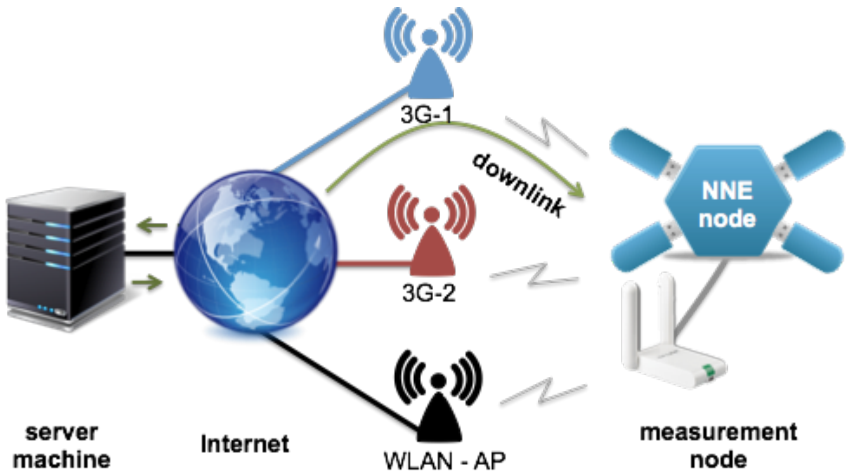
Multipath Transport over Heterogeneous Networks

Özgü Alay

Simone Ferlin-Oliveira

Thomas Dreibholz

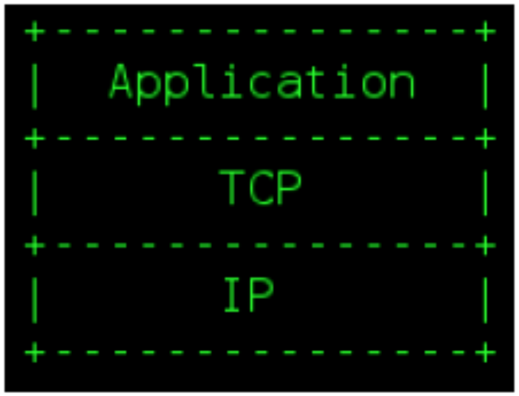
Simula Research Laboratory AS



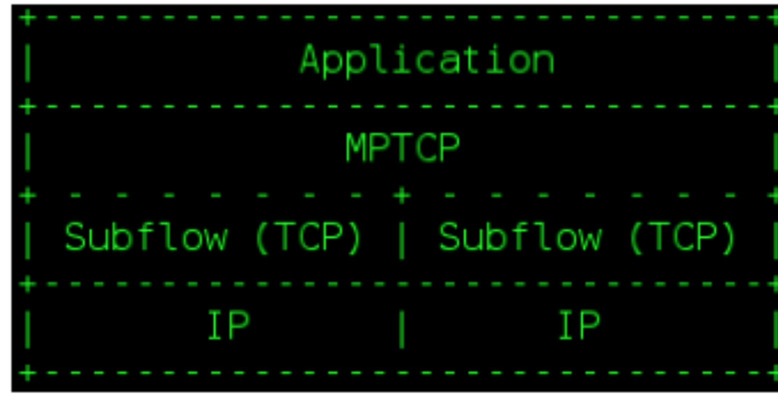
Motivation

- Multipath provides benefits
  - Increased bandwidth (resource pooling)
  - Robustness (diversity)
- MPTCP drew a lot attention recently
  - look like regular TCP for a firewall/middlebox along the subflows' path, making Multipath TCP deployable on today's Internet
- How does MPTCP works in real operational networks especially when the links are heterogeneous?
  - Goodput, application delay, buffers

Background: MPTCP



RFC6182



RFC6182

Pros: - Applications remain unmodified.

- It runs on TCP.
- It exploits reliability through network (path) diversity.

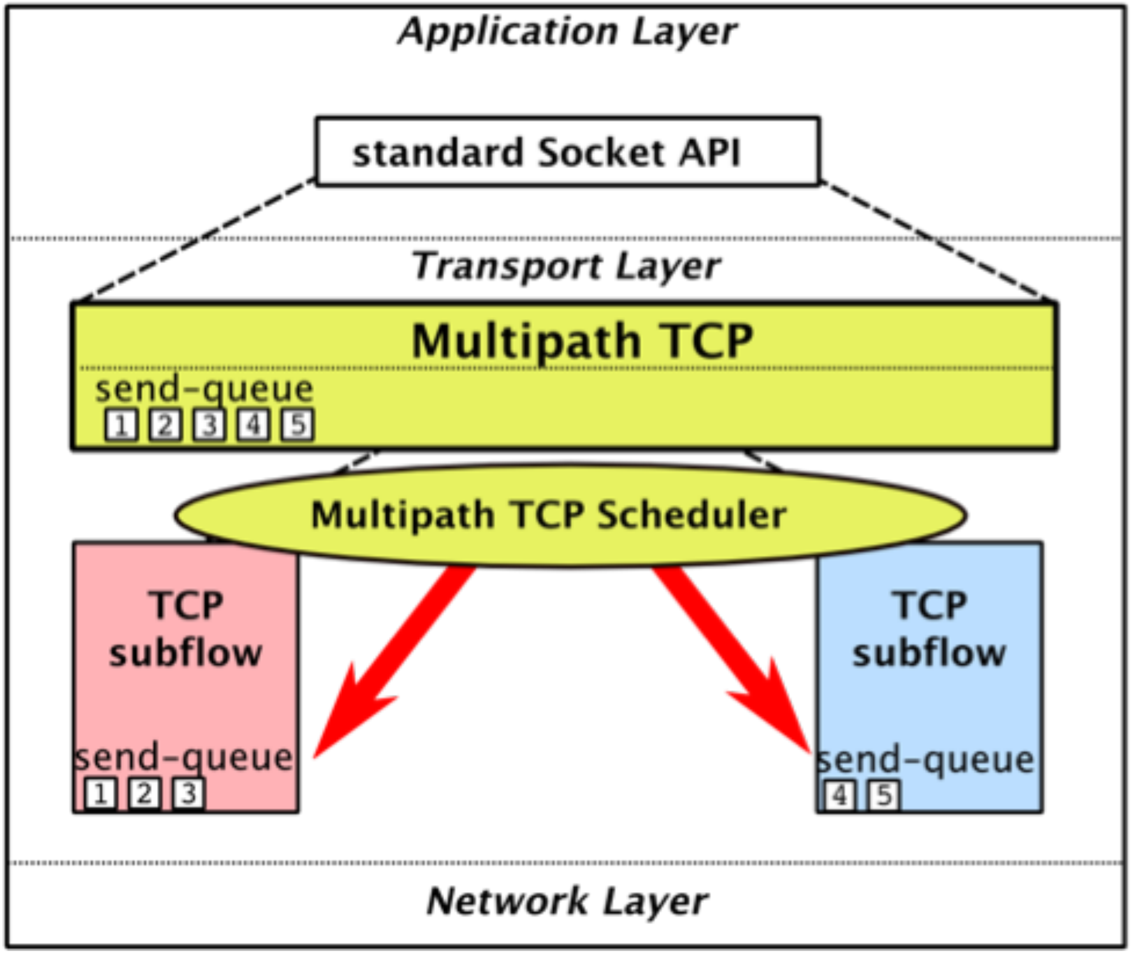
Cons: Needs more testing, requires extensions, and wide(r) OS.

Building Blocks of MPTCP

- Scheduler
  - When over which path to send a packet
- Path Management
  - when and how to set up paths (subflows);
  - how many paths (subflows) to use;
- Congestion Control

SCHEDULER

MPTCP Scheduler

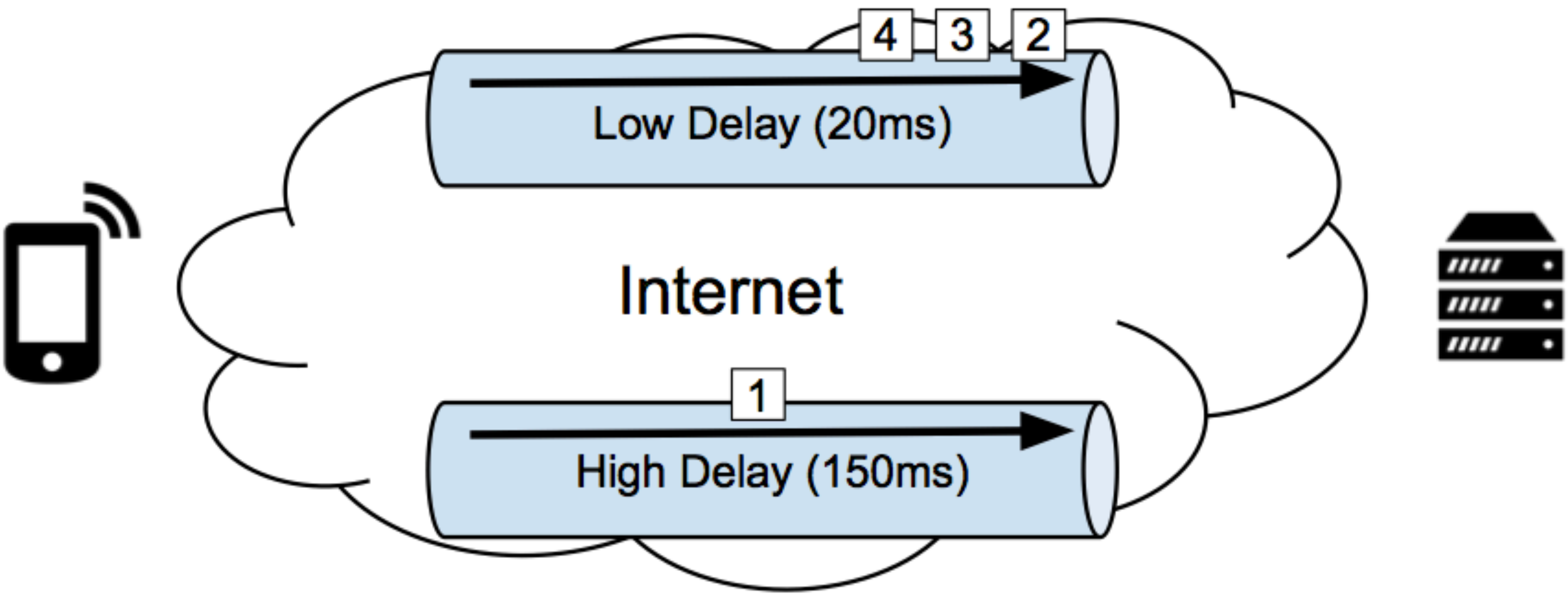


Paths can be different

- Bandwidth
- **Delay**
- Loss

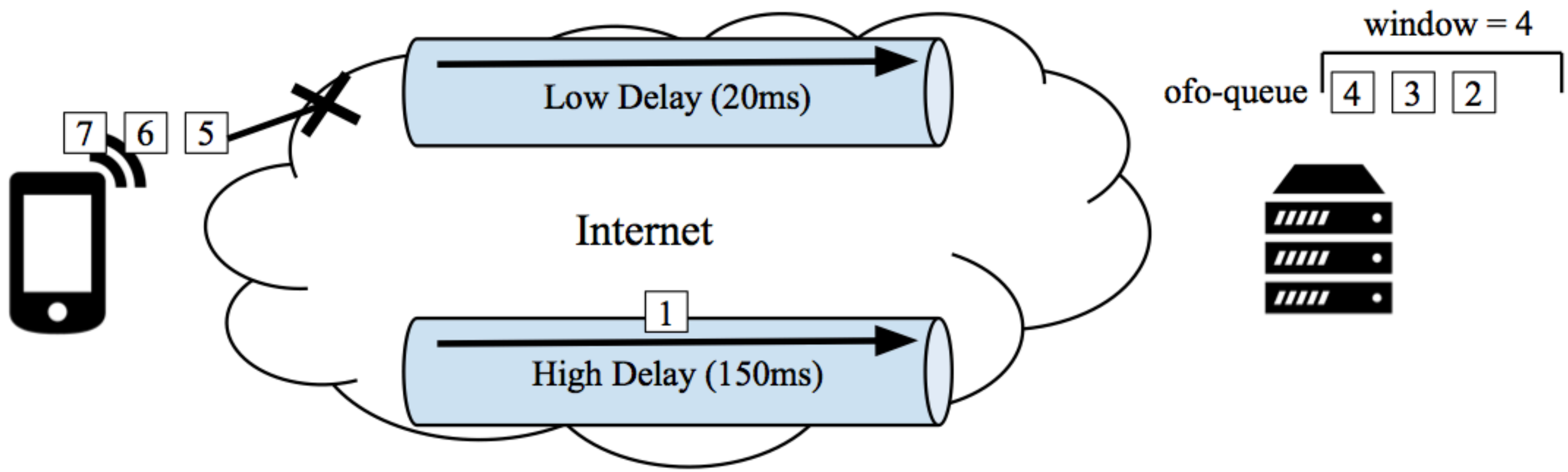
Example: WLAN vs 3G

Head of Line Blocking



- Receiver is waiting for packet #1
- Burstiness -> delay the data delivery to the application

Receive Window Limitation



- Buffer space to fully utilize all the subflows

$$\text{Buffer} = \sum_{i}^{n} \text{bw}_{\text{i}} \times \text{RTT}_{\text{max}} \times 2$$

- Reduced goodput

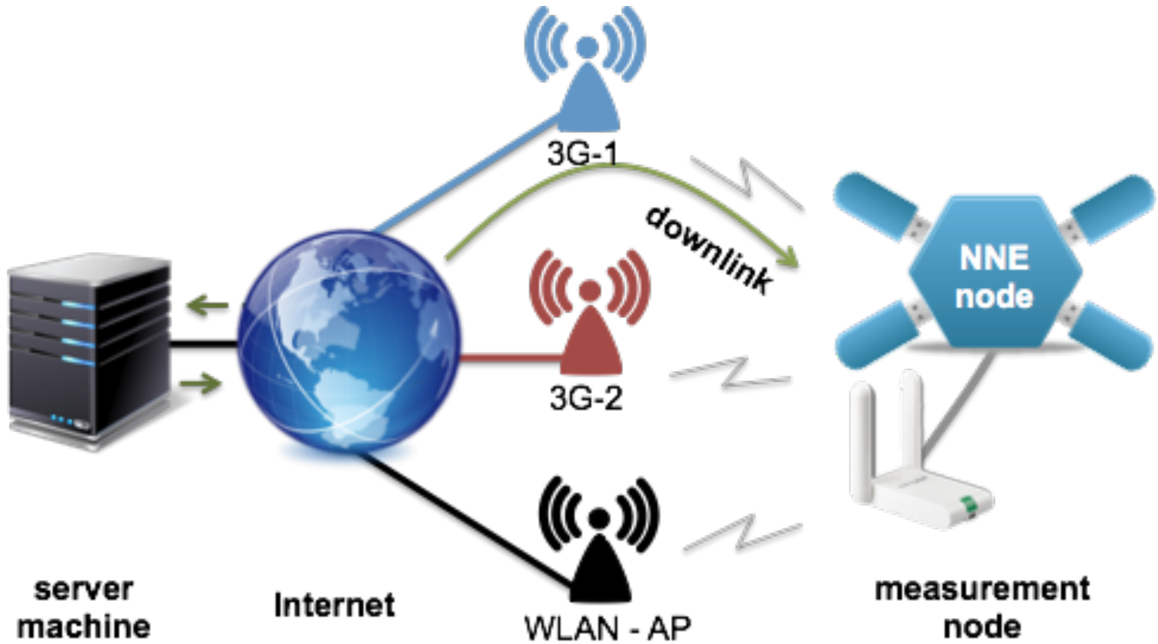
Current MPTCP Scheduler

- Lowest RTT first
- **Penalization and Opportunistic Retransmission**: Prevents out-of-order reception and, thus, receive window limitation by halving the cwnd of the *slow* subflow (*slow:* TCP connection with higher RTTs) and setting ssthresh
(only if ssthresh was already set, i.e., congestion avoidance).

How does this mechanism work in a realistic use case, e.g. smart phone with one MBB and one WLAN interface?

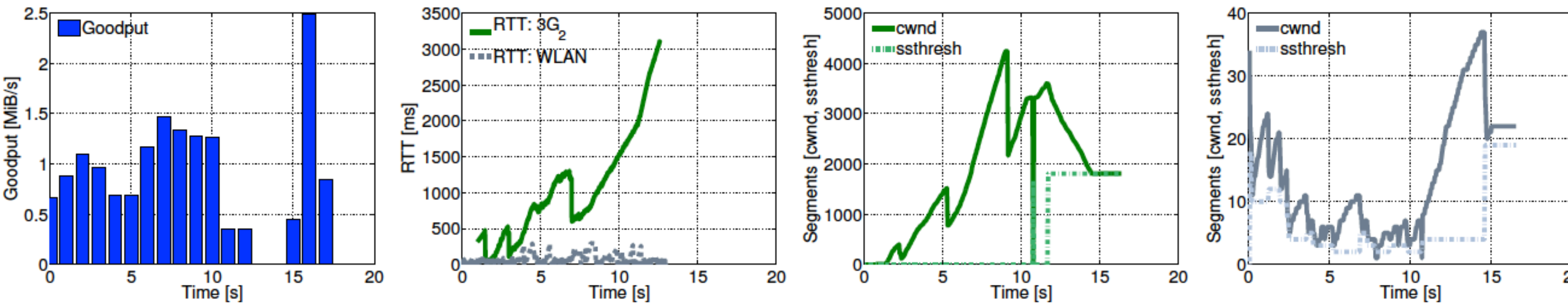
- Mobile broadband networks have massive buffers (bufferbloat)

Experimental Setup



NorNet Edge: www.nntb.no
- 2 different 3G UMTS ISPs in Norway and WLAN.
- Bulk transfer (16MiB) in downlink.
- Unbounded buffers

Impact of Bufferbloat: Example with $3G_2$ + WLAN



(a) Goodput: $3G_2$ and WLAN (b) RTT: $3G_2$ and WLAN (c) Congestion Window: $3G_2$ (d) Congestion Window: WLAN

Figure (a): Goodput gaps due to high RTTs in $3G_2$.

Figure (b): HOL blocking is caused by high RTTs.

Figure (c) and (d): - MPTCP penalizes $3G_2$ but cwnd keeps growing.

- MPTCP becomes receive-window limited.
- Capacity of WLAN is underutilized.
- **But:** $3G_2$ has higher capacity and is penalized!

MULTIPATH TRANSPORT BUFFERBLOAT MITIGATION

Multipath Transport Bufferbloat Mitigation

Algorithm 1 Per-Subflow Bufferbloat Mitigation by MPT-BM

Initialization:

$sRTT \leftarrow \infty$

$sRTT_{min} \leftarrow \infty$

RTT estimation:

$sRTT_{min} \leftarrow min(sRTT_{min}, sRTT)$

How many segments can be sent?

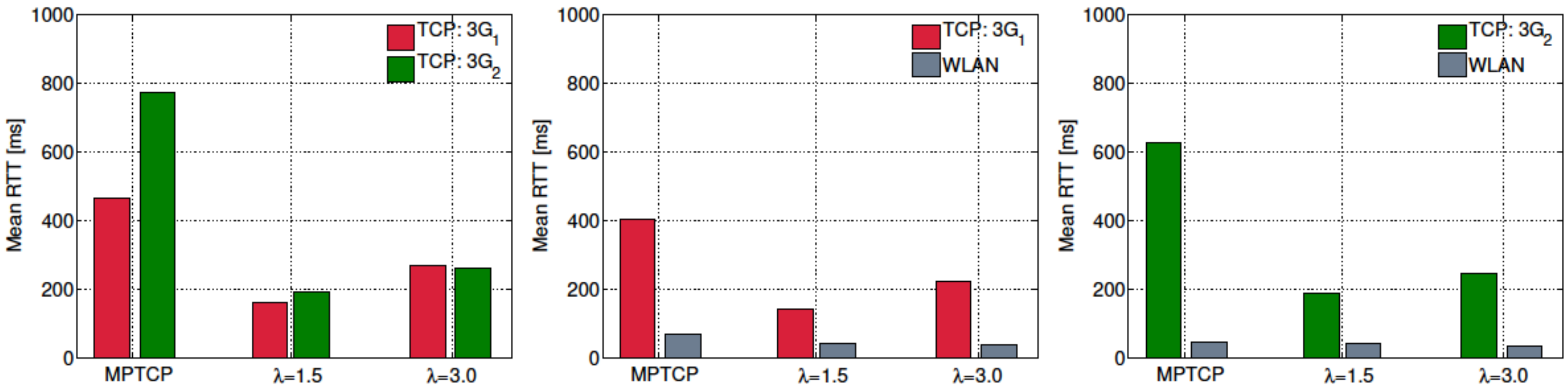
$cwnd_{limit} \leftarrow \lambda * (sRTT_{min}/sRTT) * cwnd$

$$\text{send} \leftarrow \begin{cases} max(0, (min(\text{cwnd}, \text{cwnd}_{\text{limit}}) - \text{inflight}) & (\text{RTT}_{\text{min}} \geq \Theta) \\ max(0, \text{cwnd} - \text{inflight}) & (\text{RTT}_{\text{min}} < \Theta) \end{cases}$$

Sender side :

- Monitor shift between **sRTT** and **$sRTT_{min}$** for each subflow.
- *Tolerance* shift is given by **λ** (set through sysctl)
- Caps the cwnd for each subflow by **$cwnd_{limit}$**.

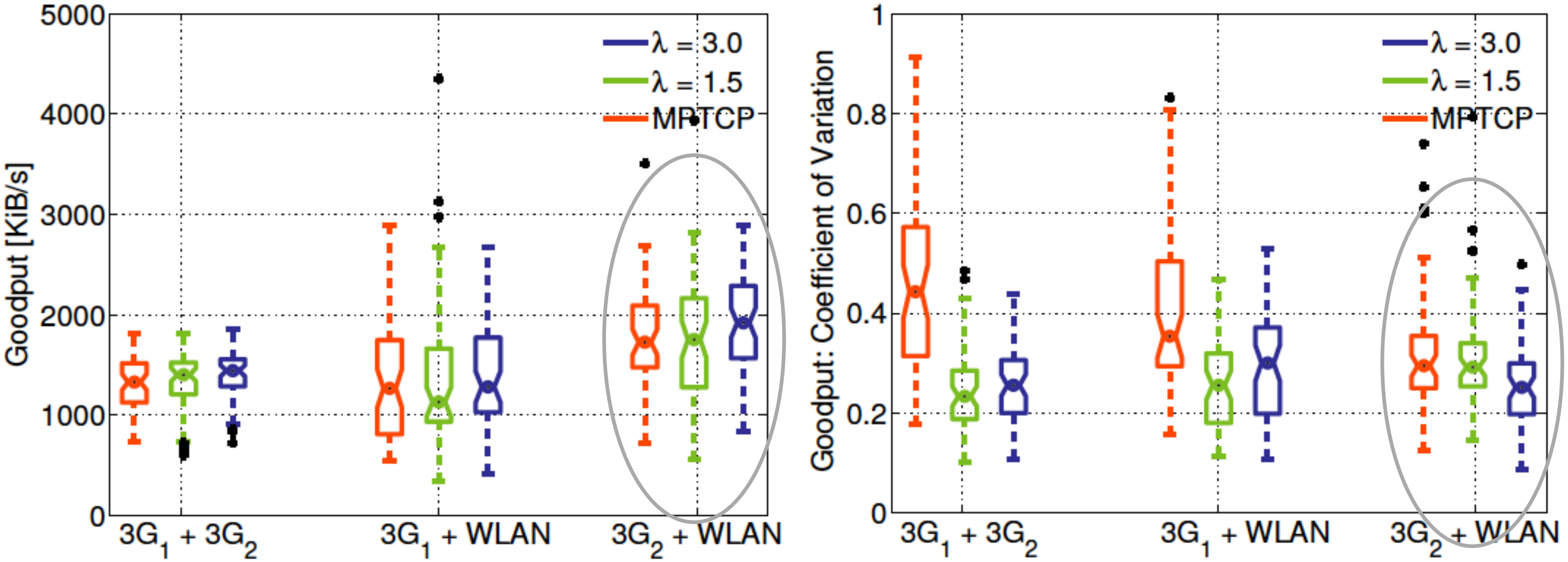
MPT-BM: RTT Capping



(a) RTT: $3G_1$ + $3G_2$ (b) RTT: $3G_1$ + WLAN (c) RTT: $3G_2$ + WLAN

- Scenarios: $3G_1$ + WLAN, $3G_2$ + WLAN and $3G_1$ + $3G_2$
- MPT-BM successfully caps the RTT for **λ=1.5** and **λ=3.0**

MPT-BM: Goodput volume and variance.

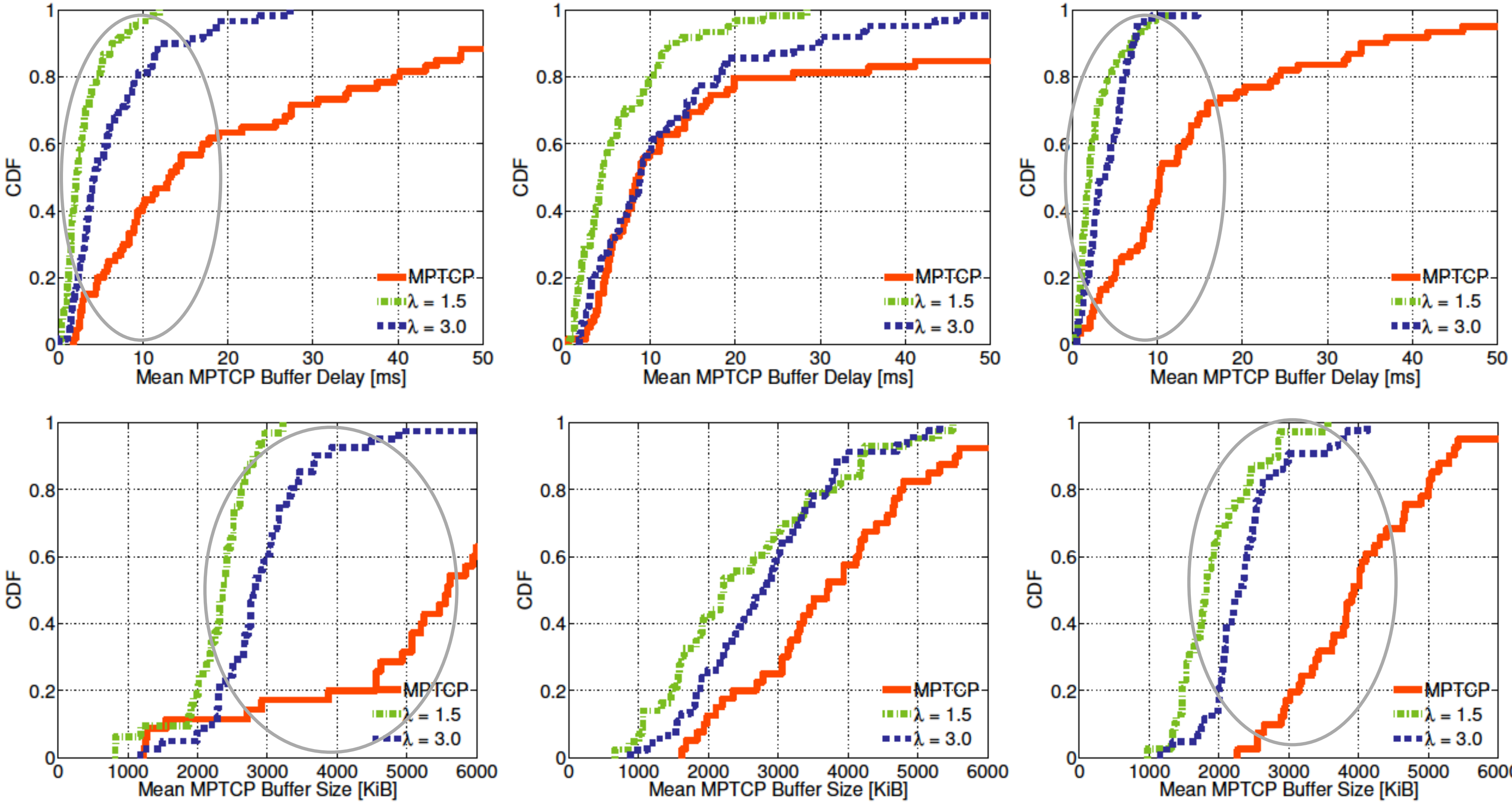


(a) Application Goodput

(b) Coefficient of Variation

- Goodput volume: Marginal improvement.
  Approx. 8 up to 15% with both **λ=1.5** and **λ=3.0.**
- Goodput variance: Coefficient of variation (σ / μ) as metric.
  Approx. 15 up to 45% with both **λ=1.5** and **λ=3.0.**

MPTCP Buffer Delay and Size



(a) $3G_1 + 3G_2$

(b) $3G_1$ + WLAN

(c) $3G_2$ + WLAN

Discussion

- MPT-BM caps RTTs successfully, hence limits the head of line blocking due to the bufferbloat
- MPT-BM provides improvements in goodput volume and quality for bulk transfer
- How about application limited traffic?

EVALUATION OF DIFFERENT SCHEDULERS FOR MPTCP

Motivation

- Comparison of different schedulers
  - Traffic: Bulk Transfer, Application Limited Traffic
  - Metrics: Aggregation benefit, goodput, application delay
- Extensive analysis
  - Mininet
  - Nornet

Evaluated Schedulers

- Round Robin (RR)
- Lowest RTT first (LowestRTT)

Extensions of LowestRTT

- Penalization and Retransmission (PR)
- Bufferbloat Mitigation (BM)

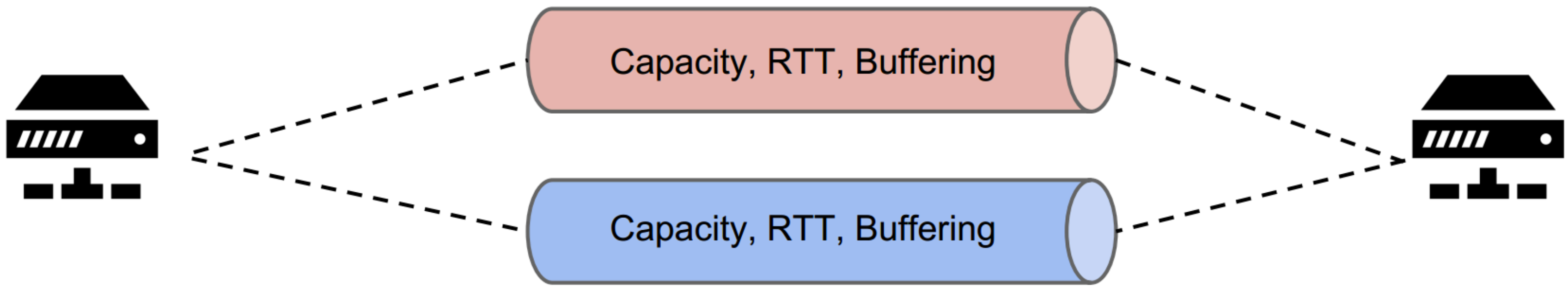
Parameters

- Aggregation Benefit and Application Delay for Bulk Transfer

$$B = \begin{cases} [-1;0) & \text{if MPTCP} < \text{TCP} \\ 0 & \text{if MPTCP} == \text{TCP} \\ (0;1] & \text{if MPTCP} > \text{TCP} \end{cases}$$

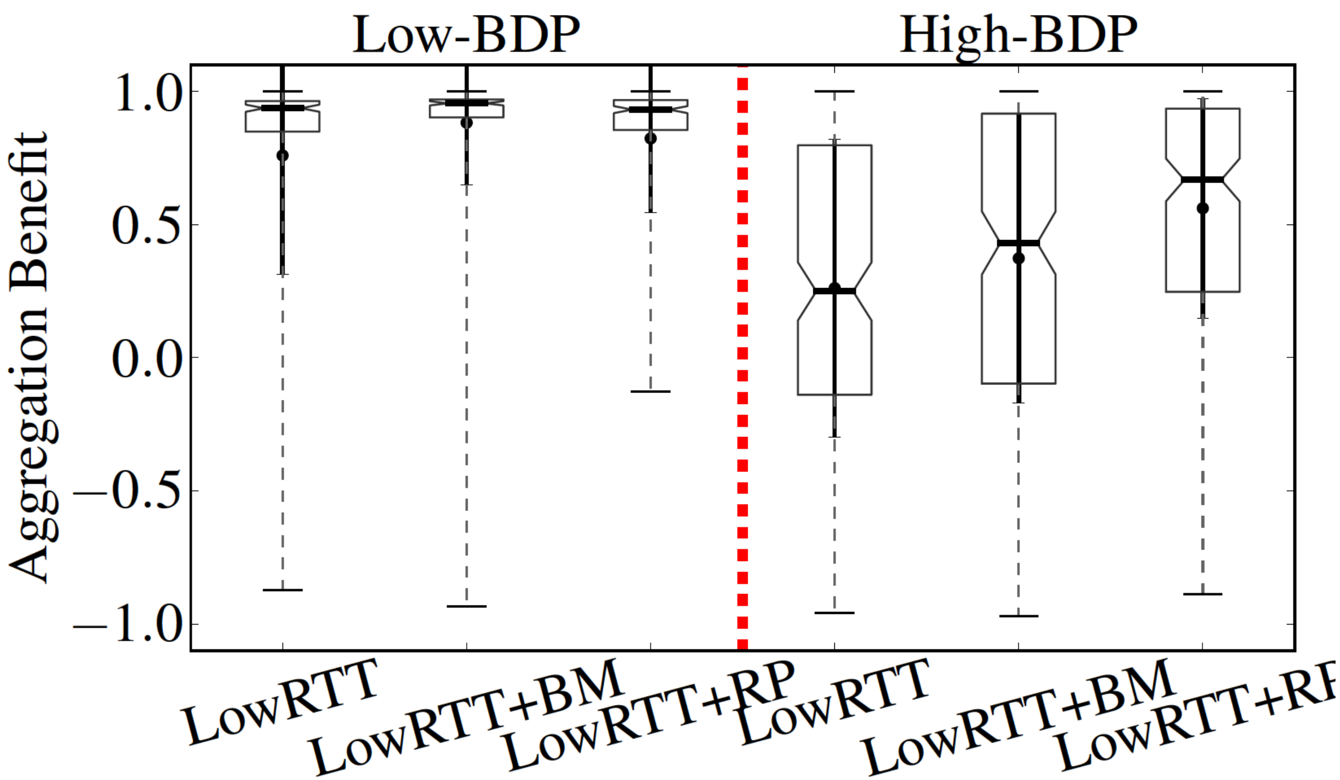
- Application Delay
  - sending at constant rate, blocks of 8KB data

Mininet Evaluation



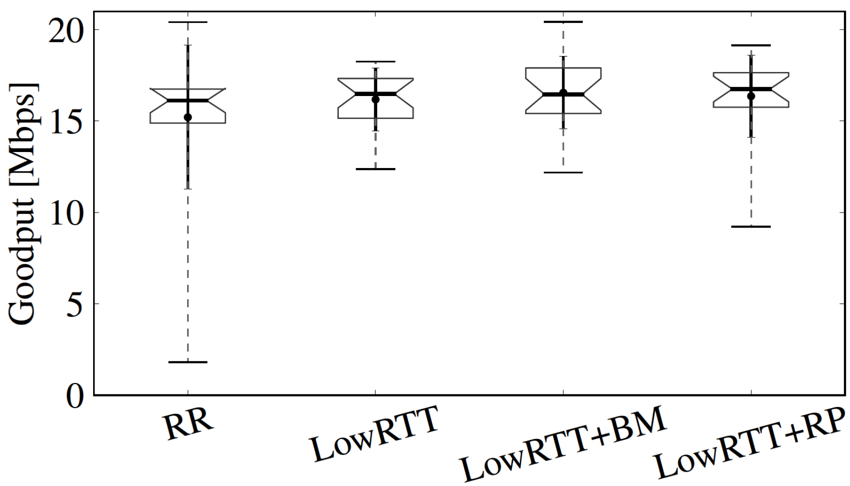
	Low-BDP	High-BDP
Capacity	0.1 to 100 Mbps	0.1 to 100 Mbps
RTT	0 to 50 ms	0 to 400 ms
Buffering	0 to 100 ms	0 to 2000 ms

Mininet Results

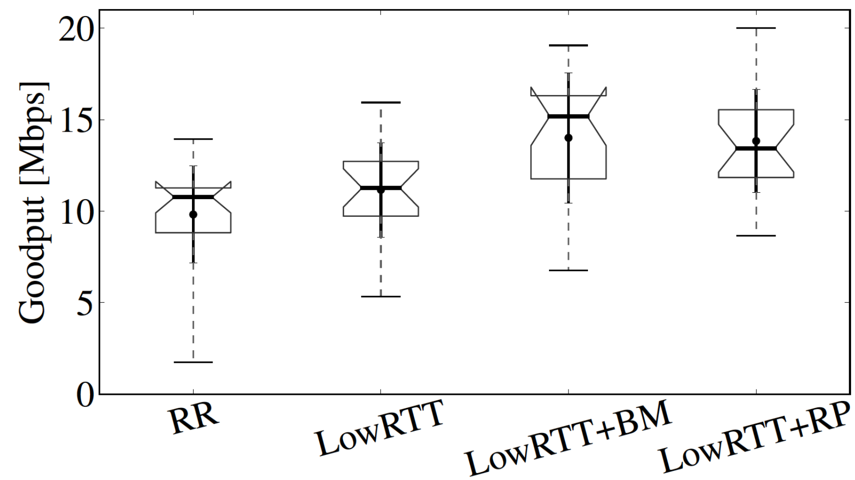


- RR is similar to LowestRTT
- In high-BDP scenarios the connection becomes receive-window limited and BM and RP show their benefits.

NorNet: Bulk & Goodput



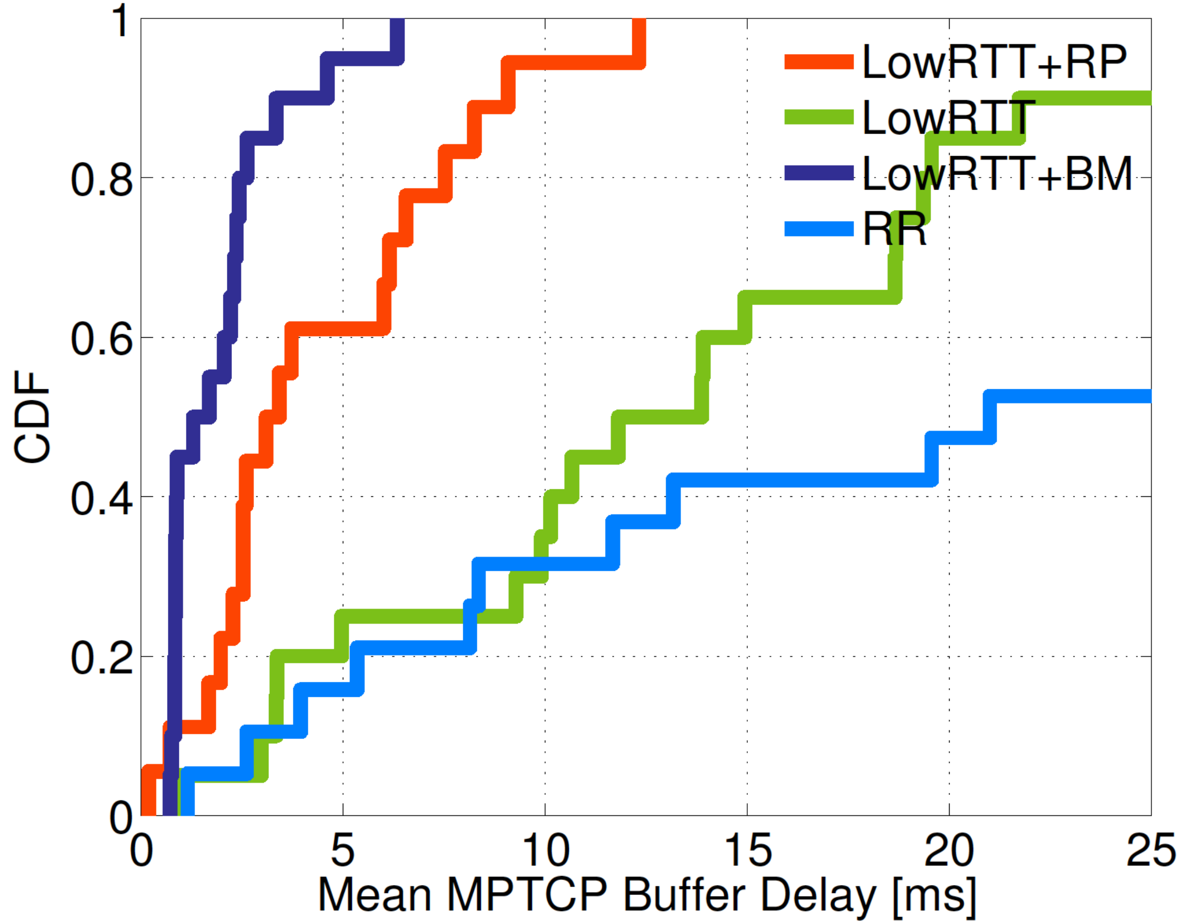
Unbounded buffers (16MB)



Bounded buffers (2MB)

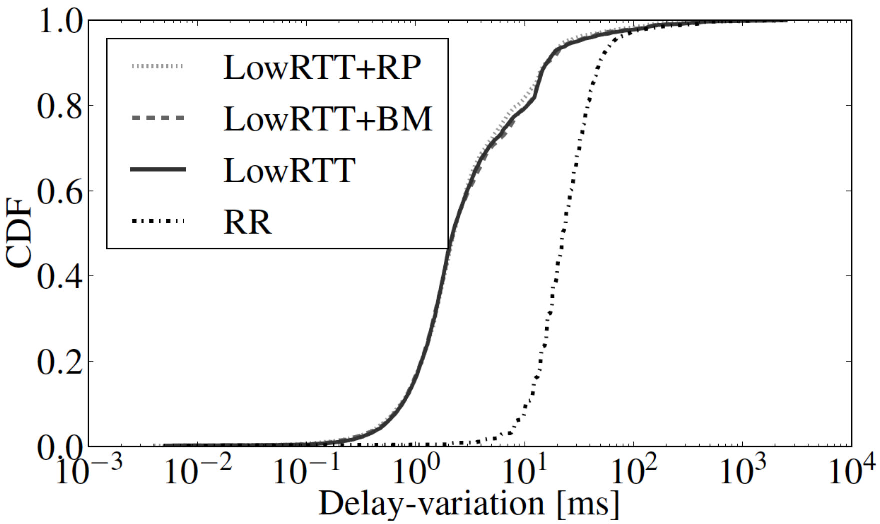
- With unbounded buffers, each scheduler achieves similar goodput.
- With bounded buffers, LowRTT+BM and LowRTT+RP achieve the best performance.

NorNet: Bulk & Buffer Delay

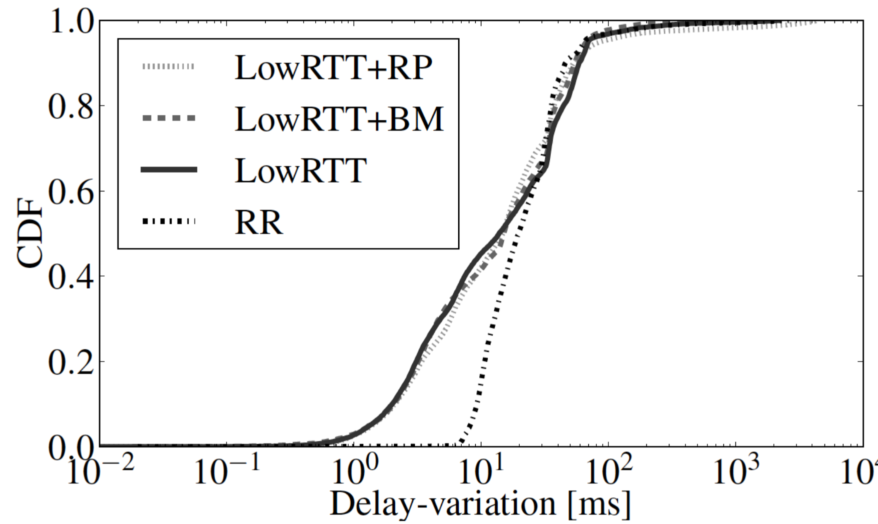


With unbounded buffers (16MB), LowRTT+BM provides the least buffering delay.

NorNet: Application Limited Traffic



(a) NorNet Edge with 500kbps.



(b) NorNet Edge with 1875kbps.

LowRTT and its extensions behave similar and they outperform RR

Discussion

- Scheduling decisions have significant impact on the delay
- For bulk transfer, LowRTT and its extensions outperform RR.
  - LowRTT+RP and LowRTT+BM provides gains over LowRTT by reducing the delay difference among paths.
  - LowRTT+BM provides significant gains compared to other schedulers especially when there is a “bufferbloated” link.
- For the application-limited traffic, LowRTT and its extensions behave similar and they outperform RR.
- Considering both the bulk transfer and application-limited flows, the best scheduler available is LowRTT extensions (LowRTT+RP and LowRTT+BM) in terms of delay performance.

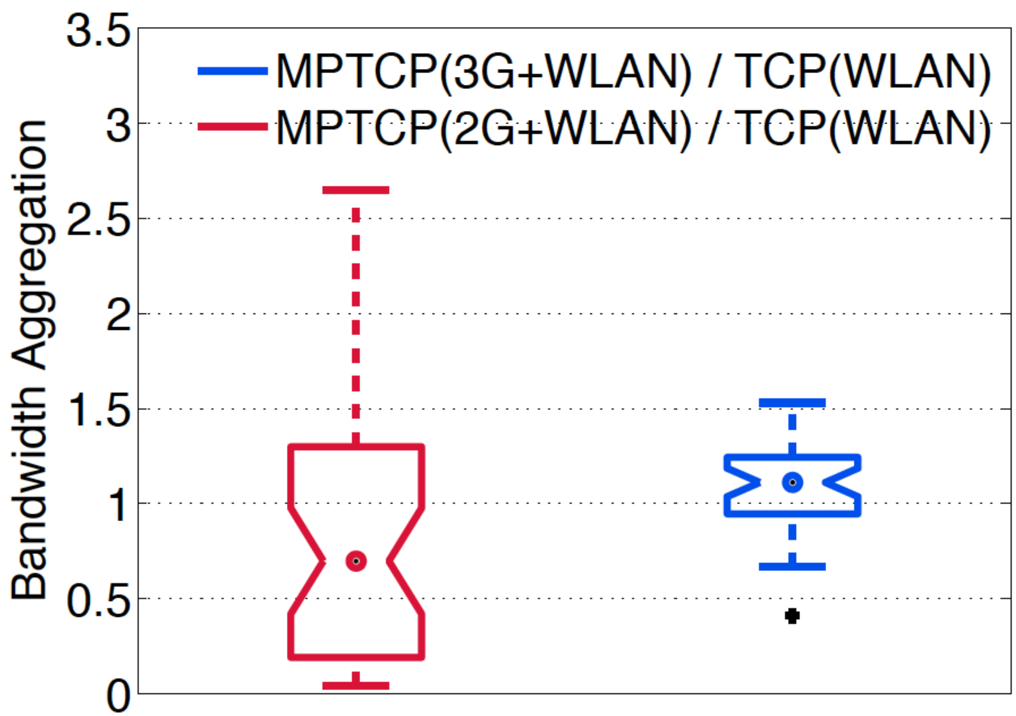
PATH MANAGEMENT

Path Management

- Current MPTCP
  - Flows are open sequentially
  - All available paths are used

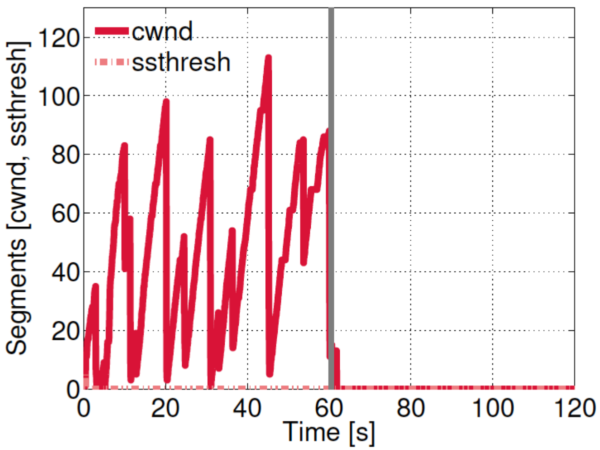
- Do we need all the paths at all time?
- Are there cases where it is better off not to open a subflow?

Bandwidth Aggregation in Real Heterogeneous Networks

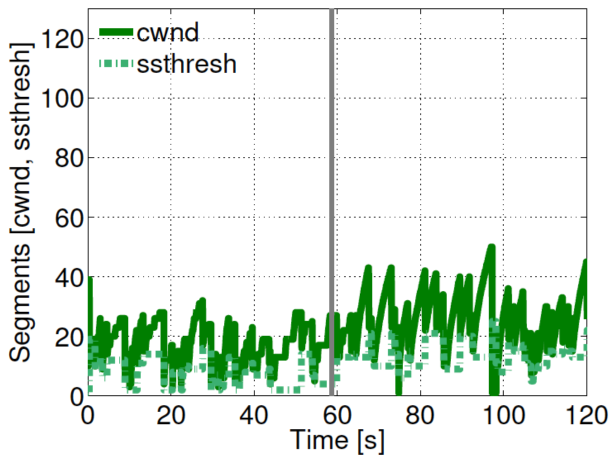


2G+WLAN is worse than WLAN alone!

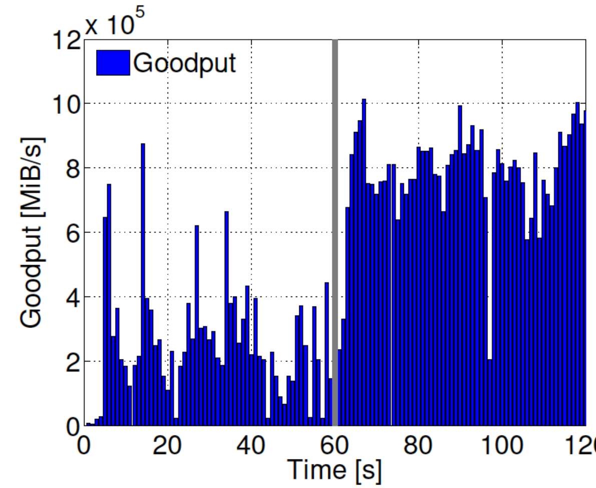
Active vs Backup State



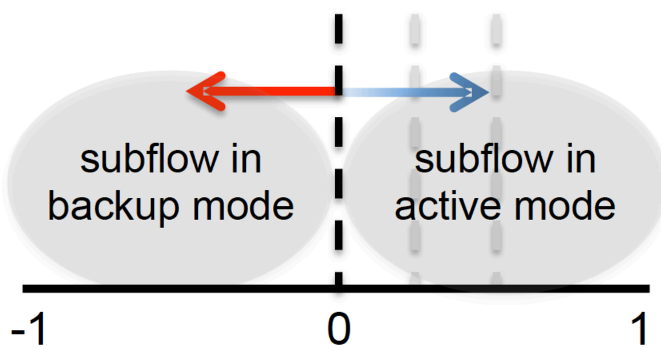
(a) Congestion Window of the 2G Subflow



(b) Congestion Window of the WLAN Subflow



(c) Resulting MPTCP Goodput



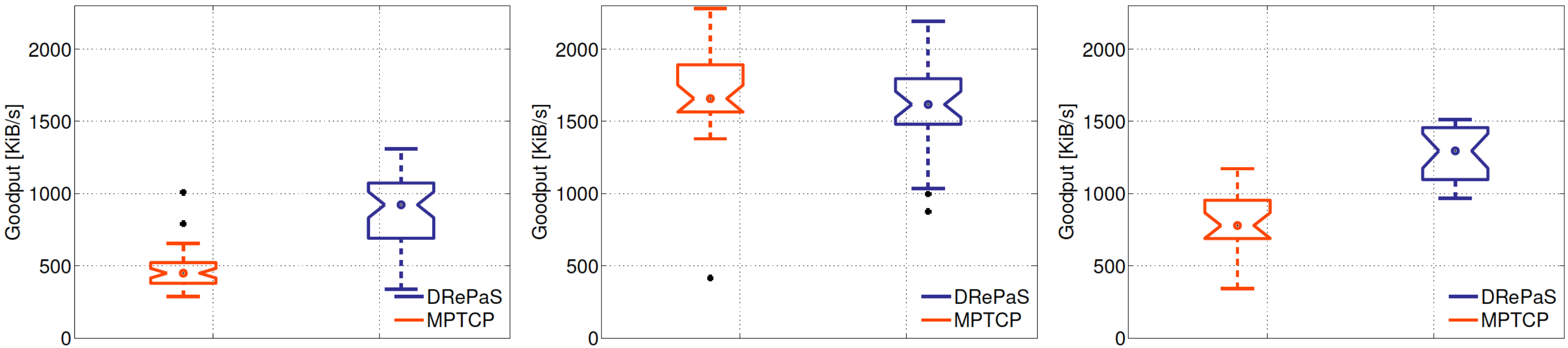
DREPAS – DYNAMIC RELATIVE PATH SCORING

- Dynamically scores the paths relative to the best path
- For each subflow
  - **Throughput** is defined as the amount of inflight data divided by smoothed RTT
  - **Factor,** the ratio of throughput to the maximal throughput, is computed (between 0 and 1)
  - **Score** is compared with a predefined threshold to determine the score which is either 1 or 0.
  - Score=1 -> active, Score =0->probing(backup)

Probing (Backup) State

- No payload is scheduled, however, the sub-flow remains established for redundancy purposes
- Probing traffic is sent to evaluate the subflow, e.g., if its QoS characteristics improve.
- Whenever the best subflow's performance decrease, the probing subflow is resumed (due to its now relatively beneficial contribution)
- The probe data can be be dynamically adapted via sysctl.
- In our experiments, we used 10 packets of 1 KiB each.

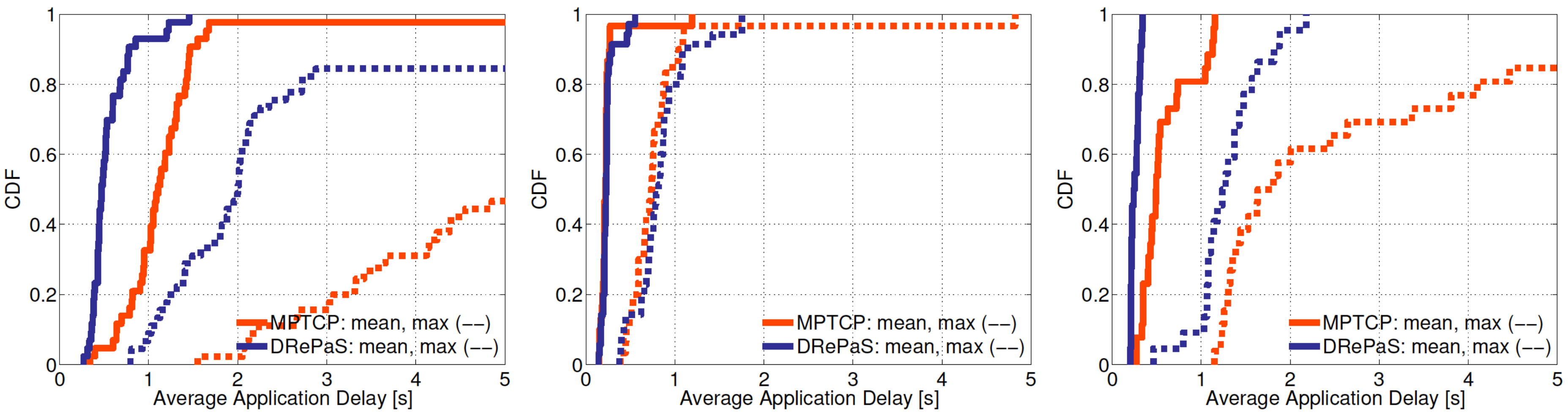
Results I



(a) 2G + WLAN (b) 3G + WLAN (c) 2G + 3G

- DREPAS improves the goodput especially when there is bufferbloated low capacity link
- Similar performance for the 3G and WLAN case

Results II



(a) 2G + WLAN (b) 3G + WLAN (c) 2G + 3G

- DREPAS improves the application delay when there is a bufferbloated low capacity link
- Similar performance for the 3G and WLAN case

Discussion

- Multi-path transport is not always beneficial under realistic conditions and parameter settings, e.g. 2G and WLAN.
- There is a need to continuously evaluate the contribution of each path to the overall performance and dynamically adapt
- DRePaS outperforms the current MPTCP implementation especially when the paths are very heterogeneous

Ongoing and Future Work

- Shared Bottleneck Detection for Multipath
  - Uncoupled congestion or Coupled congestion?
  - To find additional paths
- IPv4 and IPv6 paths are not congruent
  - Can we utilize this diversity to provide reliability and increased performance?