adaptive BC_LIMIT を組み合わせた提案方式を用いた場合の、図１と同様のモデルでの各フローのスループット推移を示す。(a)のように、全フローのスループットがほぼ等しくなり、公平性が大きく改善できるケースが多かった。ただ、(b)のように一部公平性が悪化する時間区間の存在するケースもあったが、図に示すように送信レートに差が生じた後にそれを公平な方向に回復させる形で提案方式が効果を示していることがわかった。

これらの成果は、国際会議（Best Paper Award 受賞）ならびに電子情報通信学会論文誌にて発表している。

・エンドツーエンドアプローチ

【多段ボトルネック対応 DCTCP】

データセンタにおけるエンドツーエンド型輻輳制御としては、DCTCP がよく知られている。本課題では、まず多段ボトルネックが存在する状況での DCTCP の基礎評価を、2 段のフロー(フロー1)と、それと 1 段ずつを共有する 2 つの 1 段フロー(フロー2, 3)が存在するモデルを用いて行った。DCTCP では ECN でマークされたパケット数に応じて送信側のウィンドウサイズを減少させる手法をとっている。複数のボトルネックが存在する場合には、図 5(a)に示すように、long flow はそれぞれのボトルネックでマークされる可能性があり、より多くの ECN マークを受けるため、short flow に比べ低いスループットに抑えられるという不公平性が生じている。この問題を解決するために、どのスイッチでマークされたかを識別させ、送信ホストでは各スイッチの輻輳状況を把握したうえで、最悪輻輳状態にあるスイッチの輻輳状況にあわせてウィンドウサイズを制御する、新しい手法 MB-DCTCP(DCTCP for Multiple Bottleneck)を提案した(図 4)。

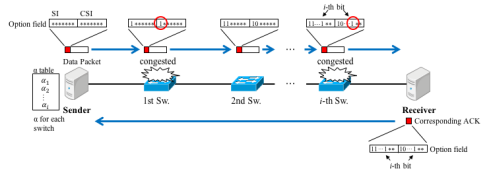


図 4 MB-DCTCP の概要

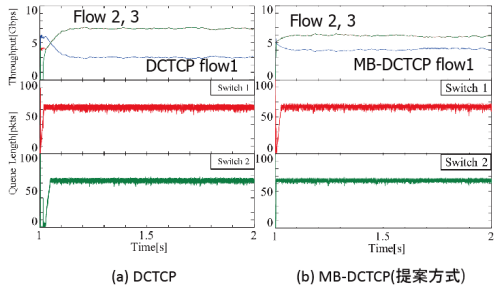


図 5 DCTCP と提案方式の比較

図 5(b)に示すように、long flow であるフロー 1 のスループットが大きく改善され、公平性が改善できている。なお、エンドツーエンド型では、RTT の増加に応じてスループットが若干下がるという特徴があるため、多少フロー1 のスループットが他に比べて低くなっているが、これは今回改善対象とするものではない。本成果は、ACM Sigcomm 主催の国際会議 CoNext の併催ワークショップにて発表を行った。

【マルチキャスト輻輳制御】

データセンタで OS 更新などを一括して行う際には、1 対多通信であるマルチキャストが効率的である。エンドツーエンド輻輳制御として、DCTCP をユニキャストフローに用いた場合、マルチキャストフローはこの DCTCP と公平な形でボトルネックリンクを共有することが求められる。本課題では、送信ホストが ECN でマーキングされたパケットを受信した際に受信ホストごとの見積もりスループットを計算し、最悪スループットを与える受信ホスト、すなわち最悪輻輳状態にある受信ホストの見積もりスループットに、送信レートを調整する方式 DCMC(Data Center Multicast Congestion control)を提案した。スループットの見積もり値の計算については、TCP のスループット式をもとに、DCTCP が同一状況（マルチキャストフローと同一状況）で得るスループットを算出した。

図 6 に示すモデルを対象に、性能評価を行い、図 7 に示すような結果を得ている。マルチキャストフローは、最悪ボトルネックである SW2-4 間のリンクを DCTCP と公平配分する

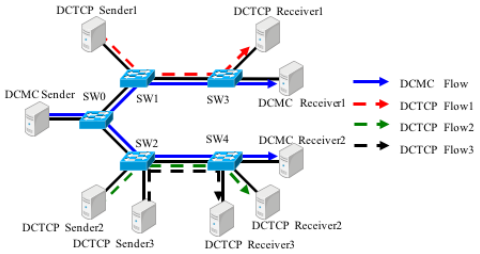


図 6 マルチキャスト評価モデル

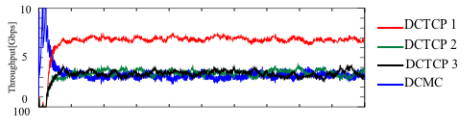


図 7 DCMC のスループット性能

スループットを得ており、SW1-3 リンクではマルチキャストが使用した 1/3 の帯域の残りを DCTCP が使用する形で、理想的なスループット配分が実現できている。これらの成果は、研究会、国際会議で発表した。

・ハイブリッドアプローチ

本研究課題では、スイッチ介在アプローチに分類される輻輳制御方式 QCN に対し、その技術課題を解決する手法としてエンドツーエンドアプローチを併用する新しい提案も行った。QCN においては、輻輳時に返送するフィードバック情報をスイッチが確率的に送信することで、制御オーバーヘッドを削減している。このため、多くのフローがボトルネックリンクを共有する際には、輻輳が発生しているにも関わらず一部のフローが確率的返送フィードバックを受け取れず、自身の送信レートを減少しないことから、ボトルネックキューの待ち行列長が大きく変動するという技術課題がある。この問題を解決する手法として、本研究課題では、エンドホストが自身の送出フレームの RTT を計測し、その増加が輻輳を暗示的に示していることを利用した遅延ベースの輻輳制御を併用する手法を提案した。紙面の都合で評価結果については割愛するが、フローの参入によりフロー数が増加し、従来方式では待ち行列長が大きく変動するケースにおいて、提案方式を用いることで待ち行列長を安定させつつ抑制できることを示した。

(2)空間軸方向で対応するトラヒック制御

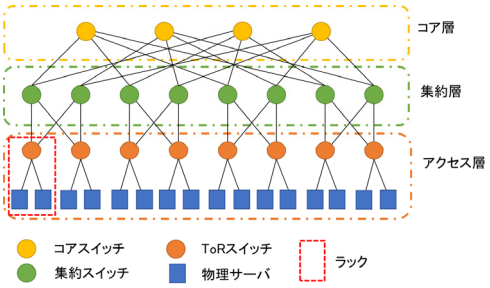


図 8 Fat-Tree モデル

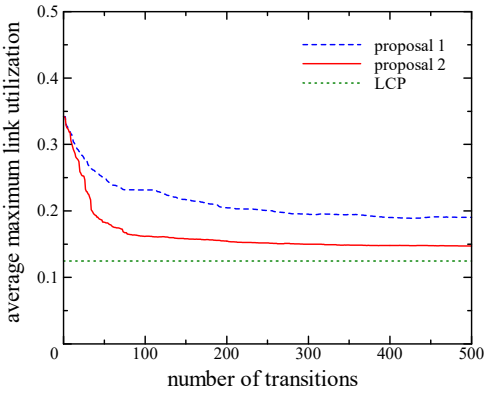


図 9 最大リンク利用率の変化

本研究課題では、仮想マシンの参加離脱によって頻繁に変動するデータセンタのトラヒック特性を考慮し、マルコフ近似を用いたデータセンタトラヒックエンジニアリング手法の提案を行った。マルコフ近似は近年提唱された分散処理最適化フレームワークである。マルコフ近似では、ネットワーク構成を時間可逆連続時間マルコフ連鎖の状態として表す。マルコフ連鎖上において評価値に応じた時間間隔で状態を遷移させていくことで、最適値の近似解を導出する。この際、各ユーザの振る舞いをネットワーク構成とすることで、各ユーザの独立した振舞いをマルコフ連鎖の状態の変化として表すことが可能であり、分散処理で最適化を行うことが可能である。また、時間とともに状態を変更させていくことで最適化を狙うものであり、動的な仮想マシンの参加離脱によるトラヒックの変動に対し柔軟に対応可能である。提案手法は，マルコフ近似を用いたトラヒックエンジニアリングにより，データセンタネットワークの最大リンク利用率の最小化を狙う。本研究課題では、探索空間が異なる二つの戦略を考え、それぞれマルコフ連鎖を形成し、最適化を行った。一方は、マルコフ近似の考え方をデータセンタトラヒックエンジニアリングにそのまま適用したものである（戦略 1)。もう一方は，マルコフ近似を応用した局所探索法であり、これはマルコフ近似の探索空間を限定することで、より良い解を素早く発見し、かつ評価値の時間平均の改善を狙うものである（戦略 2)。

本研究課題では、シミュレーション実験により提案手法の有効性を評価した。使用したネットワークモデルは図 8 に示す Fat-Tree モデルである。図 9 は、マルコフ連鎖上の遷移回数に対する最大リンク利用率を示しており、遷移回数が増加するほど、最大リンク利用率が低下していることがわかる。また、戦略 2 の方が戦略 1 よりも効果的に最大リンク利用率を削減できていることがわかる。さらに図 10 は、経過時間に対する最大リンク利用率の変化を表している。この図では、経過時間が 500 近辺で、大量の仮想マシンが参加したことを想定しており、実際に最大リンク利用率が大幅に増加していることがわかる。しかし、提案手法を用いることにより、素早く最大リンク利用率の削減が行えていることがわかる。これらの成果は、研究会、国際会議で発表した。

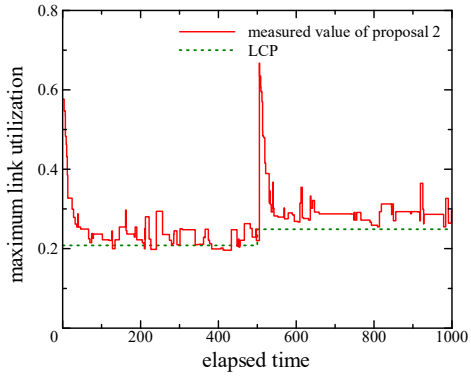


図 10 デマンド変化への対応

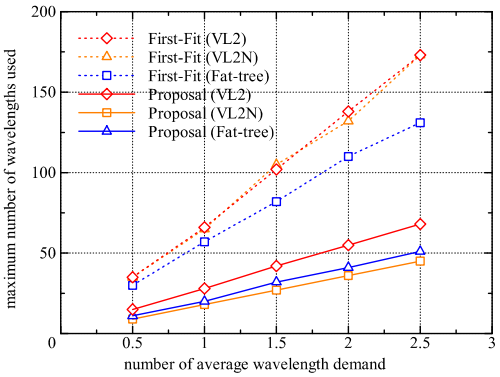


図 11 最大利用波長数

(3) 光ネットワーク設計

本研究課題では、波長資源枯渇によるボトルネックリンク発生を抑制するための光データセンタネットワーク設計手法の提案を行った。光ネットワークにおいては、送受信ノード間で波長が割当てられた経路である光パスをデータ伝送前に設定し、その光パスに沿ってデータが伝送される。光パスは、経路上の波長資源を占有し、他の光パスが同一リンク上で同じ波長を使用することはできない。また、送受信ノード間では同一の波長

を使用しなければならないという制約が存在する。光ネットワークの設計においては、これらの制約を満たしながら、経路及び波長を決定することが重要となる。またこの際、この経路及び波長選択の考え方により、静的な光パス設定と動的な光パス設定のどちらかを用いることになる。静的な光パス設定では、予め各送受信ノード間で要求されるトラヒック量が与えられており、その上で最適化問題を解くことで、設定される光パスの経路及び波長が決定される。一方、動的な光パス設定では、要求トラヒック量が与えられておらず、各送受信ノード間において時々刻々と発生する光パス設定要求に応じて、その度に経路と波長を選択して光パスを設定する。提案手法では、静的な光パス設定に着目し、ネットワーク内で使用される波長数の最大値を最小化することを目的とした最適化問題を扱った。

本研究課題では、数値実験により提案手法の有効性を示した。使用したネットワークモデルは図 8 の Fat-Tree モデルに加え、データセンタネットワークでの使用が想定される、VN2 及び VN2N-Tree を用いた。図 11 に要求トラヒック量に対する、最大利用波長数を示す。図より、提案手法を用いることで、既存手法である First-Fit よりも最大利用波長数が効果的に削減できていることがわかる。この成果は、国際会議において発表した。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計２件）

①Kenta Matsushima, Kouji Hirata, Miki Yamamoto, Fairness Improvement of Multiple-Bottleneck Flow in Data Center Networks, IEICE Trans. on Communications, 査読有, Vol.E99-B, No.7, 2016, pp.1445-1454
DOI: 10.1587/transcom.2015EBP3404

② Kenta Matsushima, Yuki Tanisawa, Miki Yamamoto, QCN/DC: Quantized Congestion Notification with Delay-based Congestion Detection in Data Center Networks, IEICE Trans. on Communications, 査読有, Vol.E98-B, No.4, 2015, pp.585-595
DOI: 10.1587/transcom.E98.B.585

〔学会発表〕（計９件）

① Kouji Hirata, Miki Yamamoto, Data Center Traffic Engineering Using Markov Approximation, IEEE ICOIN 2017, pp.173-178, 2017 年 1 月 12 日, Da Nang(Vietnam)

② Junya Akamatsu, Kenta Matsushima, and Miki Yamamoto, Equation-Based Multicast Congestion Control in Data Center Networks, APNOMS 2016, pp.1-6, 2016 年 10 月 7 日, 金沢商工会議所(石川)

③ Syuhei Okuda, Kouji Hirata, Miki Yamamoto, Improving Fairness between DCTCP and CUBIC in Datacenter Networks, 11th ISST 2016, 2016 年 7 月 27 日, 関西大学（大阪）

④ 奥田周平, 松嶋健太, 平田孝志, 山本幹, 李忠翰, 小口直樹, 田中淳, データセンタ環境下における TCP 間の公平性改善に関する一検討, 電子情報通信学会 2016 年総合大会, B-6-77, 2016 年 3 月 15 日, 九州大学(福岡)

⑤ 赤松純弥, 松嶋健太, 山本幹, データセンタにおけるエンドツーエンドマルチキャスト輻輳制御方式, 電子情報通信学会ネットワークシステム研究会, NS2015-202, 2016 年 3 月 4 日, フェニックスシーガイア(宮崎)

⑥ 平田孝志, 山本幹, マルコフ近似によるデータセンタトラヒックエンジニアリング, 電子情報通信学会ネットワークシステム研究会, NS2015-188, 2016 年 3 月 3 日, フェニックスシーガイア（宮崎）

⑦ Masaaki Takezaki, Kouji Hirata, Miki Yamamoto, Design of All-Optical Data Center Networks with Static Lightpath Establishment, 10the ISST 2015, 2015 年 9 月 1 日, Bangkok (Thailand)

⑧ Kenta Matsushima, Yuki Tanisawa, Miki Yamamoto, Fairness Improvement of Multiple-Bottleneck Flow in Data Center Networks, INFOCOMP2014, pp.103-108, 2014 年 7 月 23 日, Paris(France): Best Paper Award 受賞

⑨ 松嶋健太, 谷澤佑樹, 山本幹, データセンタ輻輳制御方式 QCN/BS の多段ボトルネック環境下での公平性改善手法, 電子情報通信学会ネットワークシステム研究会, NS2014-16, 2014 年 4 月 18 日, 石垣市民会館(沖縄)

６．研究組織

(1)研究代表者

山本　幹（YAMAMOTO, Miki）
関西大学・システム理工学部・教授
研究者番号：30210561

(2)研究分担者

平田　孝志（HIRATA, Kouji）
関西大学・システム理工学部・准教授
研究者番号：10510472