Congestion control, generalized

Ankit Singla

ETH Zürich Spring 2017

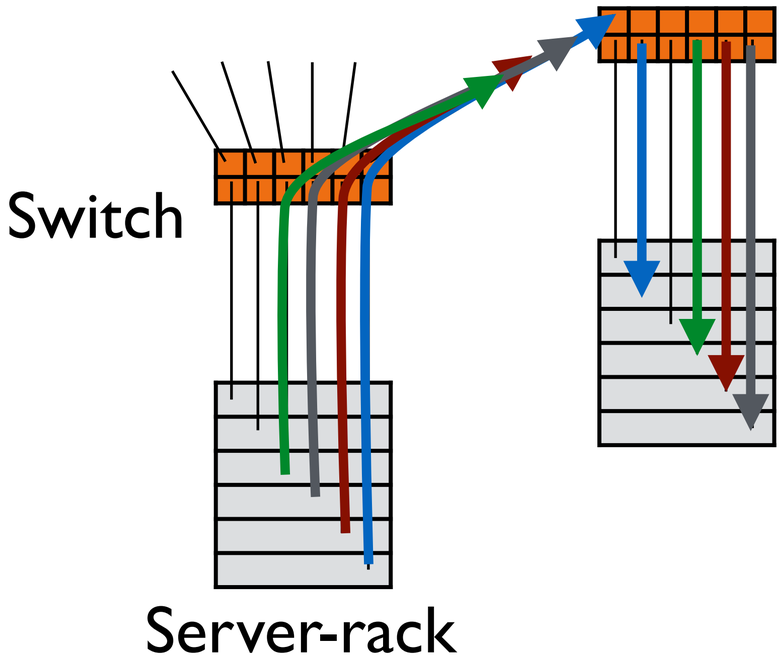
[Some slides are adapted from Mo Dong's PCC presentation at NSDI '15]

This lecture …

- TCP is not good enough (again)
- How do we fix it?!

- Intro: "Content distribution networks"
  - *Collaboration Opportunities for Content Delivery and Network Infrastructures (pages 6-25)*

Congestion



Problem: what rate to send at?

The right rate is important

Rate > available capacity …

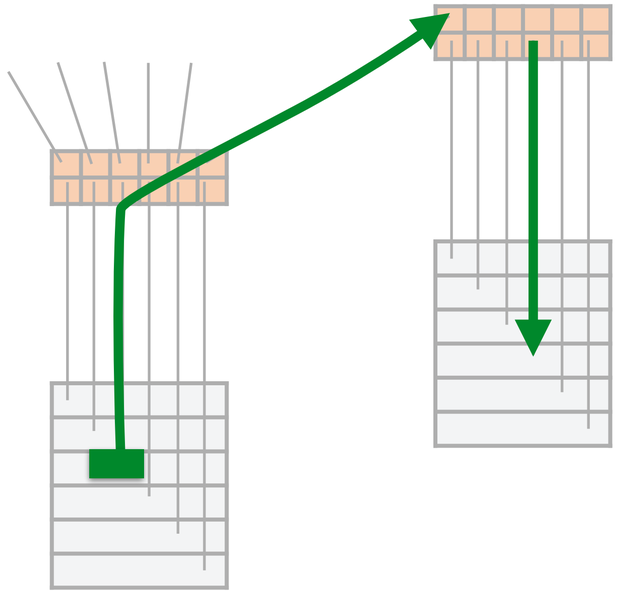
Packet loss and delay

Unfair capacity distribution

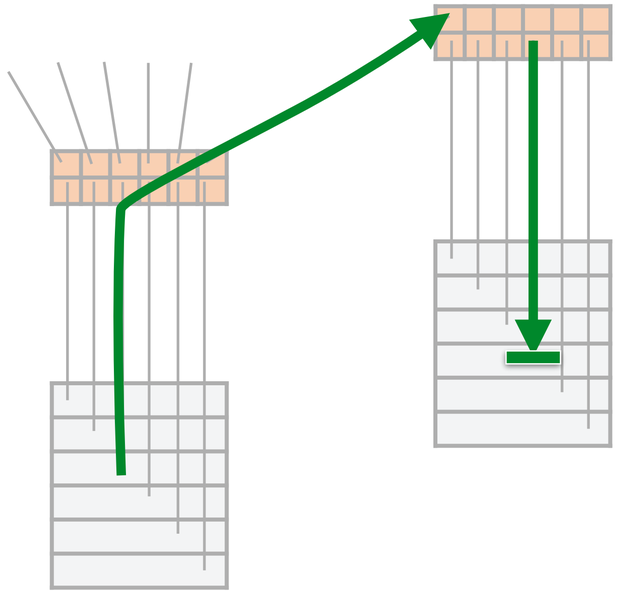
Rate < available capacity …

Inefficient use of the network

Probing the capacity

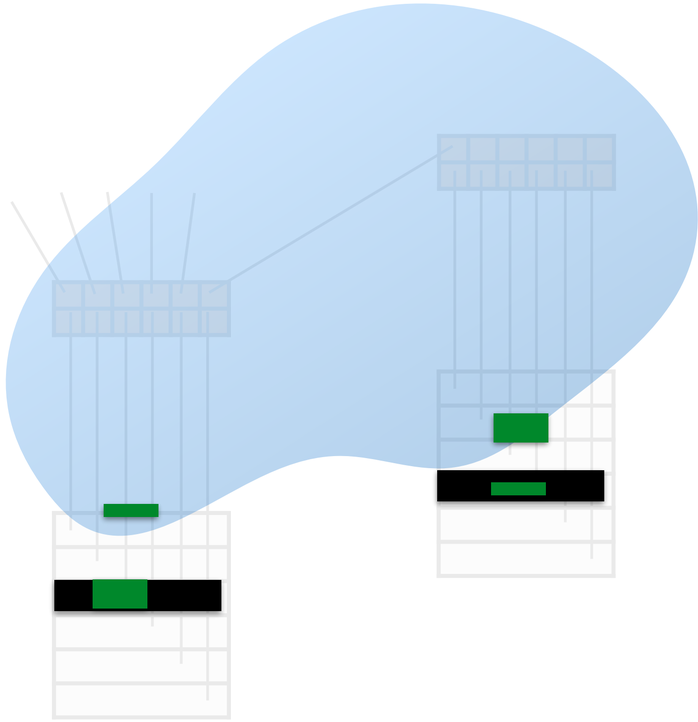


Probing the capacity



Feedback control loop

The network as a black box



1. Send packets
2. Wait
3. ACKs return in an RTT (or not)
4. Adjust sending rate

Loss-reactive congestion control

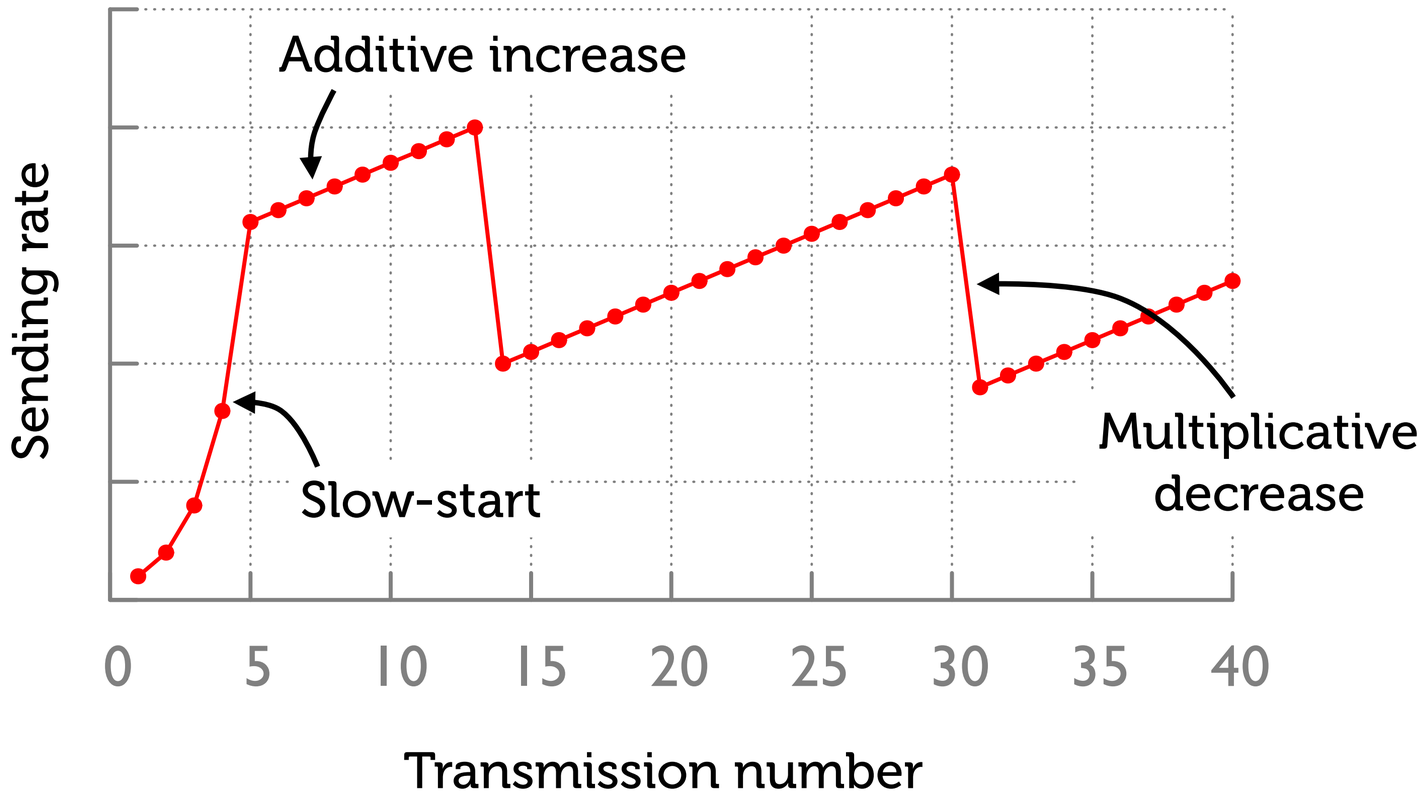
All packets ACK-ed …

Increase rate!

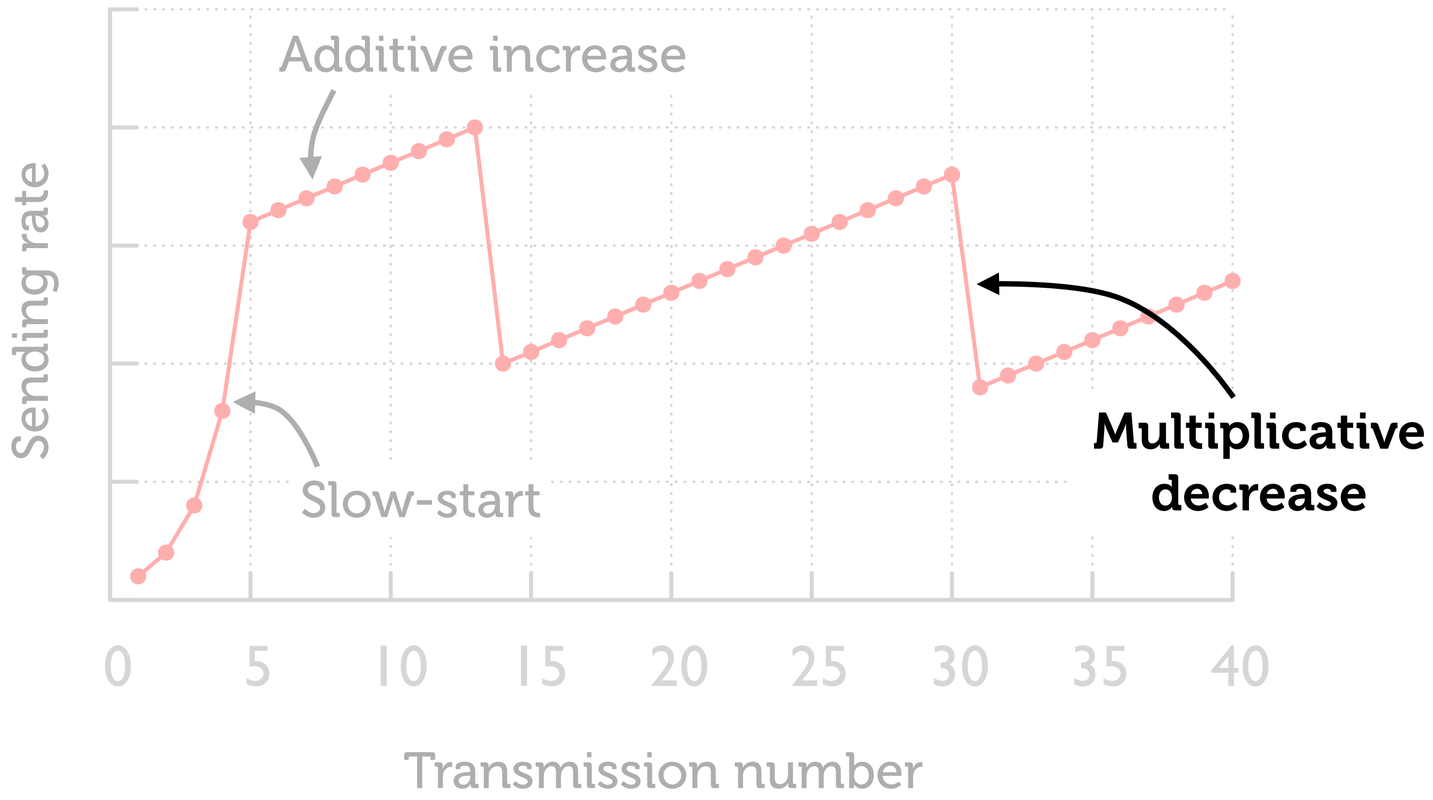
Some packets presumed lost …

Decrease rate!

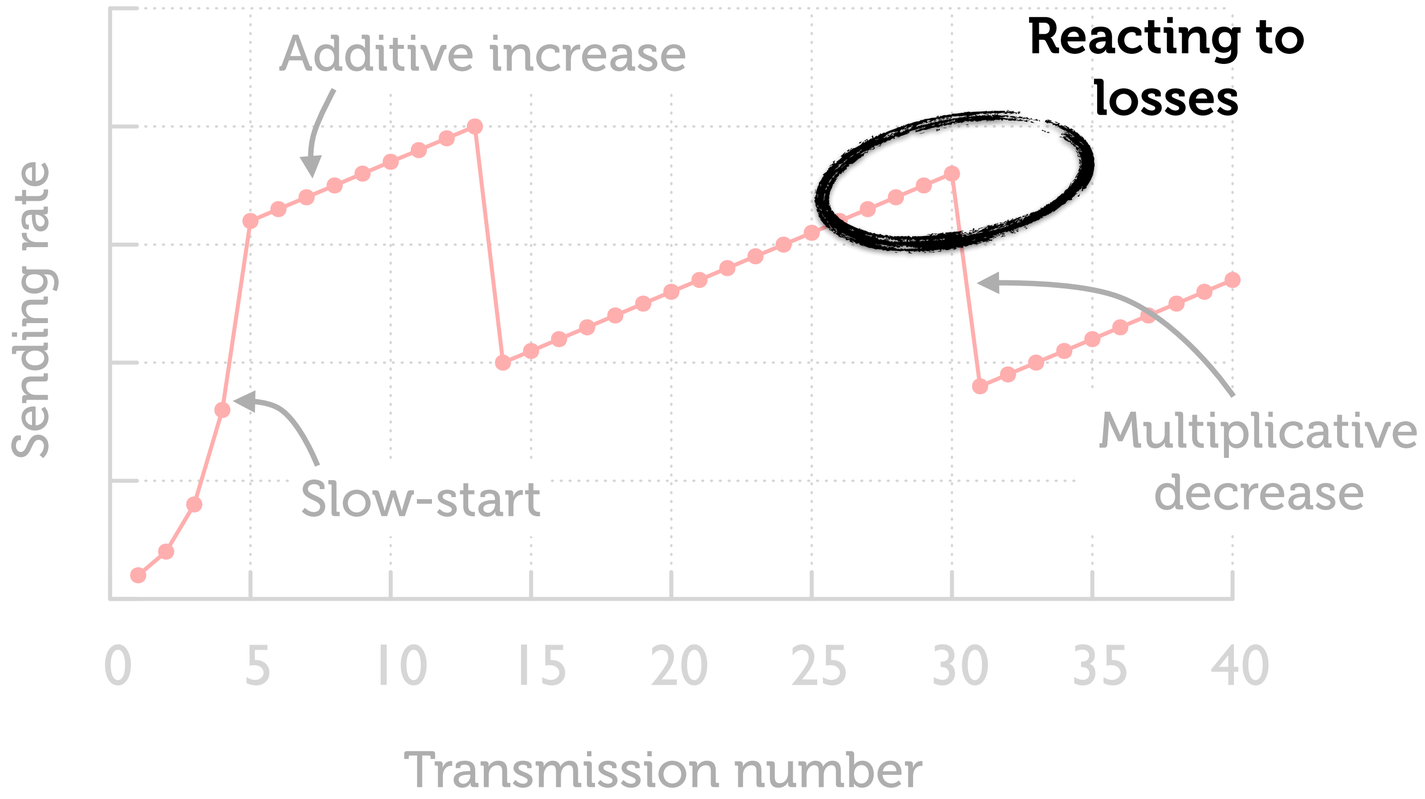
TCP (extremely briefly)



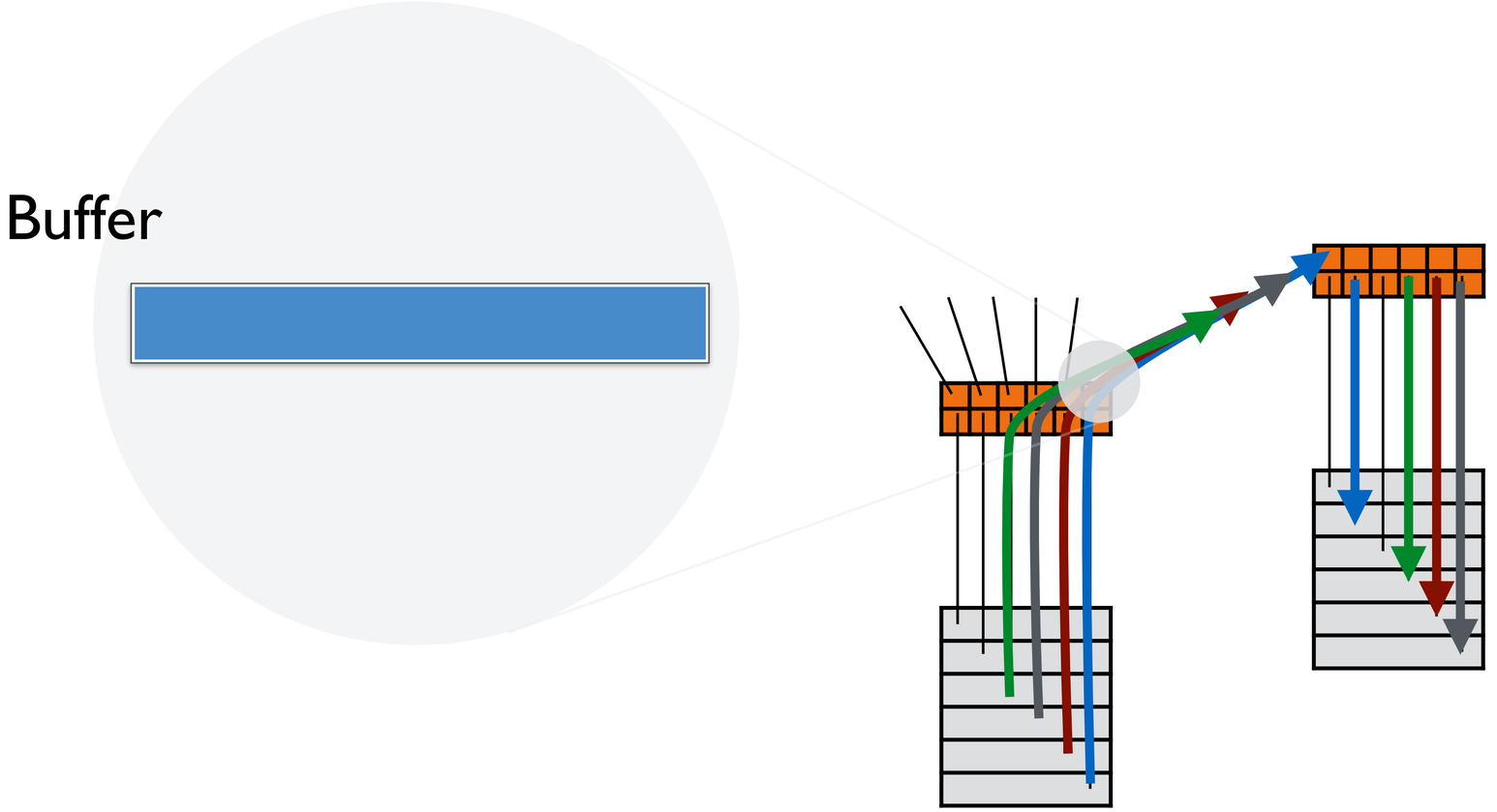
Problems with TCP



Problems with TCP



Long queues ⇒ queueing delays

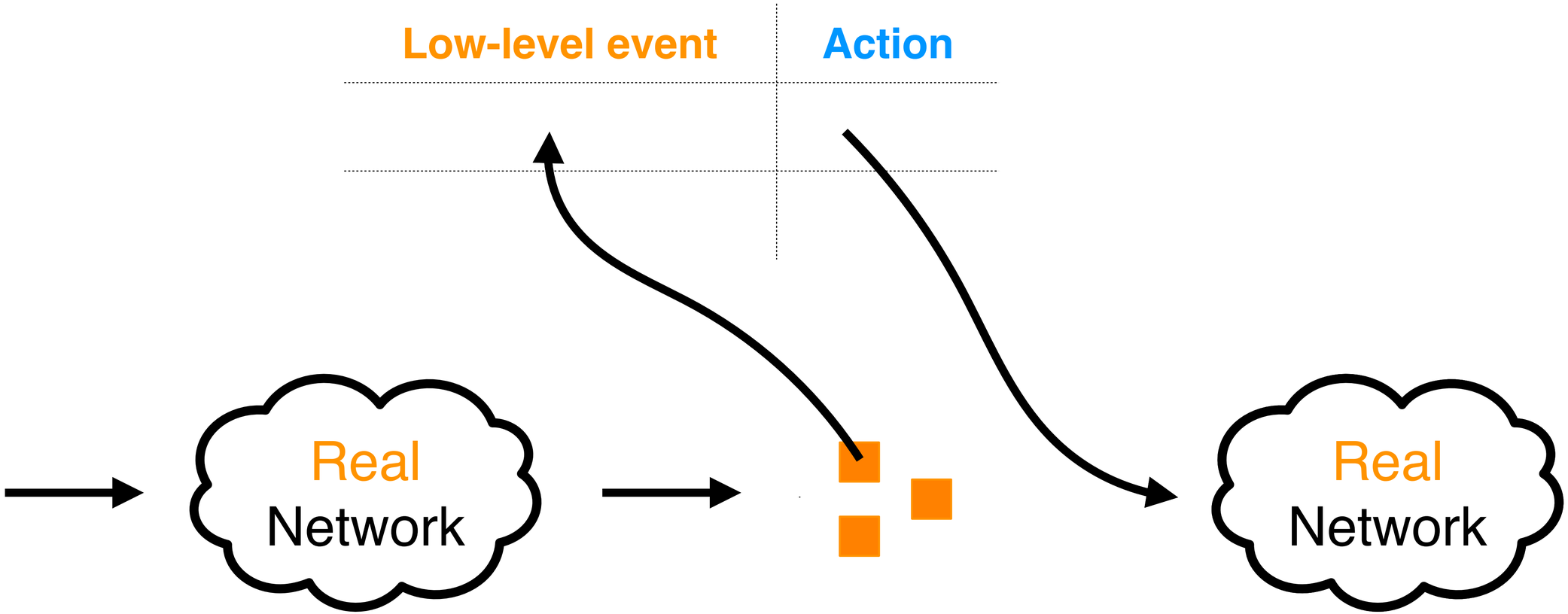


Generic (whatever)-TCP strategy

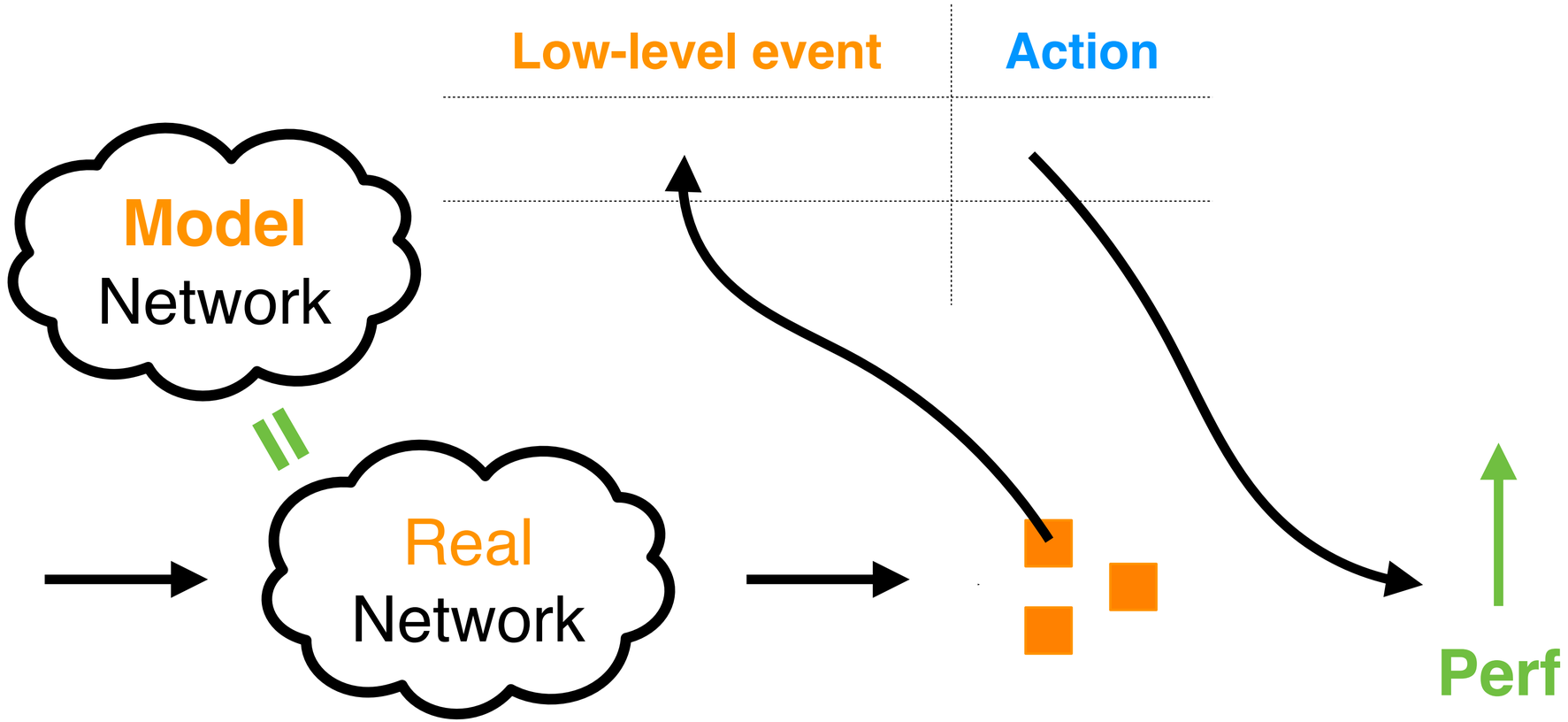
	Low-level event	Action
Reno	1 pkt loss	cwnd/2
Scalable	ACK	cwnd+1
CUBIC	Time pass 1ms	cwnd+f(t,cwn,rtt)
FAST	RTT increase x%	Reduce cwnd to f(x)%
HTCP	100 ACK	cwnd+f(cwnd)/cwnd

Hard-wired mappings: low-level events to control actions

Generic (whatever)-TCP strategy



Generic (whatever)-TCP strategy



The event-action mappings encode a model of the network

Problem: this model is often wrong!

Event	Model	Action
Packet loss	I'm causing congestion	
	Shallow buffer overflow	
	Congestion from other flows	
	Loss is random	
	…	…

High BDP	Wireless	Satellite	Inter-DC	Intra-DC
BIC	Westwood	Hybla	Illinois	ICTCP
H-TCP	Vegas	STAR	SABUL	DCTCP
Compound	Veno			
CUBIC				
FAST TCP				

Generalizing congestion control?

ACM SIGCOMM, 2013

TCP ex Machina: Computer-Generated Congestion Control

Keith Winstein and Hari Balakrishnan

Usenix NSDI, 2015

PCC: Re-architecting Congestion Control for Consistent High Performance

Mo Dong*, Qingxi Li*, Doron Zarchy**, P. Brighten Godfrey*, and Michael Schapira**

ACM Queue, 2016

BY NEAL CARDWELL, YUCHUNG CHENG, C. STEPHEN GUNN, SOHEIL HASSAS YEGANEH, AND VAN JACOBSON

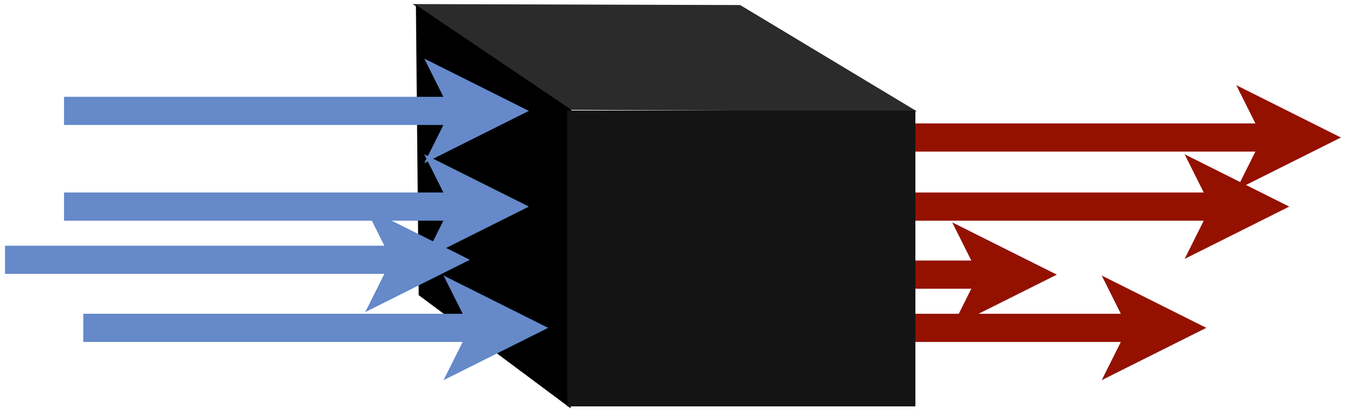
BBR: Congestion-Based Congestion Control

PCC: a black-box, learning approach

Usenix NSDI, 2015

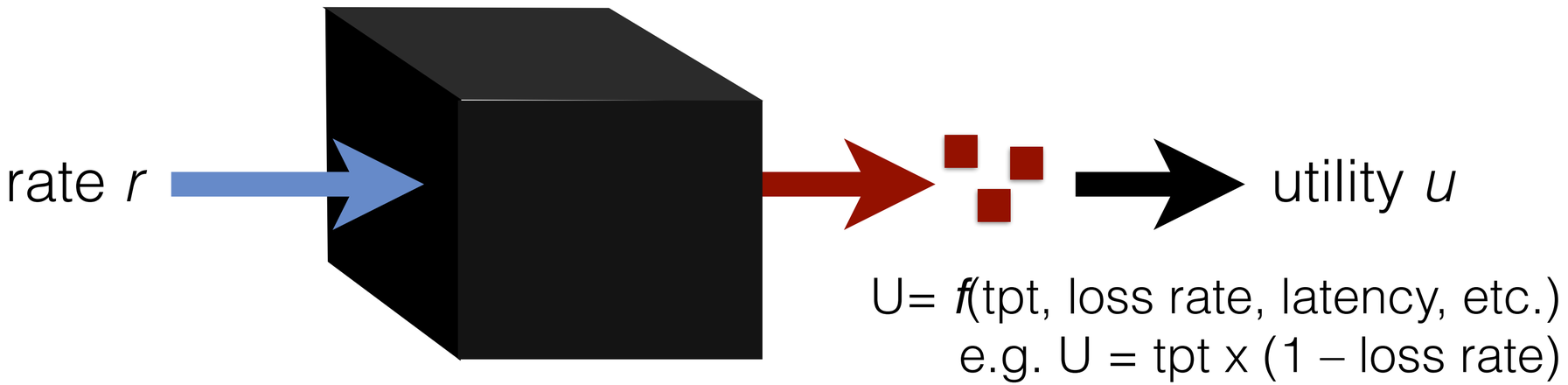
PCC: Re-architecting Congestion Control for Consistent High Performance

Mo Dong*, Qingxi Li*, Doron Zarchy**, P. Brighten Godfrey*, and Michael Schapira**

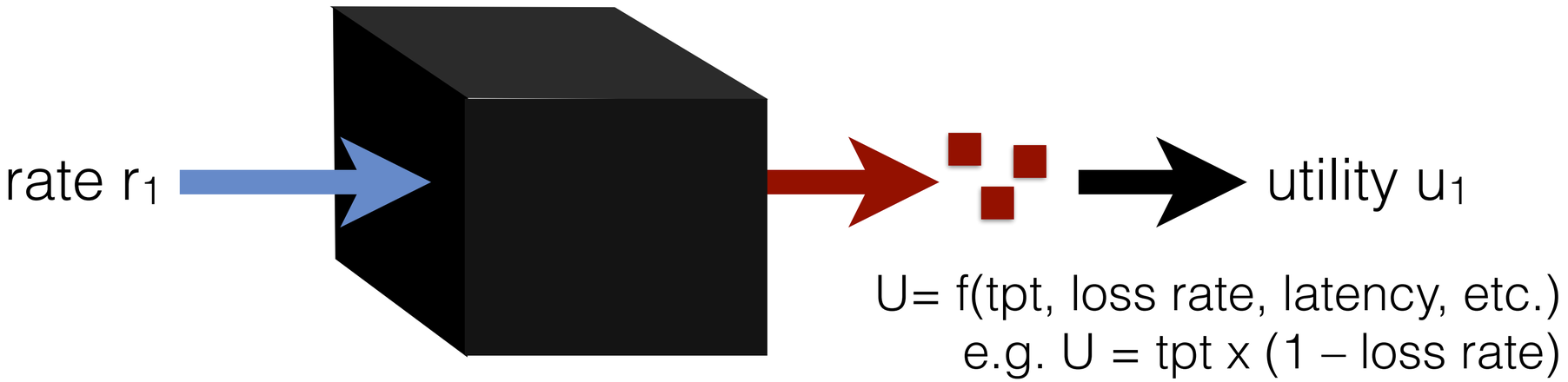


What is the right rate to send?

What is the right rate to send?

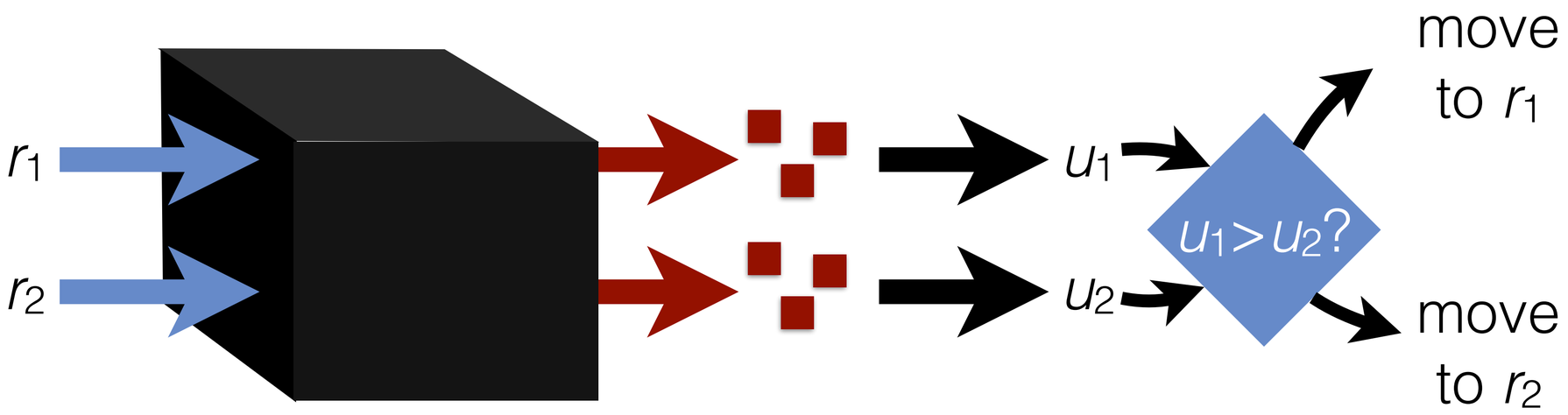


What is the right rate to send?

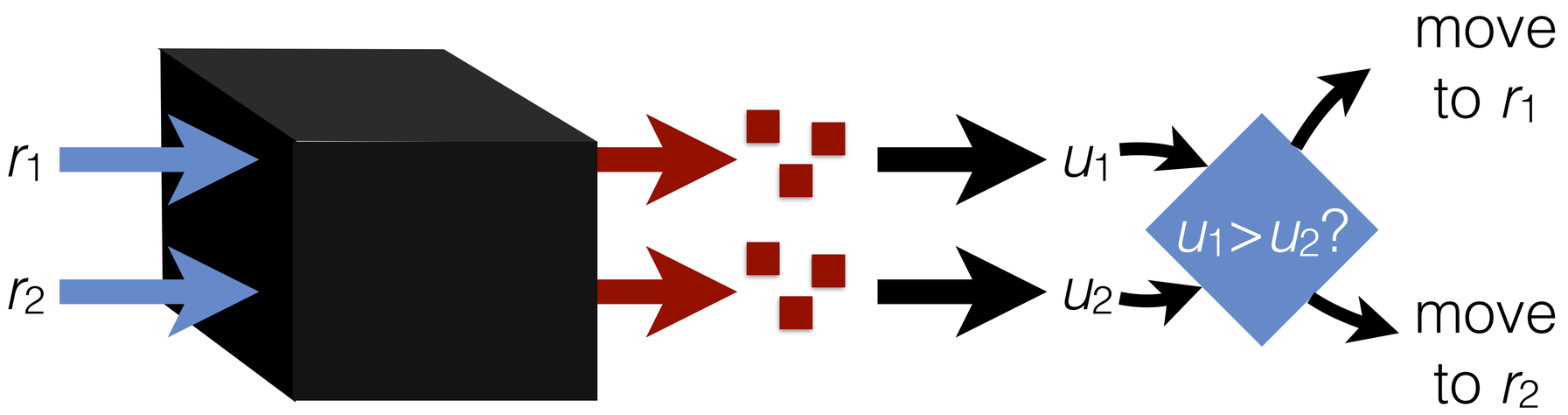


No matter how complex the network,
rate r —> utility u

PCC: control based on evidence



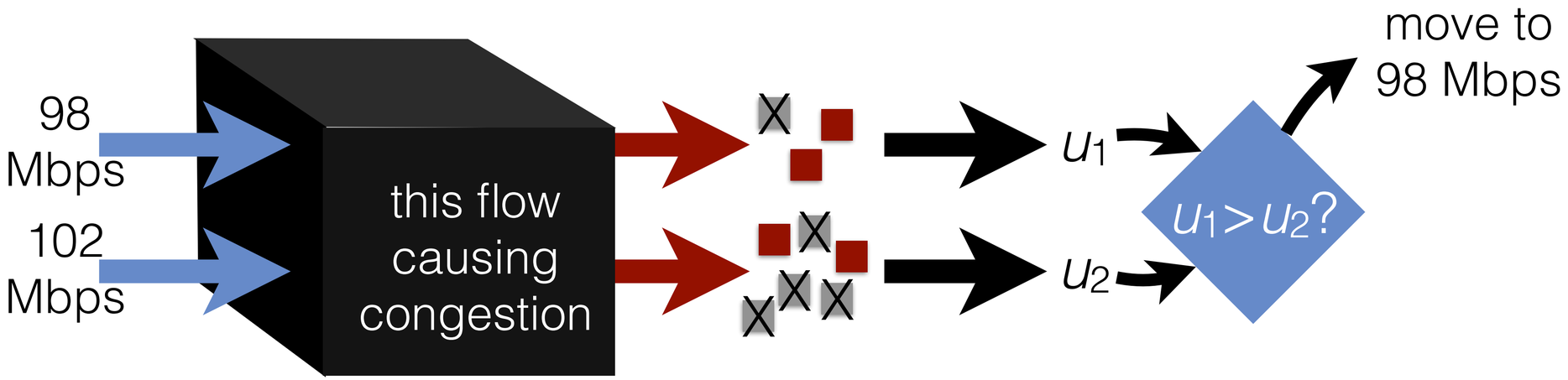
Performance-oriented congestion control



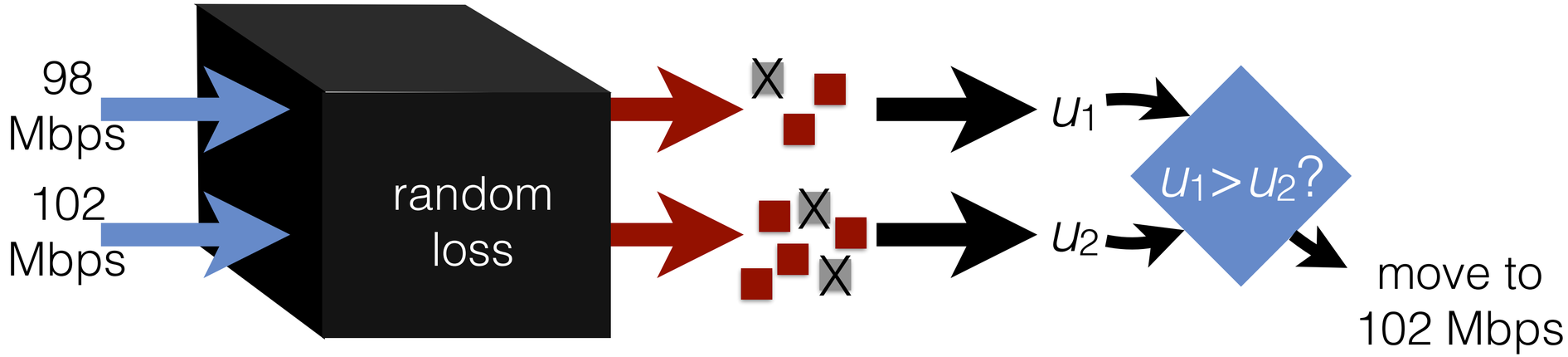
Observe real performance

Control based on empirical evidence

PCC: control based on evidence



PCC: control based on evidence

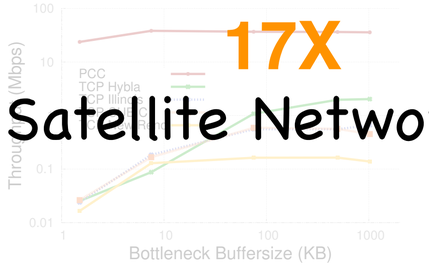


Higher performance

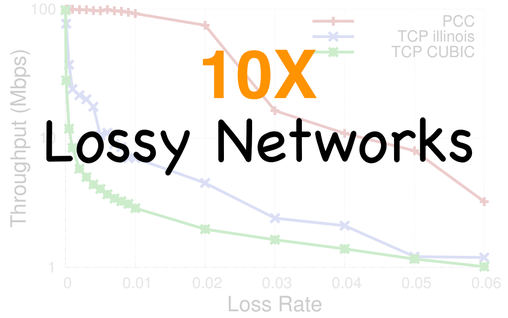
Table 1: PCC significantly outperforms TCP in inter-data center environments. RTT in msec; throughput in Mbps.

Transmission Pair	RTT	PCC	SABUL	CUBIC	Illinois
GPO → NYSERNet	12.1	791	674	129	326
GPO → Missouri	47.4	808	661	80.7	90.1
GPO → Illinois	35.4	766	664	84.5	102
NYSERNet → Missouri	47.4	816	662	108	109
Wisconsin → Illinois	9.01	801	700	547	562
GPO → Wisc.	38.0	783	487	79.3	120
NYSERNet → Wisc.	38.3	791	673	134	134
Missouri → Wisc.	20.9	807	698	259	262
NYSERNet → Illinois	36.1	808	674	141	141

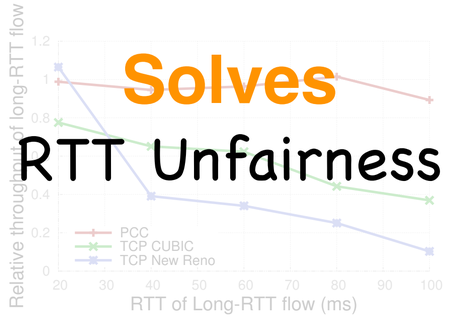
4X
InterDC



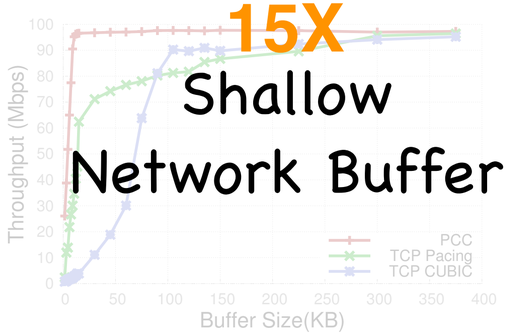
17X
Satellite Networks



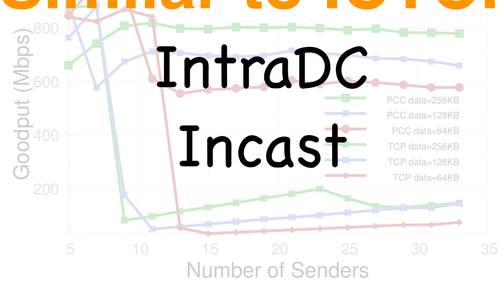
10X
Lossy Networks



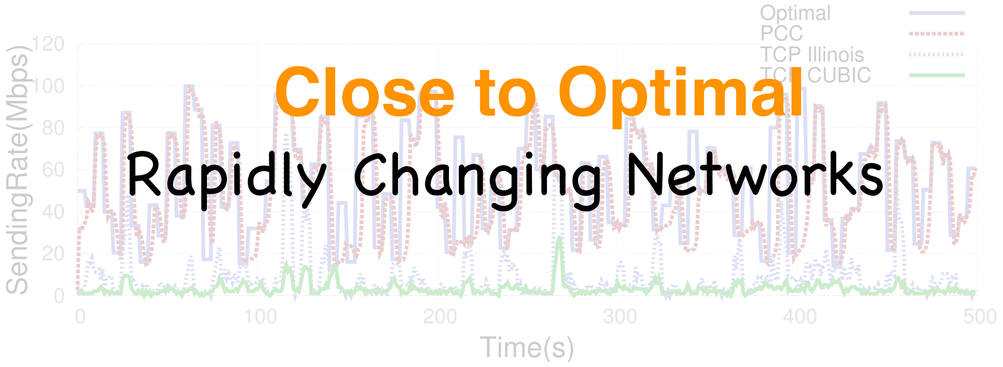
Solves
RTT Unfairness



15X
Shallow
Network Buffer



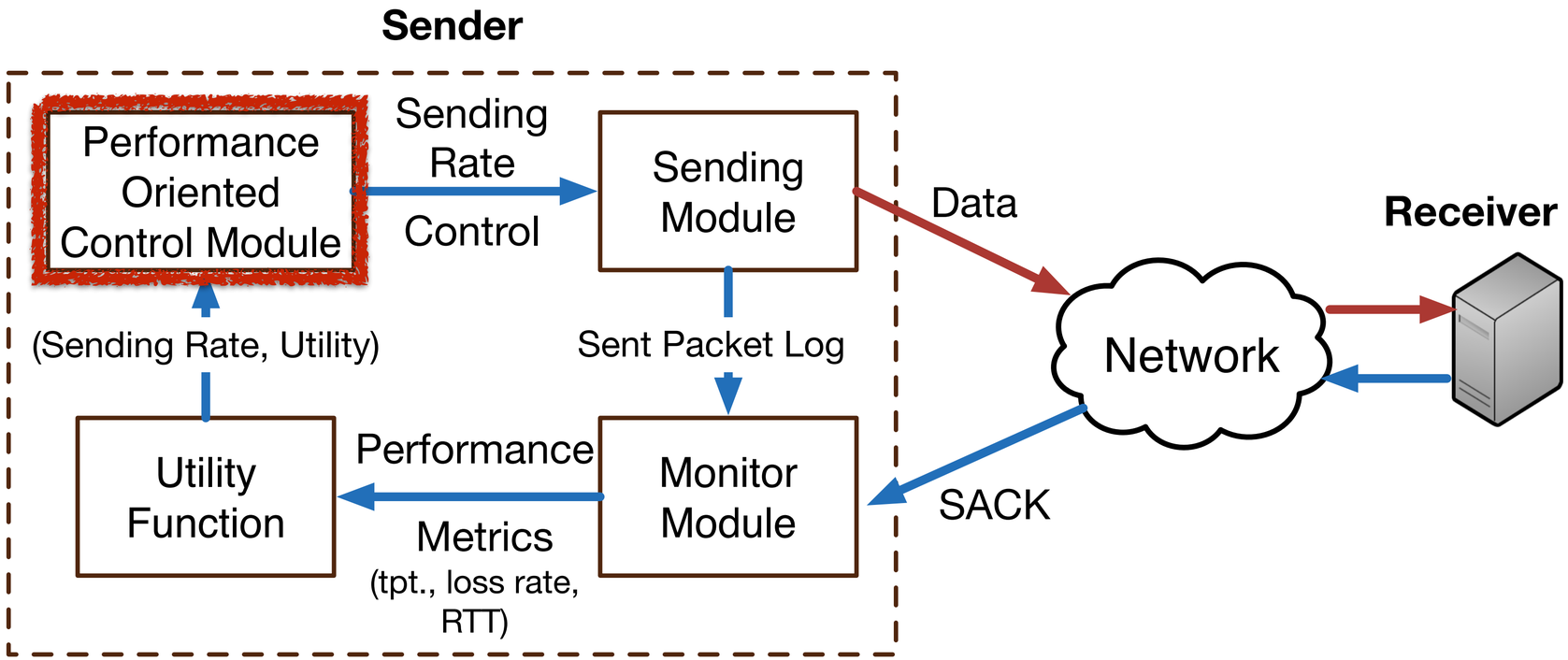
Similar to ICTCP
IntraDC
Incast



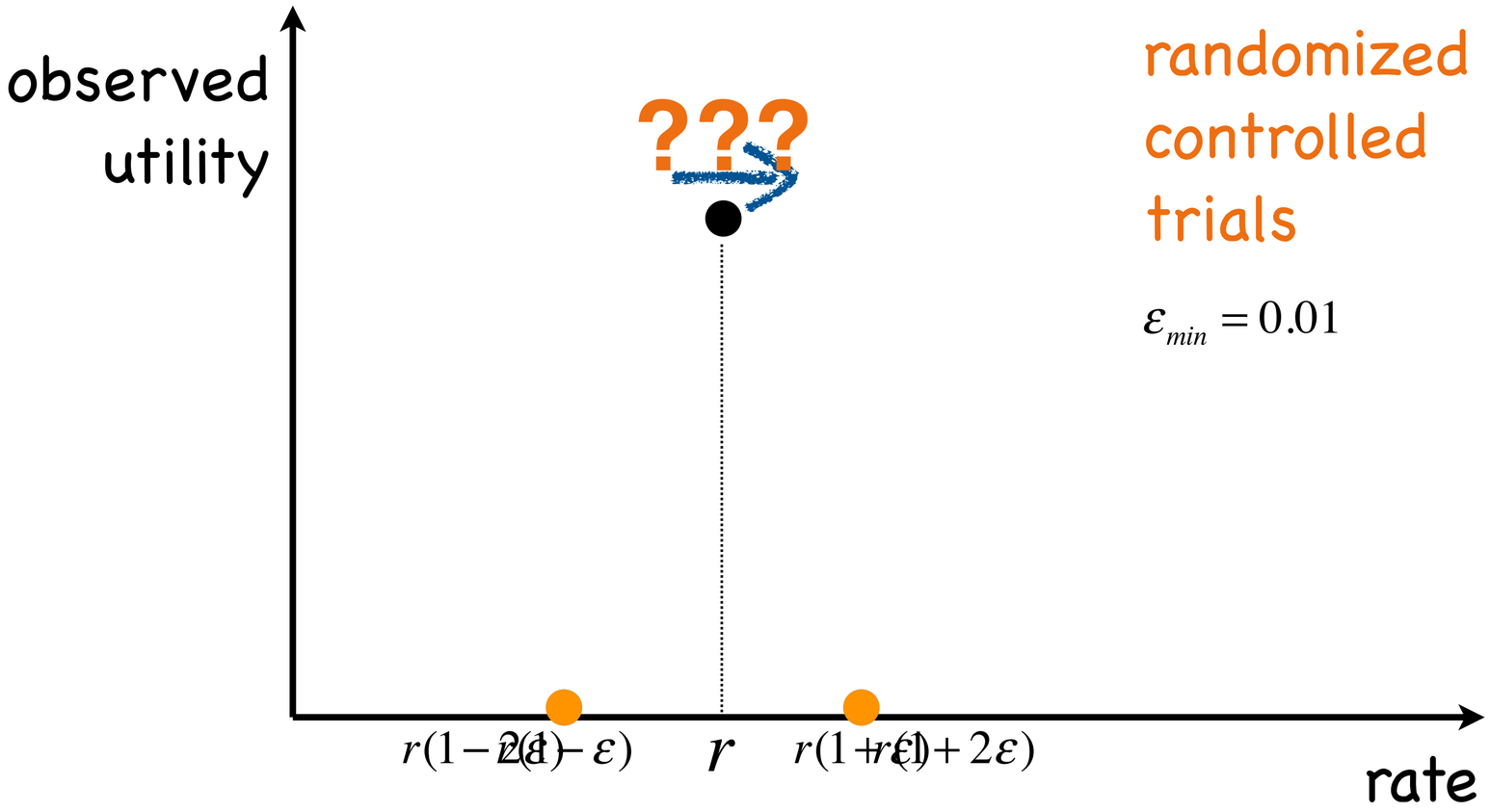
Close to Optimal
Rapidly Changing Networks

5X in median
Global Commercial Internet

Software components



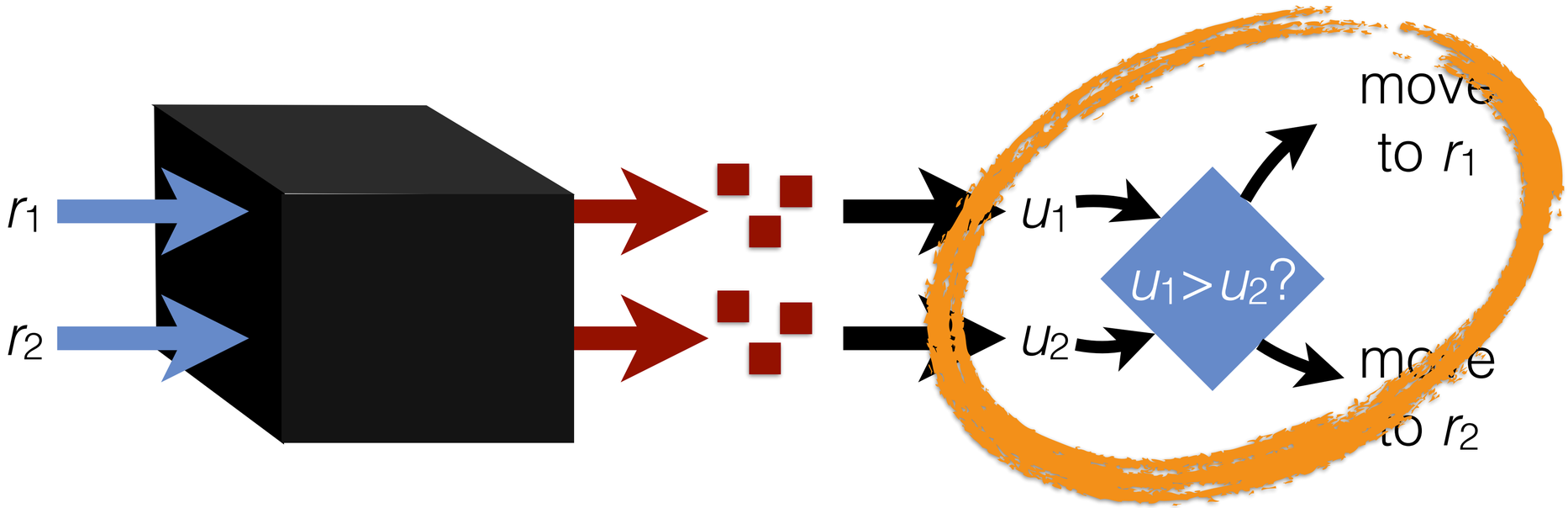
Performance oriented control



randomized
controlled
trials

$\varepsilon_{min} = 0.01$

Where is the congestion control?



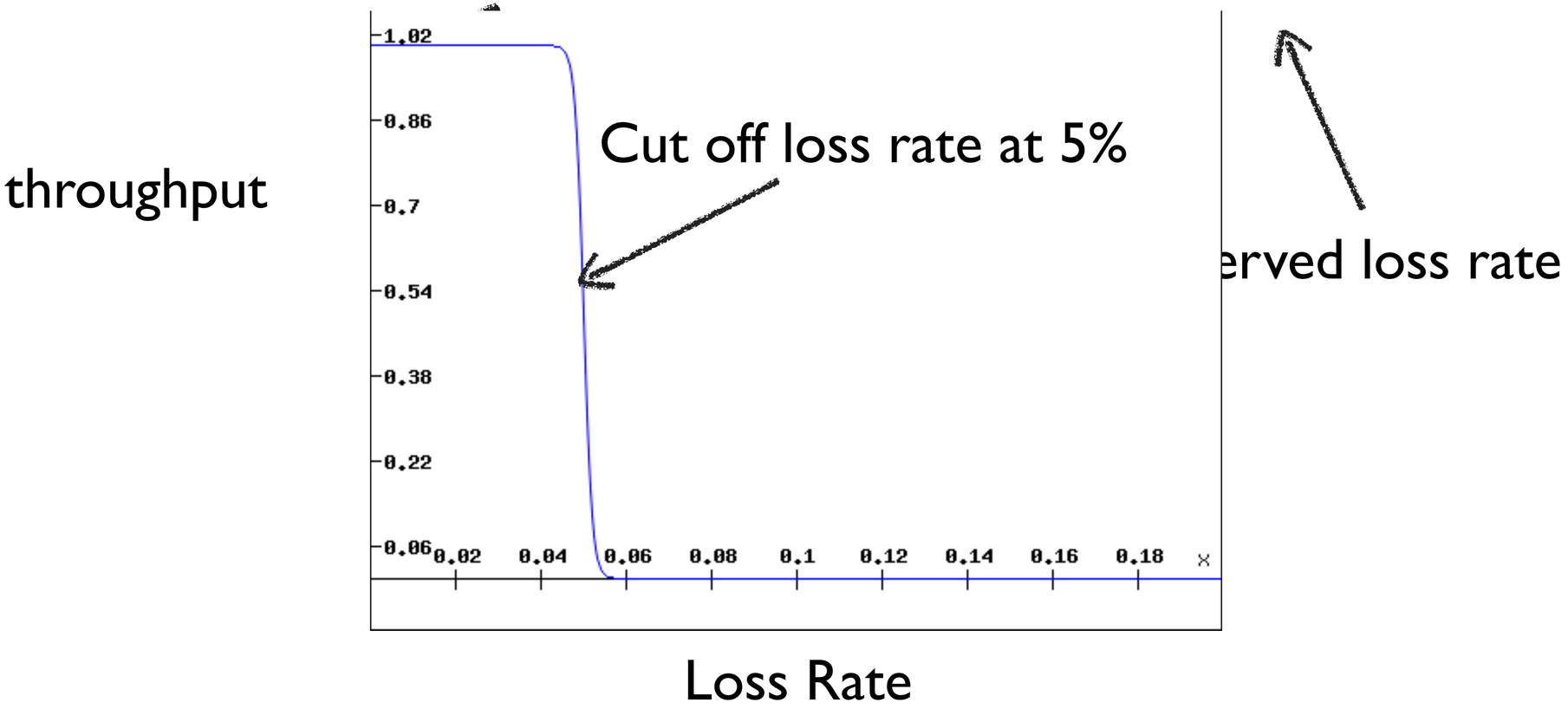
Selfishly maximizing utility ⇒ non-cooperative game

Does PCC converge to a fair Nash equilibrium?

Congestion control is in game theory

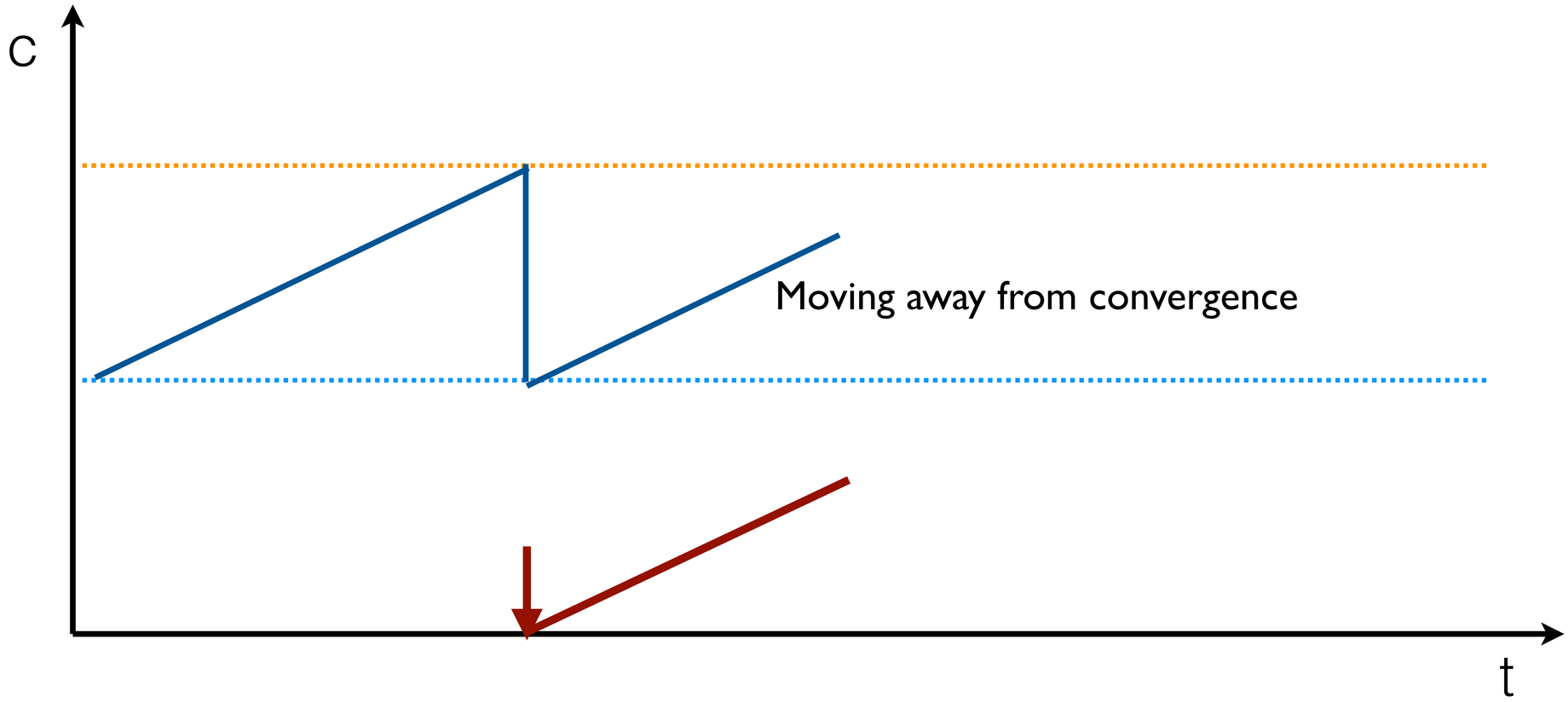
Some utility functions converge to a fair, efficient NE

$$u_i(x) = T_i - x_i * L_i$$



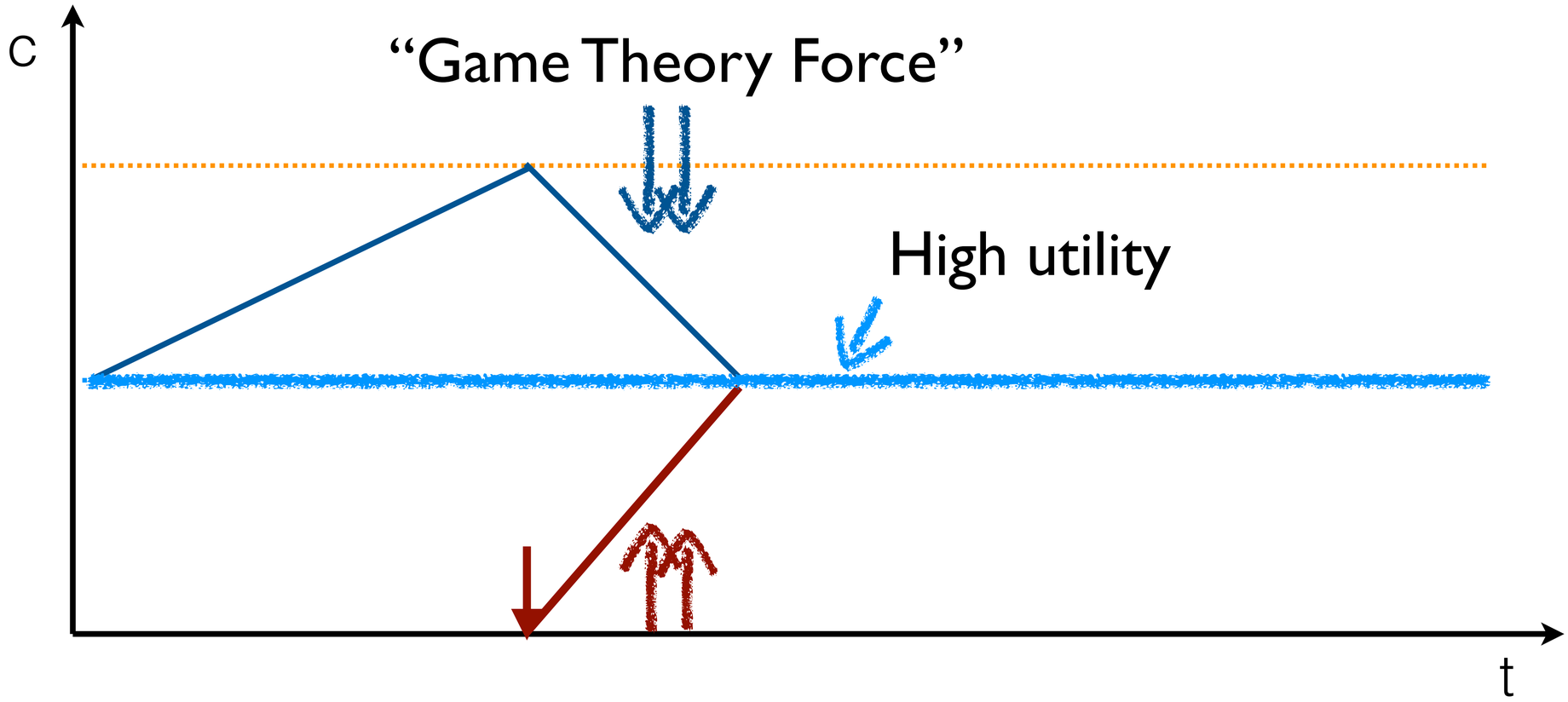
Dynamic behavior and fairness

TCP uses AIMD for asymptotic fairness

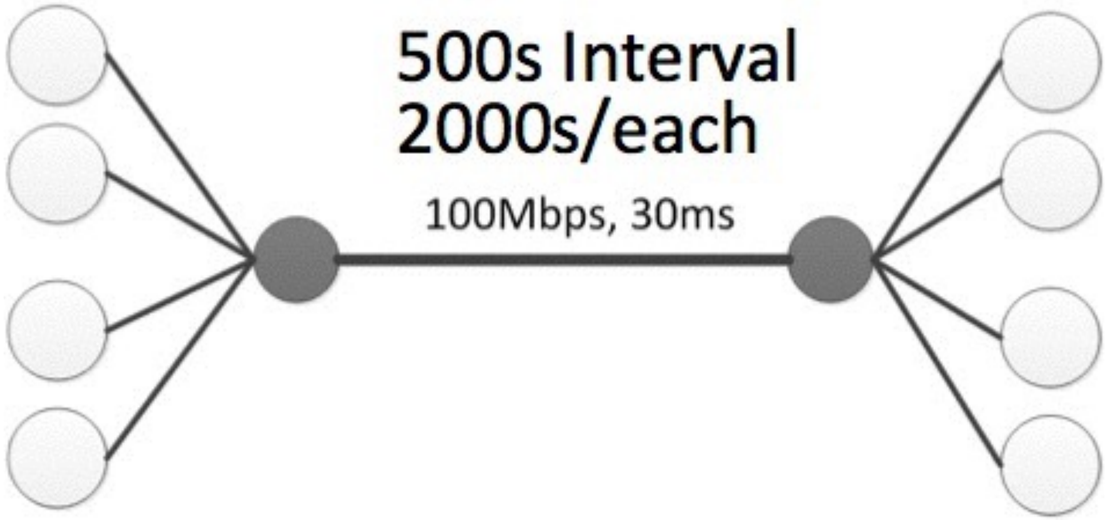


Dynamic behavior and fairness

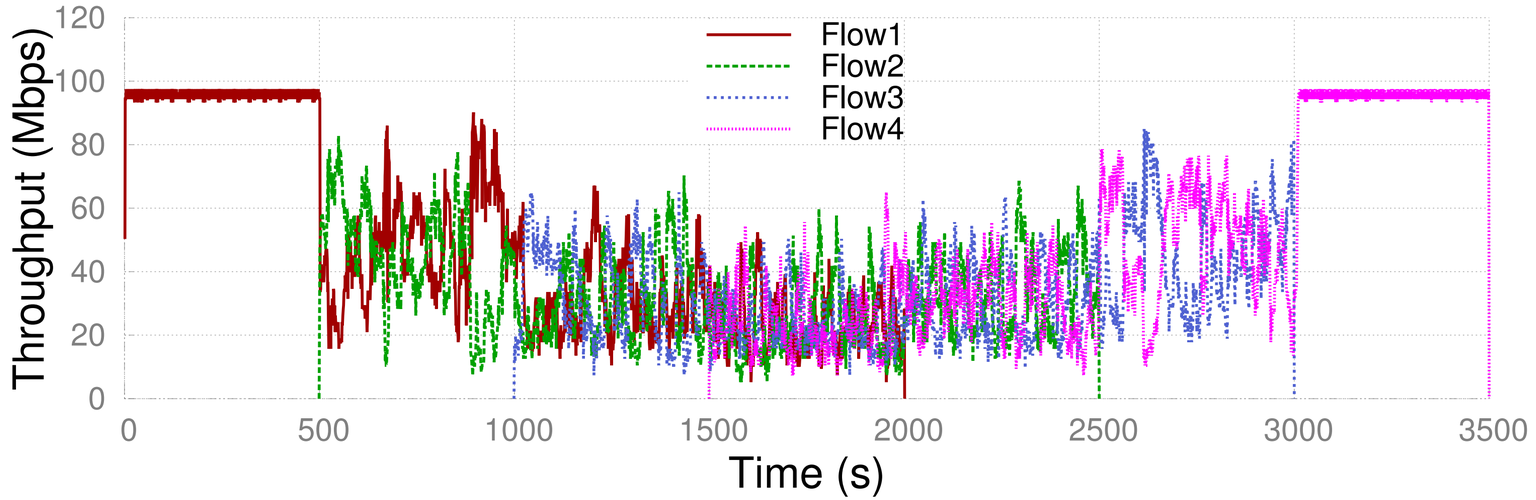
PCC does not need AIMD because it looks at real performance



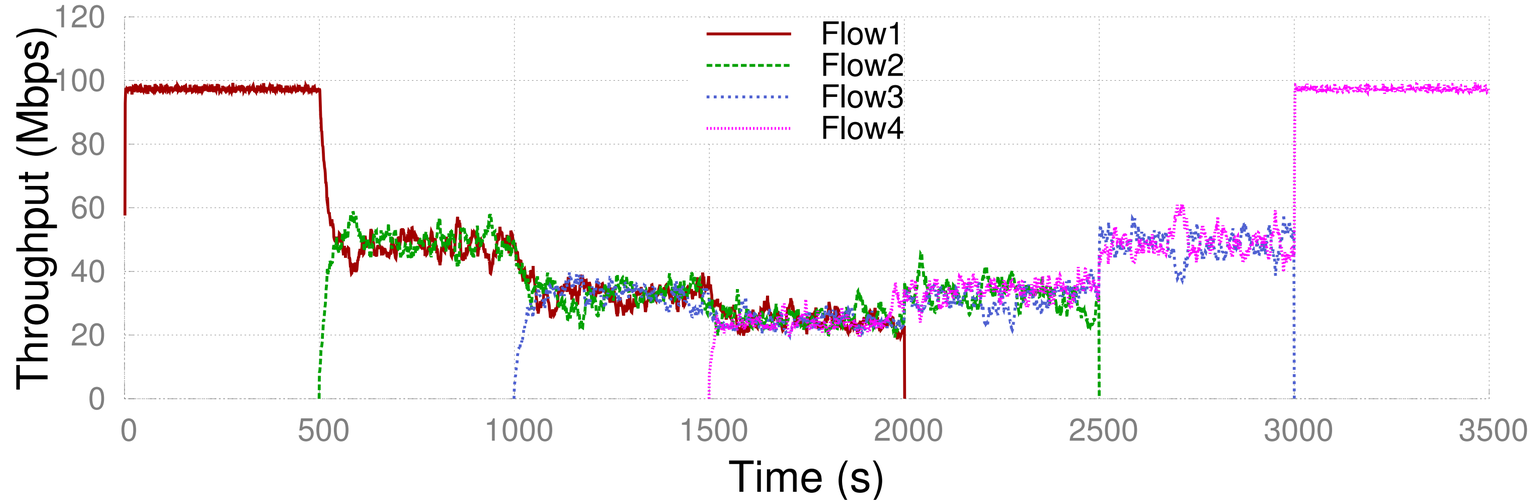
Dynamic behavior and fairness



Dynamic behavior and fairness



TCP



PCC

Deployment

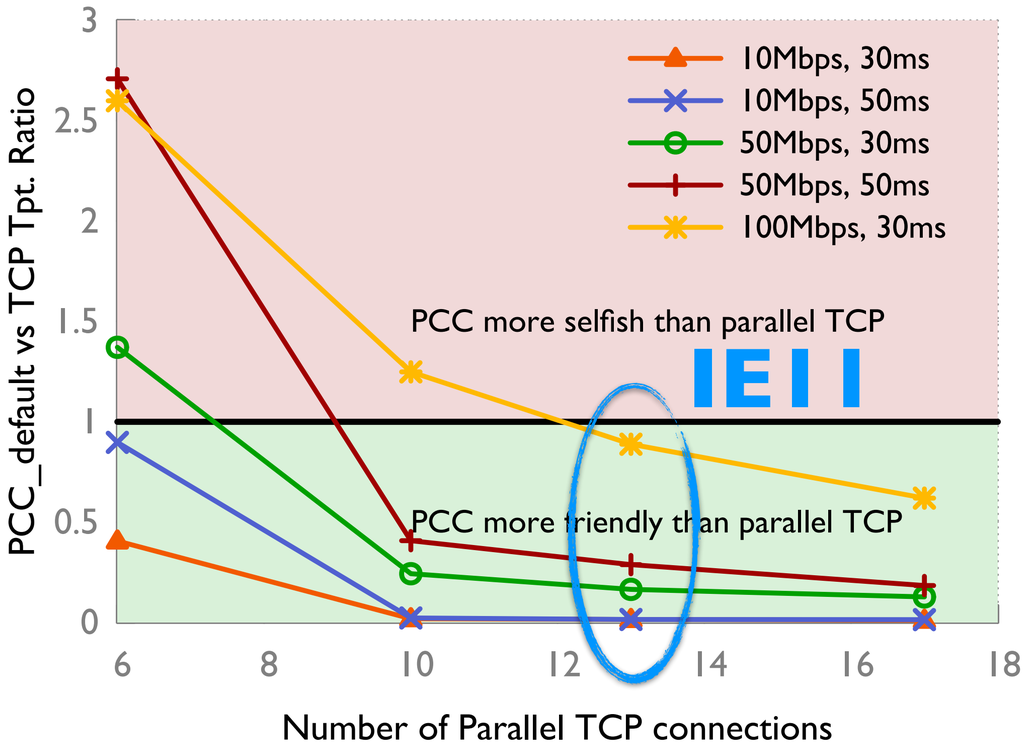
No hardwired support, packet header, protocol change needed

Where to deploy?

- CDN backbone, Inter-data center, dedicated scientific network
- In the wild?

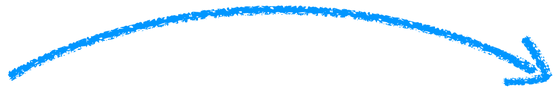
TCP friendliness

PCC's default utility function is not TCP friendly



TCP friendliness

Different utility functions could provide fairness



		30ms	60ms	90ms
$\beta = 10$	10Mbit/s	0.94	0.75	0.67
	50Mbit/s	0.74	0.73	0.81
	90Mbit/s	0.89	0.91	1.01
$\beta = 100$	10Mbit/s	0.71	0.58	0.63
	50Mbit/s	0.56	0.58	0.54
	90Mbit/s	0.63	0.62	0.88

Different utility functions

Same rate control algorithm

+

Different utility function

=

Flexibility

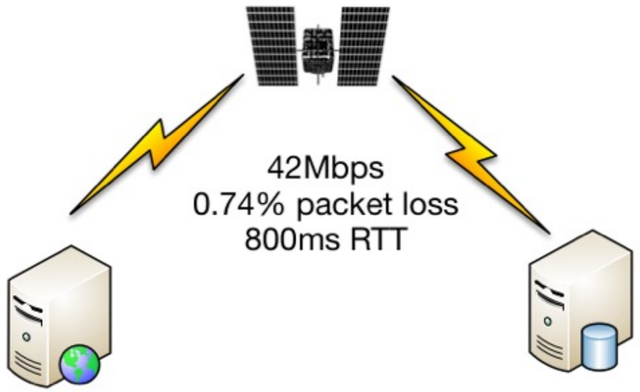
Performance: inter DC

Inter datacenter and dedicated high speed networks

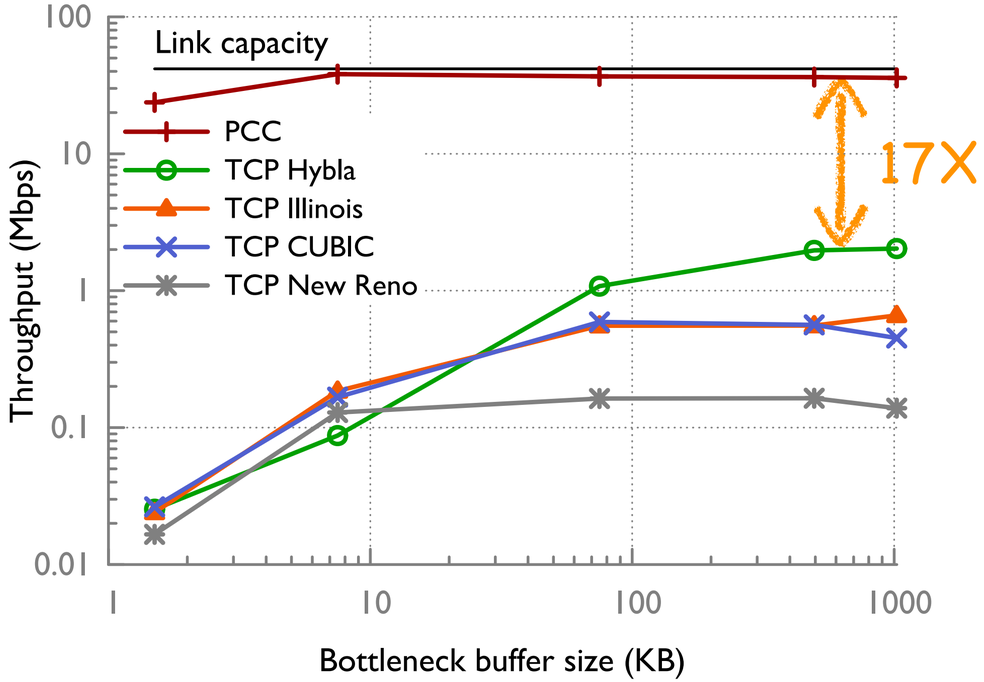
Transmission Pair	RTT	PCC	SABUL	CUBIC	Illinois
GPO → NYSERNet	12.1	818	563	129	326
GPO → Missouri	46.5	624	531	80.7	90.1
GPO → Illinois	35.4	766	664	84.5	102
NYSERNet → Missouri	47.4	816	662	108	109
Wisconsin → Illinois	9.01	801	700	547	562
GPO → Wisc.	38.0	783	487	79.3	120
NYSERNet → Wisc.	38.3	791	673	134	134
Missouri → Wisc.	20.9	807	698	259	262
NYSERNet → Illinois	36.1	808	674	141	141

123%

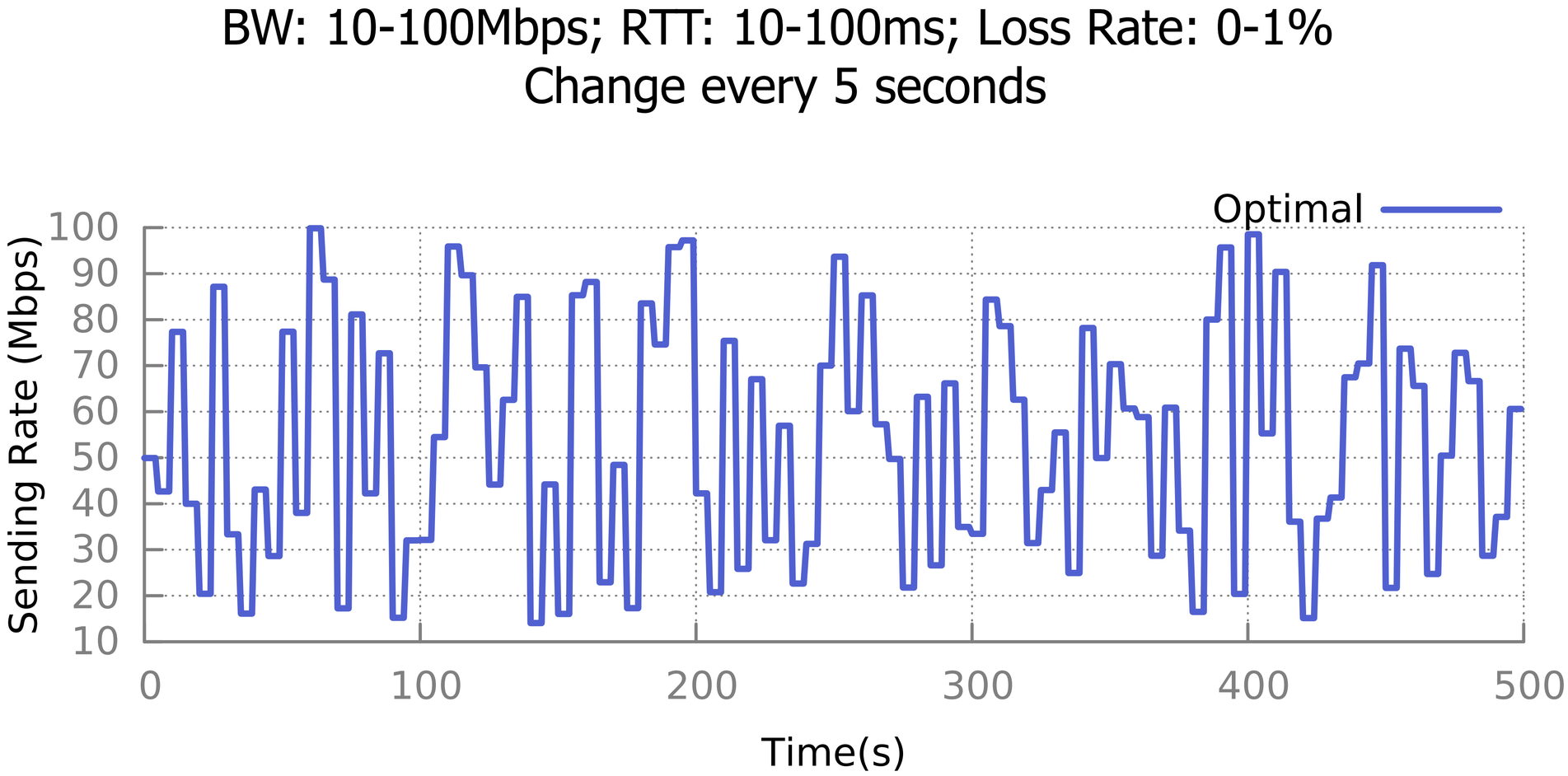
Performance: satellite network



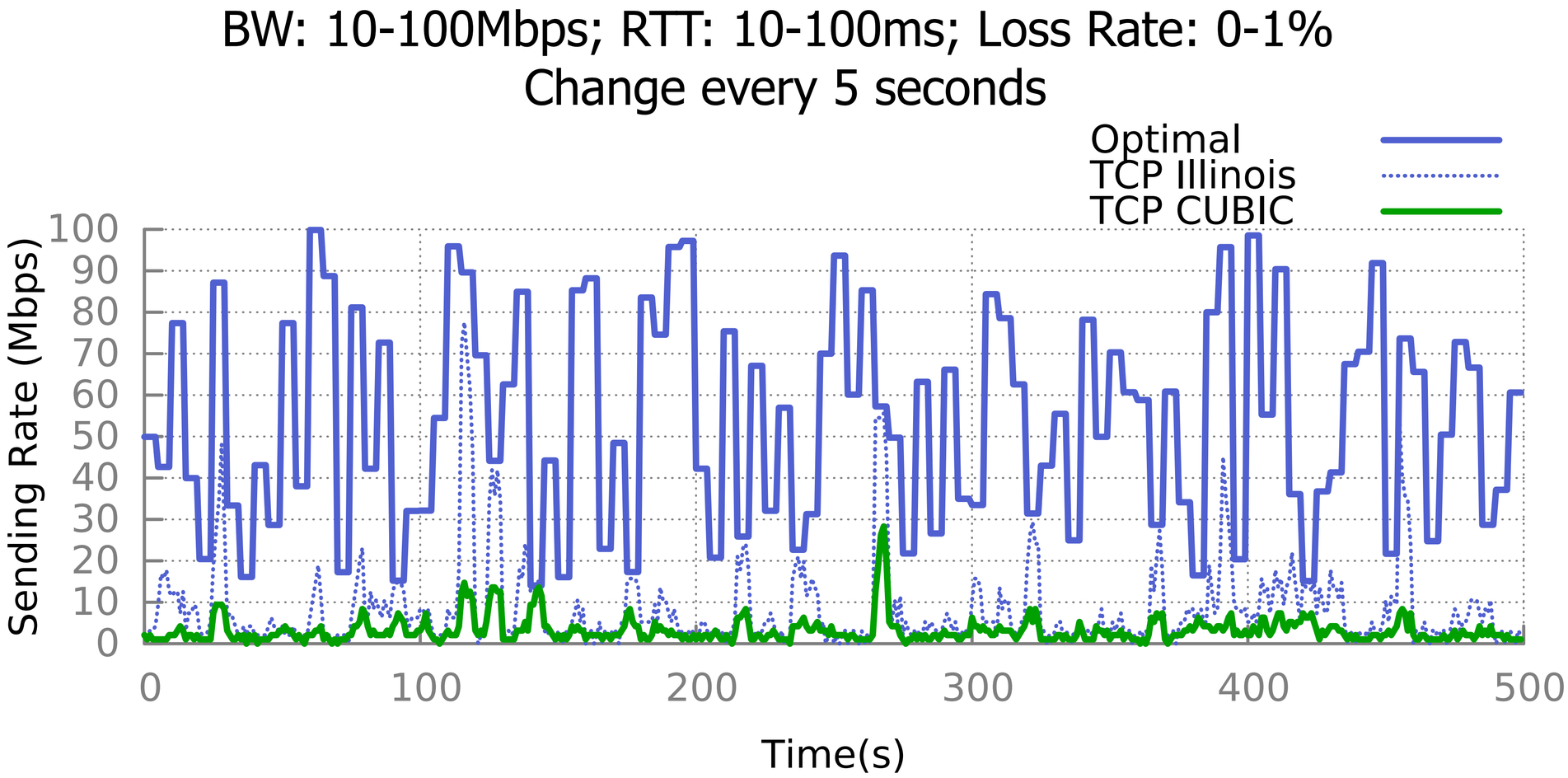
WINDS System



Performance: rapidly changing networks

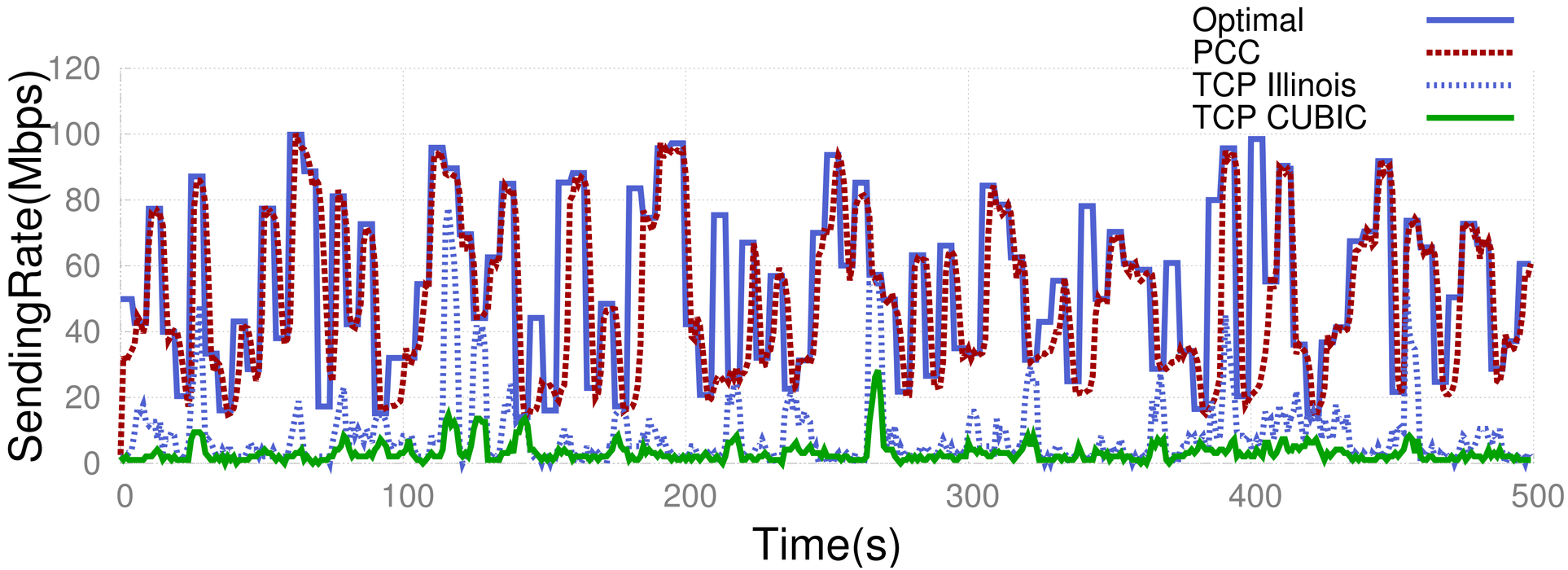


Performance: rapidly changing networks



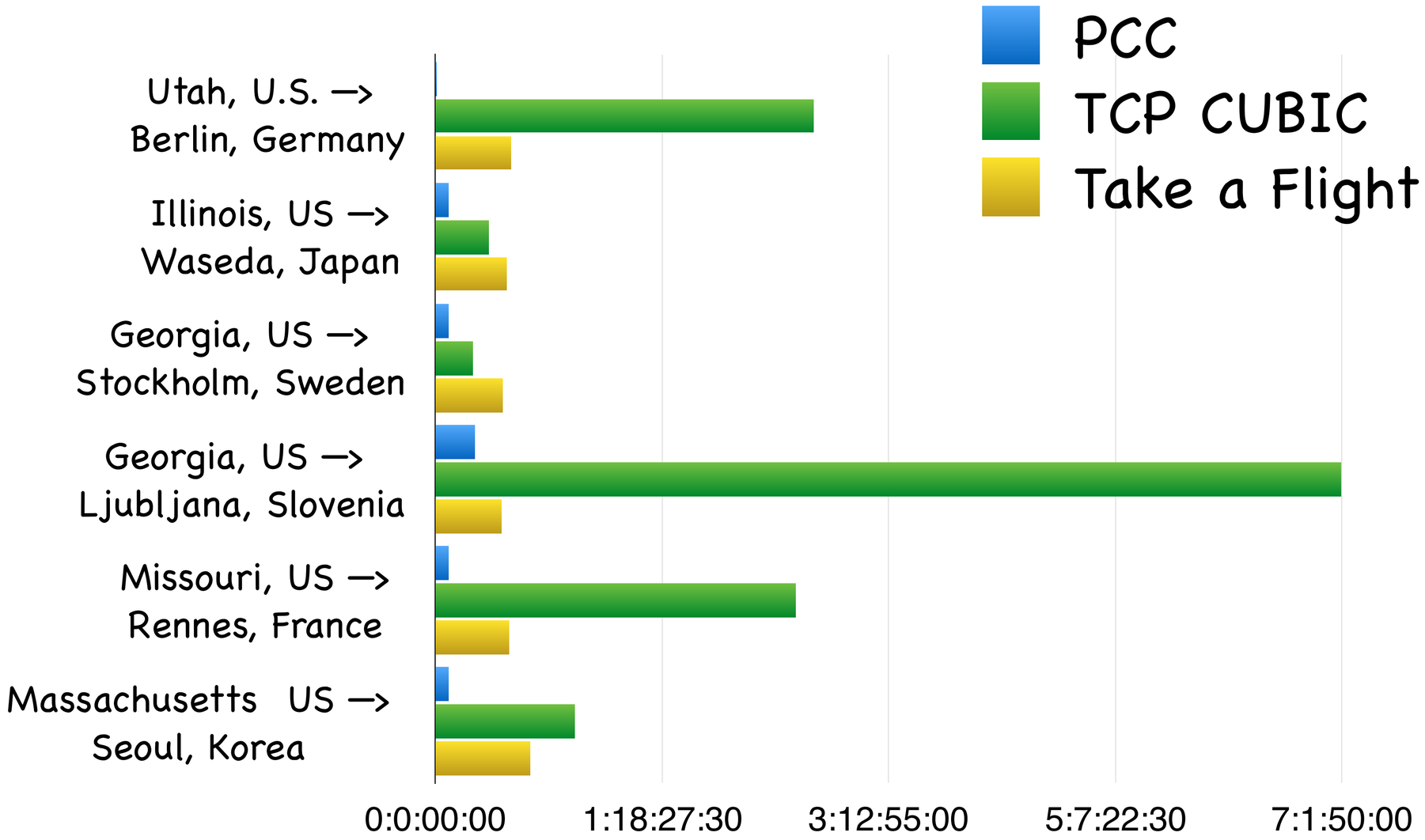
Performance: rapidly changing networks

BW: 10-100Mbps; RTT: 10-100ms; Loss Rate: 0-1%
Change every 5 seconds




<1.2X
1.2X-2X
2X-5X
5X-10X
10X - 100X
>100X

PCC vs TCP vs Take a flight: 100 GB



Long list of things in paper …

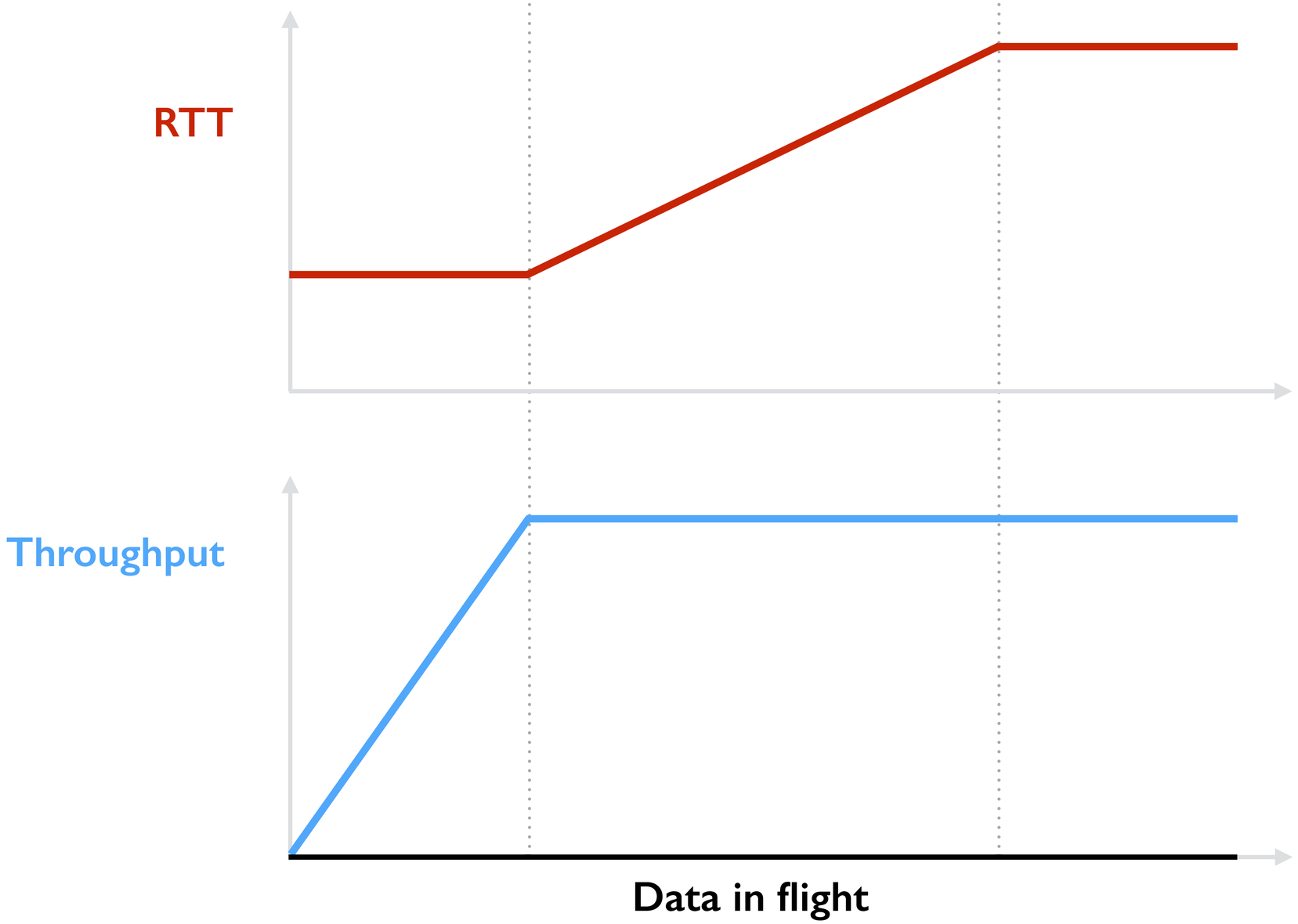
- More stories about the fact that TCP is broken
- Proof of Nash Equilibrium and Convergence
- Concrete Implementation of PCC
  - Performance monitoring
  - Details of learning control algorithm
  - Implementation designs and optimizations
- Performance Evaluation
  - Inter data center networks
  - small buffer networks
  - Reactiveness and stability tradeoff
  - Jain index fairness
  - Benefit of Randomized Control Trials
  - Details of TCP friendliness evaluation
  - Emulated satellite networks
  - Emulated datacenter networks
  - Cure RTT unfairness
  - Does not fundamentally harm short flow FCT
  - Evaluation in the wild vs non-TCP protocols
- Flexibility by pluggable utility function

11 more!

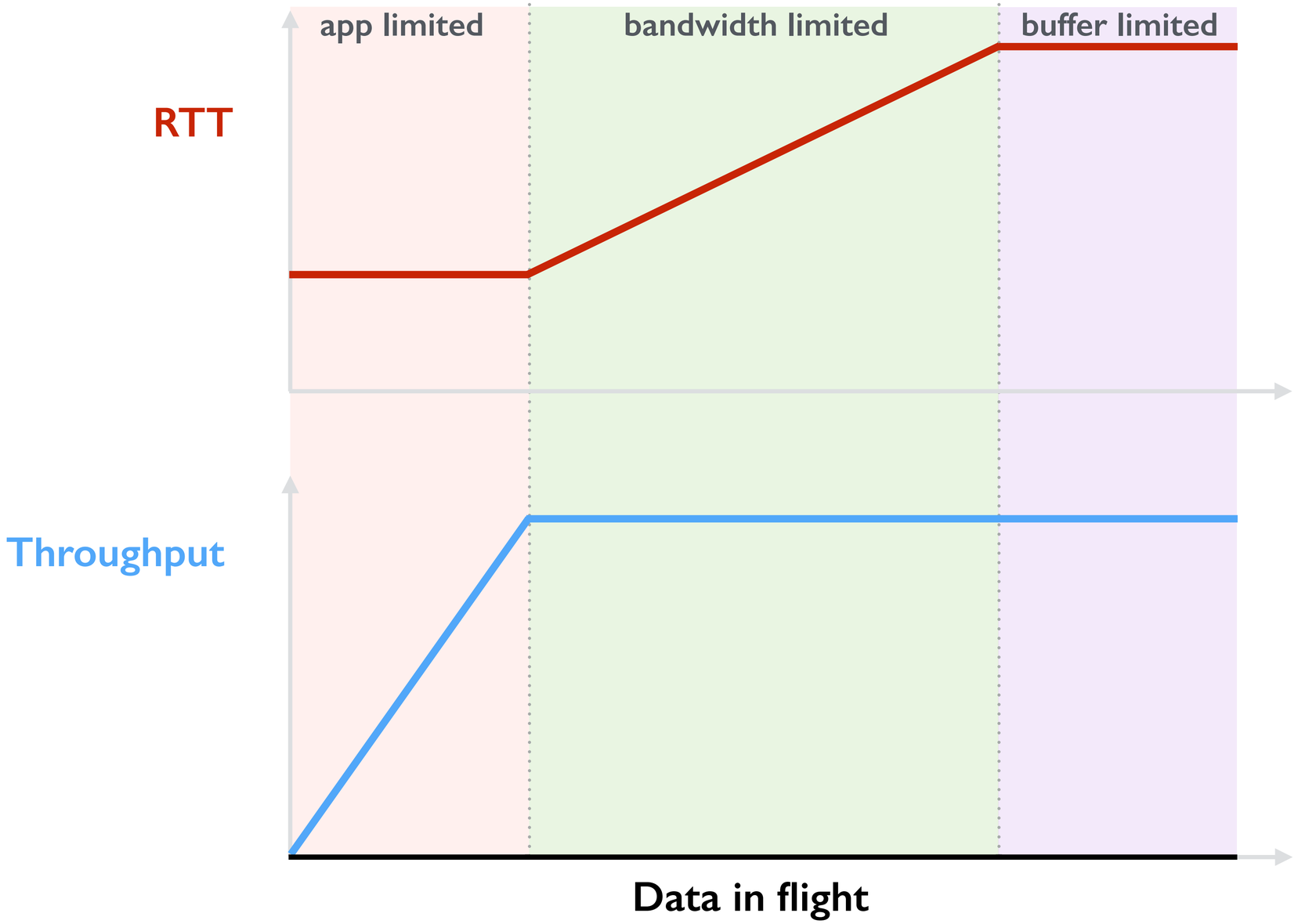
BBR: “congestion-based congestion control”

How do we operate at (nearly) zero congestion?

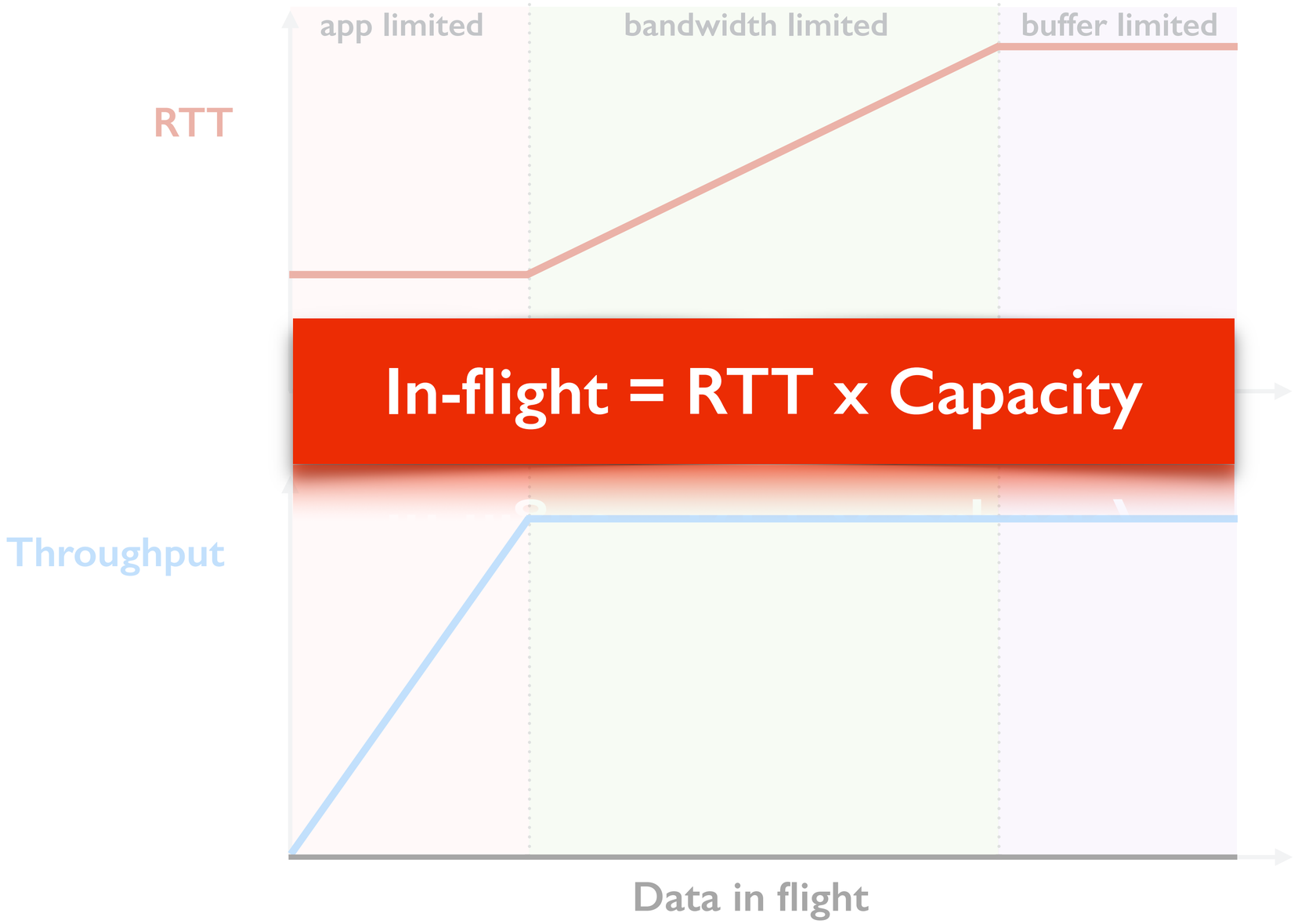
The target operational point



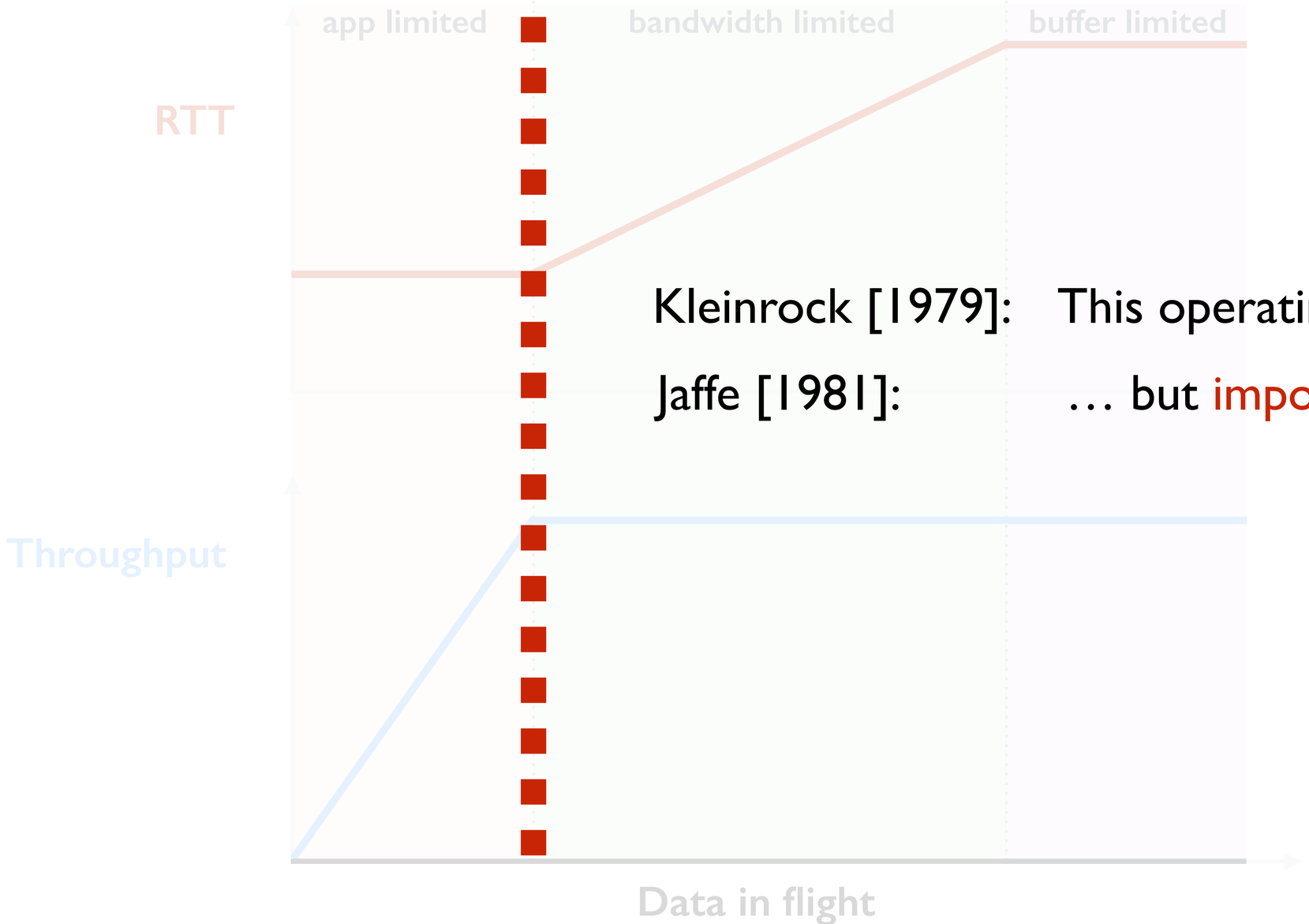
The target operational point



The target operational point



The target operational point



Kleinrock [1979]: This operating point is optimal!

Jaffe [1981]: … but impossible for distributed algorithms

How do we estimate RTT and Capacity?

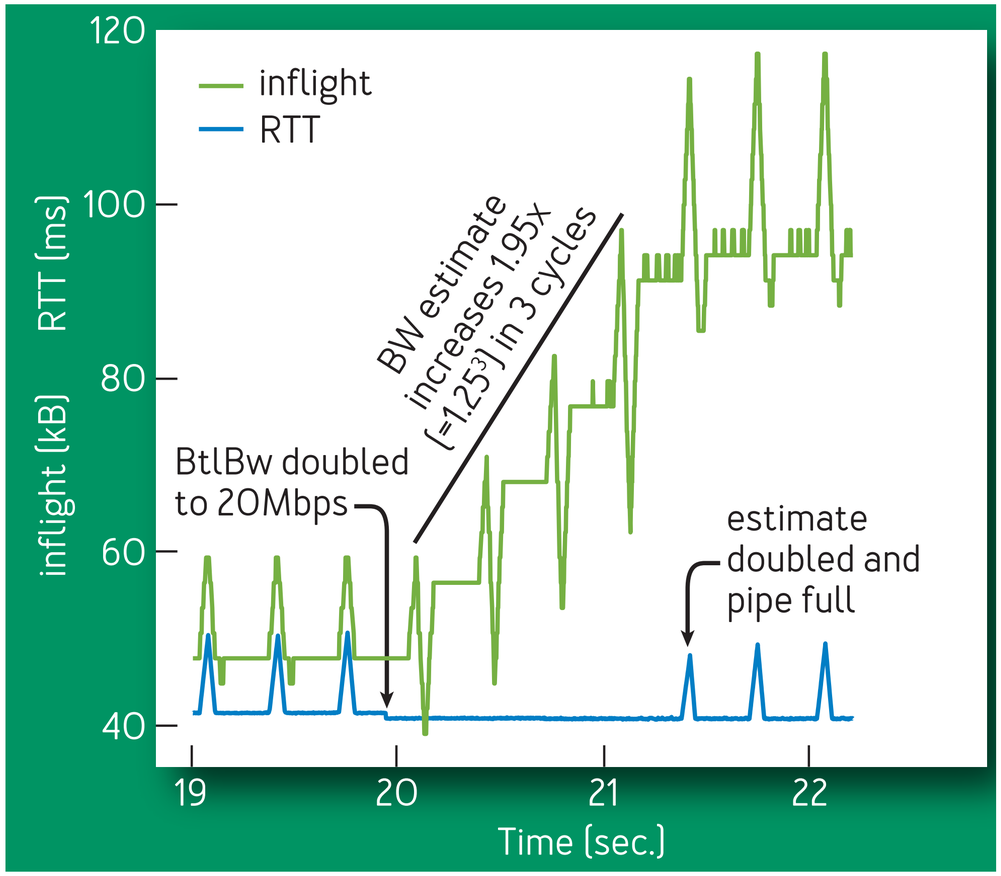
BBR: two target conditions

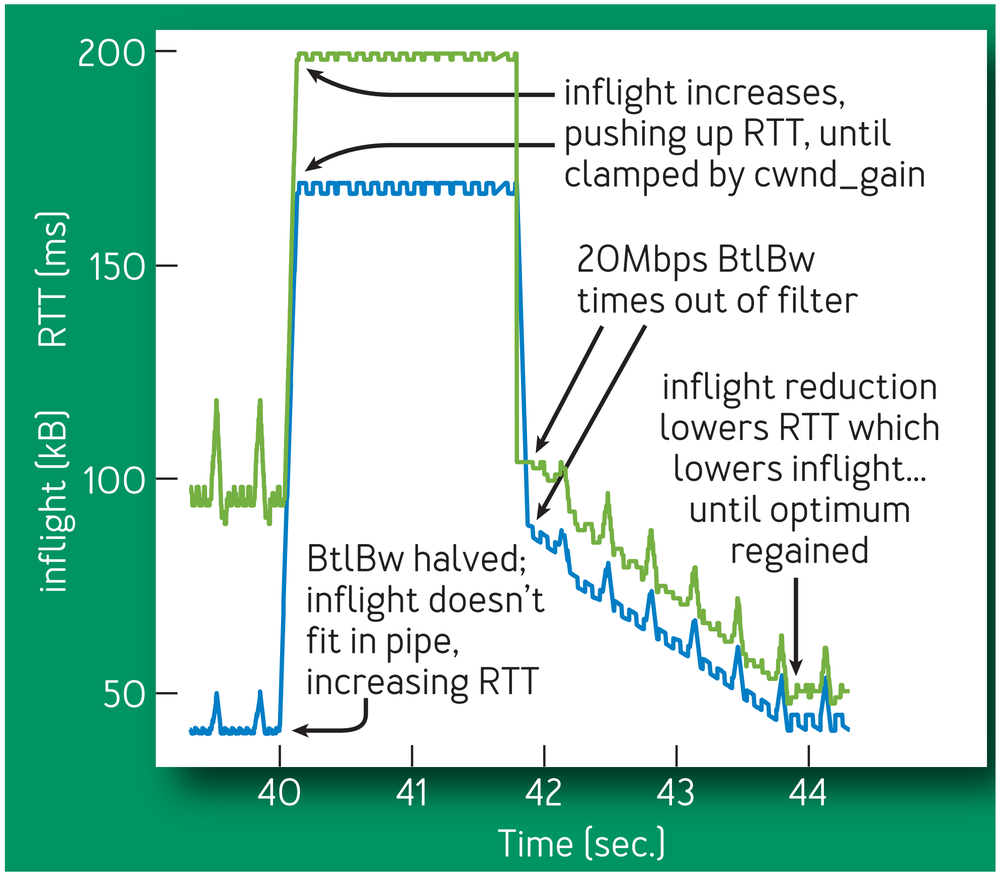
- Full pipe
  - `in-flight = min-RTT * bottleneck-bw`
- Rate balance
  - `buffer fill rate = emptying rate`

BBR: how it works

- `BtlBw`
  - Probe around [1, 1.25, 0.75] of current
- `RTProp`
  - (occasionally) drop rate to see an empty queue
- Needs pacing of packets!

BBR: how it works



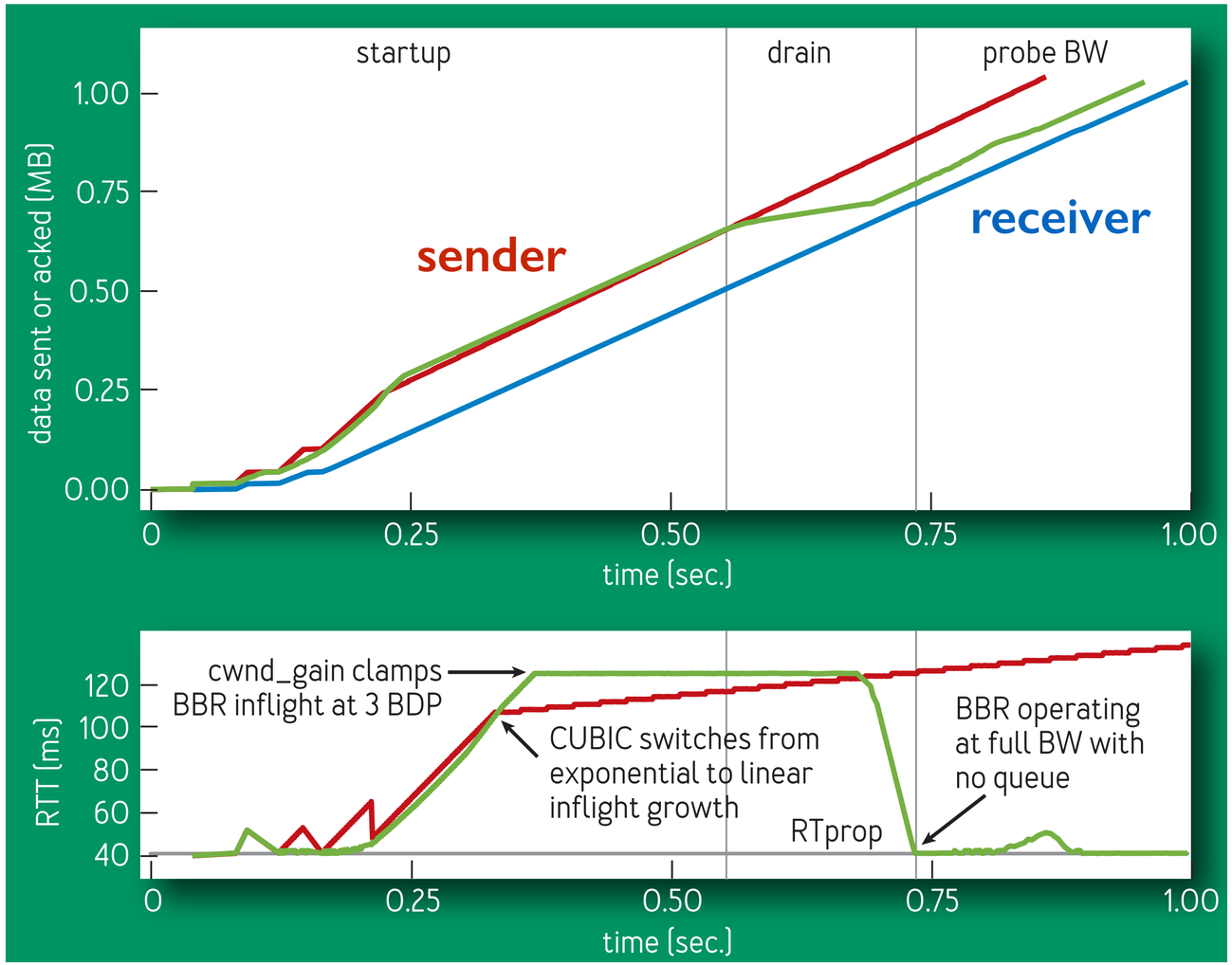


BBR Congestion-Based Congestion Control, ACM Queue '16
Neal Cardwell, Yuchung Cheng, C. Stephen Gunn, Soheil Hassas Yeganeh, Van Jacobson

BBR: what about the startup phase?

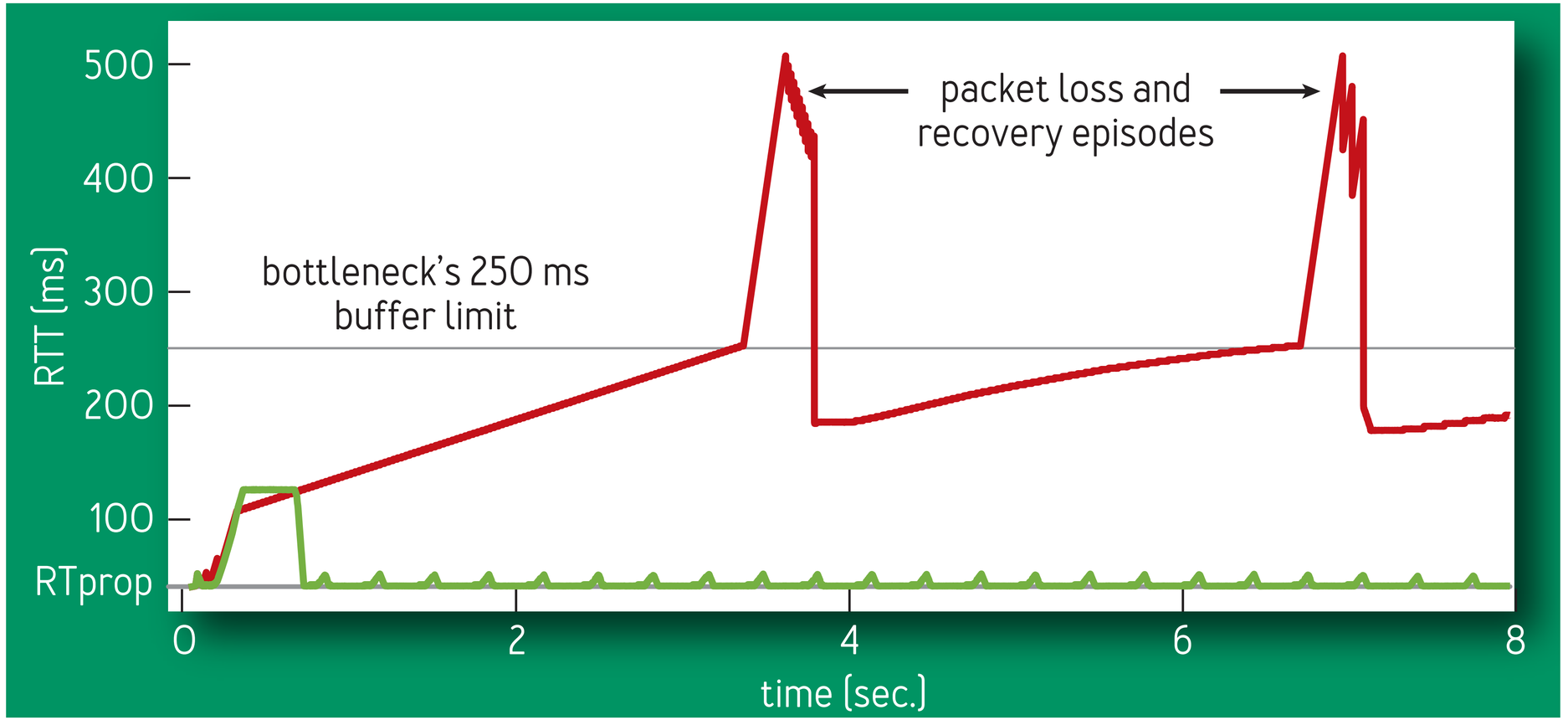
- `BtlBw`
  - like slow-start until the estimate plateaus (not just until loss)
- Drain queue created in startup

BBR: how it works



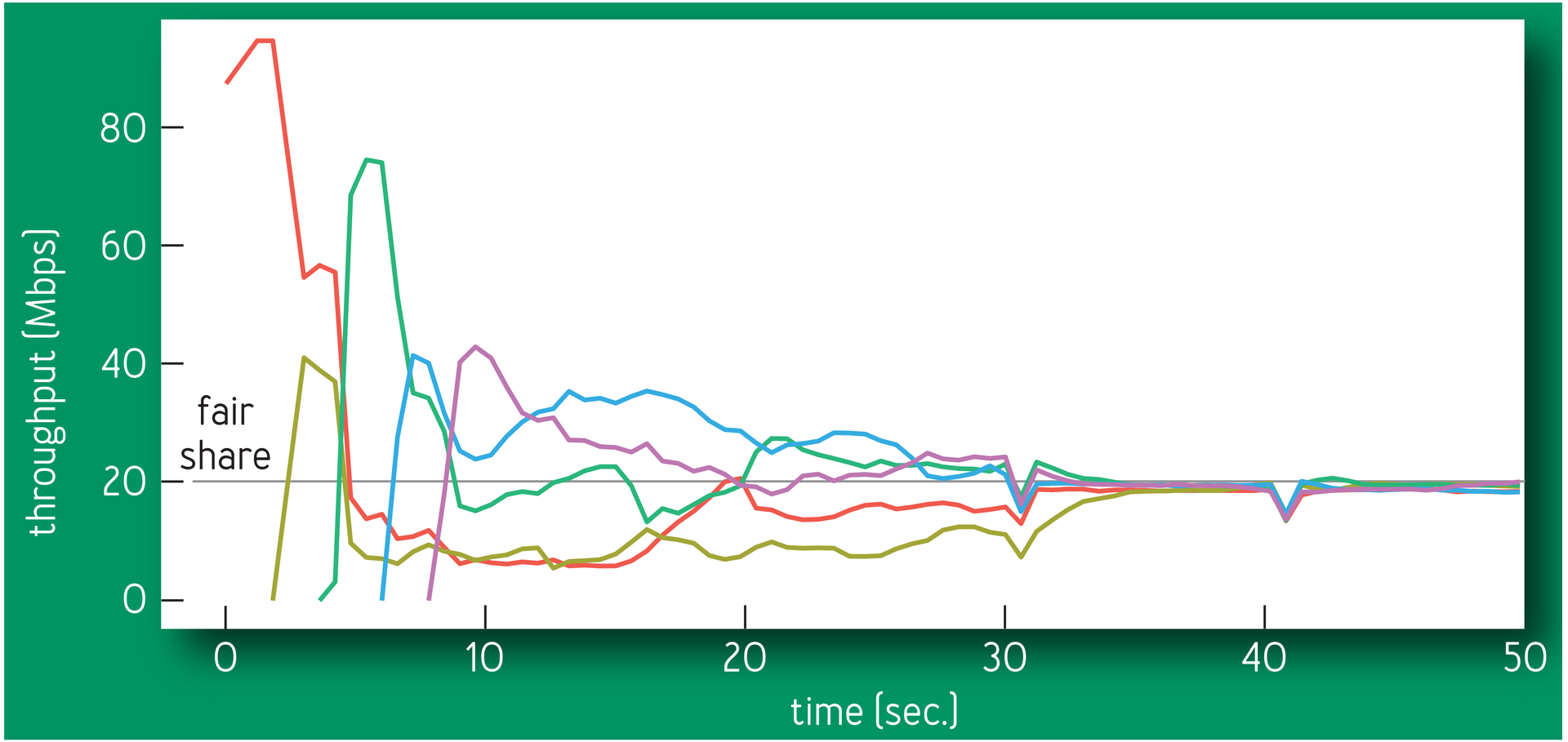
BBR Congestion-Based Congestion Control, ACM Queue '16
Neal Cardwell, Yuchung Cheng, C. Stephen Gunn, Soheil Hassas Yeganeh, Van Jacobson

BBR: how it works



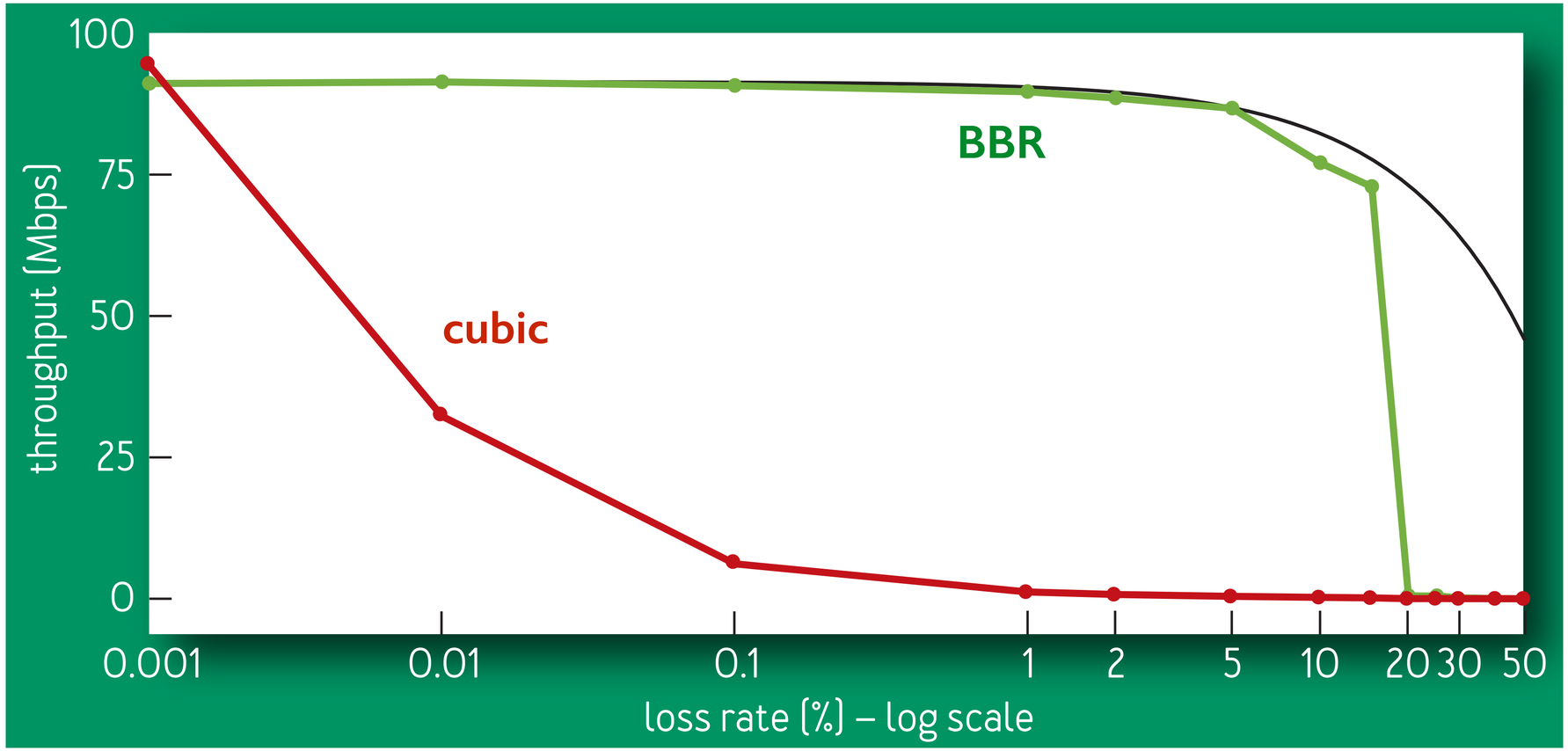
BBR Congestion-Based Congestion Control, ACM Queue '16
Neal Cardwell, Yuchung Cheng, C. Stephen Gunn, Soheil Hassas Yeganeh, Van Jacobson

BBR: with many flows



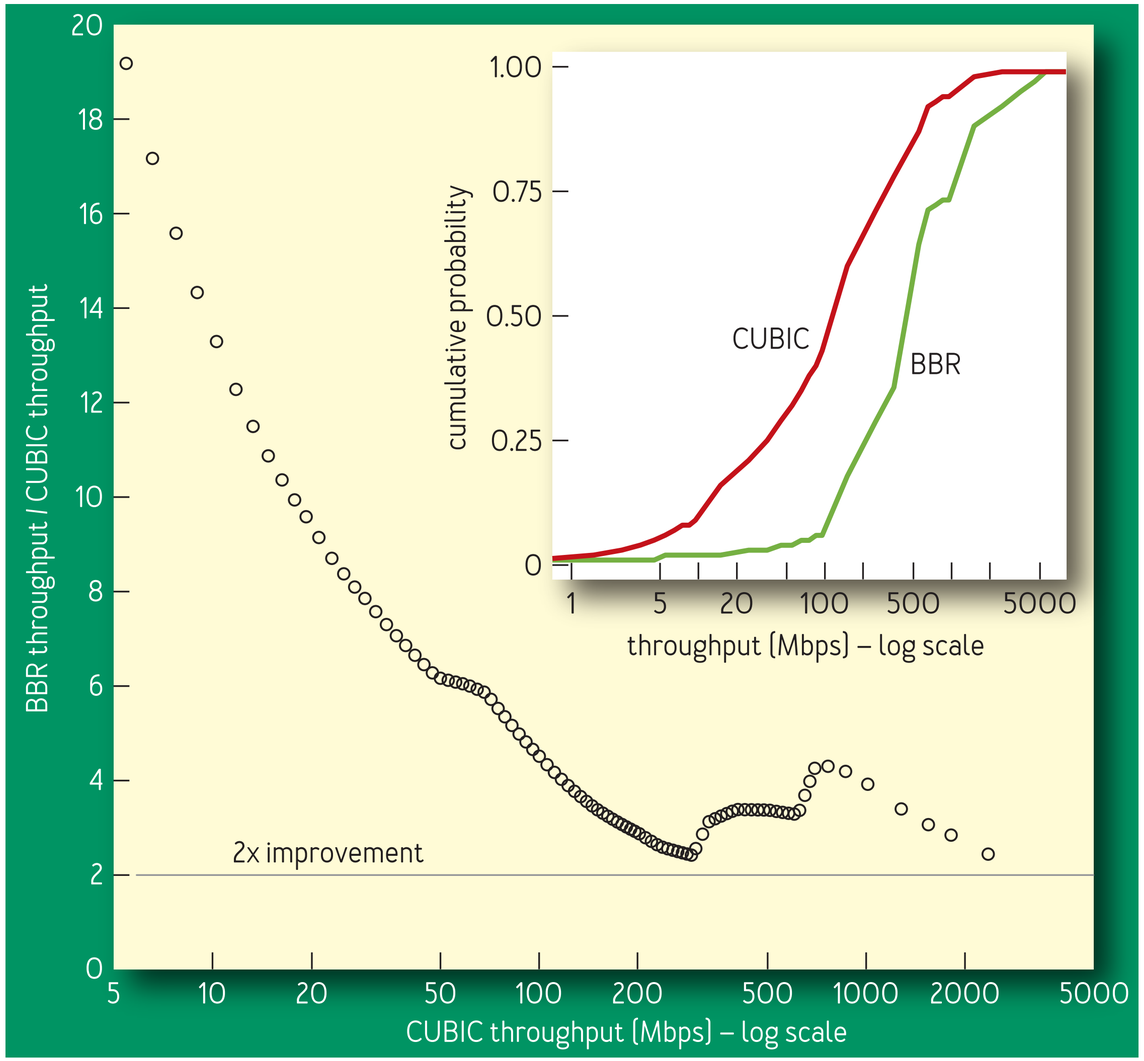
BBR Congestion-Based Congestion Control, ACM Queue '16
Neal Cardwell, Yuchung Cheng, C. Stephen Gunn, Soheil Hassas Yeganeh, Van Jacobson

BBR: under loss



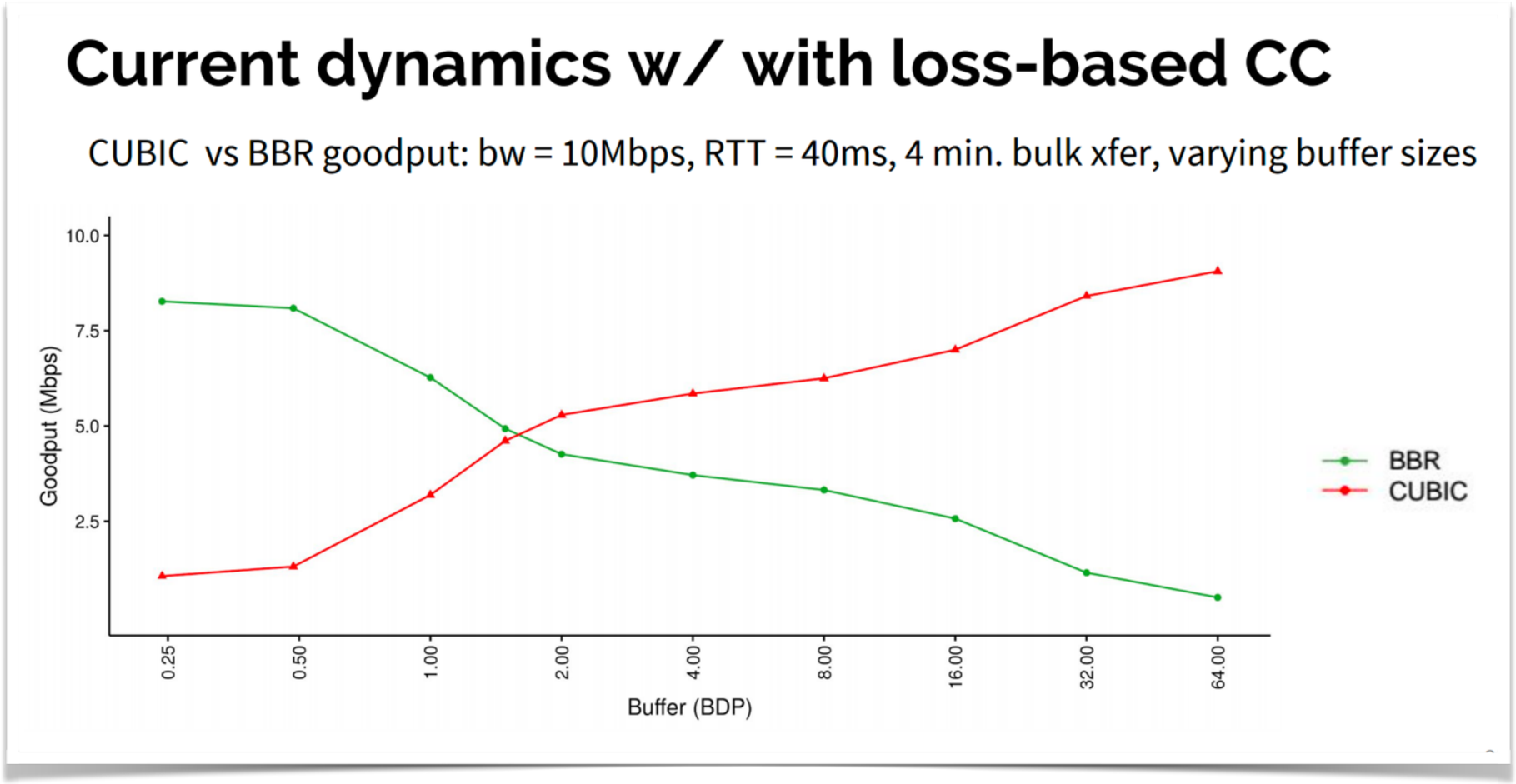
BBR Congestion-Based Congestion Control, ACM Queue '16
Neal Cardwell, Yuchung Cheng, C. Stephen Gunn, Soheil Hassas Yeganeh, Van Jacobson

BBR: in actual deployment (probes)



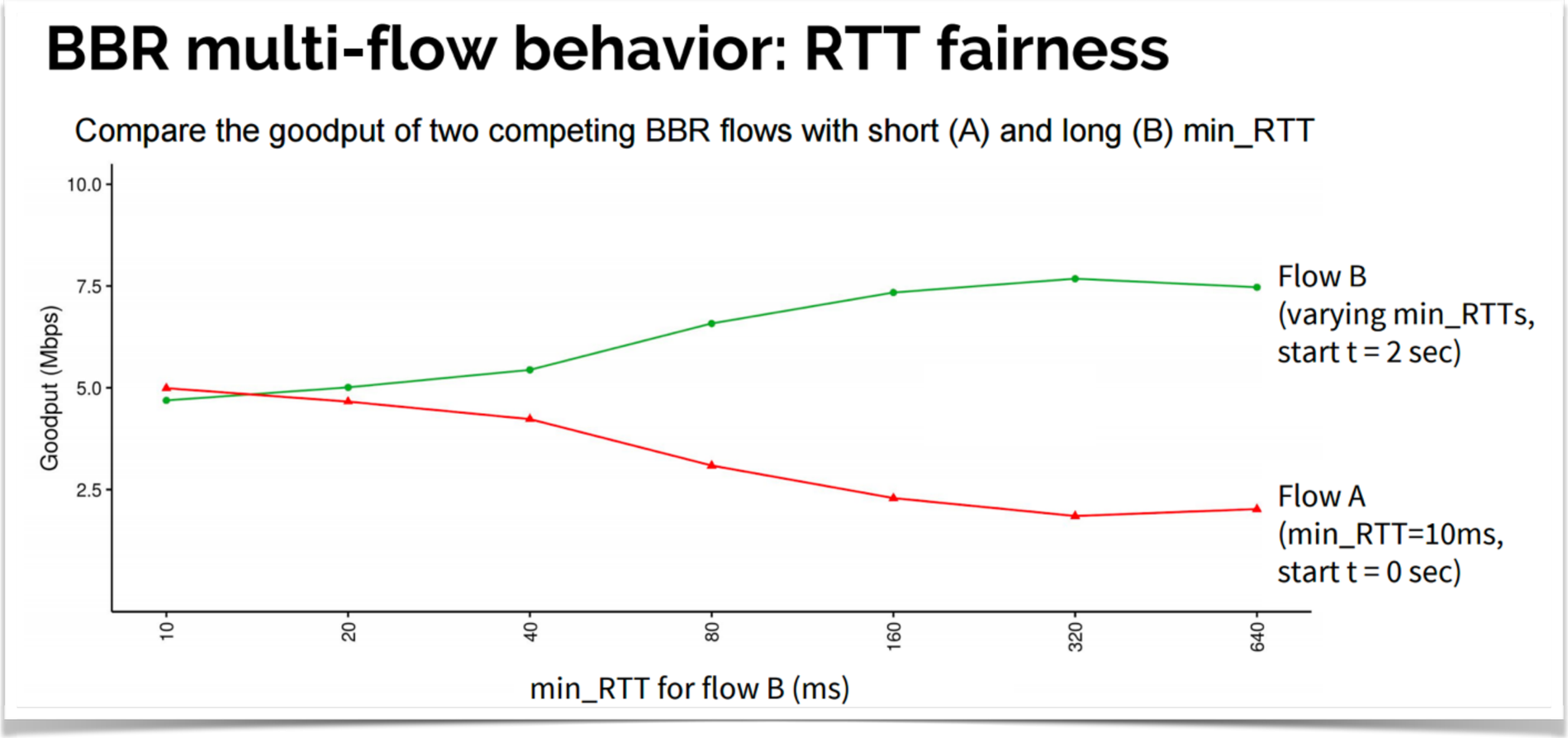
BBR Congestion-Based Congestion Control, ACM Queue '16
Neal Cardwell, Yuchung Cheng, C. Stephen Gunn, Soheil Hassas Yeganeh, Van Jacobson

BBR: still not devoid of problems!



BBR Congestion Control, IETF 97 (2016)
Talk by Neal Cardwell

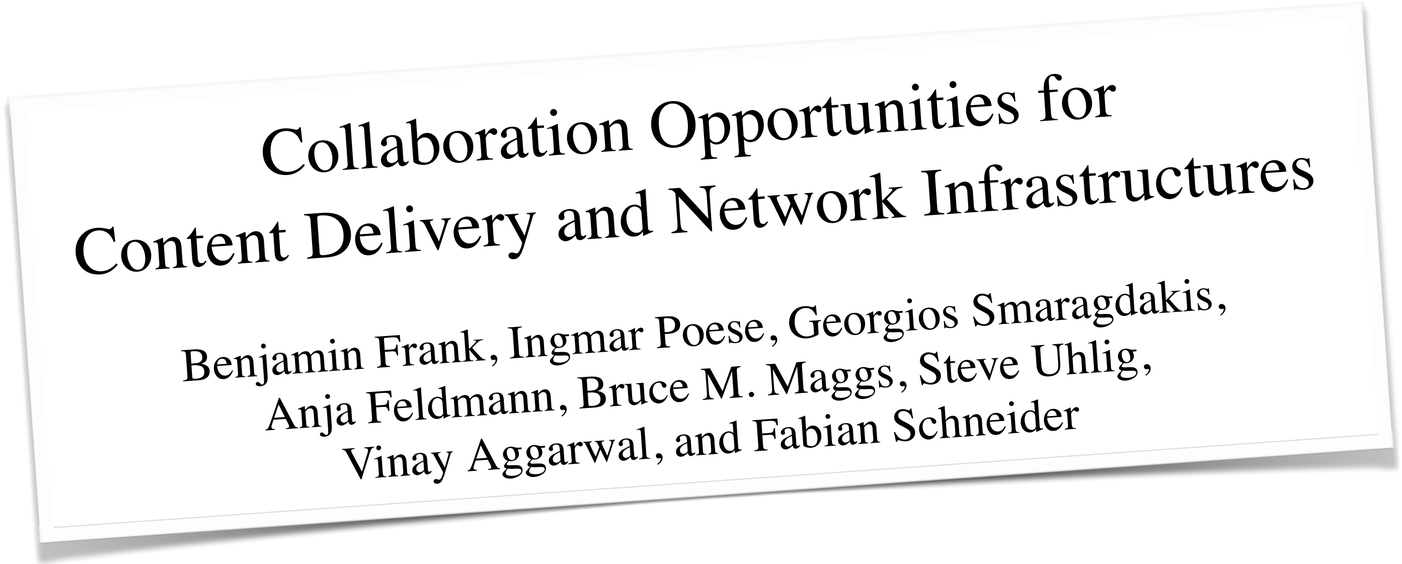
BBR: still not devoid of problems!



BBR Congestion Control, IETF 97 (2016)
Talk by Neal Cardwell

Weekly reading guide

Content distribution networks


Collaboration Opportunities for
Content Delivery and Network Infrastructures

Benjamin Frank, Ingmar Poese, Georgios Smaragdakis,
Anja Feldmann, Bruce M. Maggs, Steve Uhlig,
Vinay Aggarwal, and Fabian Schneider

- (Only pages 6-25)
- Any time you're browsing, it's likely you're talking to a CDN

Course evaluations

Please fill in the survey!