

# データセンター環境におけるショートフロー通信改善手法の一提案

藤居 翔吾<sup>†</sup>      田崎 創<sup>‡</sup>      関谷 勇司<sup>‡</sup>

<sup>†</sup> 東京大学大学院工学系研究科

<sup>‡</sup> 東京大学情報基盤センター

クラウド型のサービスの性質により、今日のデータセンターではデータセンター内のトラフィックが増大しており、複数の経路を持つネットワーク環境を応用して、通信性能向上を目指す取り組みがされている。しかし、様々なニーズを抱えるトラフィックが混在している中で、レイテンシ志向なサイズの小さいフローに対し、既存の MPTCP 実装では性能を劣化させる問題が報告されている。そこで本論文では、この問題に対し、並列分散処理アプリケーションが稼働しているクラスター PC のトラフィックを観測する事で、単一 NIC(Network Interface Card) への通信集中によるキュー負荷の影響があることを検討し、各ノードが複数の物理キューを持つネットワーク環境において、通信経路を切り替えることによりキューの負荷を分散させる手法を提案した。この手法による効果について、中継スイッチとエンドノードに対してスループット、フロー完結時間の二つのメトリックを用いた予備検証を行い、改善手法が与える影響を考察する。

## A proposal method for shortflow traffic in datacenter network

Shogo Fujii<sup>†</sup>      Hajime Tazaki<sup>‡</sup>      Yuji Sekiya<sup>‡</sup>

<sup>†</sup>The University of Tokyo, Graduate School of Engineering

<sup>‡</sup>The University of Tokyo, Information Technology Center

As increasing the amount of traffic in a datacenter by cloud service, the effective network for utilization of massive computer clusters has been studied. Recently, Multipath TCP(MPTCP) has been tackled this problem. MPTCP can achieve the effective consumptions of the resources with multipath, but a researcher reported MPTCP causes the delay of flow completion for short flows in such a multipath network environment. In this paper, I presented measurements of the distribute processing cluster PCs and reveal impairment mechanisms that lead to that latencies, rooted in single NIC(Network Interface Card) with intensive load traffic, and proposed the method balancing the load of queue for multiple queue, in such a MPTCP datacenter model. I verified the effect of the method for switch and end-node with two metrics, FCT(Flow completion time) of short flows and throughput of background flows, and considered the effects of the load balancing method as preliminary experiment.

### 1 まえがき

今日の一般家庭のインターネット接続環境がギガビット級の速度に達しようとしている中、多様な端末がインターネットに接続できるようになり、大量かつ多種多様なデータの取得が可能となった。特にトラフィックデータ量の増加傾向は顕著で、18~24ヶ月単位で総データ容量が2倍になるという予測がされている [1]。また Facebook では、300 ペタバイト以上のデータ量を保有しており、1日あたりに1ペタバイトのデータを解析している [5]。このように近年では、ビッグデータの活用が着目され、例えばウェブ検索エンジン、SNS(Social Networking Service)などのデータセンターを用いたクラウド型サービスにおいて、リアルタイムに近いレスポンスを返すような場面で使われ始めている。そのようなクラウドサービスには近年、より高

いユーザーエクスペリエンスが要求されており、Amazon では100[ms]の遅延により売り上げが1%下がる、といった報告 [2]があるように、例えばeコマースサイトでの商品購入や、インターネット広告のコンバージョンのようなユーザの意思決定へのレスポンス遅延の影響は深刻な問題である [3]。そのため、大規模データをより高速に処理することが求められており、データセンターではサーバ運用台数が増加の一途を辿っている。そうした中で、可用性、計算性能、低コストの三つの要件がデータセンターの抱える課題となっている [4]。特に計算性能について、大量の計算機資源から最大限の性能を引き出すためには、従来の仕組みではデータセンター内トラフィックに対して一部の資源にトラフィックが集中する問題に対応できないため、計算機資源を有効活用するための研究が盛んに行われている [6-11, 21]。そのようなスケーラビリティ

拡大には、ネットワークトポロジー、アプリケーション、プロトコルに対する三つのアプローチがある。

ネットワークトポロジーを改良するアプローチでは、従来の単純な階層構造では、データセンター内で発生するトラフィックに対して帯域が最大限割り当てられない [7]。そのため、近年ではそのようなトラフィックに対してスイッチを多段に構成することで帯域を有効利用するトポロジーが提案されており、マルチパス環境を実現している [7]。

大量データの処理速度を改良するアプローチでは、並列分散処理のために partition-aggregate 計算モデルが提案されている。MapReduce [6] 等の並列分散処理フレームワークは、これに従っており、今日の大規模クラウドサービスにおいて必要不可欠である。

プロトコルを改良するアプローチでは、従来の TCP を拡張した Multipath TCP (MPTCP) [13] をデータセンターネットワークに用いる提案がされている [7]。MPTCP を用いることにより、OS の制御によって複数の NIC(Network Interface Card)、複数の経路を同時に利用し、スループットを向上させることが期待されている。

しかし、並列分散処理フレームワークを用いることで発生する、大量のフローサイズの小さいクエリーフローが遅延を引き起こし、MPTCP を用いたマルチパス環境で、さらなる性能障害を引き起こす問題が報告された [9, 18]。

このような背景から、データセンターにおけるマルチパス環境でのショートフロー遅延の問題は並列分散処理性能の面で深刻な問題であり、本研究は、低レイテンシなネットワークによりコンスタントに性能の出せるデータセンターの実現を目指している。そこで本論文では、データセンターにおけるショートフロー遅延の問題解消のため、二つの検証を行った。

一つは、実環境の並列分散処理アプリケーションが稼働しているクラスター PC を用いて、トラフィックの観測、解析を行った。クラスター PC の定常状態時とジョブを実行している時の二種類のトラフィックを解析することで、並列分散処理アプリケーションが生成するトラフィックパターンの特徴および、汎用的なネットワーク機器を用いたデータセンターにおけるショートフロー遅延が生じる背景について、ボトルネックとなりうるネットワーク環境を検討し、その原因を示した。

二つ目は、マルチパス環境における、複数の NIC、複数の経路を利用した経路切り替えによる改善手法を提案する。これは、複数の経路を利用し、スイッチやエンドノードの持つ複数のキューへと負荷を分散させることで、単一のキューへ負荷が集中することで生じていたキューイングやプロトコル処理の遅延を抑える事ができる、という仮定に基づいた提案手法である。この手法により、並列分散処理におけるショートフローのような低レイテンシが求められる通信においては、他のフローの影響を受けずにすぐに割込み処理が行われ、受信処理の CPU 負荷の分散を期待しており、その効果を実機を用いた実験で検証した。そして今後の課題として、経路切り替えのアルゴリズムの検討をしていく。

## 2 関連研究

本章では、これまでに報告されている複数経路利用によるフロー完結時間短縮化技術について簡潔に述べ、その優位性や問題点を示す。

2010 年に Alizadeh らによって、データセンターネットワーク特有のトラフィックパターンに特化して、パラメータを決定するアルゴリズムが提案された [8]。データセンター特有のバースト性のあるトラフィックが引き起こす問題点として、キューの生成によるキューイング遅延、キュー溢れによるパケットロス、スイッチのバッファに掛かる負荷がある。これらの問題に対し、経由するスイッチにおいて、ECN(Explicit Congestion Notification) によってエンドノードに輻輳を通知し、キューの大部分が占有される前にウィンドウサイズを動的に変化させ、キューサイズを小さく保つアルゴリズムによって、大部分のキューの伝送時間を短縮することを可能にした。しかし、大規模計算資源を想定したトポロジーにおける検証がされておらず、また各ネットワークデバイスに細かなチューニングを必要とするため、大規模データセンターでは運用面での問題がある。さらに、サイズの大きいフローの割合の大きいトラフィックの中では、その影響によりウィンドウサイズを小さくする制御が働いたため、サイズの小さいフローの完結時間への効果が小さい、と報告されている [11]。

2012 年に Zars らによって、複数レイヤー間でトラフィックを監視し、しきい値を設定することによるフロー完結時間の短縮化技術を提案した [10]。今日のデータセンター内ネットワークのような、サイズの異なるフローが混在するネットワークにおいては、サイズが小さいフロー

がサイズの大きいフローに圧迫され、伝送遅延が大きくなる問題があったが、この提案手法では、データリンク層からアプリケーション層までの各層が、相互にトラフィックを監視する機能をスイッチに実装し、優先度をつけ、バッファサイズを調整することで、フロー完結時間の劣化を抑えることを可能にした。しかし、実験では Click [12] を用いて実装を行っており、現実世界での全てのネットワーク機器の置き換えが必要となるので、実現は難しい。

以上で述べたことをまとめると、近年のデータセンターネットワークに対して、以下のような要件が考えられる。

- 大規模計算機を有効活用するトポロジーの利用
- 分散処理の際に発生する大量のサイズの小さいフローの送信時間の短縮
- 特殊な実装やデバイスを用いず、汎用的でシームレスな運用の実現

### 3 データセンターネットワーク

本章では、データセンターネットワーク環境を構成する技術に関して、その概要を述べる。

#### 3.1 マルチパス環境を実現するネットワークトポロジー

従来のデータセンターモデルでは、階層的に二分木トポロジーを形成しており、トラフィックの大部分がデータセンター外の通信には有効であった [7]。しかし、今日のようなデータセンター内で生じるトラフィックが大半を占める場合、上の階層にある広帯域の経路を使わない通信が増え、帯域の割当が不適切となる。そこで近年の研究 [7] では、トラフィックがデータセンター内に集中した時の問題を、物理的なアプローチとして、トポロジーを工夫する事で解消を試みている。

FatTree [7] では、コアスイッチを複数用いる事で、物理パスの最大帯域を供給する。また、汎用的な性能のスイッチを多数用い、データセンター内のエンドノード同士の通信で複数の経路を実現する事で、冗長化、ネットワーク資源の効率的な利用、低コストを実現している。これら複数の経路をリンクエラー時の冗長性を持たせる目的だけでなく、性能向上に活用することが求められており、MPTCP ではスループットに対し、性能向上を実現する [9]。

既存の研究では、複数の経路を用いることで複数の帯域を同時に利用できる事に着目したも

のが主であったが、本研究では、NIC を追加する事で複数のキューが利用できる点に着目し、ショートフローの改善を目指す。

#### 3.2 トラフィックシナリオ

大量の計算機資源を有効活用するためには、並列分散処理フレームワークを用いられ、多数の処理ノードと分散処理の制御をする管理ノードから構成されている partition-aggregate 構造をとり、管理ノードからクエリーが発行され、処理ノードがそれを受け取り、レスポンスを返す。このとき、トラフィックパターンが (1) *Query traffic*, (2) *Short message traffic*, (3) *Background traffic* の 3 つに分類される [8]。

**Query traffic.** Query traffic とは、大規模計算処理を分割して並列処理を開始する際に、aggregator ノードから処理ノードへ具体的な処理を割り当てるためのトラフィックである。Query traffic の特徴は、非常に小さいフローサイズ (2KB~20KB) で、フローの役割上、処理全体の遅延に非常に強く影響を及ぼす事である。そのため、アプリケーション性能を考慮すると、低レイテンシでの通信が求められている。また並列分散処理システムの構成上、Query traffic は ms~ $\mu$ s 単位で生成され、バースト性があるといえる [8]。

**Short message traffic.** Short message traffic とは、処理ノードの動作を制御するためのトラフィックである。Short message traffic の特徴は、フローサイズは 50KB~1MB で、Query traffic と同様に処理全体の遅延に影響を及ぼすという事である。しかし、Query traffic ほどのフロー数は生成されず、生成間隔も秒単位である。

**Background traffic.** Background traffic は、各処理ノードへ更新データを送信するトラフィックである。Background traffic の特徴は、フローサイズが 1MB~50MB と大きいことにある。さらに、その生成間隔は大きい。また、Background traffic での更新データは、処理精度の向上に寄与するが、処理に必須ではないので、処理全体の遅延にはつながらない。

つまり、分散処理開始時に生成される Query traffic が遅延すると、処理全体に対し遅延を引き起こすので、Query traffic を含むショートフローのフロー完結時間は極めて重要なメトリックである。

また、Alizadeh らは、実際のデータセンターのトラフィックでは、レイテンシ志向なショートフローとスループット志向なロングフロー、そし

てバースト性のある Query traffic が混在していると報告している。さらに、Background traffic のフロー数自体は少ないが、全体のトラフィック量の大部分が Background traffic によって占められているという特徴がある [14]。

### 3.3 実トラフィック解析

この節では、クラウドサービスを想定したトラフィックの一例として並列分散処理アプリケーションを用いた二種類のトラフィックの測定結果を示す。測定結果からトラフィックの特徴を示す事で、従来の TCP で構成されたクラスターの抱えるボトルネックと複数のキュー、複数の経路を持つマルチパス環境におけるデータセンターモデルの利点をそれぞれ示し、提案手法の設計指針とする。

測定環境には、管理ノード 1 台 (Master)、処理ノード 10 台の計 11 台のクラスター PC を用いた。管理ノードは 10Gbps イーサネットリンクで Top of Rack (ToR) スイッチに接続されている。

このクラスター PC で Presto [5] によりインタラクティブなレスポンスを返す、分散 SQL データベースを実現しており、§3.2 で示した三種類のトラフィックが混在している。トラフィックの測定には、管理ノードのインターフェースを用いて、tcpdump [22] によるパケットレベルの測定を行った。

**定常状態:** 管理ノードに対し、ジョブ命令を一切与えていない中で約 10 時間程度トラフィックを測定した。図 1 に定常時のフローサイズの累積分布を示す。この分布から、80% 以上のフローが 10KB 以下であるようにショートフローの数が全体のトラフィックの大部分を占めていることがわかる。一方で通信量に着目すると、フロー数は比較的少ないがフローサイズの大きいトラフィックが大半を占めている。

次に、図 2 に管理ノードへのトラフィックの影響を示す。この分布が示すように、各処理ノードから管理ノードへのトラフィックの割合が大きく、それぞれフローサイズも大きい。一方で、管理ノードから各処理ノードへのトラフィックについては、比較的フローサイズの小さいトラフィックの割合が大きい。

さらに、図 3 に時間毎の同時接続数の分布を示す。図 3 中の長時間通信は通信時間が全測定時間の 90% 以上であるフロー数を表している。この分布が示すように、各処理ノードから管理ノードへのトラフィックの同時接続数が多く、積極的に通信が行われている。また、短い通信時

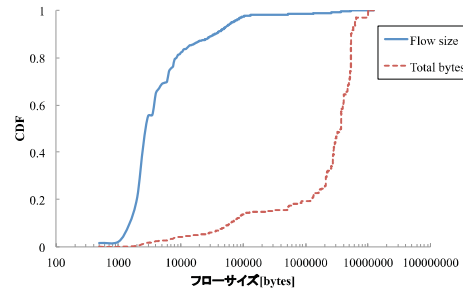


図 1 Presto クラスターの定常時のトラフィック分布

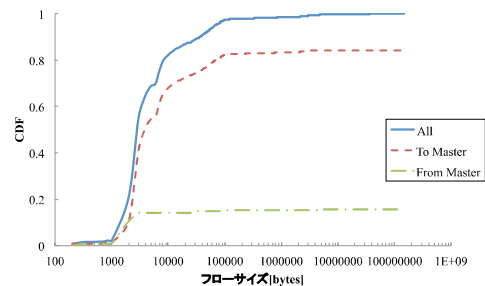


図 2 管理ノードから見た定常時のトラフィック累積分布

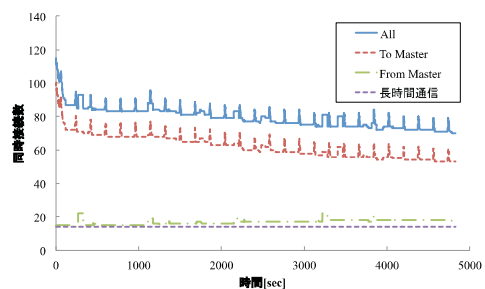


図 3 定常時トラフィック:同時接続数の分布

間でスパイク性のある中で、長時間通信を行うフローが固定的に存在している。

**並列分散処理実行時:** 管理ノードに対し、約 1 分間程度で完了する SQL ジョブを与えた中でジョブが完遂するまでの間トラフィック測定を行った。SQL ジョブには、“select \* from \$テーブル where \$条件”を実行し、全ての処理ノードにジョブを与えられるようにした。図 4 にジョブ実行時のフローサイズの累積分布を示す。この分布が示すように、ショートフローの数が全体のトラフィックの大半を占めるが、定常状態と比べると、全体的にフローサイズ大きいトラフィックが増えている。実際、80% 以上のフロー



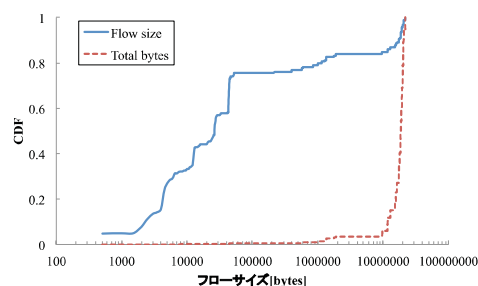


図4 Presto クラスターのジョブ実行時のトラフィック分布

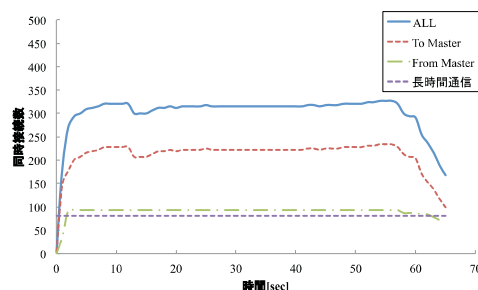


図6 ジョブ実行時トラフィック:同時接続数の分布

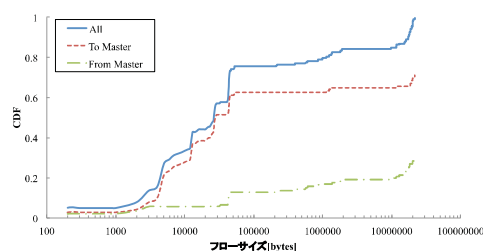


図5 管理ノードから見たジョブ実行時のトラフィック累積分布

が110KB以下であるように、ショートフローの割合が小さくなった。同様に通信量に着目すると、フロー数は比較的少ないがフローサイズの大きいトラフィックが大半を占めるという事が分かる。

次に、図5に管理ノードへのトラフィックの影響を示す。この分布が示すように、各処理ノードから管理ノードへのトラフィックの割合が大きく、フローサイズは小さいものが多いことが分かる。しかし、図2の定常時のトラフィックと比べると、管理ノードから各処理ノードへのトラフィックの割合が大きくなっている。

さらに、図6に時間毎の同時接続数の分布を示す。この分布が示すように、ジョブ実行中は全体的にフローの数は増え、とりわけ管理ノードから各処理ノードへのトラフィックの割合が大きくなっている。さらに、ジョブ終了後も同時接続数が大きく変化していないことから、長時間通信を行うフローが固定的に存在している。また、各処理ノードから管理ノードへのトラフィックに着目すると、ジョブ開始時に接続数が大きく増えていることから、バースト性があるトラフィックであるといえる。

これらの分布から、クラウド型サービスを想定したトラフィックの特徴として以下の事が述

べられる。

- 定常時もジョブ実行時も同様に、管理ノードへ送信されるトラフィック量は多い
- 長い時間通信を行うフローが固定的に存在している
- ジョブ実行時の処理ノードから管理ノードへのトラフィックには、フローサイズも小さく、バースト性がある

また、これらの特徴から、管理ノードへのトラフィックの集中、ショートフローのバースト性、そして長時間通信を行う Background traffic の問題が生じていると考えられる。従って、管理ノードに対するトラフィックとして大きく二つのパターンを検討する必要がある。

1. ジョブ開始時のバースト性のあるショートフロートラフィック
2. アプリケーション性能に直接影響しない Background traffic が通信している中で、低レイテンシ通信が求められているショートフローの通信

こうしたトラフィックは主に、複数の処理ノードから単一の管理ノードへの通信で発生する。そして、中継スイッチ、エンドノードともに単一のNICキューへとトラフィックが集中する事で、ボトルネックになり、ショートフローの遅延が問題となる [24]。

### 3.4 性能障害

次に、これらのトラフィックパターンが引き起こす可能性のある二ヶ所のボトルネックについて検討する。

#### 3.4.1 スイッチ - 性能障害

現在のスイッチ機器では複数のフローを多重に扱うための共有メモリを持ち、共有メモリプールからMMU(Memory Management Unit)によって各インターフェースが利用できるメモ

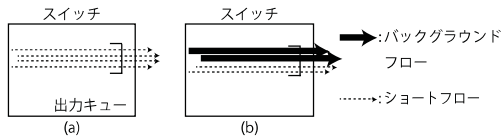


図7 中継スイッチで引き起こすボトルネック

リ量を動的に割り当てる事で、複数の通信を公平に処理する事を目指す [20]. しかし、比較的安価なスイッチでは制御できるメモリ量が制限されているため、様々な性能障害を引き起こす [20].

### Incast

図7(a)に示すように、短期間に一つのインターフェースへとフローが集中した場合、用意されているキューが溢れ、最悪の場合パケットロスを引き起こす。こうしたトラフィックは、§3.2で示した partition-aggregate 構造によるもので、リクエストを受けた処理ノードが同期して一斉にレスポンスを返すことにより、そのレスポンスを集約して受け取るノードが接続しているスイッチでのインターフェースのキューサイズが大きくなり、遅延、パケットロスを生じる。

### Queue buildup

§3.2で示したように、並列分散処理のレスポンスには直接影響しない Background traffic は、スイッチバッファにパケットロスを引き起こすほどの影響を及ぼし、そのインターフェースがボトルネックとなる可能性がある。図7(b)に示すように、Background traffic と Query traffic が同じインターフェースを利用する場合に、サイズの大きいフローによるショートフローのキューイング遅延が生じる。このとき、Query traffic には *Incast* とは異なり、バースト性は必要ではない。

### 3.4.2 エンドノード - 性能障害

今日の GbE (Gigabit Ethernet) 通信において、割込み処理は大きなボトルネック要因の一つである。例えば、1GbE において 64 バイトフレームの最大受信可能数は、毎秒約 150 万であり、1 パケット受信する度に割り込み処理を行うと、CPU リソースが枯渇する。そのため、割込み処理の回数を抑えることが必要であるが、その分レイテンシが上がる可能性があり、互いのトレードオフを適切に対処し高い性能を得る必要がある。また、今日の多くの CPU はマルチコアであり、CPU リソースを効率的に利用する事が求められている。

### 割込み処理

パケット受信の際の NIC によるハードウェア割

込みは、即座に受信処理を行う事ができ、キューイングの遅延を小さくすることができる。しかし、割込み処理が増えれば、その分オーバーヘッドが大きくなり、OS の性能が劣化する。割込み処理を扱う代表的な仕組みとして、ポーリング、interrupt coalescing がある。

ポーリングは NIC の割り込みを使わず、タイマーにより定期的に NIC の受信キューを監視することで、割り込み負荷を軽減するソフトウェア技術である。しかし、NIC でパケットを受けてから即座に処理できない為、遅延が発生する場合がある。現在の Linux カーネルにおいては、NAPI により、通信量が多く高負荷時にはポーリングが作用する [15]

interrupt coalescing は、複数のパケット、あるいは一定期間待ってからをまとめて一度で割り込ませる事で、割込み回数を減らすハードウェア技術である。しかし、ポーリングと同様、即座に処理できない為、遅延が発生する場合がある。

### プロトコル処理

マルチコア環境においても基本的には一つの NIC の受信処理は 1 つの CPU でしか行えない。そのため、ハードウェアへのアプローチとして、1 つの NIC に複数の受信キューを持たせて、受信処理をそれぞれの CPU へ分散させている、Receive Side Scaling (RSS) がある [16]. しかし、一般に複数受信キューを持ち、RSS 機能がある NIC は高価である [23]. そのため、一つしか受信キューを持たない NIC であっても、複数の CPU を分散させるソフトウェア技術として、RPS (Receive Packet Steering) がある [17]. これらの技術により、CPU の複数のコアをより効率良く利用する事ができる。また、プロトコル処理やアプリケーションでの処理については、RPS 等で複数の CPU へと分散させる事ができるが、その際の割込み処理についてはオーバーヘッドが生じる可能性がある。

## 4 提案手法

データセンターにおけるマルチパスネットワークの研究動向と、並列分散処理アプリケーションが生成する特有のトラフィックパターンが引き起こす機能障害をふまえて、改善手法を提案する。

提案手法の動機となったのは、MPTCP を用いたデータセンターネットワークモデルである [9]. このネットワークモデルでは、エンドノードが複数の NIC を持ち、一つのフローの通信でそれらを同時に利用することでスループット

トを向上させる．このような複数経路の効率的利用には、IP ベースのルーティングで実現しており、それぞれのエンドノードが持つ IP アドレスのペアにより、通信経路が決定する．現在の MPTCP の実装では、TCP コネクション確立後に互いの IP アドレスを交換し、新しいサブフローを形成する仕組みになっているため、サイズの小さいフローの通信では、サブフローを形成するまでに通信が完結する．以前の我々の解析では、このコネクション確立の際に遅延が生じることが分かっており [19]、どの経路を利用するかによって、性能性能が大きく変わる．提案手法の目的は、低レイテンシなネットワークを実現し、レイテンシ指向であるショートフローをなるべく遅延を抑えながら通信を完結する事である．これに対して、NIC を追加で設け、物理的に複数のキューを用意する事で、パケット処理の負荷を分散させる事で実現できると仮定し、その手法として、複数の NIC によるマルチパス環境でのトラフィック制御を提案する．

提案するトラフィック制御では、スイッチ、エンドノードに対してそれぞれのキューの混雑具合を考慮し、経路を決定する．本質的な狙いは、フローサイズに従って通信が終えるということであり、具体的には、フローサイズの大きいバックグラウンドフローが通信している中でショートフローが発生した場合、バックグラウンドフローが占有しているキューを避けた経路で通信することでフロー完結時間を抑えるというシナリオを実現するものである．そのために、複数経路のうちロングフローレーン (LFL)、ショートフローレーン (SFL) を設置し、優先度によって切り分けて通信を行う．

#### 4.1 キュー負荷分散

提案するトラフィック制御では、大きく 3 つのフェーズから構成されている．

**(1)Initial phase:** 提案手法ではフロー生成時にはショートフローであると仮定し、最も優先度が高く通信することができる．そこで、通信開始時点でエンドノードと中継スイッチに対して、SFL のうち最も混雑していない組み合わせを選択し、通信を行う．SFL が複数ある際には、クライアント側で通信しているフロー数によって分散する．

**(2)Judging phase:** 次に、そのフローがショートフローであるかどうかを判断する．ショートフローでないと判断された場合には、優先度が低い通信であると判断され、次のフェーズで LFL での通信へと切り替えられる．その際

に用いるメトリックとして通信占有時間を用い、しきい値を設定する．

**(3)Switching phase:** 提案手法では MPTCP を用いて、複数の経路を用いた通信が可能であると想定する．Judging phase において、ショートフローでないと判断された時、輻輳制御を用いて最初に割り当てられた通信経路のウィンドウサイズを下げ、経路の切り換えを行う．

§3.3 にて示した二つのトラフィックにパターンに対して、既存の単一 NIC、単一キューの環境では、スイッチ、エンドノードで通信が集中することでボトルネックとなり、通信性能を劣化させてしまう．提案手法では、通信経路を切り替え、複数の物理キューに対し負荷分散させる事で、これらのボトルネックを解消できる．

## 5 検証実験

この章では、実機での実験を用いて、低レイテンシでの通信が求められるショートフローに対して、バックグラウンドトラフィックが利用しているインタフェースを回避し、適切に経路を選ぶ事で、単一キューへの通信負荷の問題は解消され、ショートフローのフロー完結時間 (FCT) が改善できるという仮説の検証、また複数のキュー、複数の経路を利用した経路切り替えによる改善手法に対する予備実験を行う．具体的には、中継スイッチとエンドノードへのそれぞれの単一キューの負荷について、複数の NIC を用いて分散させ、その効果を検証する．

### 5.1 実験環境

#### (1) 中継スイッチに対する負荷実験

ネットワークトポロジーには、2 段で構成されたトポロジーを用いる．図 8 に、用いたトポロジーを示す．ベンチマークトラフィックについては、二つのペアに対してエンドノード同士の 1 対 1 通信を用いている．一方のペアに対しては、シミュレーションを実行している間、継続してデータ転送 (バックグラウンドトラフィック) を行う．他方のペアに対しては、TCP による 70Kbyte のデータ転送 (ショートフロー) を毎 10ms 一様生起させ、転送完了までにかかった時間 FCT (Flow Completion Time) を計測する．ショートフローのルーティングに関しては、図 8 に示す 3 つのパターンを用いて中継スイッチへのキューイング負荷の影響を検証する．

#### (2) エンドノードに対する負荷実験

ネットワークトポロジーには、二つの NIC を持った二つエンドノード同士を L2 スwitchを

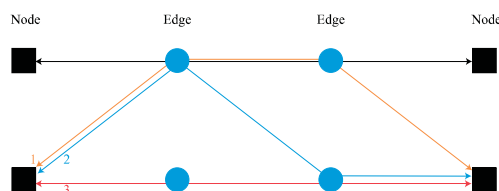


図8 中継スイッチへのNIC 負荷実験トポロジー

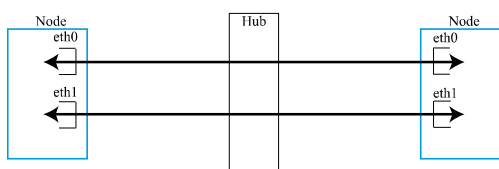


図9 エンドノードへのNIC 負荷実験トポロジー

| 項目         | スペック                 |
|------------|----------------------|
| OS         | Linux 3.13.0         |
| CPU        | Intel Xeon CPU L3426 |
| メモリー       | 4GByte               |
| NIC 対応ドライバ | e1000                |
| スイッチ-実験1   | Catalyst 2940(100T)  |
| スイッチ-実験2   | GS905L V2(1000T)     |
| リンク        | 1GbE                 |

表1 実験環境パラメータ

介してそれぞれのNIC毎に接続した。図9に、用いたトポロジーを示す。ルーティングに関しては、それぞれの対をなすNIC同士が通信を行う。ベンチマークトラフィックについては、エンドノード同士の1対1通信を用い、ショートフローとバックグラウンドフローを通信させる。バックグラウンドフローについては、ショートフローが通信しているNICペアと同じものを使って共有して通信させるパターンとショートフローが通信しているNICペアとは異なるペアのNICを用いて通信を行うパターンの2パターンについて検証する。ショートフローは、TCPによる70Kbyteのデータ転送を毎10ms一様生起させ、転送完了までにがかった時間を計測している。バックグラウンドトラフィックは、シミュレーションを実行している間、継続してデータ転送を行う。

表1に用いた機器の詳細を示す。

## 5.2 実験結果

### (1) 中継スイッチに対する負荷実験

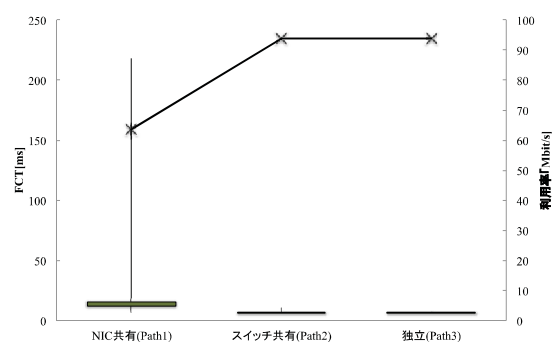


図10 中継スイッチに対する負荷実験での70kb ベンチマークトラフィックに対するフロー完結時間とリンク利用率

図10に上記の実験環境での結果として、70KBのショートフローのFCTとバックグラウンドフローの経路利用率を示す。FCTの箱ひげ図の上端には、95パーセンタイル値、下端には最小値を用いている。最大値でなく95パーセンタイル値を採用したのは、特に遅延した下位5パーセントに着目することで、遅延した割合の大きさを比較するためである。このメトリックにより、コンスタントにアプリケーション性能の出せるデータセンターネットワークの実現への指針となる。この結果から、ショートフローの通信が中継スイッチにおいて、バックグラウンドフローと経路およびインターフェースを共有した影響で、一部のフローが大きく遅延し、その分散が大きくなっている事が分かる。一方で、同じスイッチでインターフェースは共有しなかったフローに対しては大きな影響はなかった。これは、単一NICキューに対して二つのトラフィックが集中したことによる遅延の影響であると考えられる。その影響からバックグラウンドフローに対しては、スループットが下がっている事が分かる。

これらの事から、アプリケーション性能に直接影響しないバックグラウンドフローが通信している中で、低レイテンシ通信が求められているショートフローの通信をする際、中継スイッチでの利用するインターフェースが競合する場合、単一のキューに対しトラフィックが集中し、受信処理の割込みのオーバーヘッドや、プロトコル処理の遅延の影響が生じたと考えられる。

### (2) エンドノードに対する負荷実験

図11に上記の実験環境での結果として、70KBのショートフローのFCTとそれぞれの経路の利用率を示す。FCTの箱ひげ図の上端には、95パーセンタイル値、下端には最小値を用いてい



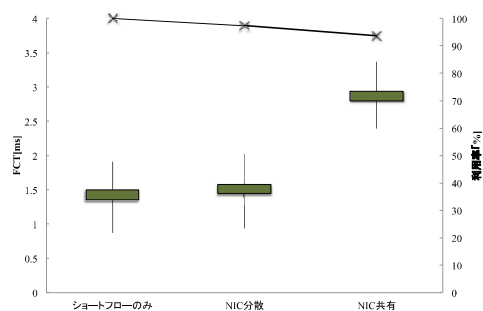


図 11 エンドノードに対する負荷実験での 70kb ベンチマークトラフィックに対するフロー完結時間とリンク利用率

る。この結果から、ショートフローの通信が、エンドノード間通信において、バックグラウンドフローと経路およびインターフェースを共有した影響で、FCT の分散が大きくなっている事が分かる。一方で、経路、インタフェースは競合しなかったものの、バックグラウンドフローとショートフローが同時に通信を行ったことで若干の遅延の影響が生じ、分散が大きくなっている。これは、単一 NIC に対して二つのトラフィックが集中したことによる負荷分散の効果が得られたが、プロトコル処理以降の部分で、複数のフローが同時に通信を行った事に対するオーバーヘッドが生じたと考えられる。またバックグラウンドフローに着目すると、インターフェースを共有した場合においては、ショートフローだけでなくバックグラウンドフローにも遅延が生じ、スループットが低下している。

これらの事から、アプリケーション性能に直接影響しないバックグラウンドフローが通信している中で、バースト性のあるショートフロートラフィックの通信をする際、エンドノードに対して、利用するインタフェースが競合する場合、単一の NIC に対しトラフィックが集中する事で、受信処理の割込みのオーバーヘッドや、プロトコル処理の遅延の影響があると考えられる。

### 5.3 考察

これらの解析結果から、エンドノード、スイッチに対する機能障害が引き起こる要因について述べ、今後の改善手法の検討を行う。大量の計算機資源をいかに効率的に利用するか、という課題を今日のデータセンターは抱えており、並列分散処理アプリケーションを用いる事が一般的である。今の並列分散処理システムが partition-aggrigation 構造である限り、管理ノードや多段のクラスター構成であればアグリゲーター

ノードに対して、処理ノードからのトラフィックが集中する問題は発生する。その結果、Queue buildup や Incast のような単一キューへの負荷集中の問題が中継スイッチやエンドノードに対して生じ、CPU 性能を効率的に引き出せず、並列分散処理の性能が劣化する。

こうした遅延の影響を軽減する為には、混雑時の通信量を抑える制御を行う、あるいは混雑時にも空いているリソースを効率良く利用する事が必要である。MPTCP による既存の計算機資源に対して複数 NIC を用いて性能向上を目指すように、複数のフローを通信する際に異なる物理インターフェースを利用する事で、マルチコアを持つ CPU の効率的な利用につなげられる。すなわち、複数のキューに対して通信を分散させるようなトラフィック制御により、例えばレイテンシ志向なショートフローとスループット志向なバックグラウンドフローのような役割の異なるトラフィックを共存させ、最適な通信の実現が可能であることが実験により明らかとなった。そのような物理的に複数の NIC によりマルチキューが介在する汎用的な機器で構成されているネットワークの中で、トラフィックをどのように制御するかという点については、スイッチやエンドノードの OS スタック等のどこで制御をするか、またどのようなアルゴリズムでそれを実現するかを検討する必要がある。

### 5.4 Directing result

経路混雑時のトラフィック制御について、今後の指針となる一つの結果を示す。提案手法では、適切な経路を選んで通信を開始をする必要がある。経路を決める際のメトリックとしては様々なものが考えられ、その中の一つとして経路の利用率がある。経路利用率がショートフローの通信に対してどのような影響があるのか示した結果を図 12 に示す。実験環境には、中継スイッチに対する負荷実験と同様であり、バックグラウンドフローの負荷の割合を変化させた。この結果から、70 ~ 80 % の経路利用率の場合、遅延するショートフロー発生する事が分かり、適切な経路選択には利用率も一要因として考慮する必要がある。

## 6 あとがき

本論文では、データセンターにおけるショートフロー遅延の問題を解決する為に、二つの検証を行った。

一つは、実環境の並列分散処理アプリケーションが稼働しているクラスター PC を用いて、実ト

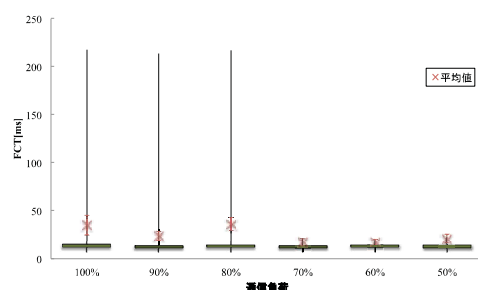


図 12 スイッチに対する負荷実験でのバックグラウンドフローの影響

ラフィックの解析を行うことで、ジョブ開始時のバースト性のあるショートフローとラフィックと Background traffic が通信している中でのショートフローの通信のトラフィックパターンが存在し、汎用的なシングルキュー NIC を用いたネットワーク機器においては機能障害を引き起こす原因となることを検討した。

二つ目は、MPTCP を用いたデータセンターモデルにおける、複数のキュー、複数の経路を利用した経路切り替えによる改善手法を提案した。この改善手法により、NIC キューが他のフローの影響を受けずにすぐに割り込み処理を行い、レイテンシ志向なショートフローへの受信処理の遅延を軽減する事ができた。また、スループット志向であるバックグラウンドフローに対しても、通信性能を向上させた。

今後はネットワークの通信状況を考慮しながら、通信経路を切り替え、通信状況を最適化するアルゴリズムの検討を行い、コンスタントにアプリケーション性能の出せるデータセンターの実現を目指す。

## 参考文献

- [1] 日本アイ・ビー・エム株式会社. IBM 第 1 章大容量データのバックアップ, <http://www-06.ibm.com/systems/jp/storage/column/backup/01.html>
- [2] Jim Liddle. Amazon found every 100ms of latency cost them 1% in sales, August 2008. <http://blog.gigaspace.com/amazon-found-every-100ms-of-latency-cost-them-1-in-sales/>
- [3] R. Kohavi et al. Practical Guide to Controlled Experiments on the Web: Listen to Your Customers not to the HiPPO. KDD, 2007.
- [4] J. Hamilton. On designing and deploying Internet-scale services. In USENIX LISA, 2007.
- [5] Facebook. Presto: Interacting with petabytes of data at Facebook, <https://www.facebook.com/notes/facebook-engineering/presto-interacting-with-petabytes-of-data-at-facebook/10151786197628920>
- [6] Dean, Jeffrey, and Sanjay Ghemawat. "MapReduce: simplified data processing on large clus-

ters." Communications of the ACM 51.1 (2008): 107-113.

- [7] Al-Fares, Mohammad, Alexander Loukissas, and Amin Vahdat. "A scalable, commodity data center network architecture." ACM SIGCOMM Computer Communication Review. Vol. 3
- [8] Alizadeh, Mohammad, et al. "Data center tcp (dctcp)." ACM SIGCOMM Computer Communication Review 40.4 (2010): 63-74.
- [9] Raiciu, Costin, et al. "Improving datacenter performance and robustness with multipath TCP." ACM SIGCOMM Computer Communication Review. Vol. 41. No. 4. ACM, 2011.
- [10] Zats, David, et al. "DeTail: Reducing the flow completion time tail in datacenter networks." ACM SIGCOMM Computer Communication Review 42.4 (2012): 139-150.
- [11] Alizadeh, Mohammad, et al. "pFabric: Minimal near-optimal datacenter transport." Proceedings of the ACM SIGCOMM 2013 conference on SIGCOMM. ACM, 2013.
- [12] Kohler, Eddie, et al. "The Click modular router." ACM Transactions on Computer Systems (TOCS) 18.3 (2000): 263-297.
- [13] Ford, Alan, et al. TCP Extensions for Multipath Operation with Multiple Addresses: draft-ietf-mptcp-multiaddressed-03. No. Internet draft (draft-ietf-mptcp-multiaddressed-07). Roke Manor, 2011.
- [14] Benson, Theophilus, Aditya Akella, and David A. Maltz. "Network traffic characteristics of data centers in the wild." Proceedings of the 10th ACM SIGCOMM conference on Internet measurement. ACM, 2010.
- [15] J. Salim, When NAPI Comes to Town, Proceedings of Linux 2005 Conference, UK, August 2005.
- [16] Microsoft corporation. scalable networking with rss, 2005.
- [17] Herbert, T. rps: receive packet steering, september 2010. <http://lwn.net/Articles/361440/>.
- [18] Vasudevan, Vijay, et al. "Safe and effective fine-grained TCP retransmissions for datacenter communication." ACM SIGCOMM Computer Communication Review. Vol. 39. No. 4. ACM, 2009.
- [19] 藤居 翔吾, 田崎 創, 関谷 勇司, "MultiPath TCP 適用時のデータセンターネットワークでのフローサイズが与える影響に関する一考察", 電子情報通信学会, 信学技法, vol. 113, no. 364, IA2013-65, pp. 47-52, 2013.
- [20] P. Agarwal, B. Kwan, and L. Ashvin. Flexible buffer allocation entities for traffic aggregate containment. US Patent 20090207848, August 2009.
- [21] S. Floyd and V. Jacobson. The synchronization of periodic routing messages. IEEE/ACM ToN, 1994.
- [22] tcpdump, <http://www.tcpdump.org/>
- [23] intel, <http://www.intel.com/content/www/us/en/network-adapters/gigabit-network-adapters/ethernet-server-adapters.html>
- [24] Lu, Guohan, et al. "Using cpu as a traffic co-processing unit in commodity switches." Proceedings of the first workshop on Hot topics in software defined networks. ACM, 2012.