MMPTCP: A Novel Transport Protocol for Data Centre Networks

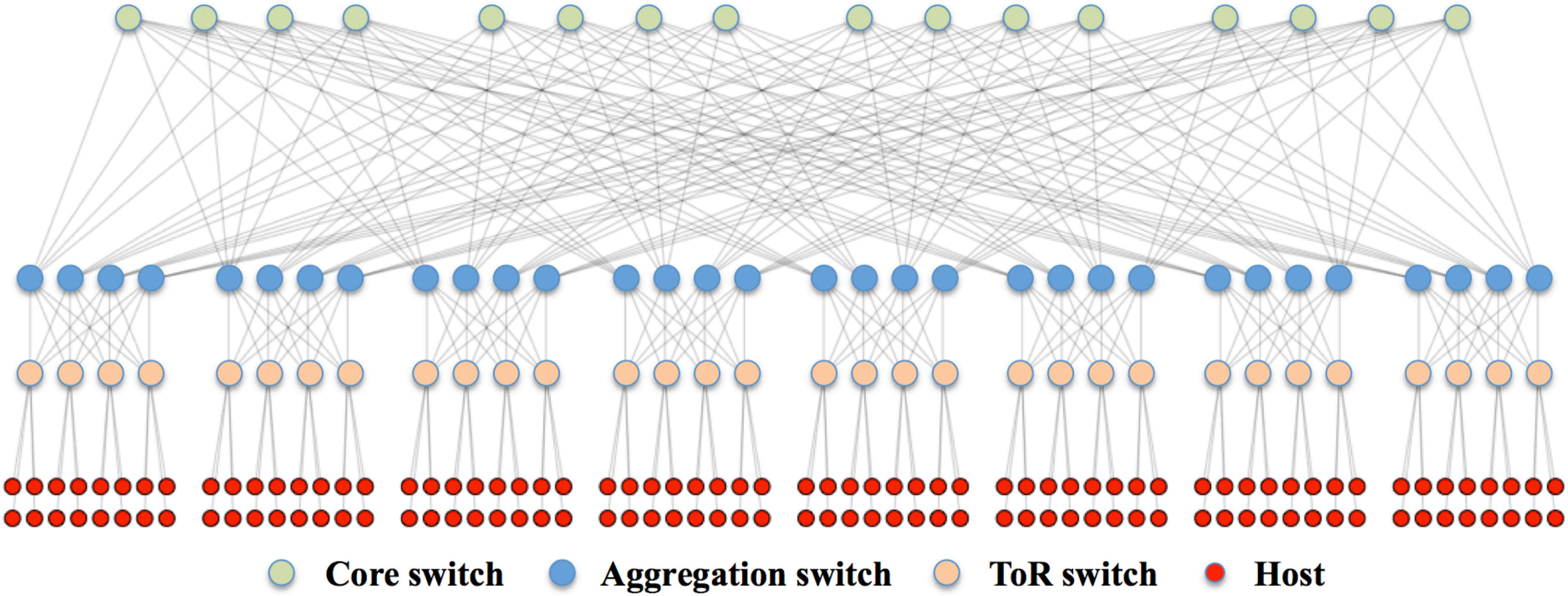
Morteza Kheirkhah

FoSS, Department of Informatics, University of Sussex

Modern Data Centre Networks

FatTree

University of Sussex



- It provides full bisection bandwidth between all pair of servers.
- It provides dense interconnectivity in the network.
- It relies on ECMP routing to use its path diversity.

Data Centre Network Properties

1. **Short flow dominance.**
   - 99% of flows are short flows (flow size < 100MB).
   - Majority of short flows are query flows with deadline in their flow completion time (flow size < 1MB – e.g. 50KB)
   - 90% of total bytes come from long flows (size > 100MB).

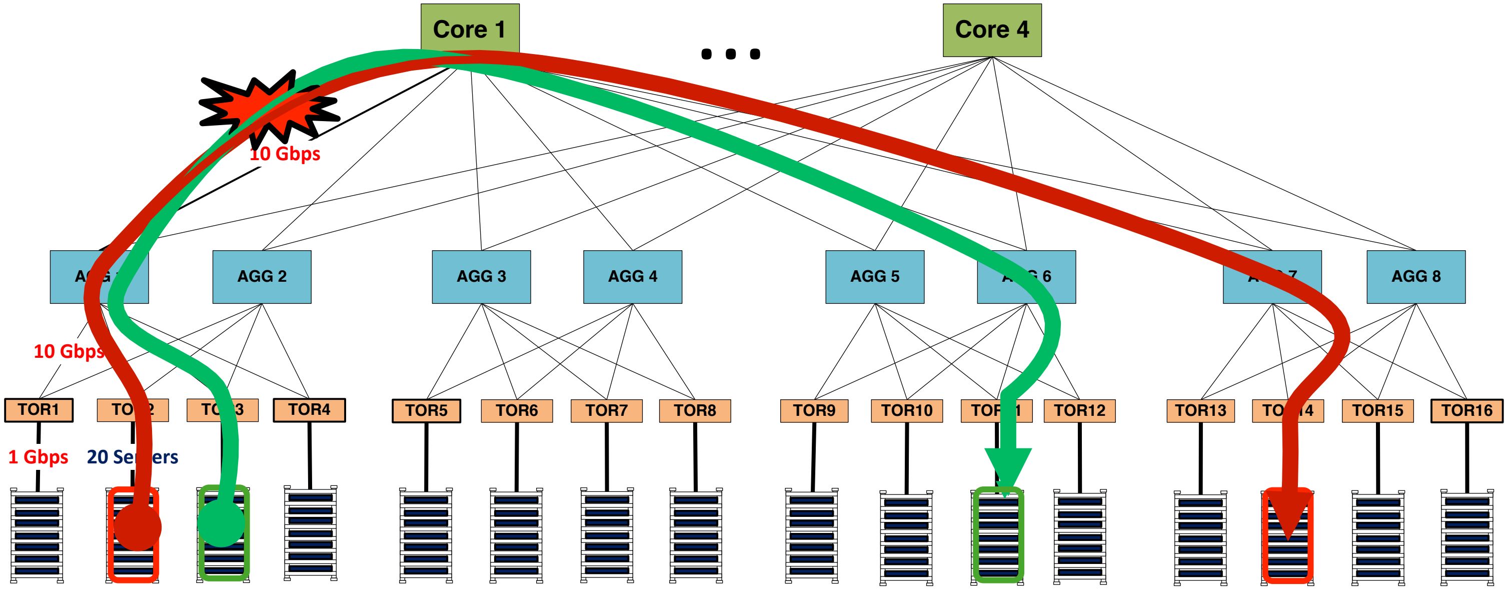
2. **Traffic pattern is very bursty, volatile and unpredictable**
   - Bursty traffic pattern is originated from short flows.

3. **Low latency and high bandwidth communication.**
   - Latency is in the order of microsecond (100-250).
   - Minimum link capacity is 1Gbps.

Congestion in FatTree

Equal-Cost Multi-Path (ECMP) Issue

A key limitation of ECMP is two or more long-lived TCP flows can collide on their hashes and end up on the same output port, creating avoidable congestion.



Possible Approach to ECMP Issue

1. **MPTCP**

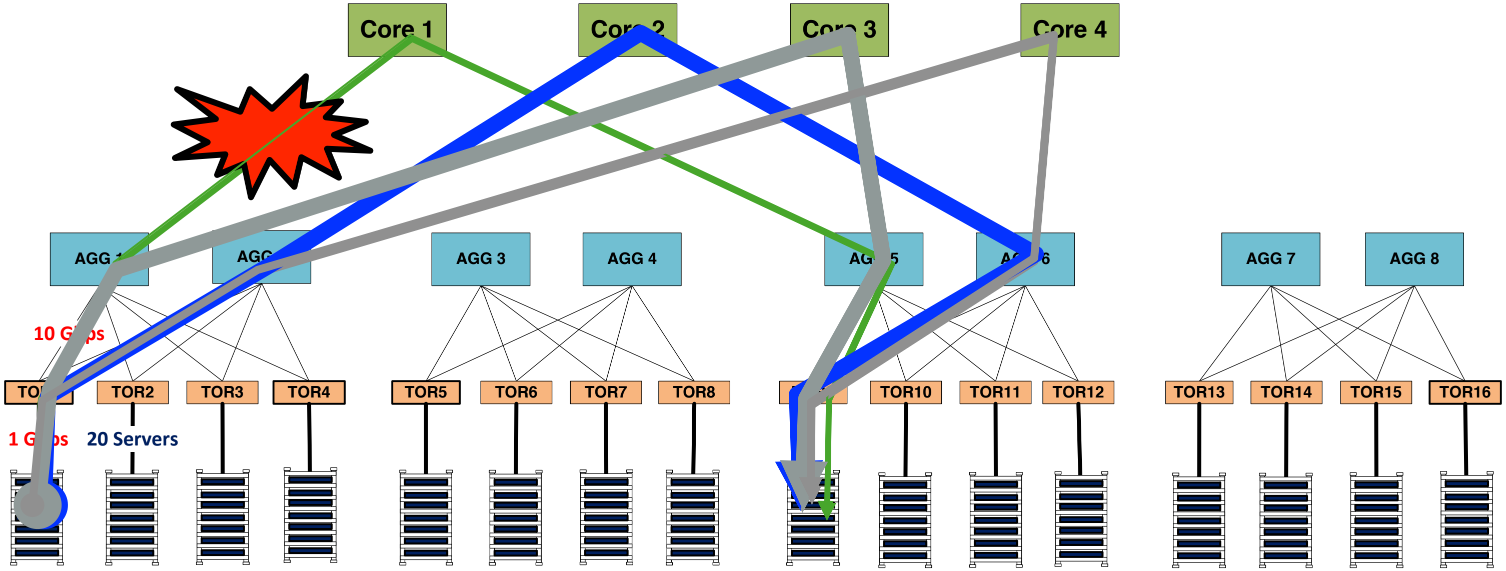
2. **MMPTCP (Our Solution)**

Possible Approach to ECMP Issue

Multi-Path TCP (MPTCP)

University of Sussex

MPTCP with four subflows. Each subflow looks similar to single-path TCP



Possible Approach to ECMP Issue

Multi-Path TCP (cont)

Pros:

- It can handle hotspots in the network core gracefully.
- It significantly increases overall network throughput.

Cons:

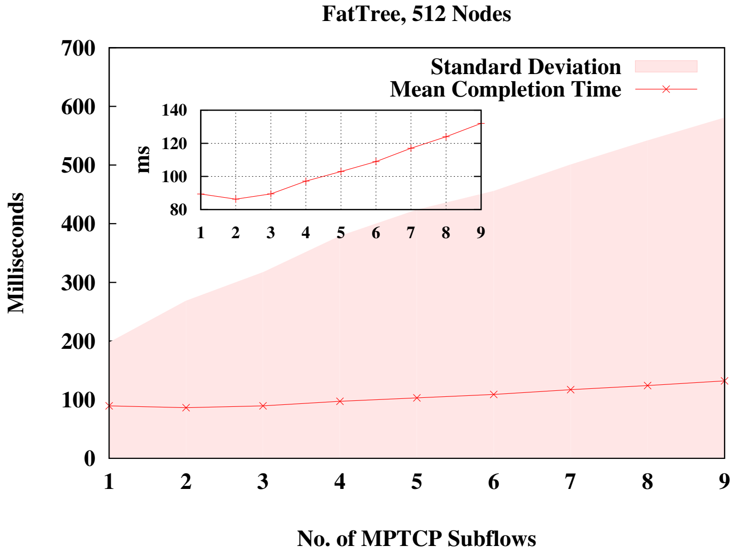
- It is not good for handling short flows.
- It is not good for competing with TCP short flows.

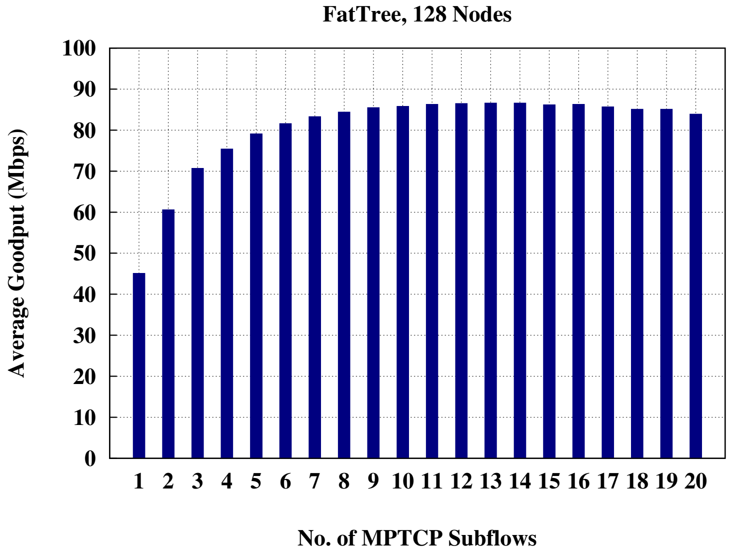
Possible Approach to ECMP Issue

Multi-Path TCP (cont)

- **Simulation Setup**: A 4:1 oversubscribed FatTree topology with 512 nodes, running a Permutation traffic matrix, 33% of nodes send continuous traffic (long MPTCP flows with 8 subflows) and the remainder send short MPTCP flows (70KB) based on a Poisson arrival (250 arrival/sec in average, assigned by a central short flow scheduler).
- **Result:** *MPTCP is not good for short flows due to small window and timeout problem in subflows.*

- **Simulation Setup:** A 1:1 oversubscribed FatTree topology with 128 nodes running a Permutation traffic matrix of long MPTCP flows.
- **Result**: *MPTCP is extremely good for long flows (8 subflows seems to be the right number for achieving high overall network throughput in a full bisection bandwidth topology)*





Possible Approach to ECMP Issue

1. MPTCP

2. **MMPTCP (Our Solution)**

Possible Approach to ECMP Issue

MMPTCP goals

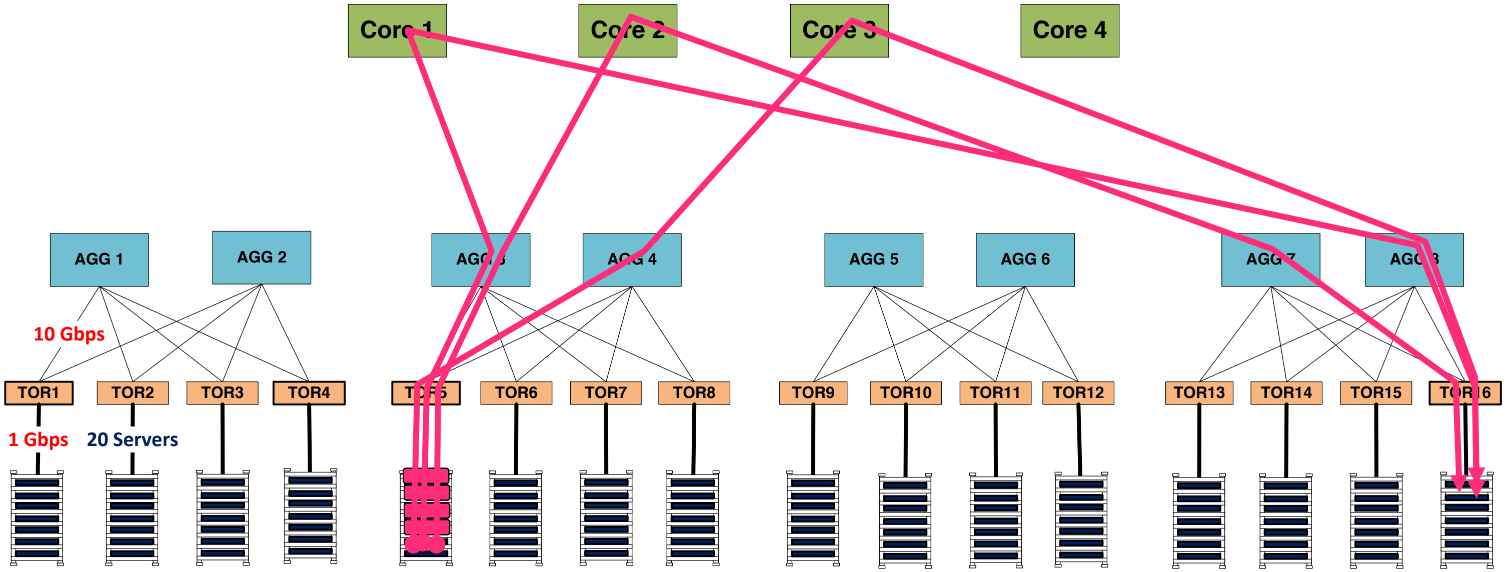
- High bandwidth for long flows
- Low latency for short flows
- High burst tolerance

Possible Approach to ECMP Issue

MMPTCP (our idea)

MMPTCP connection starts with a single subflow (i.e. TCP) and randomises its traffic on per-packet basis (packet scattering) until it delivers a certain data volume e.g. 1 MB (to cover short/query flows).

It then opens several subflows (e.g. 8) for rest of the connection, during which the MPTCP congestion control governing data transmission (the initial subflow would be deactivated at this point).



Possible Approach to ECMP Issue

MMPTCP (our idea)

Pros:

- It handles bursty traffic patterns gracefully by diffusing them throughout the network.
- It decreases flow completion time for short flows.
- It increases network throughput for large flows.

Cons:

- Packets get reorder in the initial phase.

MMPTCP and Packet Reordering

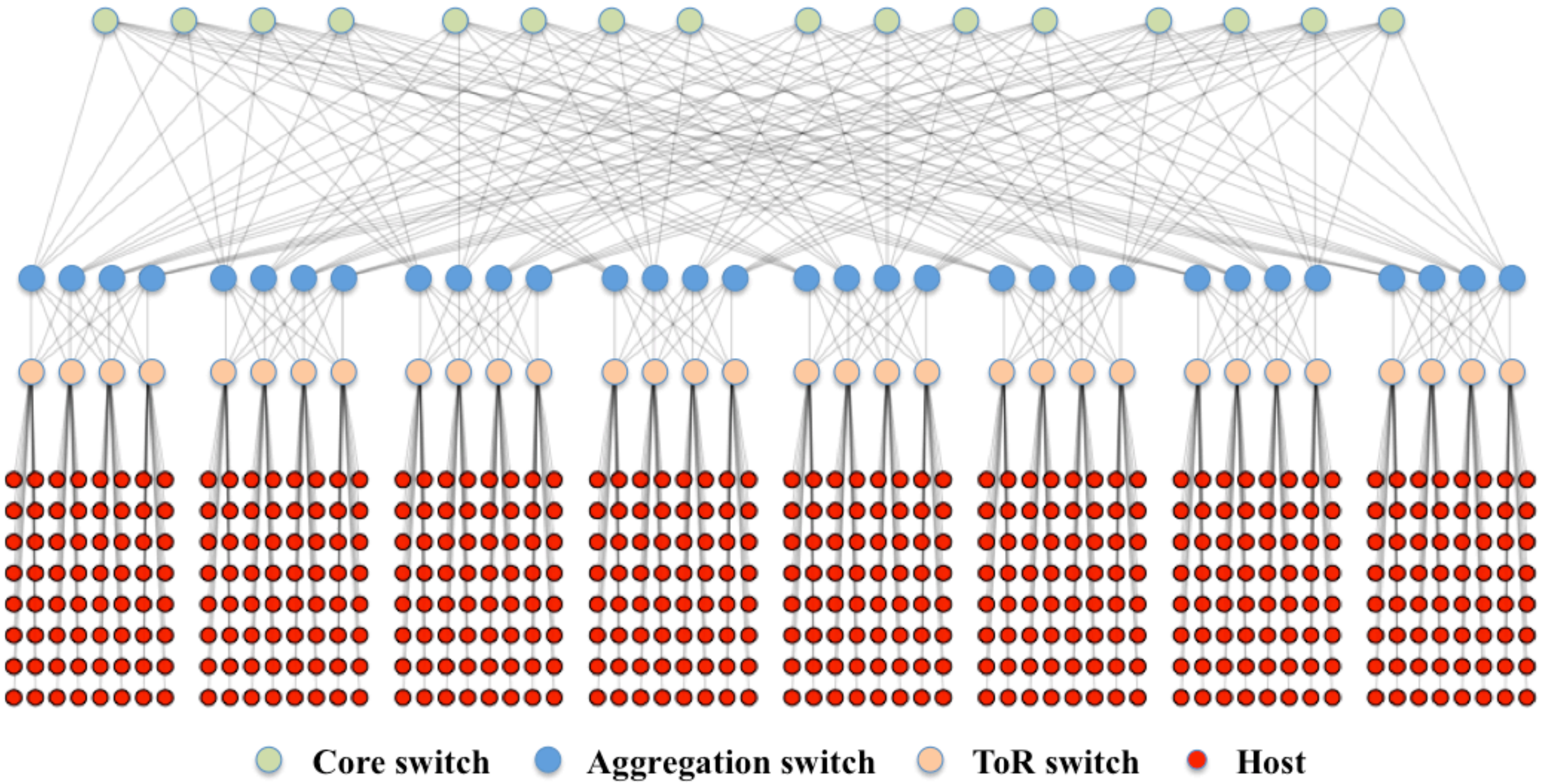
Three aspects need to be considered:

- **Preventing**, **detecting** and **mitigating** spurious retransmissions due to out of order packets.
- Solutions: RR-TCP (needs DSACK), Eifel (only for detection and mitigation) and so on.

- Quick solution with TCP NewReno - just preventing spurious retransmission by adjusting TCP dupack threshold based on topology-specific information.

Evaluation

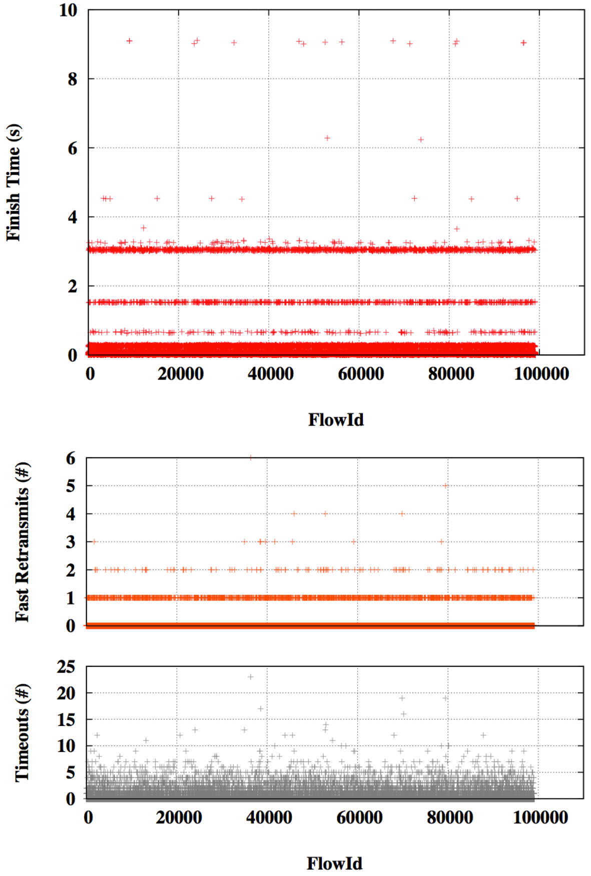
FatTree topology

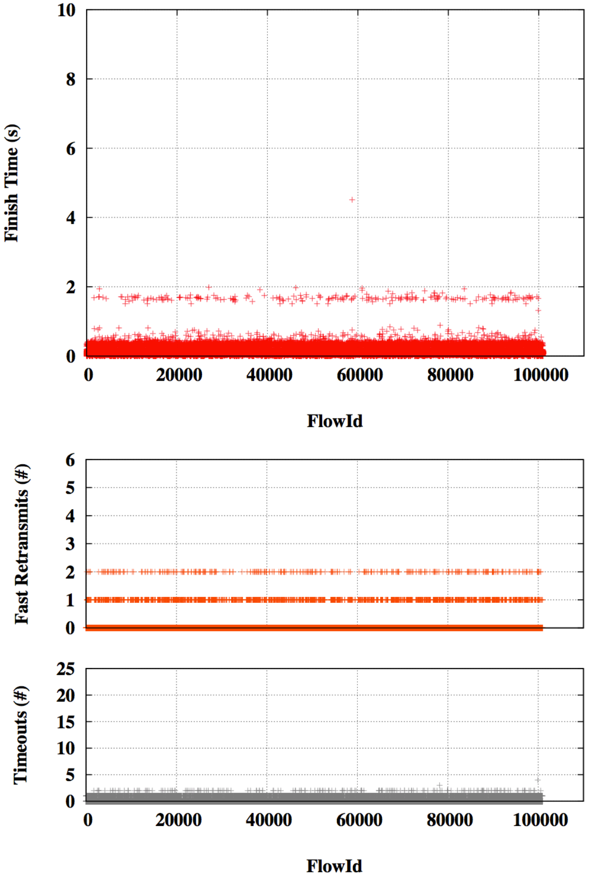


Simulation Setup: A 4:1 oversubscribed FatTree topology with 512 nodes, running a Permutation traffic matrix, 33% of nodes send continuous traffic (long MPTCP flows with 8 subflows) and the remainder send short flows (70KB) based on a Poisson arrival (250 arrival/sec in average, assigned by a central short flow scheduler).

Evaluation

MMPTCP vs. MPTCP with 8 subflows





MPTCP with 8 subflows

Mean flow completion time: 125ms

Standard deviation: 425ms

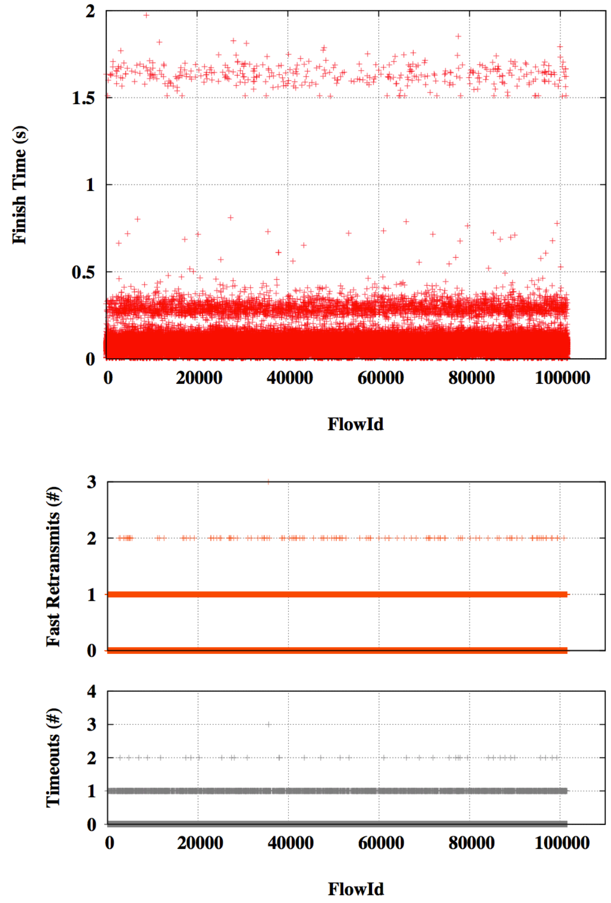
MMPTCP

Mean flow completion time: 116ms

Standard deviation: 101ms

Evaluation

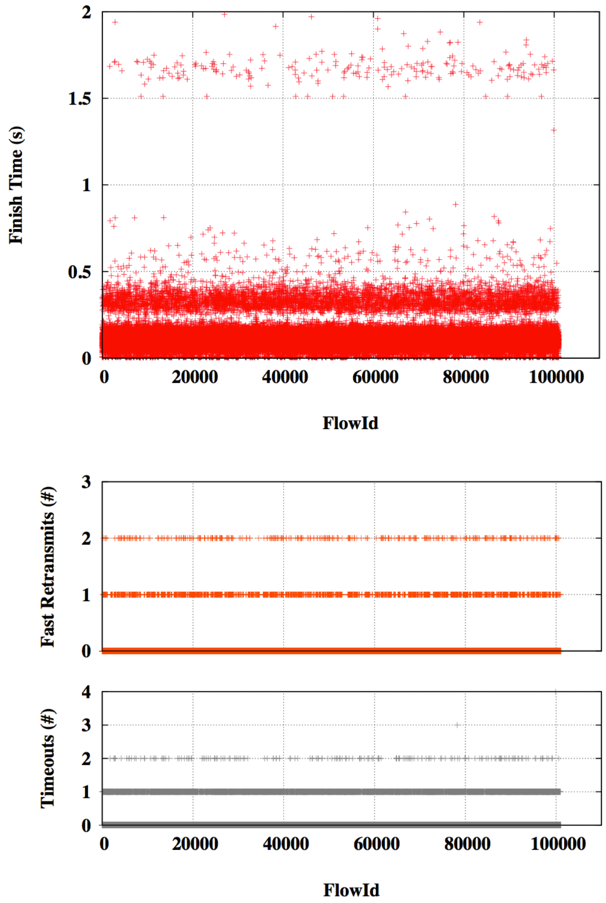
MMPTCP vs. MPTCP$_{SFTCP}$



MPTCP$_{SFTCP}$

Mean flow completion time: 89.2ms

Standard deviation: 108.9ms



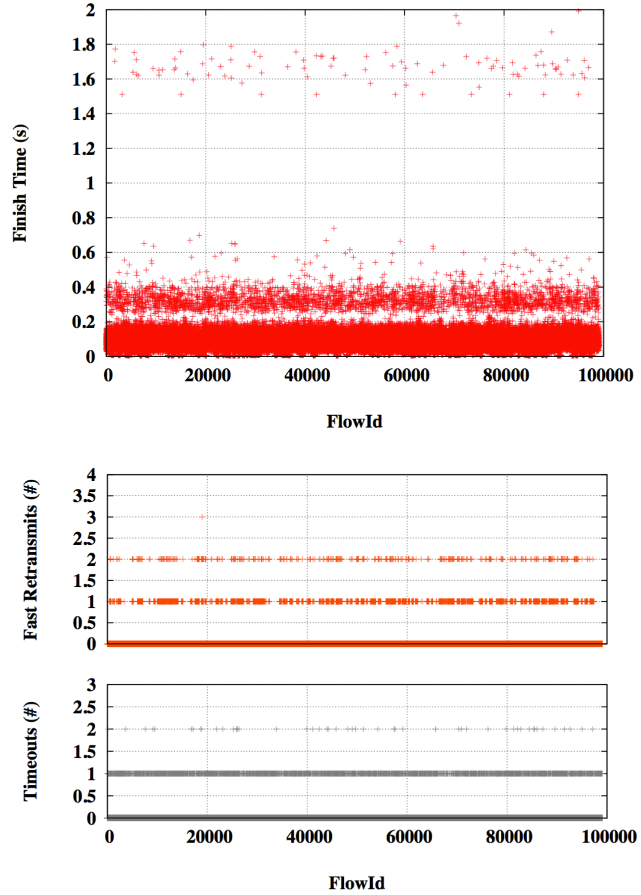
MMPTCP

Mean flow completion time: 116ms

Standard deviation: 101ms

Evaluation

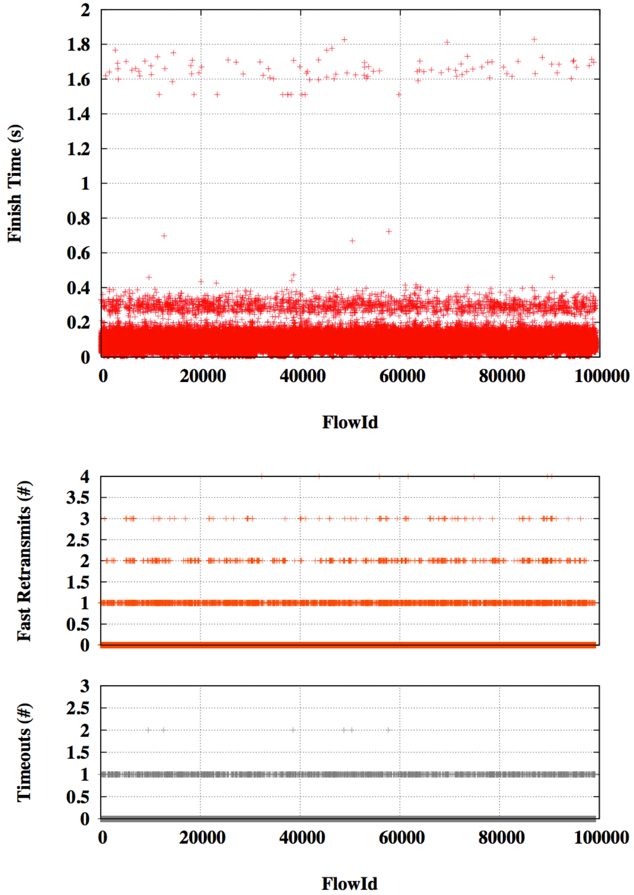
MMPTCP vs. MMPTCP$_{LT}$ (in a 2:1 FatTree)



MMPTCP

Mean flow completion time: 98.9 ms

Standard deviation: 74.8 ms



MMPTCP$_{LT}$

Mean flow completion time: 89.1 ms

Standard deviation: 67.2 ms

Future Direction

- We realized that employing TCP congestion control during the initial phase of MMPTCP is an overkill approach. We believe a right congestion control (e.g. DCTCP-like) could significantly improve the performance of MMPTCP for short flows.

- MMPTCP is capable of utilising multi-homed network topologies. In this way, TCP Incast could potentially be eliminated because MMPTCP is capable of delivering all flows via all available network interface devices.

- Advance QoS features have become increasingly available in data centre switches. Our hypothesis is that if packets of the initial phase of MMPTCP are marked high priority and routed through a different queues, then MMPTCP effectively and seamlessly helps latency sensitive short flows to complete their data delivery faster and potentially meet their deadline.

MPTCP implementation in ns-3:

https://github.com/mkheirkhah/mptcp

Publication:

Kheirkhah, M., Wakeman, I. and Parisis, G. Short vs. Long Flows: A Battle That Both Can Win, In Proceedings of ACM SIGCOMM 2015, London, UK.

Thank you!

Question?