P4air Congestion Control

Belma Turkovic and Fernando Kuipers

TU Delft

Background

- New transport protocols – MPTCP, MPQUIC
- New congestion control algorithms – BBR (2016), TCP LoLa (2017), …
- QUIC enables quick development of new transport features
- Congestion control algorithms typically developed in isolation -> fairness issues

TU Delft

Outline

1. Evaluation of multi-path congestion control using MPTCP and MPQUIC
2. Evaluation of different AQMs and queue management techniques available in the Linux kernel
3. Investigating how a P4 switch can be used to identify different congestion control algorithms – P4air

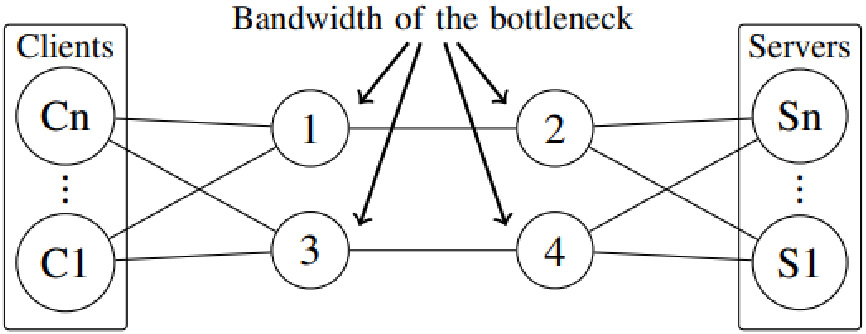
TUDelft

MPTCP and MPQUIC

TU Delft

Experimental Evaluation – Setup

- The bandwidth between nodes 1/3 and 2/4 (bottleneck) is limited
- Delays on links between nodes 3 and 4, as well as node 2 and nodes S_i were artificially increased using Linux TC
- Transport protocol: MPTCP and MPQUIC



TU Delft

Baseline performance

- BBR is not able to utilize the second link fully
- MPTCP flows achieved a higher sending rate than MPQUIC flows

Protocol	Group	Algorithm	Link 1 Average throughput $[Mbps]$	Link 2 Average throughput $[Mbps]$
MPTCP	Coupled	LIA	98.37	98.72
		OLIA	98.11	98.64
		BaLIA	97.96	98.39
		Wvegas	98.81	98.72
	Loss-based	Reno	98.45	98.66
		BIC	98.79	98.85
		Cubic	98.67	98.71
		HS-TCP	98.78	98.86
		HTCP	98.37	98.72
		Hybla	97.03	98.02
		Westwood	98.75	98.86
	Delay-based	Vegas	98.84	98.75
		LoLa	98.81	98.81
	Hybrid	Veno	97.97	98.54
		Illinois	98.15	98.62
		YeAH	98.78	97.96
		BBR	97.45	45.48
		DCTCP	97.25	98.42
MPQUIC	Coupled	OLIA	73.45	73.48
	Loss-based	Cubic	69.51	72.33

Difference in bandwidth of the two links

- MPQUIC was sensitive to large differences in bandwidth available on both paths

Protocol	Group	Algorithm	Link 1(20Mbps) Average throughput $[Mbps]$	Link 2(100Mbps) Average throughput $[Mbps]$
MPTCP	Coupled	LIA	19.78	98.86
		OLIA	19.79	98.89
		BaLIA	19.80	98.90
		Wvegas	19.78	98.81
	Loss-based	Reno	19.79	98.93
		BIC	19.80	98.90
		Cubic	19.78	98.85
		HS-TCP	19.79	98.92
		HTCP	19.79	98.87
		Hybla	19.78	98.87
		Westwood	19.74	98.86
	Delay-based	Vegas	19.79	98.88
		LoLa	19.79	98.85
	Hybrid	Veno	19.79	98.84
		Illinois	19.78	98.87
		YeAH	19.79	97.90
		BBR	19.78	72.43
		DCTCP	19.78	98.86
MPQUIC	Coupled	OLIA	10.76	65.50
	Loss-based	Cubic	18.23	25.09

TU Delft

Difference in RTTs between two sub-flows

- Sub-flow on the link that had the higher RTT experienced a drop in the sending rate

Protocol	Group	Algorithm	Link 1 (100ms) Average throughput $[Mbps]$	Link 2 (0ms) Average throughput $[Mbps]$
MPTCP	Coupled	LIA	56.50	94.73
		OLIA	64.51	96.95
		BaLIA	91.90	97.21
		Wvegas	93.56	93.16
	Loss-based	Reno	67.80	96.78
		BIC	24.13	95.84
		Cubic	42.44	92.04
		HS-TCP	92.97	96.90
		HTCP	78.80	94.97
		Hybla	93.07	96.77
		Westwood	53.55	55.55
	Delay-based	Vegas	95.47	92.86
		LoLa	92.66	94.50
	Hybrid	Veno	93.31	96.82
		Illinois	90.89	97.56
		YeAH	95.90	95.40
		BBR	91.19	43.28
		DCTCP	82.39	96.01
MPQUIC	Coupled	OLIA	25.20	85.73
	Loss-based	Cubic	17.34	88.30

TUDelft

Fairness

- Inter/RTT-fairness issues present in traditional TCP (or QUIC) are also present between different MPTCP/MPQUIC sub-flows

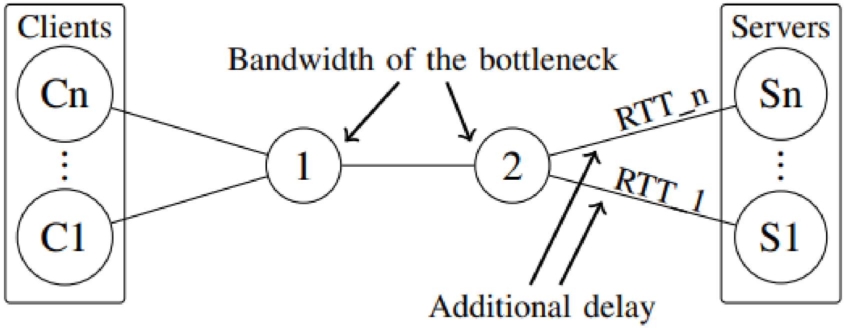
TU Delft

Evaluation of AQMs

TU Delft

Experimental Evaluation – Setup

- The bandwidth between nodes 1 and 2 (bottleneck) is limited
- On the output link (from node 1 to 2) an AQM mechanism was configured with the use of Linux TC
- Transport protocol: TCP
- AQMs: CoDel, FQ_CoDel, RED, PIE, RED, RED+SFQ



TUDelft

Inter-fairness

- AQMs mostly target loss-based algorithms, significantly improving their fairness properties
- The performance of hybrid algorithms decreased or remained the same

	Reno	BIC	Cubic	HS-TCP	HTCP	Hybla	Westwood	Vegas	LoLa	Veno	Illinois	YeAH	BBR	DCTCP
Reno	0.99	0.85	0.97	0.94	0.99	0.83	0.87	0.53	0.55	0.78	0.89	0.88	0.57	0.99
BIC	0.85	0.98	0.80	0.95	0.84	0.66	0.67	0.53	0.67	0.61	0.66	0.68	0.59	0.84
Cubic	0.97	0.80	0.99	0.88	0.96	0.87	0.88	0.53	0.56	0.82	0.88	0.89	0.58	0.96
HS-TCP	0.94	0.95	0.88	0.98	0.92	0.72	0.70	0.53	0.58	0.66	0.76	0.73	0.60	0.90
HTCP	0.99	0.84	0.96	0.92	0.99	0.81	0.86	0.52	0.56	0.78	0.88	0.88	0.57	0.99
Hybla	0.83	0.66	0.87	0.72	0.81	0.99	0.97	0.52	0.56	0.92	0.98	0.96	0.58	0.83
Westwood	0.87	0.67	0.88	0.70	0.86	0.97	1.00	0.52	0.54	0.93	0.95	0.97	0.58	0.87
Vegas	0.53	0.53	0.53	0.53	0.52	0.52	0.52	1.00	0.67	0.52	0.52	0.52	0.65	0.53
LoLa	0.55	0.67	0.56	0.58	0.56	0.56	0.54	0.67	0.80	0.54	0.54	0.56	0.79	0.56
Veno	0.78	0.61	0.82	0.66	0.78	0.92	0.93	0.52	0.54	0.98	0.92	0.92	0.60	0.78
Illinois	0.89	0.66	0.88	0.76	0.88	0.98	0.95	0.52	0.54	0.92	0.99	0.98	0.58	0.90
YeAH	0.88	0.68	0.89	0.73	0.88	0.96	0.97	0.52	0.56	0.92	0.98	0.99	0.62	0.88
BBR	0.57	0.59	0.58	0.60	0.57	0.58	0.58	0.65	0.79	0.60	0.58	0.62	0.94	0.58
DCTCP	0.99	0.84	0.96	0.90	0.99	0.83	0.87	0.53	0.56	0.78	0.90	0.88	0.58	0.99

	Reno	BIC	Cubic	HS-TCP	HTCP	Hybla	Westwood	Vegas	LoLa	Veno	Illinois	YeAH	BBR	DCTCP
Reno	0.96	0.91	0.96	0.94	0.95	0.94	0.96	0.60	0.87	0.89	0.82	0.94	0.66	0.96
BIC	0.91	0.90	0.91	0.89	0.90	0.87	0.90	0.70	0.82	0.82	0.84	0.88	0.62	0.90
Cubic	0.96	0.91	0.96	0.94	0.96	0.94	0.96	0.69	0.87	0.88	0.79	0.95	0.65	0.96
HS-TCP	0.94	0.89	0.94	0.94	0.94	0.94	0.95	0.60	0.82	0.89	0.76	0.94	0.65	0.94
HTCP	0.95	0.90	0.96	0.94	0.95	0.94	0.96	0.60	0.83	0.89	0.79	0.94	0.66	0.95
Hybla	0.94	0.87	0.94	0.94	0.94	0.95	0.96	0.61	0.77	0.93	0.76	0.95	0.66	0.94
Westwood	0.96	0.90	0.96	0.95	0.96	0.96	0.96	0.59	0.85	0.92	0.80	0.96	0.70	0.96
Vegas	0.60	0.70	0.69	0.60	0.60	0.61	0.59	1.00	0.67	0.62	0.67	0.61	0.66	0.60
LoLa	0.87	0.82	0.87	0.82	0.83	0.77	0.85	0.67	0.82	0.79	0.88	0.85	0.68	0.84
Veno	0.89	0.82	0.88	0.89	0.89	0.93	0.92	0.62	0.79	0.94	0.67	0.94	0.66	0.89
Illinois	0.82	0.84	0.79	0.76	0.79	0.76	0.80	0.67	0.88	0.67	0.99	0.76	0.63	0.81
YeAH	0.94	0.88	0.95	0.94	0.94	0.95	0.96	0.61	0.85	0.94	0.76	0.96	0.70	0.94
BBR	0.66	0.62	0.65	0.65	0.66	0.66	0.70	0.66	0.68	0.66	0.63	0.70	0.91	0.66
DCTCP	0.96	0.90	0.96	0.94	0.95	0.94	0.96	0.60	0.84	0.89	0.81	0.94	0.66	0.95

RTT-fairness

- AQMs usually decrease the already bad RTT-fairness properties of most algorithms

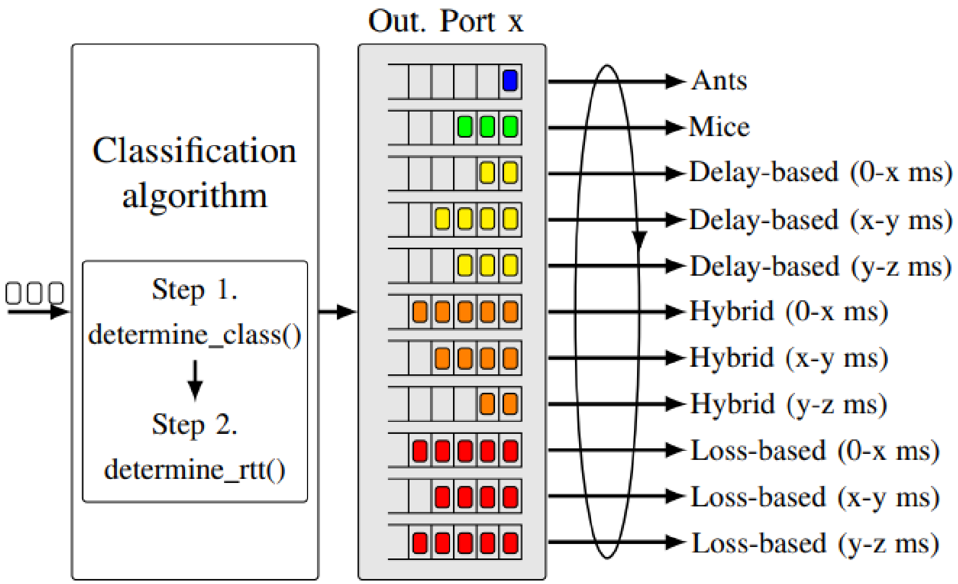
	Reno	BIC	Cubic	HS-TCP	HTCP	Hybla	Westwood	Vegas	LoLa	Veno	Illinois	YeAH	BBR	DCTCP
0	0.99	0.99	0.99	0.98	0.99	0.99	0.99	0.99	0.97	0.97	0.97	0.98	0.97	0.98
20	0.85	0.74	0.84	0.79	0.94	0.89	0.86	0.83	0.73	0.91	0.92	0.86	0.56	0.84
40	0.74	0.68	0.80	0.70	0.89	0.89	0.77	0.82	0.59	0.85	0.82	0.78	0.54	0.75
60	0.69	0.66	0.72	0.67	0.88	0.94	0.71	0.78	0.59	0.83	0.77	0.74	0.55	0.69
80	0.67	0.63	0.75	0.62	0.87	0.95	0.69	0.82	0.59	0.80	0.74	0.74	0.56	0.66
100	0.66	0.63	0.73	0.59	0.84	0.95	0.65	0.80	0.62	0.79	0.80	0.73	0.56	0.65
120	0.60	0.60	0.82	0.59	0.82	0.96	0.63	0.82	0.56	0.78	0.82	0.74	0.58	0.60
140	0.59	0.59	0.78	0.57	0.80	0.95	0.61	0.85	0.57	0.76	0.83	0.71	0.57	0.60
160	0.60	0.58	0.76	0.56	0.79	0.95	0.59	0.72	0.55	0.75	0.83	0.69	0.58	0.60
180	0.59	0.56	0.78	0.56	0.74	0.95	0.59	0.81	0.55	0.72	0.79	0.82	0.58	0.59
200	0.58	0.55	0.73	0.54	0.70	0.95	0.57	0.77	0.54	0.74	0.78	0.90	0.59	0.58

	Reno	BIC	Cubic	HS-TCP	HTCP	Hybla	Westwood	Vegas	LoLa	Veno	Illinois	YeAH	BBR	DCTCP
0	0.96	0.93	0.94	0.94	0.94	0.94	0.95	0.95	0.93	0.94	0.94	0.94	0.94	0.94
20	0.75	0.76	0.75	0.75	0.76	0.73	0.94	0.89	0.69	0.73	0.85	0.87	0.57	0.76
40	0.66	0.69	0.68	0.63	0.65	0.72	0.84	0.91	0.58	0.63	0.66	0.89	0.56	0.65
60	0.61	0.64	0.64	0.60	0.67	0.73	0.75	0.84	0.60	0.62	0.64	0.82	0.57	0.61
80	0.58	0.61	0.62	0.58	0.64	0.73	0.71	0.84	0.65	0.62	0.60	0.79	0.56	0.58
100	0.57	0.61	0.63	0.57	0.64	0.76	0.71	0.84	0.60	0.59	0.60	0.78	0.57	0.59
120	0.57	0.60	0.61	0.57	0.64	0.76	0.66	0.85	0.58	0.59	0.66	0.74	0.58	0.57
140	0.56	0.58	0.61	0.56	0.63	0.76	0.65	0.88	0.57	0.58	0.60	0.71	0.60	0.56
160	0.56	0.57	0.62	0.56	0.63	0.77	0.61	0.87	0.55	0.57	0.61	0.73	0.62	0.57
180	0.55	0.58	0.62	0.55	0.61	0.74	0.58	0.81	0.55	0.58	0.60	0.71	0.64	0.55
200	0.55	0.58	0.59	0.56	0.61	0.75	0.62	0.96	0.55	0.60	0.60	0.70	0.66	0.55

P4air

P4air

- All flows on a switch are classified into 2+3k groups:
  - ant flows
  - mice flows
  - k loss-based flows
  - k delay-based flows
  - k hybrid-based flows



TU Delft

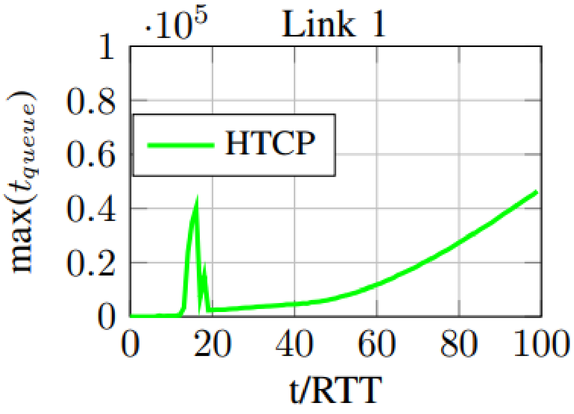
Metrics

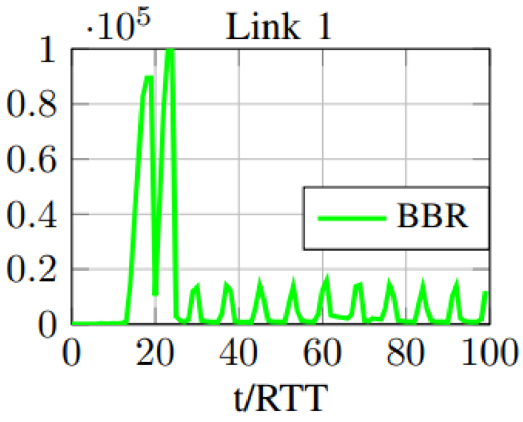
- Each switch keeps track of the following statistics for each RTT interval:
  - number of processed packets,
  - number of processed ACK packets,
  - maximum queuing delay experienced in the current RTT interval,
  - number of dropped packets as the difference between packets processed in the ingress and egress pipeline

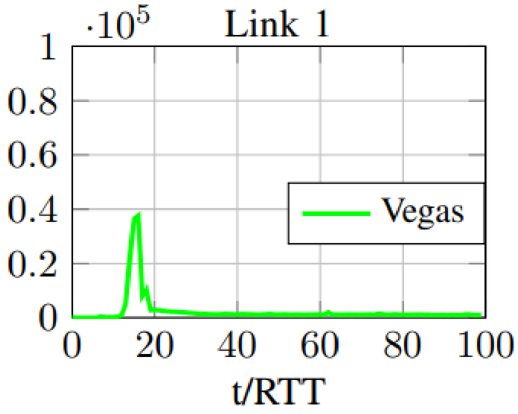
TUDelft

Example: Maximum queuing delay

- Loss-based algorithm is building a queue, without reacting to it
- Delay-based algorithms avoid the queue-buildup
- BBR algorithm periodically builds a small queue



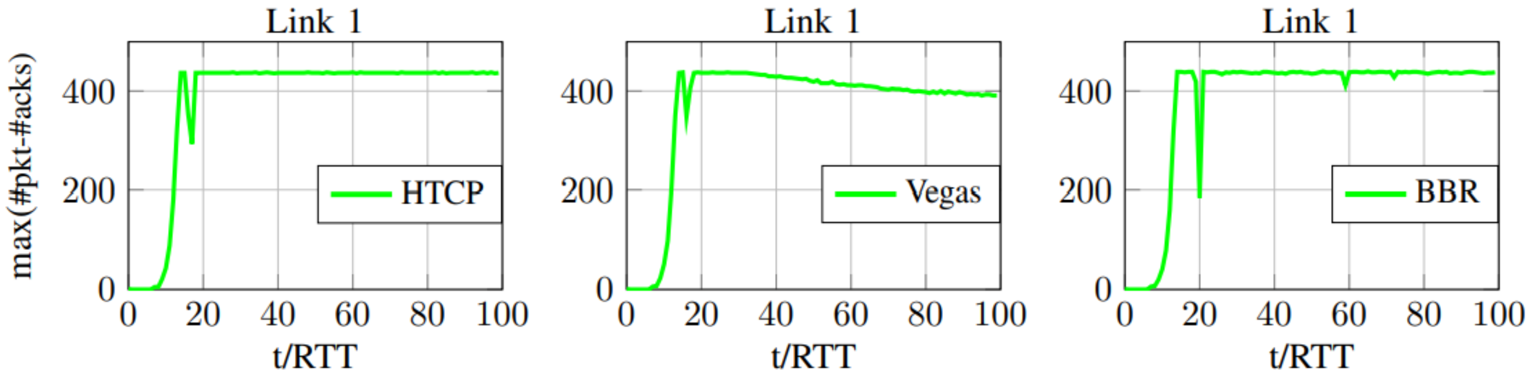




Detecting the end of the slow-start

Two possible indicators:

- The first loss
- When the number of packets sent in one RTT stops increasing



TU Delft

Conclusion

- Programmable switches can be used to determine the type of congestion control algorithms used by a flow
- In the future, we plan to investigate if applying different actions to different groups could increase the fairness among different congestion control groups

TUDelft

Questions/Comments/Suggestions?

- Contact Info:
  - Belma Turković (B.Turkovic-2@tudelft.nl)
  - Fernando Kuipers (F.A.Kuipers@tudelft.nl)

TU Delft