

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

東京大学工学系研究科

藤居翔吾

Outline

- ・ 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
- 1. 研究背景
- 2. 関連研究
- 3. データセンターネットワーク
- 4. 再現シミュレーション
- 5. 結論

研究背景

ビッグデータ : データの爆発的増加が

Facebookでは1ペタバイトの
1日に1ペタバ

スケール

ク

ク増加

**MPTCPはサイズの小さい
フローには悪影響を及ぼす**



Multipath TCP(MPTCP)でデータセンターネットワークを改善!!

[3]<https://www.facebook.com/notes/facebook-engineering/presto-interacting-with-petabytes-of-data-at-facebook/10151786197628920>

なぜ、ショートフローに着目するのか？

分散・並列処理技術：ビッグデータ、大量の計算資源の活用

分散・並列処理では大量のショートフローを生成してしまう!!

データセンタートラフィックの80%がショートフロー[16].

ショートフローは大規模計算処理の高速化のために極めて重要な要素

MPTCPを用いてデータセンターネットワークを改善する上でショートフローの問題は重要な問題

[16]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.

Motivated work

Raiciu, Costin, et al. "Improving datacenter performance and robustness with multipath TCP." *ACM SIGCOMM Computer Communication Review*. Vol. 41. No. 4. ACM, 2011.

Motivation:

- 今日の密なデータセンターネットワーク資源の有効利用

Target :

- MPTCPで複数のパスを利用してスループットを改善

Achievement :

どのトポロジーにおいてもMPTCPはスループットを改善した

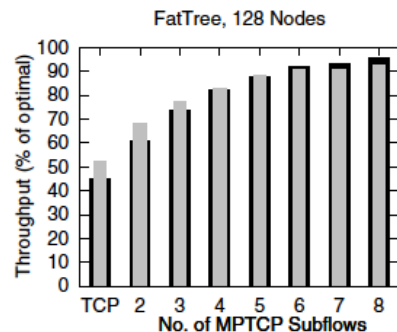


Fig1. Utilization on FatTree [10]

課題:

MPTCPがTCPによるショートフロー完結時間に悪影響を及ぼした

Table1. The effect of short flows completing for 70KB

| Algorithm | フロー完結時間 (mean/stdev) |
|-----------------|-------------------------|
| SINGLE-PATH TCP | 78 ± 108 ms |
| MPTCP | 97 ± 106 ms |

[10]Raiciu, Costin, et al. "Improving datacenter performance and robustness with multipath TCP." *ACM SIGCOMM Computer Communication Review*. Vol. 41. No. 4. ACM, 2011.

関連研究

Zats, David, et al. "DeTail: Reducing the flow completion time tail in datacenter networks." *ACM SIGCOMM Computer Communication Review* 42.4 (2012): 139-150.

Motivation:

- ユーザエクスペリエンスのために、Webページ表示時間を保証する

Target :

- ショートフローのバースト性によるパケットロスが減らし、遅延を抑える

Achievement :

実装したスイッチでトラフィックを監視し
バッファを動的制御

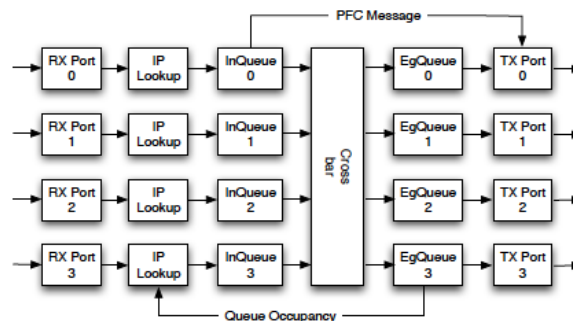


Fig2. proposed switch architecture[11]

Result :

ショートフローに対し99パーセンタイル
の完結時間を40%改善

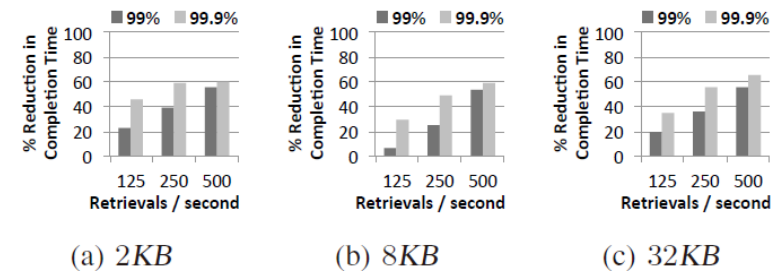


Fig3. Microbenchmarks for all-to-all workload

[11]Zats, David, et al. "DeTail: Reducing the flow completion time tail in datacenter networks." *ACM SIGCOMM Computer Communication Review* 42.4 (2012): 139-150.



研究について

目標：既存のネットワークと大量の計算資源でビッグデータを処理する

データセンターネットワークの要求案件

1. 大量の計算資源を有効活用するトポロジー
2. シームレス性：特殊な実装、デバイスを用いずに性能向上
3. アプリケーション性能向上を目的とした改善

アプローチ：

1. FatTreeトポロジー
2. MPTCPを利用
3. ショートフローの完結時間を改善
 - 報告されたショートフロー問題の再検証
 - MPTCP改善の方向性を示す



データセンターネットワーク 構成要素

データセンターネットワーク構成要素 トポロジー

従来の階層構造

- 大量の計算資源を抱えにくい
- 帯域の割当てが不適切
- データセンター内のトラフィックに対応できない

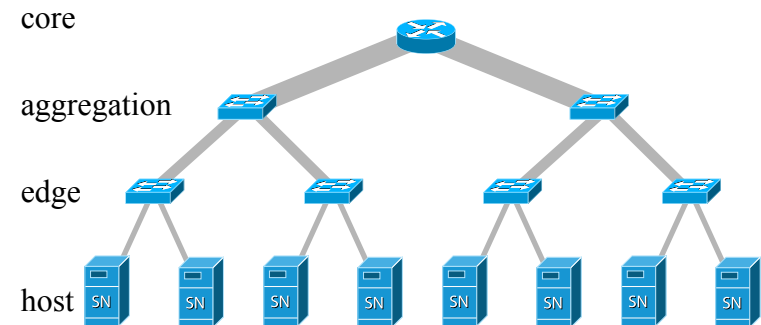


Fig4. Hierarchical topology

近年のトポロジー

- 帯域の効率的な利用
- ホスト間通信での複数のパス

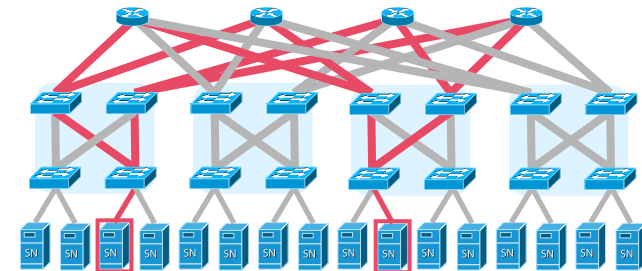


Fig5. FatTree topology[6]

複数のパスを冗長化だけでなく、**性能向上**へ

[6]Al-Fares, Mohammad, Alexander Loukissas, and Amin Vahdat. "A scalable, commodity data center network architecture." ACM SIGCOMM Computer Communication Review. Vol. 38. No. 4. ACM, 2008.

データセンターネットワーク構成要素 プロトコル

Multipath TCP(2011)

- シームレス性：既存のTCPを拡張, 一つのコネクションで複数のパスを利用
- 複数のパスを同時に利用する事で, スループットを改善

RTT, MSSに差がない場合

$$\omega_{total} \text{ の増加量} \approx \alpha \times MSS$$

$$\omega_r \text{ の増加量} \approx \alpha \times \frac{\omega_i}{\omega_{total}}$$

$$\alpha = \omega_{total} \times \frac{\max_r \frac{w_r}{RTT_r^2}}{\left(\sum_r \frac{w_r}{RTT_r}\right)^2}$$

[14] Raiciu, C., M. Handley, and D. Wischik. "Coupled congestion control for multipath transport protocols." draft-ietf-mptcp-congestion-01 (work in progress) (2011).

- データ送信量は最も性能の良いパスに依存し, 各サブフローのウィンドウサイズを増加させていく
- 性能の悪いパスを利用する割合を下げる

データセンターネットワーク構成要素 アプリケーション

分散・並列処理

大量の処理ノードと数台の管理ノードから構成される

基本的に分散・並列処理技術はPartition-aggregate計算モデルに従う

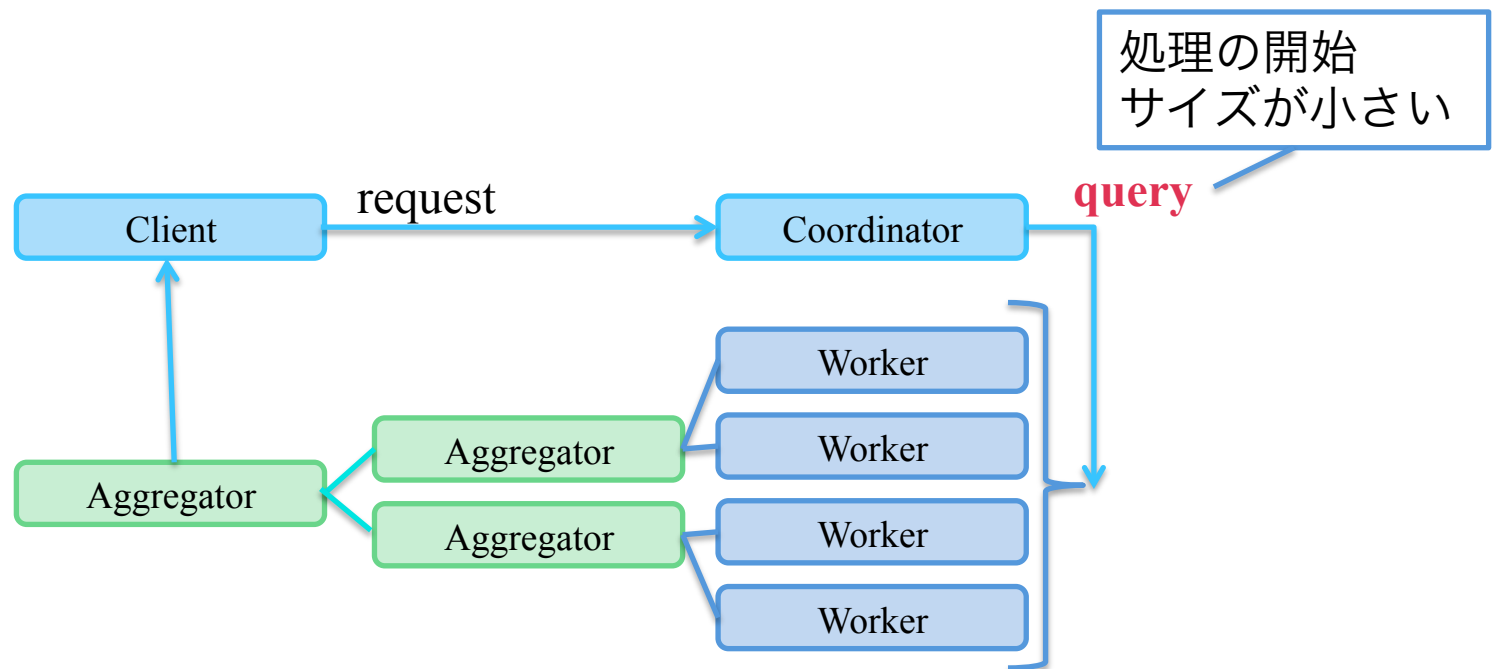


Fig6. partition-aggregate model of Presto[3]

[3] 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>



再現シミュレーションの目的

- 実際のデータセンターネットワークでは, 様々な大きさのフローが混在している.
- 特に, MPTCPはサイズの大きいフローのスループット改善には貢献する.

MPTCPによりサイズの小さいフローへの影響はないのか?

再現シミュレーションにより深い考察



再現シミュレーション

再現シミュレーション -概要

再現シミュレーション環境

トポロジー: FatTree, oversubscribed 4 : 1

70KBの通信の完結時間を測定

ランダム性

- 通信ノードをどう選ぶか
- 50回シミュレーションを実行

シミュレーター

- ns-3 dceを使用
- 再現元論文:htsimあるいはflow-level simulatorを使用

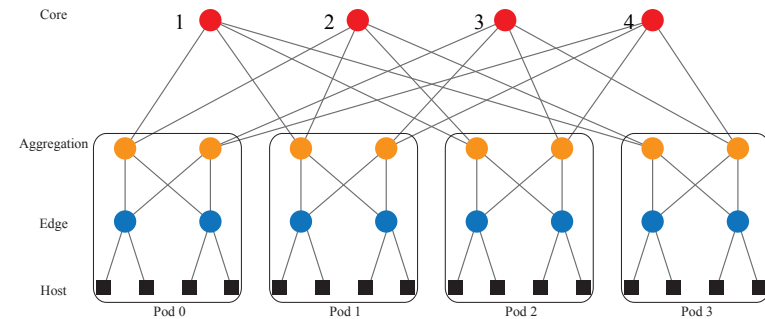


Fig7-1. Network topology on simulation

任意に設定したパラメータ

| Parameter | Value |
|------------------|---------|
| nodes | 16 |
| core-aggr | 400Mbps |
| aggr-edge | 200Mbps |
| edge-host | 100Mbps |
| RTT | 0.5ms |
| Buffer | 100KB |

再現シミュレーション -トラフィックパターン

1. トラフィック:33%のノードがデータを流し続けるバックグラウンドトラフィック(**TCP or MPTCP**)
2. 残りのノードが70KBの通信を平均200msポアソン生起 (**TCP**)

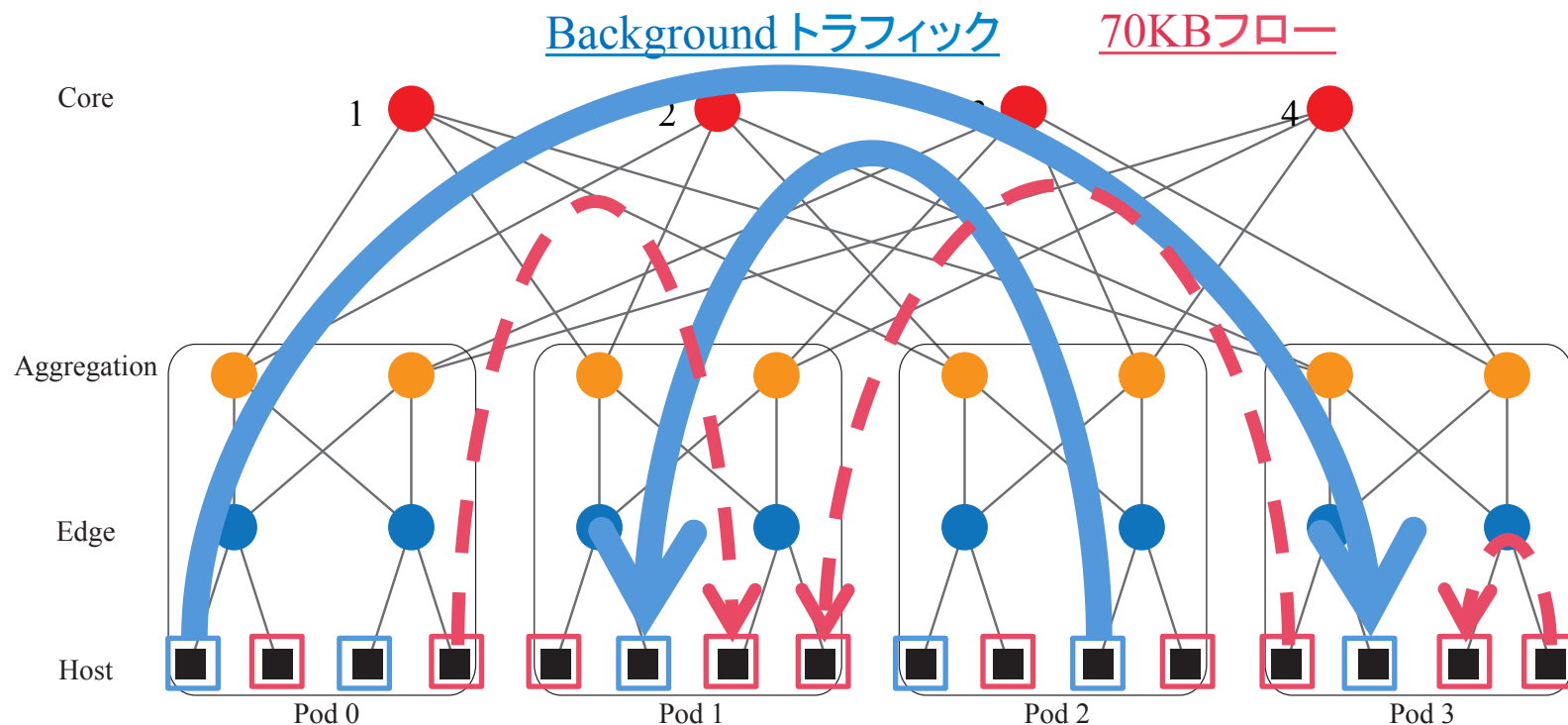



Fig7-2. Network topology on simulation



再現シミュレーション - パラメータの検証

Round Trip Time(RTT) : 0.5ms

同データセンター内のRTTは一般的に1ms以下[18]

Buffer: 100KB

帯域遅延積 : $BDP[\text{byte}] = \text{bandwidth}[\text{bps}] \times RTT \div 8$

$$100[\text{KB}] = 400[\text{Mbps}] \times 0.5[\text{ms}] \div 8 \times 4$$

帯域とノード数

今のデータセンターネットワークでは400Mbpsよりも広帯域

今回のシミュレーションでは16ノードに対し, 帯域をチューニングし,
結果を再現した

[18] Vasudevan, Vijay, et al. "Safe and effective ne-grained TCP retransmissions for datacenter communication." ACM SIGCOMM Computer Communication Review. Vol. 39. No.4. ACM, 2009.

再現シミュレーション - TCP vs. MPTCP

Table1. Results on reproducing simulation

| Algorithm | フロー完結時間 [ms] | 標準偏差[ms] | 99パーセン タイル | フロー完結時間 [ms] | 標準偏差[ms] |
|--------------------|-----------------|----------|---------------|-----------------|----------|
| SINGLE-PATH TCP | 78.4 | 122.5 | 266.7 | 78 | 108 |
| MPTCP | 91 | 140.6 | 510.5 | 97 | 106 |

Table2. Reported results

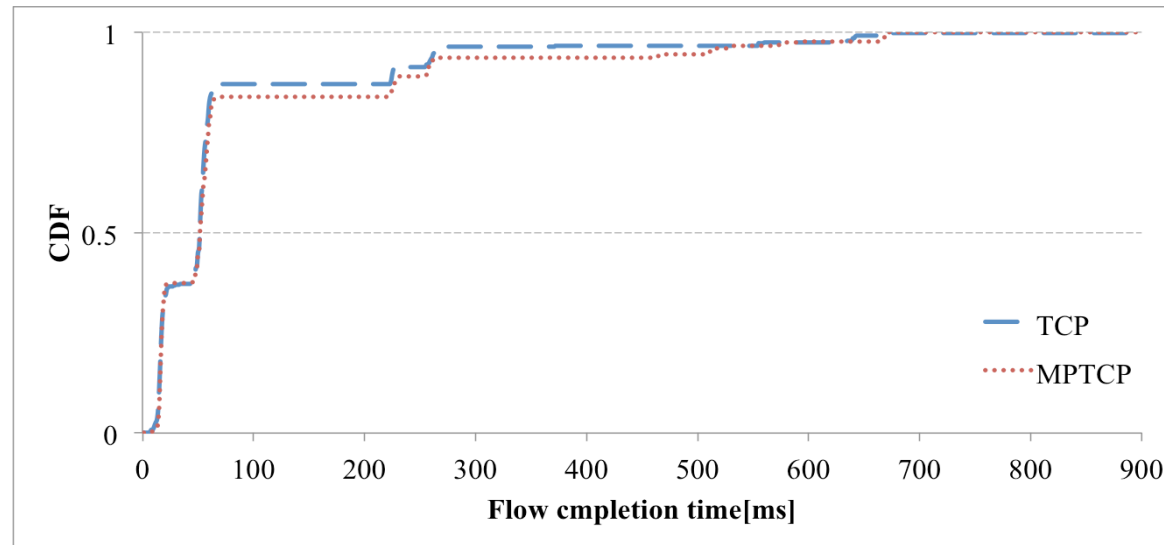


Fig8. CDF of flow completion time on reproduction experiment

- MPTCPによるバックグラウンドトラフィックが送信側のバッファを圧迫し、70KBフローでの遅延を引き起こした

再現シミュレーション - 結果

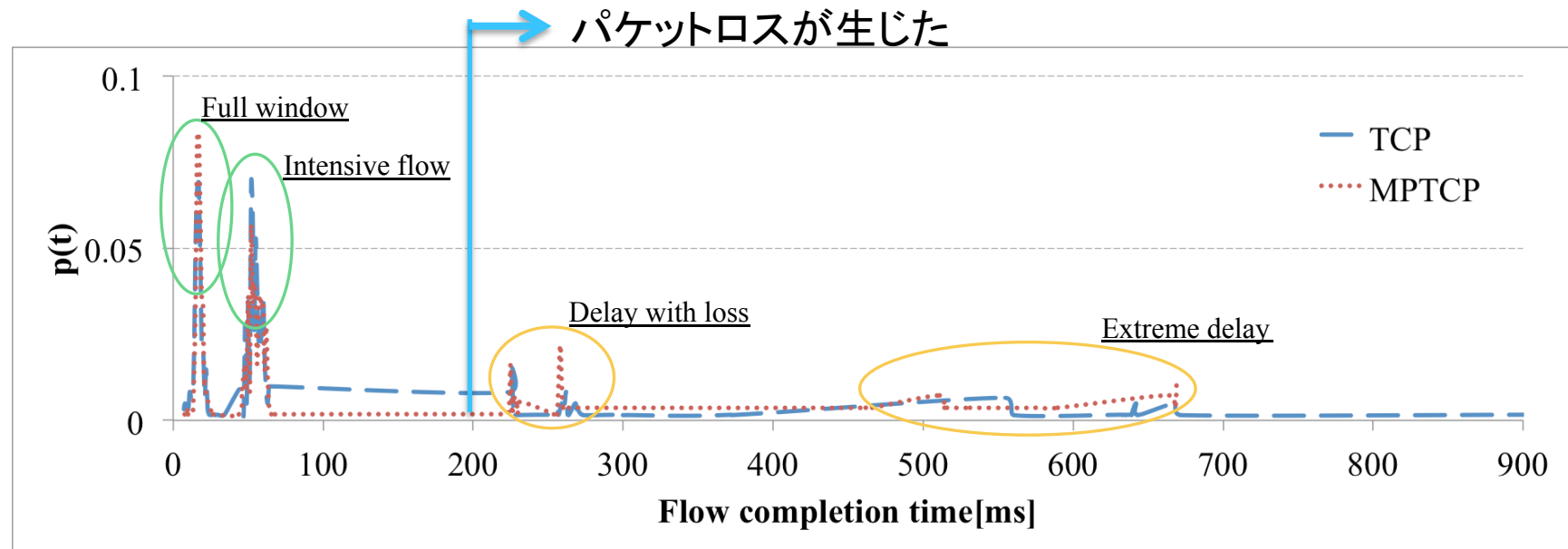


Fig9. Frequency of flow completion time on reproduction experiment

- 完結時間の分布から4つのパターンが現れた
- MPTCPはパケットロスを生じる割合が大きかった

Delay with loss vs. Extreme delay

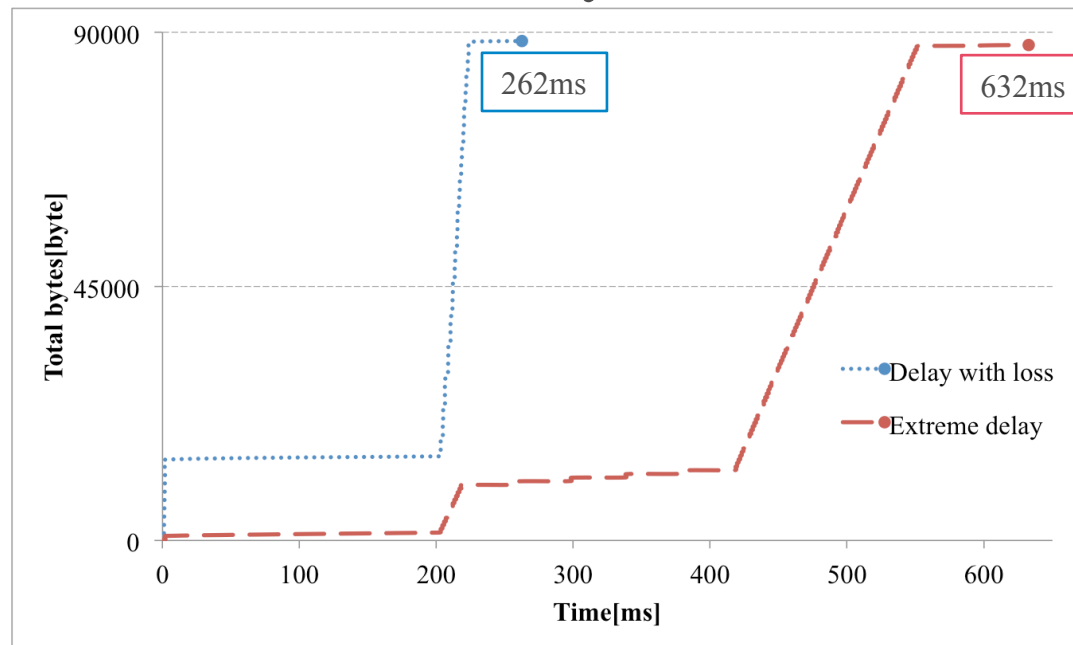


Fig10. Comparison between Delay with loss and Extreme delay

- トラフィックが数10[ms]間隔で発生したとき, パケットロスを起こしていた.
- MPTCPでは特にコネクション確立直後にパケットロスを起こす割合が高く, 送信量を減らす制御が発生した



結論

TCPとMPTCPによるバックグラウンドトラフィックの影響の比較

- MPTCPによるバックグラウンドトラフィックがパケットロスを多く引き起こした
 - MPTCPは混雑していないパスではウィンドウサイズをより増加させる
 - 短い間隔でショートフローが発生したとき, バッファの圧迫を促進させた

MPTCPはTCPに対しパケットロスを引き起こし遅延する



今後の課題

- MPTCPによるバックグラウンドトラフィックが帯域を占有し、遅延を引き起こす問題を解決する必要がある

新たな輻輳制御の提案：ウィンドウサイズの増加の制御

MPTCP実装の改善：輻輳が起きているパスを回避







追加シミュレーション

追加シミュレーション - 概要

シミュレーション環境 – トラフィック

バックグラウンドトラフィック

50% 処理ノードに対し, データを流し続ける

ショートフロー

- **Query トラフィック** : 全ての処理ノードに対し, 16KB~2KBのデータを平均200[ms]でポアソン生起
- **Short message** : 全ての処理ノードに対し, 1MB~50KBのデータを平均500[ms]でポアソン生起

ランダム性

バックグラウンドトラフィックを通信するノードの選び方

1000 回のシミュレーション

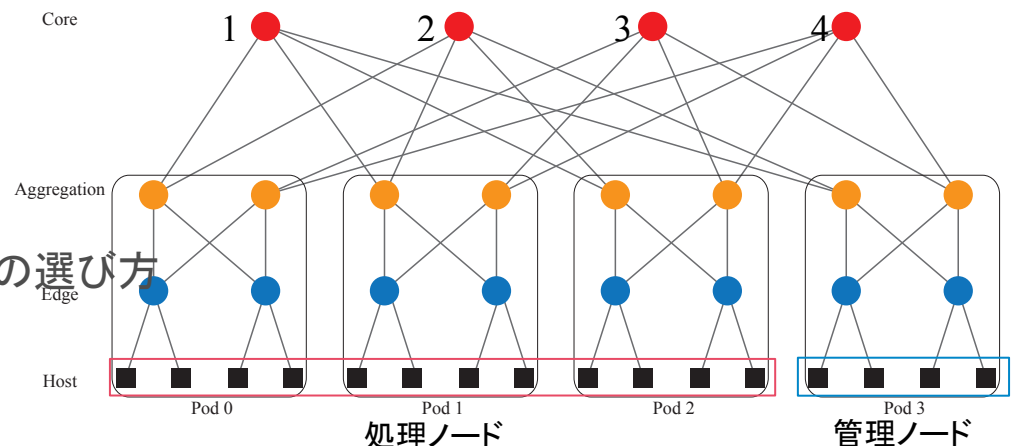


Fig11. Environment of additional simulation

追加シミュレーション -バックグラウンドトラフィック なし

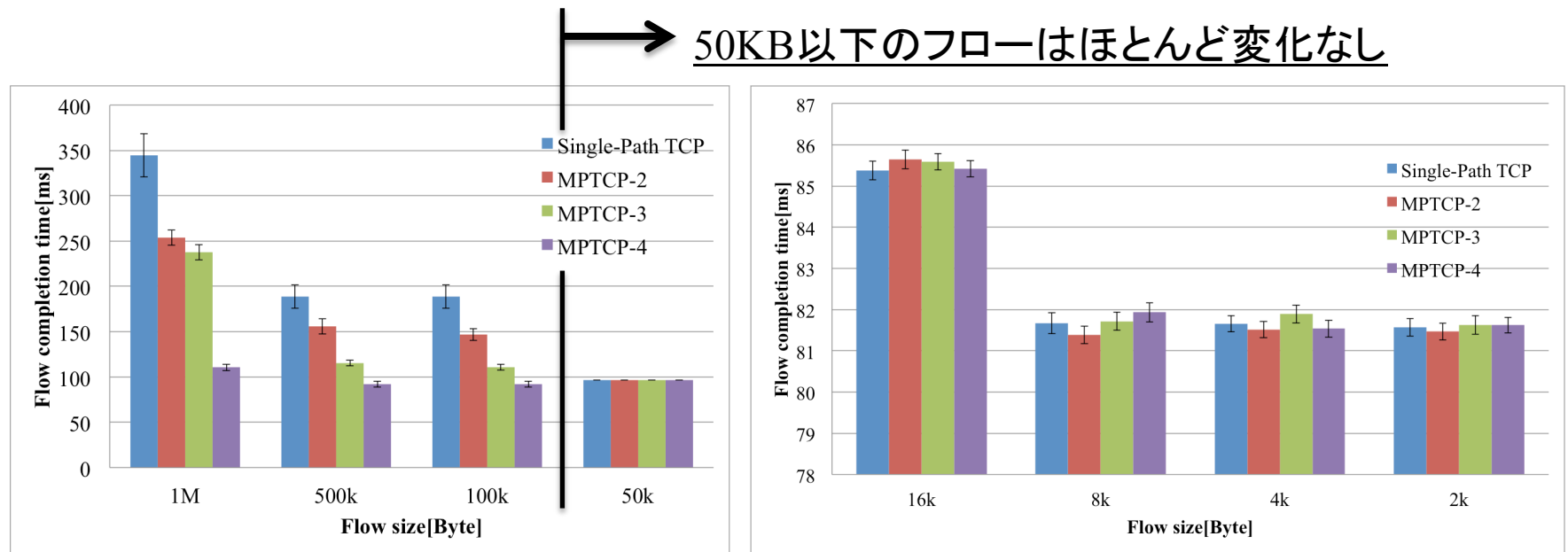


Fig12. Flow completion time without background traffic

- 50KBより大きいフローでは, MPTCPの効果が出た
- 50KBより小さいフローでは, TCPと同じ挙動だった

追加シミュレーション -バックグラウンドトラフィック あり

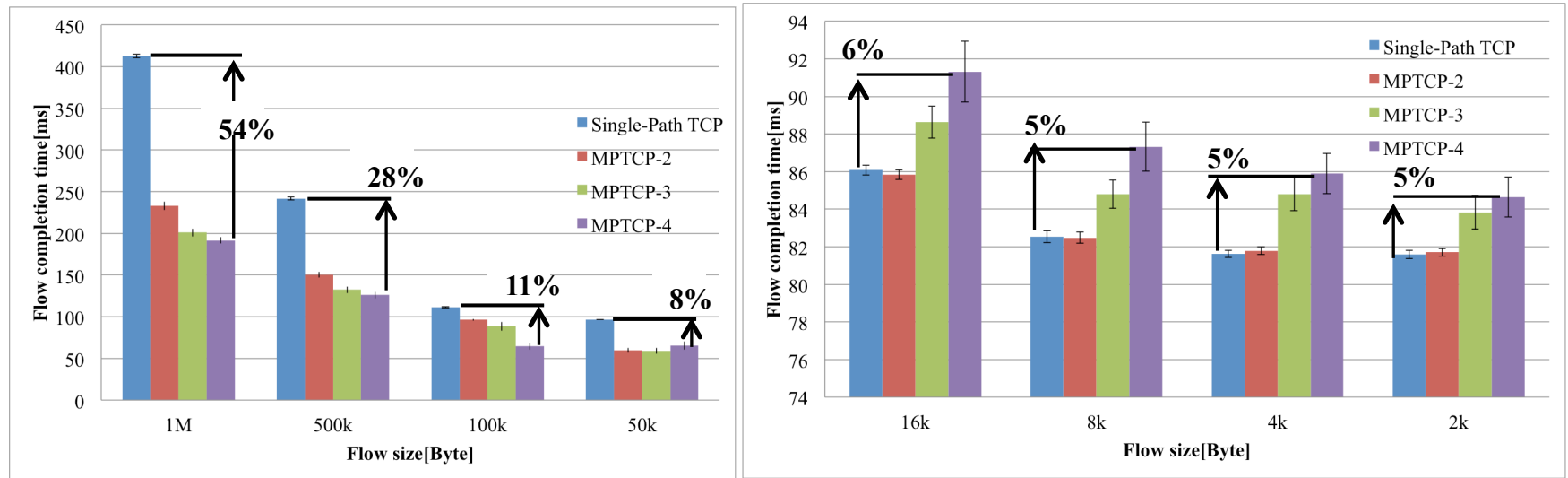



Fig13. Flow completion time with background traffic

- 50KBより小さいフローでは, MPTCPによるバックグラウンドトラフィックの影響を受け, 遅延を生じた.



追加シミュレーション - まとめ

- 50KBより大きいフローに対し, MPTCPにより完結時間を短縮した
- 50KBより小さなフローに対してはTCPと同じ挙動だった
 - MPTCPが働くしきい値のようなものがあるのか(検討点)
- 結果的にバックグラウンドトラフィックの影響を受けて遅延を生じた

- MPTCPによるバックグラウンドトラフィックが帯域を占有し, 遅延を引き起こす問題を解決する必要がある

プロトコルレベルでのアプローチ

Congestion control: ウィンドウサイズの増加の制御

MPTCPの改善: ショートフローを直接作用させる

必要な事:

ショートフローかバックグラウンドトラフィックなのかを識別
帯域を適切に割当てるアルゴリズム