PCC: Re-architecting Congestion Control for Consistent High Performance

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

NSDI 2015

Outline

- **Background**
- Design
- Implementation & Evaluation
- Review

Some of the slides below are taken from the authors' original slides, which are publically available on USENIX website


TCP
Est. 1988

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

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

...

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

...

Who can be not happy with TCP ?


are
can be not happy with TCP ?

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

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

Point Solutions

High BDP	Wireless	Satellite	Inter-DC	Intra-DC
BIC H-TCP Compound CUBIC FAST TCP	Westwood Vegas Veno	Hybla STAR	Illinois SABUL	ICTCP DCTCP
10X	10X	17X	4X	

Unstable, RTT Unfair, Bufferbloat, Crash on Changing Networks, ...

Point Solutions

+

Performance

Far from Optimal


TCP fails to achieve
consistent high performance

Why is it so hard?

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

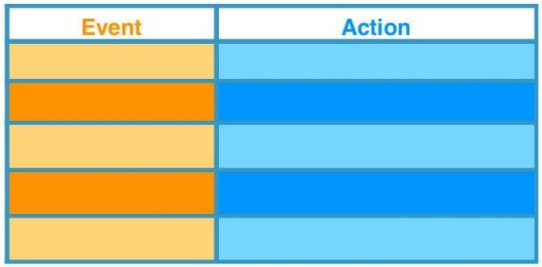
	Event	Action
Reno	1 pkt loss	cwnd/2
Scalable	ACK	cwnd+1
CUBIC	Time pass 1ms	cwnd+f(t,cwn,rtt)
FAST	RTT increase x%	Reduce cwnd to f(x)%
HTCP	100 ACK	cwnd+f(cwnd)/cwnd

Hardwired Mapping

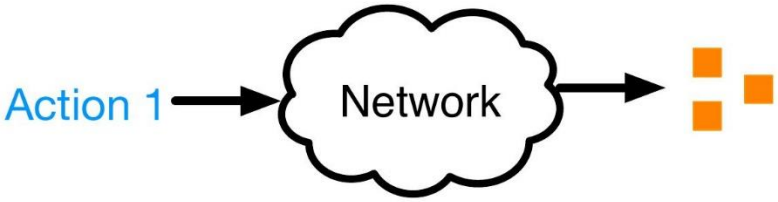
	Event	Action
Reno	1 pkt loss	cwnd/2
Scalable	ACK	cwnd+1
CUBIC	Time pass 1ms	cwnd+f(t,cwn,rtt)
FAST	RTT increase x%	Reduce cwnd to f(x)%
HTCP	100 ACK	cwnd+f(cwnd)/cwnd

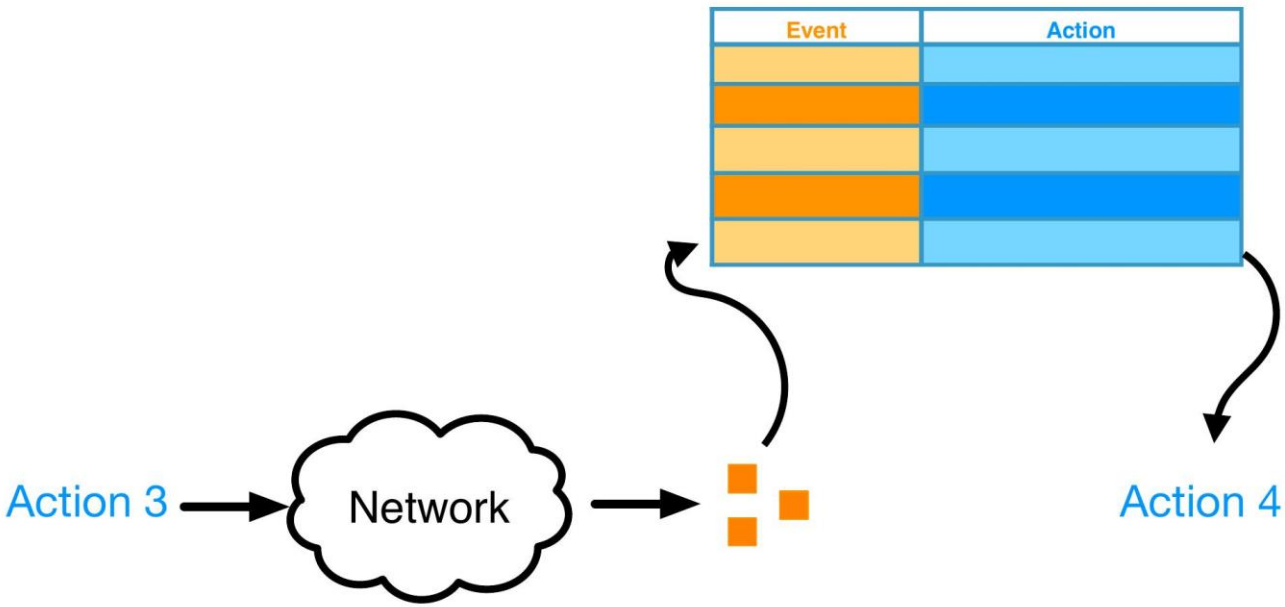
Hardwired Mapping

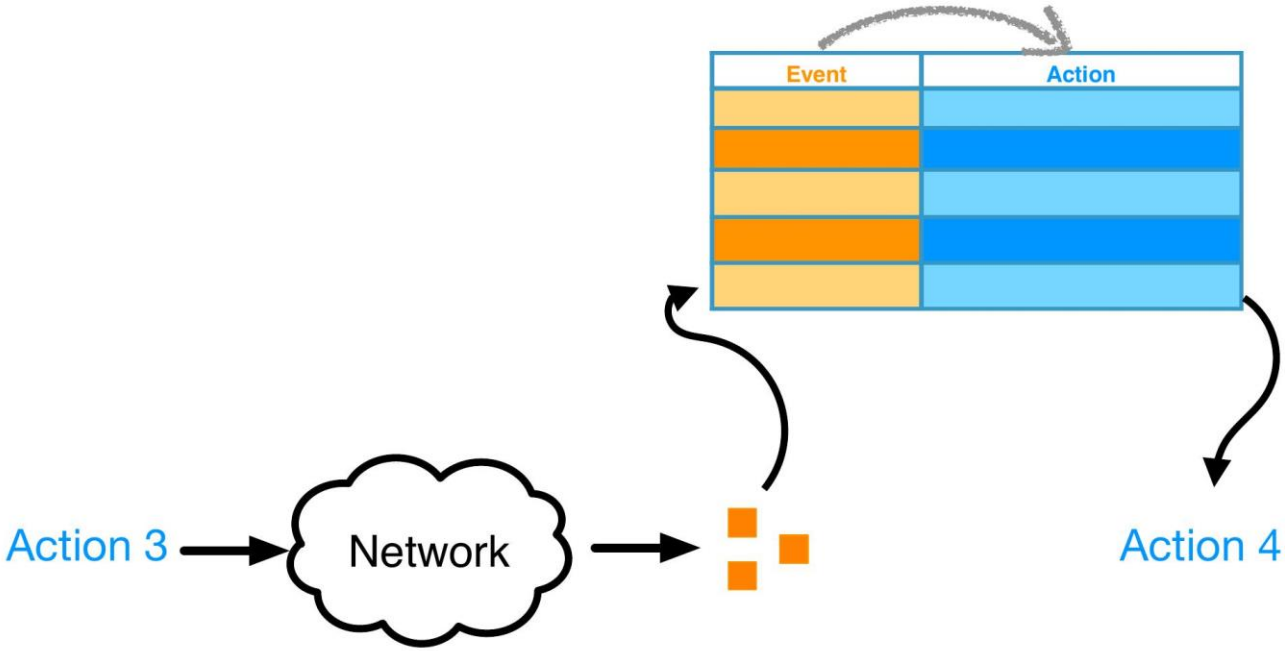
Event	Action

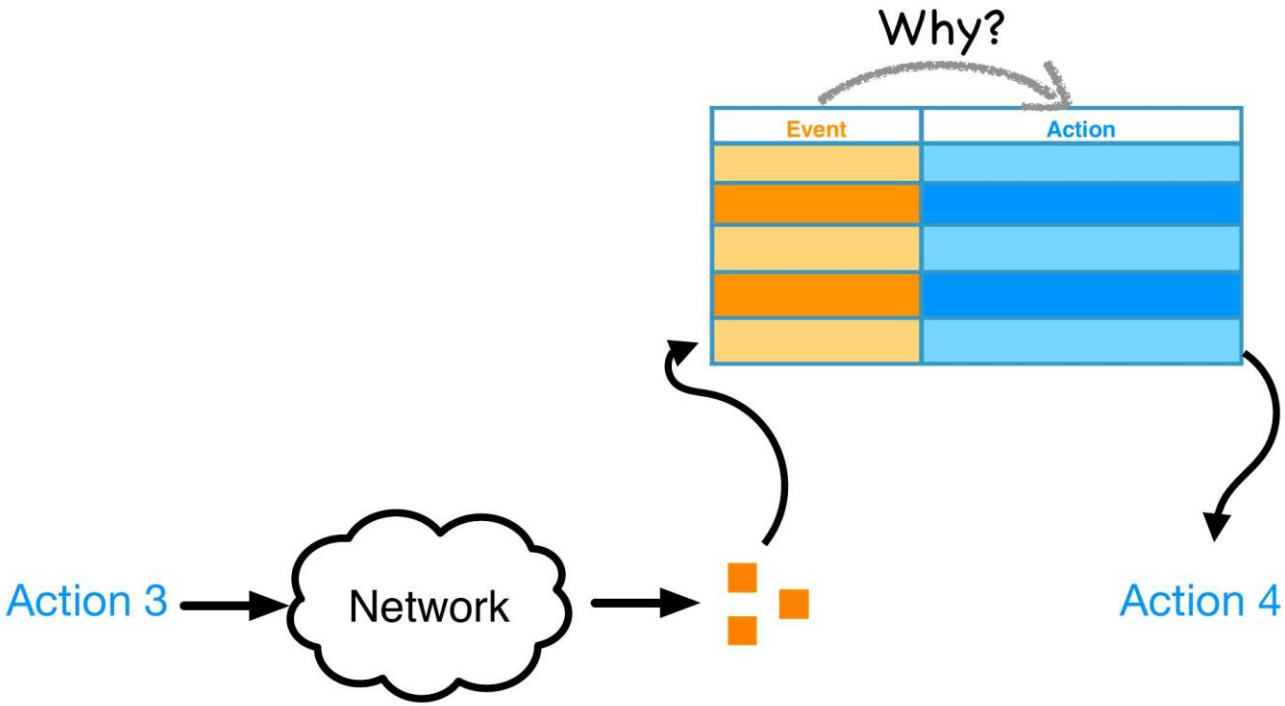

Event
Action

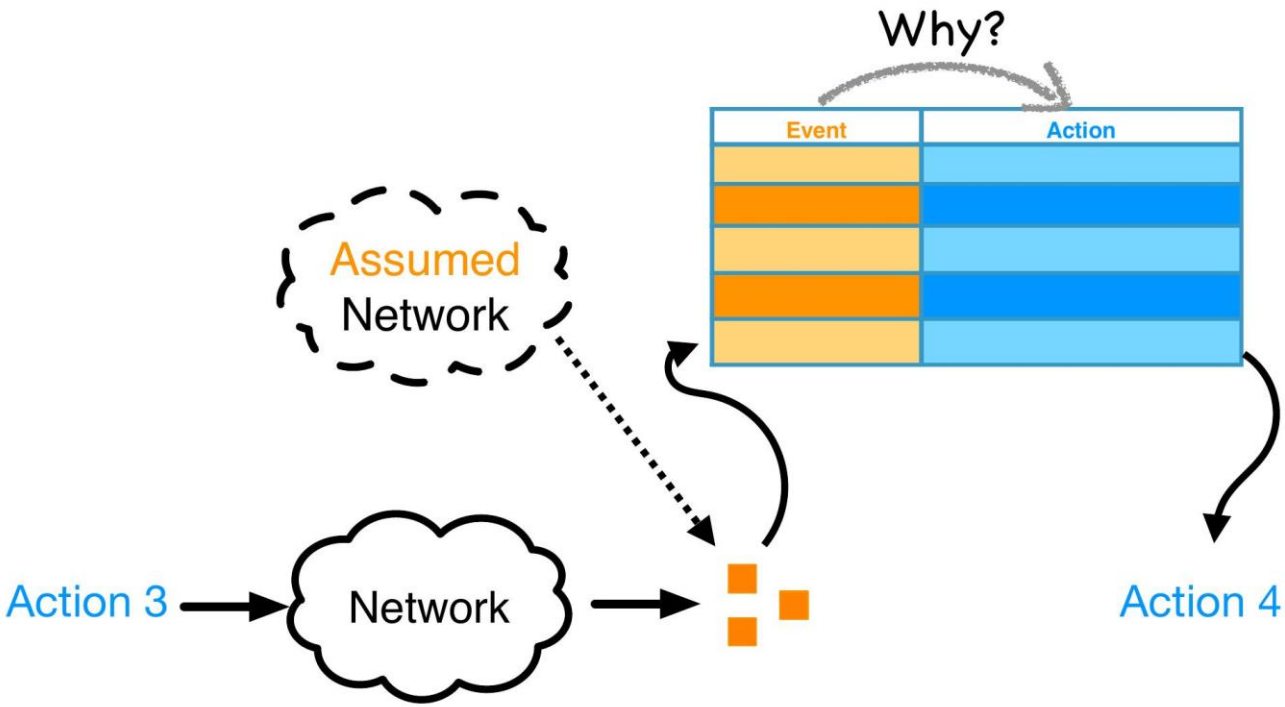
Event	Action

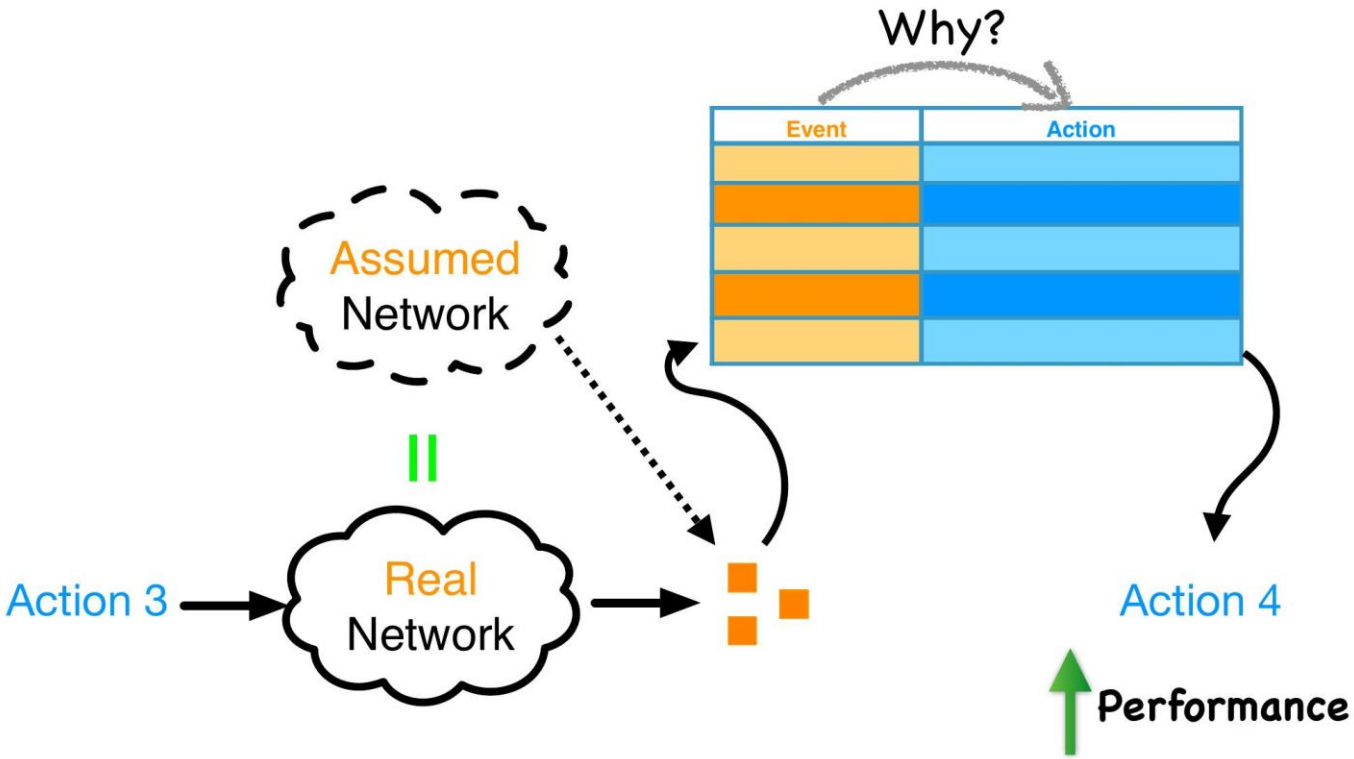


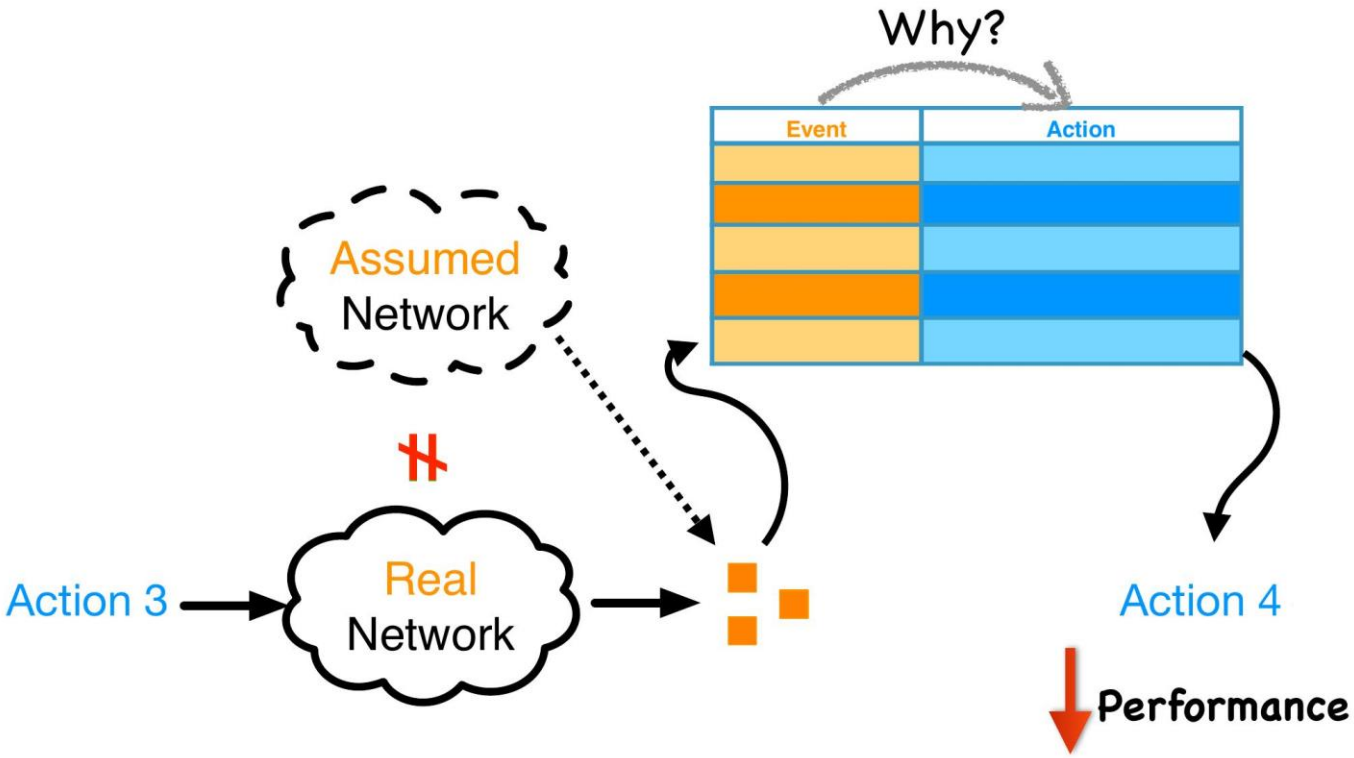

Event
Action
Action 3
Network
Action 4


Event
Action
Action 3
Network
Action 4


Why?
Event
Action
Action 3
Network
Action 4


Why?
Event
Action
Assumed
Network
Action 3
Network
Action 4


Why?
Event
Action
Assumed
Network
Real
Network
Action 3
Action 4
Performance


Why?
Event
Action
Assumed
Network
Action 3
Real
Network
Action 4
Performance

Flow f sends at R

Event	Action

Flow f sends at R

Event	Action
Packet Loss	

Flow f sends at R

Event	Action
Packet Loss	

Flow f sends at R

Event	Action
Packet Loss	

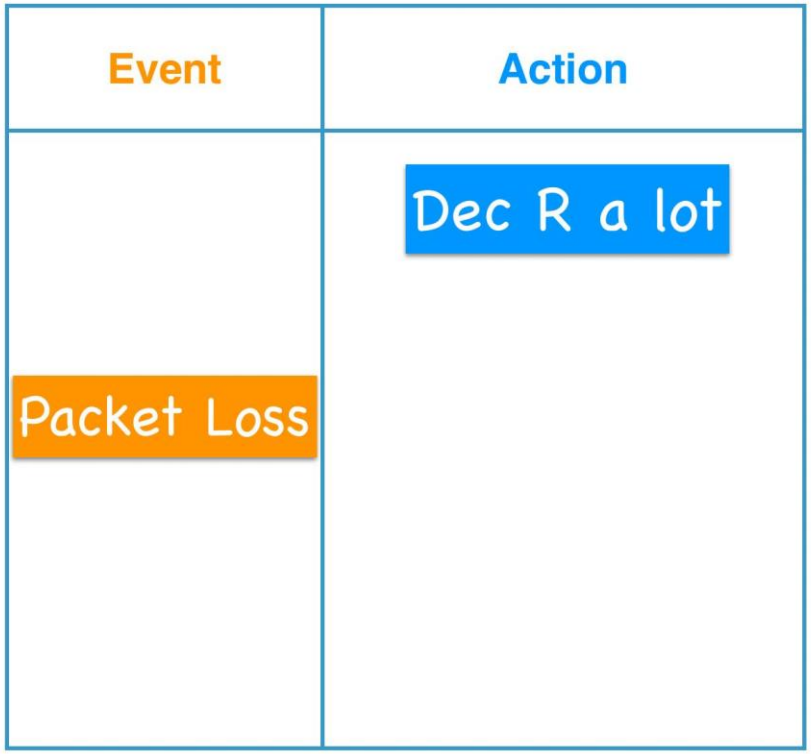
f causes
most congestion

Flow f sends at R

Event	Action
	Dec R a lot
Packet Loss	

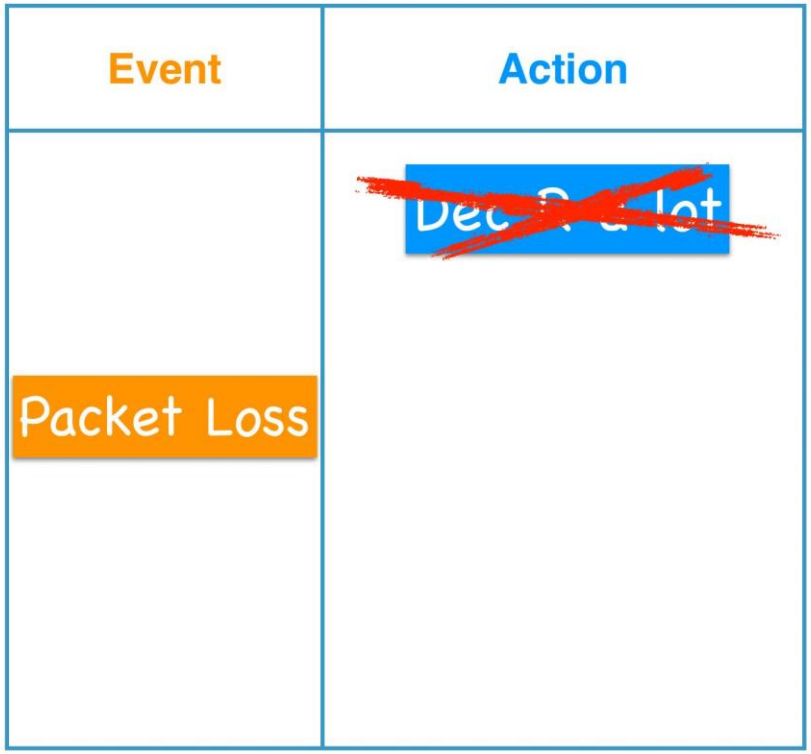
f causes
most congestion

Flow f sends at R



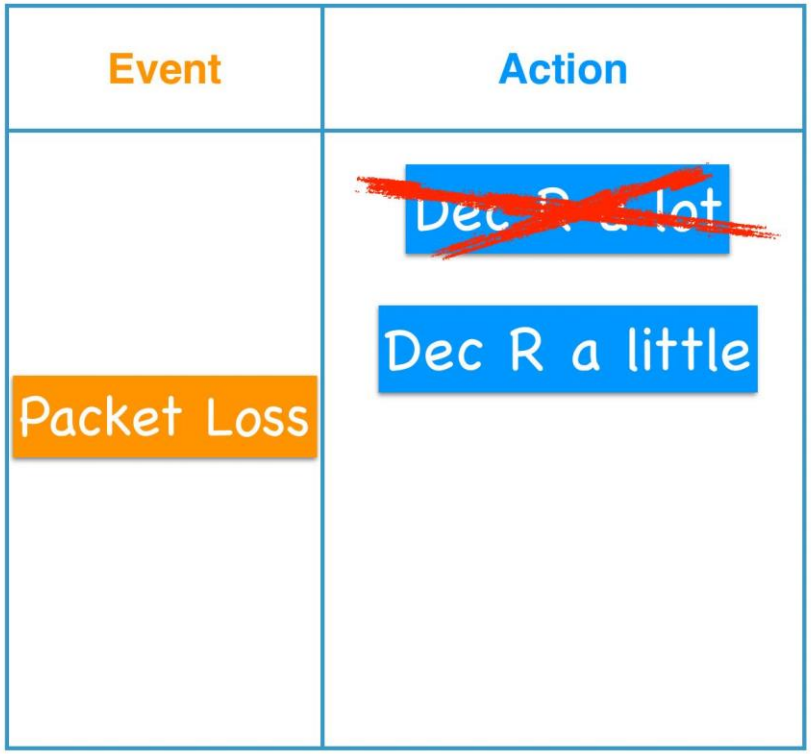
shallow buffer overflow

Flow f sends at R



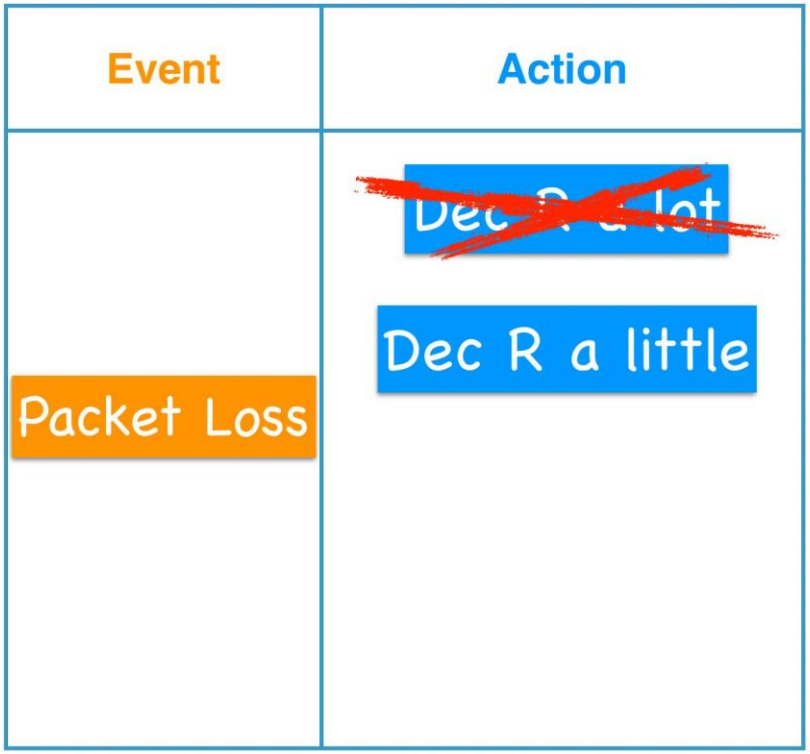
shallow buffer overflow

Flow f sends at R



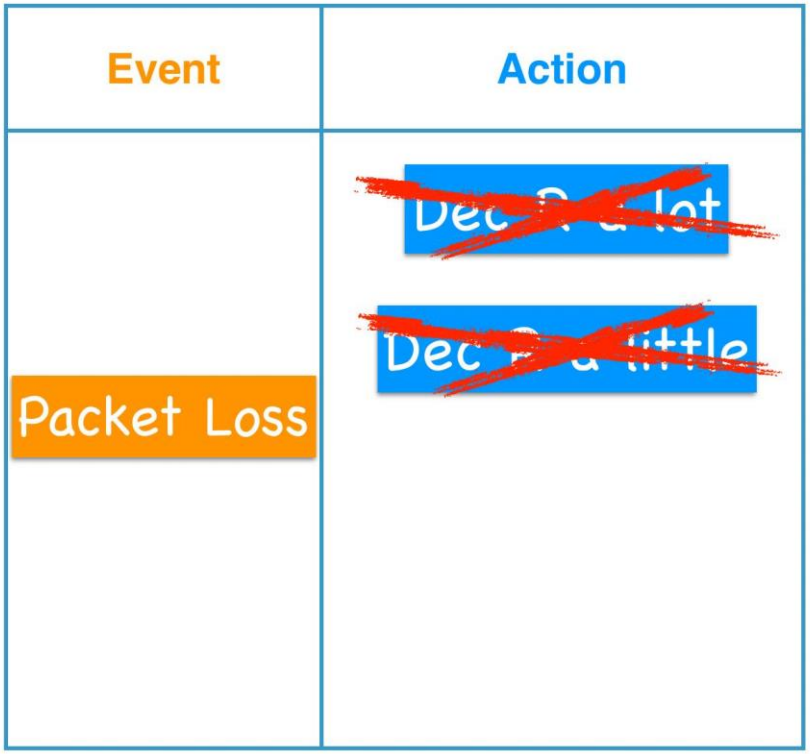
shallow buffer overflow

Flow f sends at R



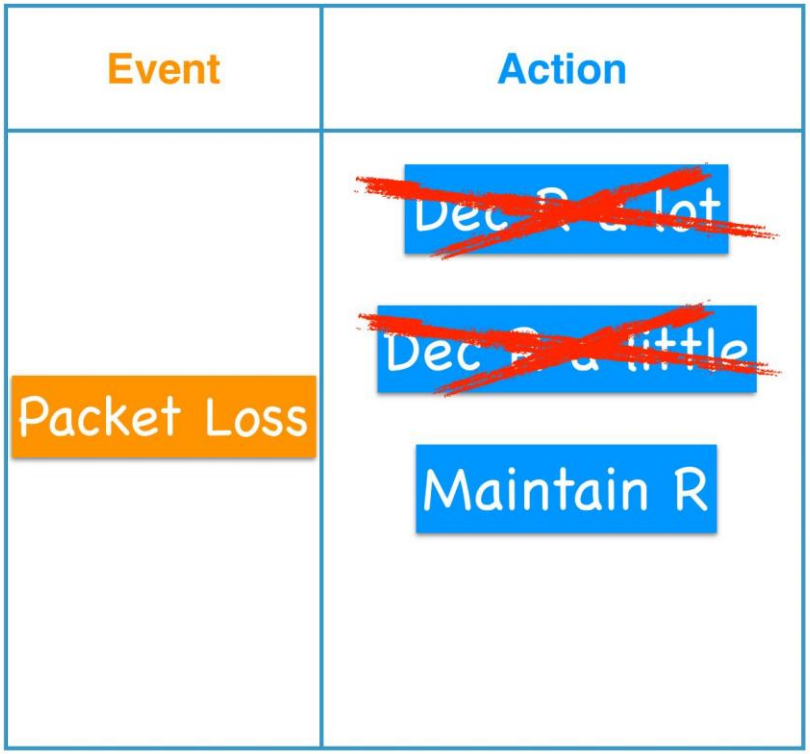
other high rate flow causing congestion

Flow f sends at R



other high rate flow causing congestion

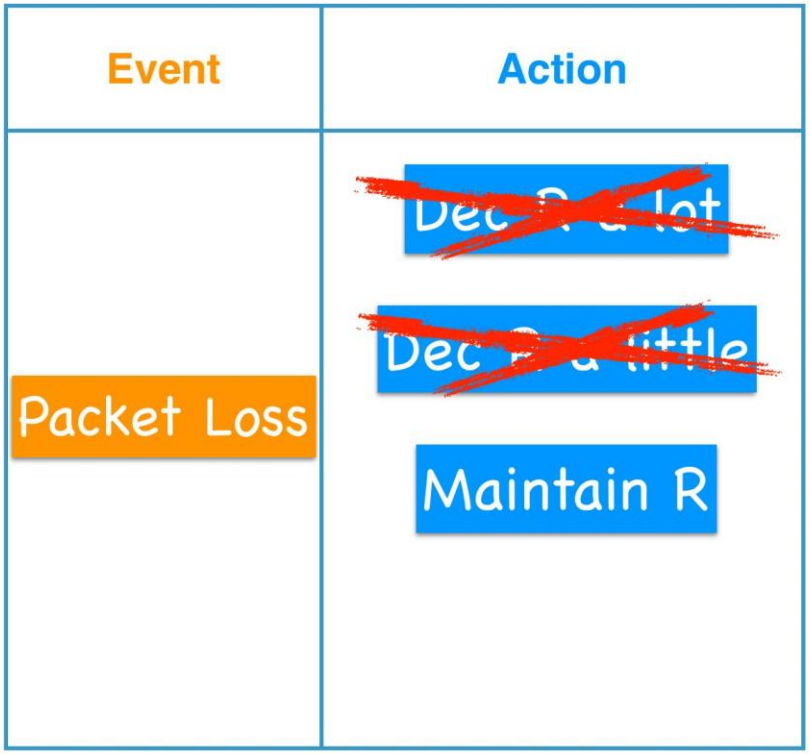
Flow f sends at R



Network

other high rate flow causing congestion

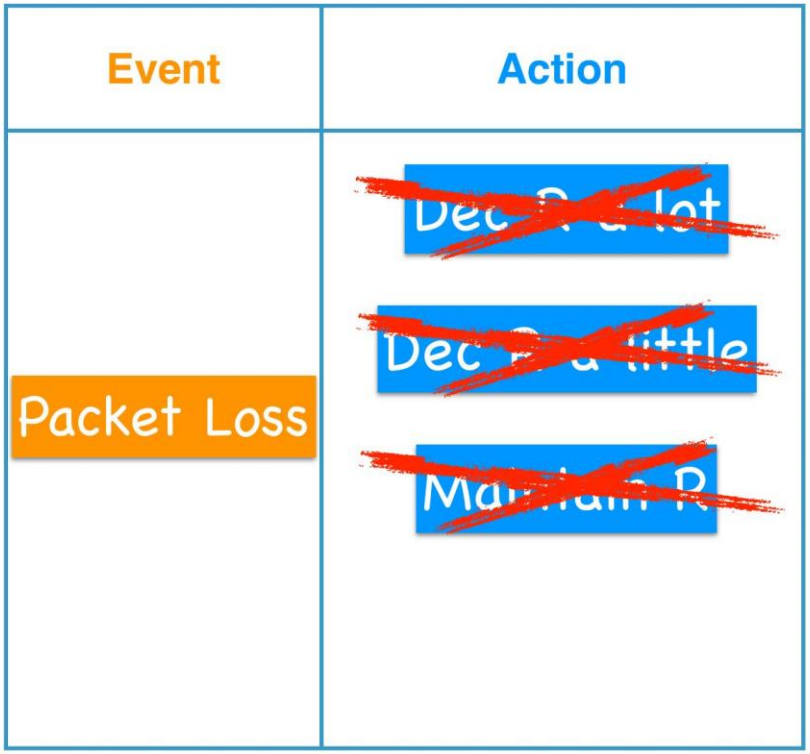
Flow f sends at R



Network

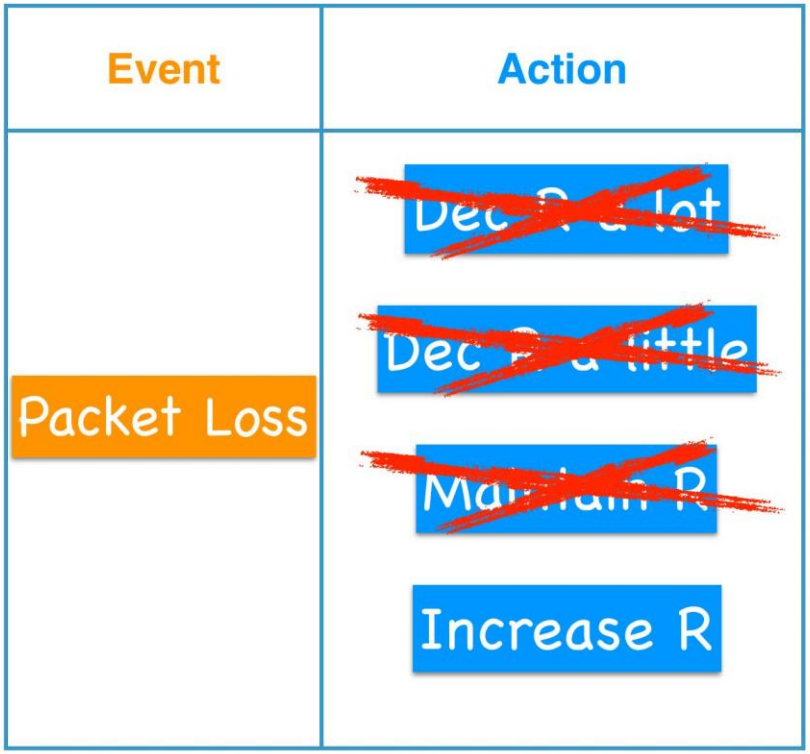
loss is random

Flow f sends at R



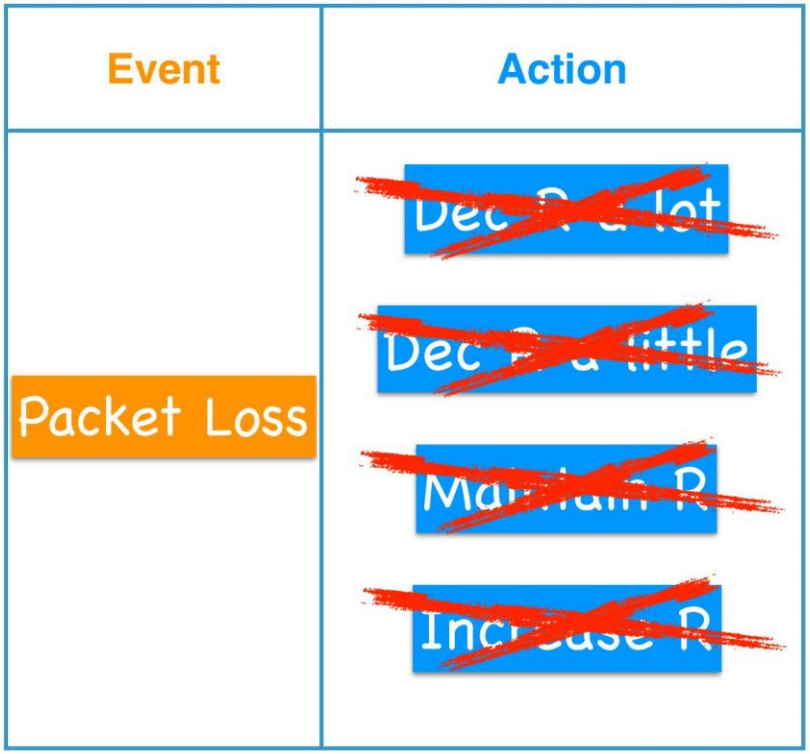
loss is random

Flow f sends at R



loss is random

Flow f sends at R

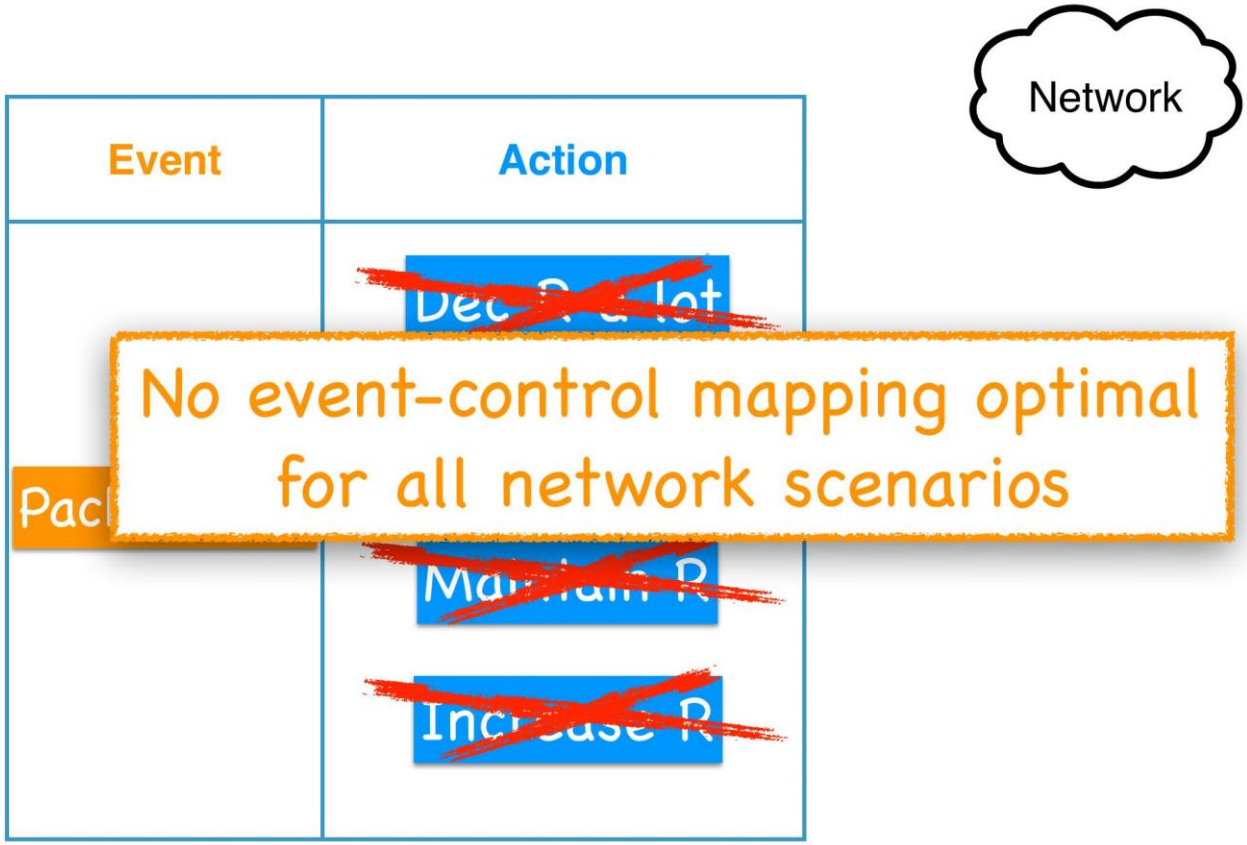


loss is random

Flow f sends at R

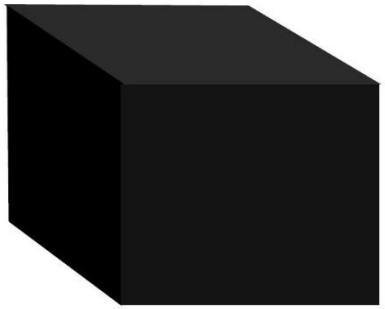
Event	Action
Packet Loss	~~Dec~~
	~~Dec~~
	~~Ma~~
	~~Inc~~

Flow f sends at R

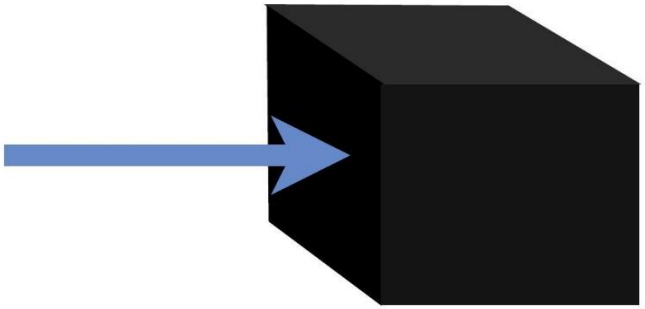


What is the right rate to send?

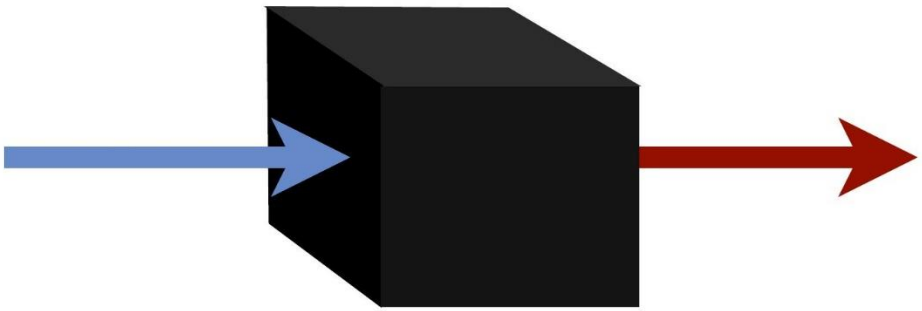
What is the right rate to send?



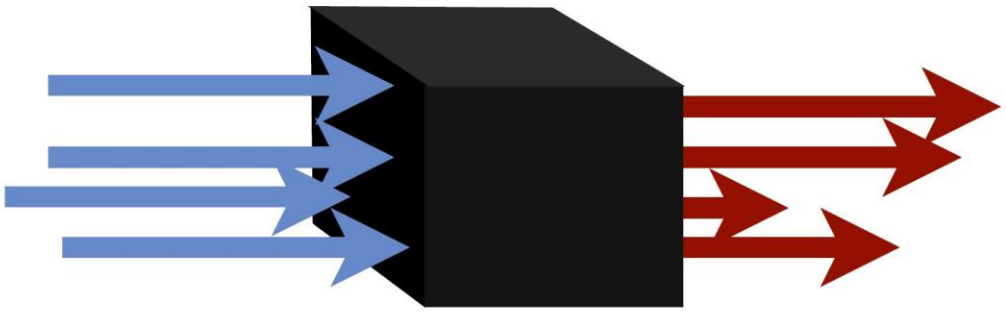
What is the right rate to send?



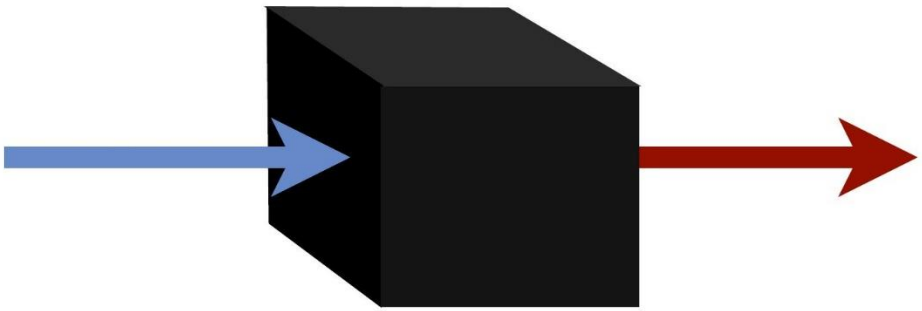
What is the right rate to send?



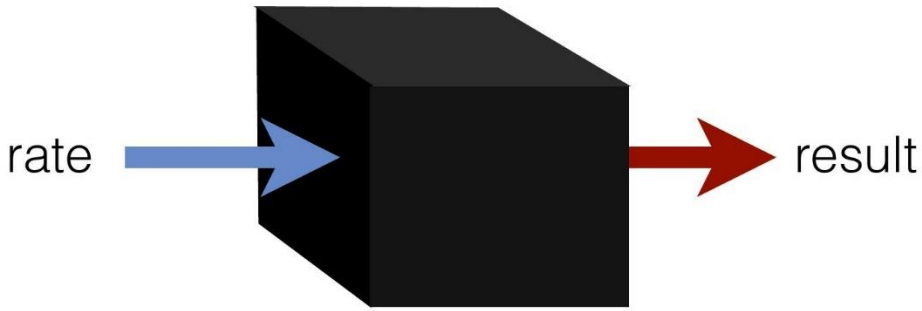
What is the right rate to send?



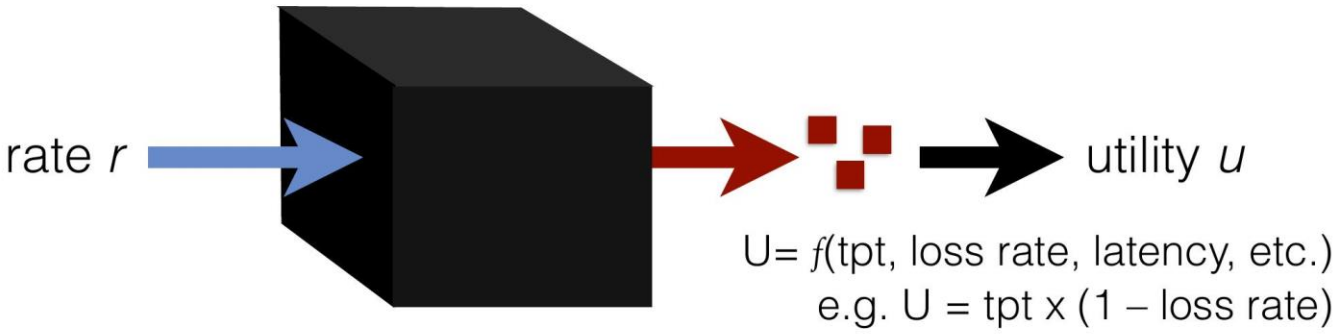
What is the right rate to send?



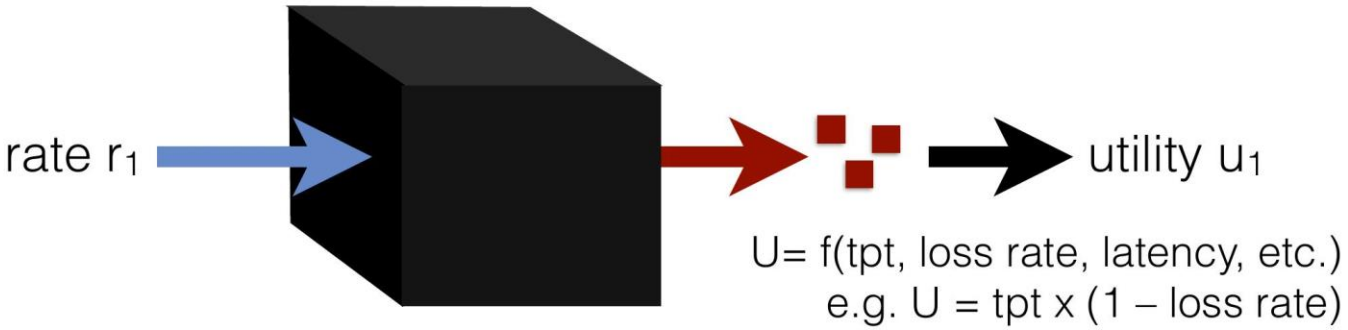
What is the right rate to send?



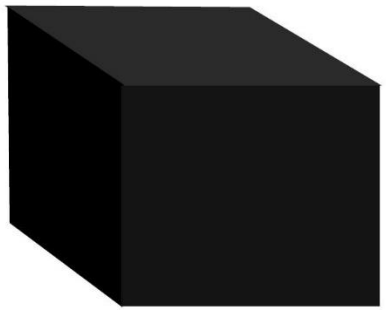
What is the right rate to send?

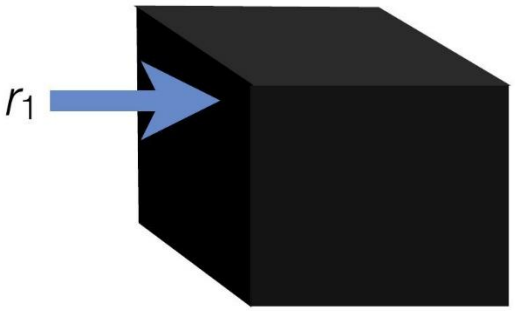


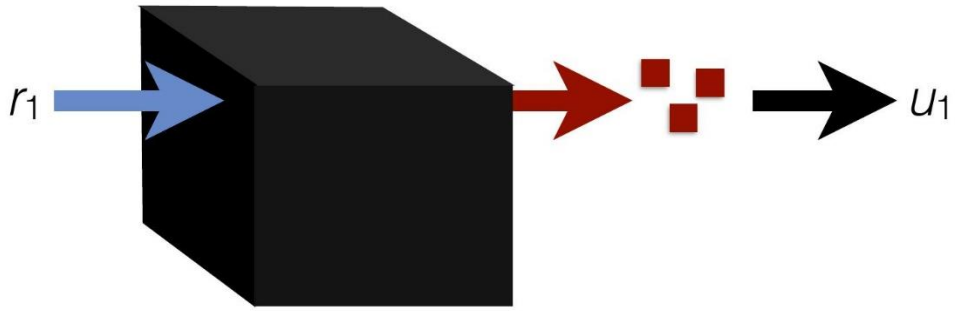
What is the right rate to send?

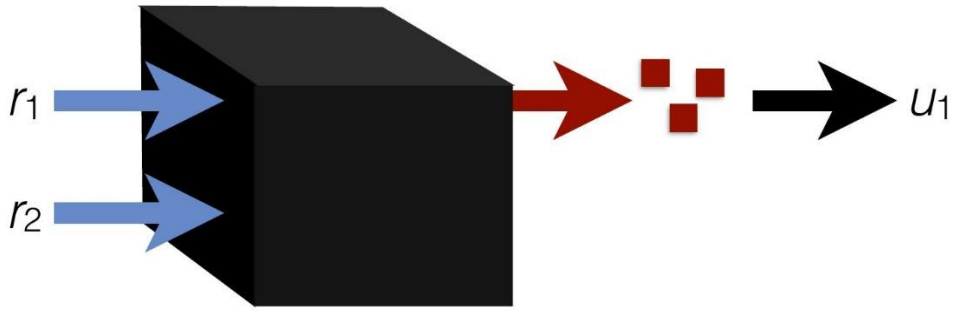


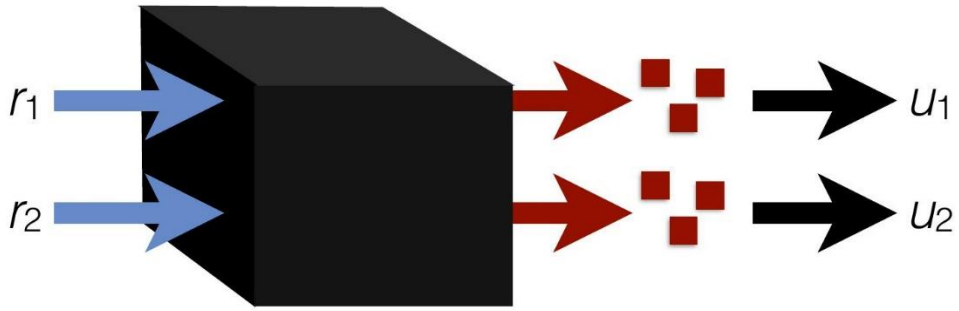
No matter how complex the network,
rate r —> utility u

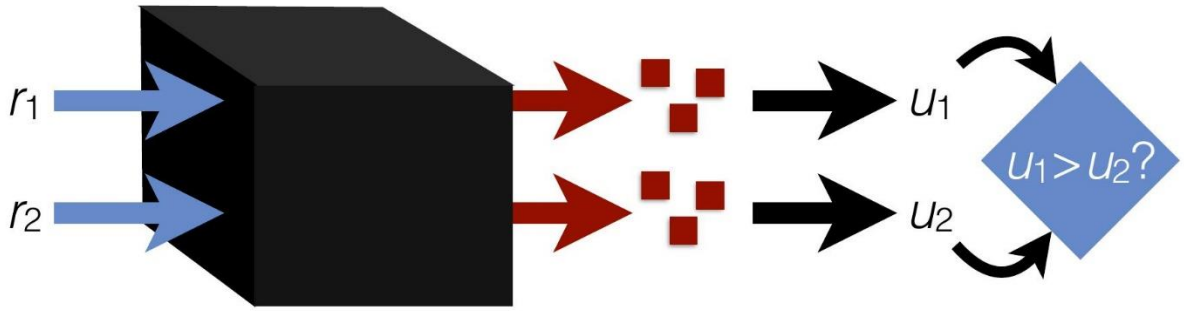


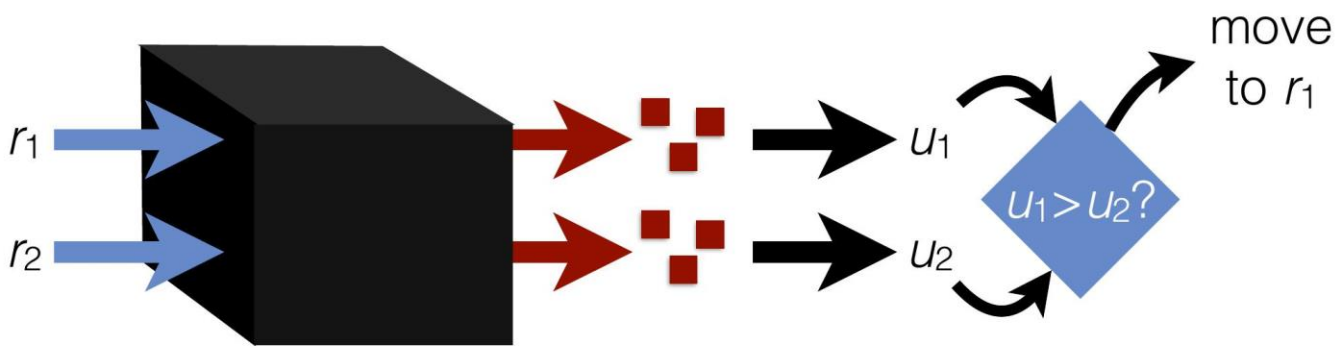

r_1

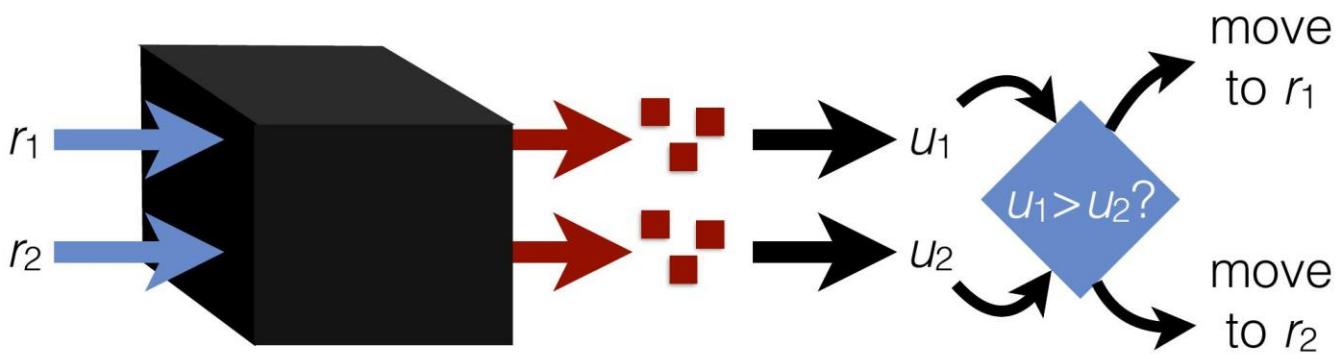

r_1
u_1


r_1
r_2
u_1

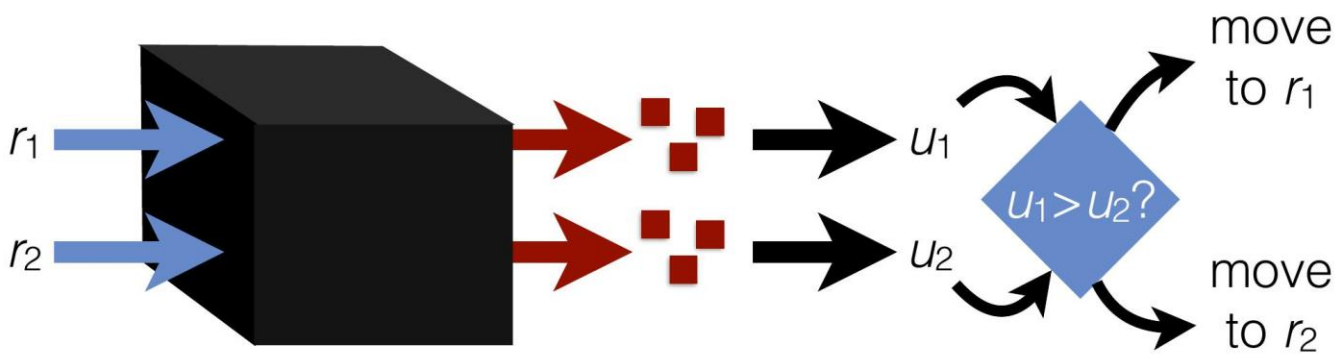

r_1
r_2
u_1
u_2


r_1
r_2
u_1
u_2
$u_1 > u_2$?

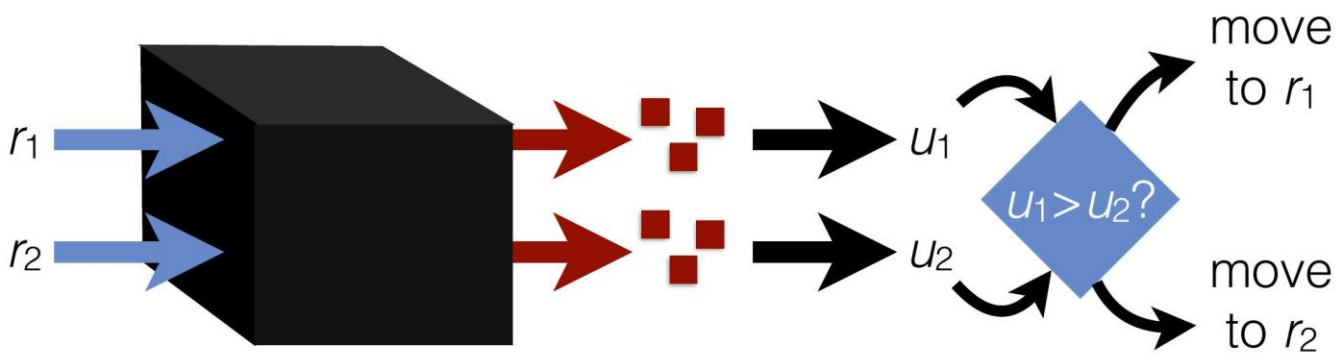

r_1
r_2
u_1
u_2
$u_1 > u_2$?
move
to r_1


r_1
r_2
u_1
u_2
$u_1 > u_2$?
move to r_1
move to r_2

Performance-oriented Congestion Control

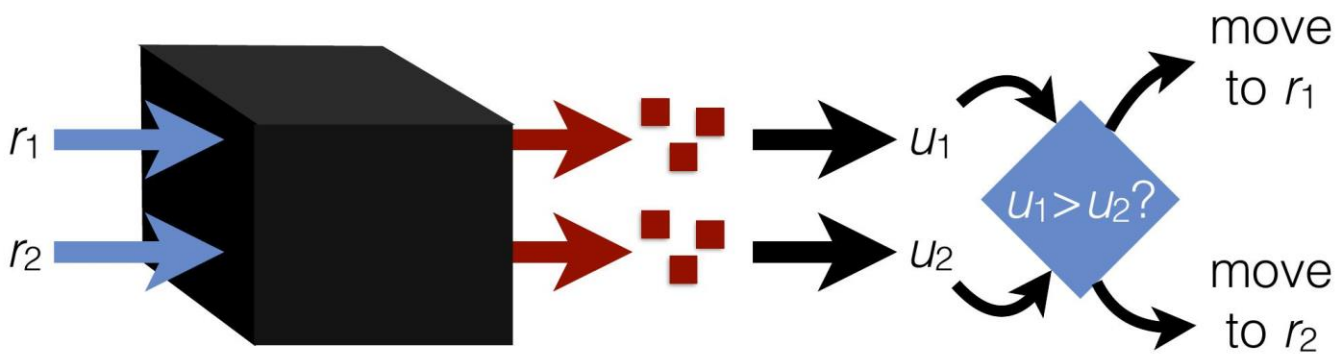


Performance-oriented Congestion Control



Observe real performance

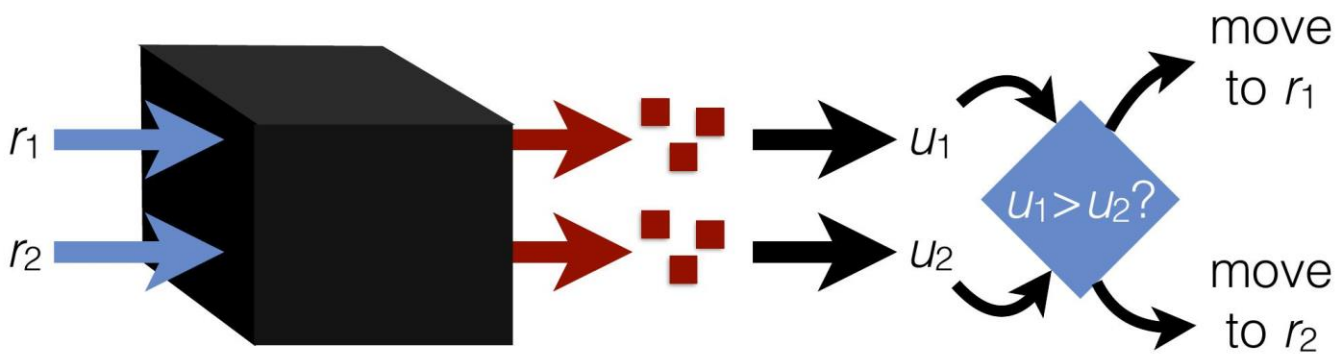
Performance-oriented Congestion Control



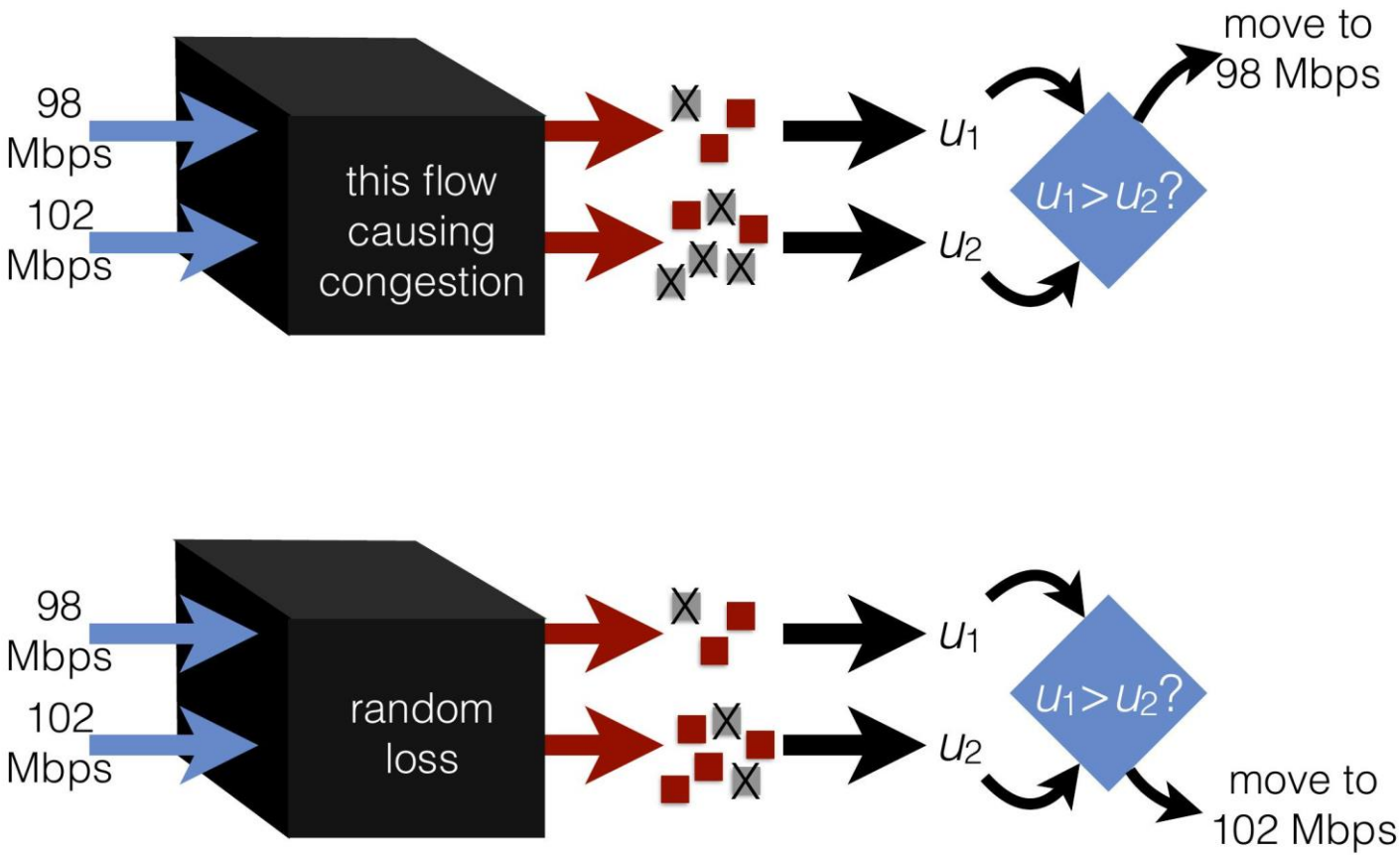
Observe real performance

Control based on empirical evidence

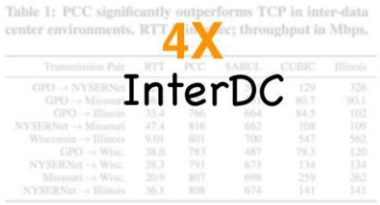
Performance-oriented Congestion Control



Observe real performance	Control based on empirical evidence	yields **Consistent high performance**

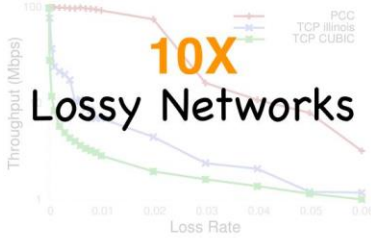

98
Mbps
102
Mbps
this flow
causing
congestion
u_1
u_2
$u_1 > u_2$?
move to
98 Mbps
98
Mbps
102
Mbps
random
loss
u_1
u_2
$u_1 > u_2$?
move to
102 Mbps

Consistent High Performance

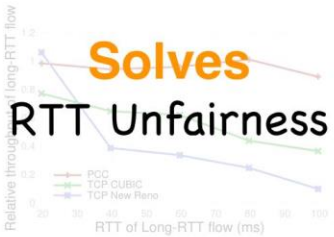


4X
InterDC

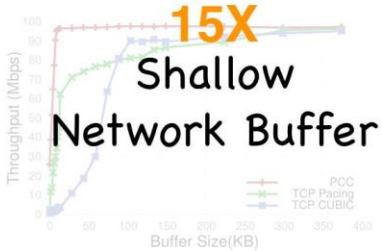
17X
Satellite Networks



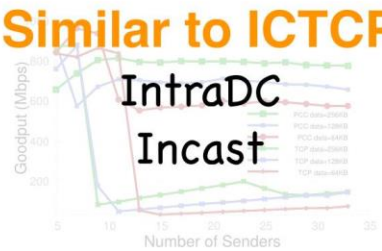
10X
Lossy Networks



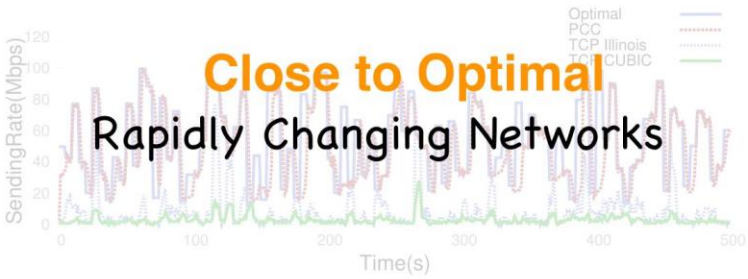
Solves
RTT Unfairness



15X
Shallow
Network Buffer



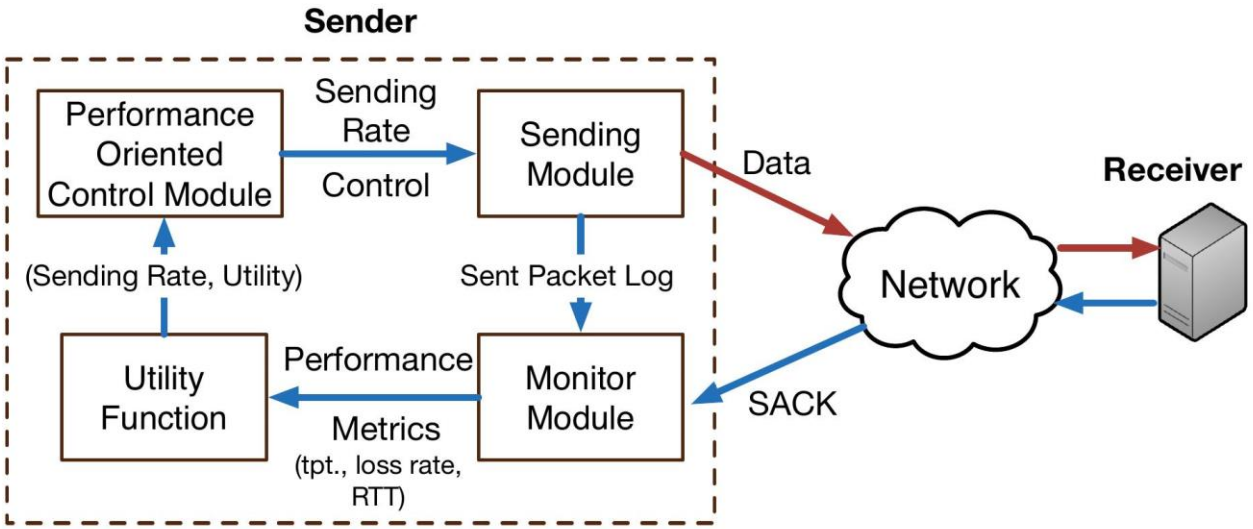
Similar to ICTCP
IntraDC
Incast



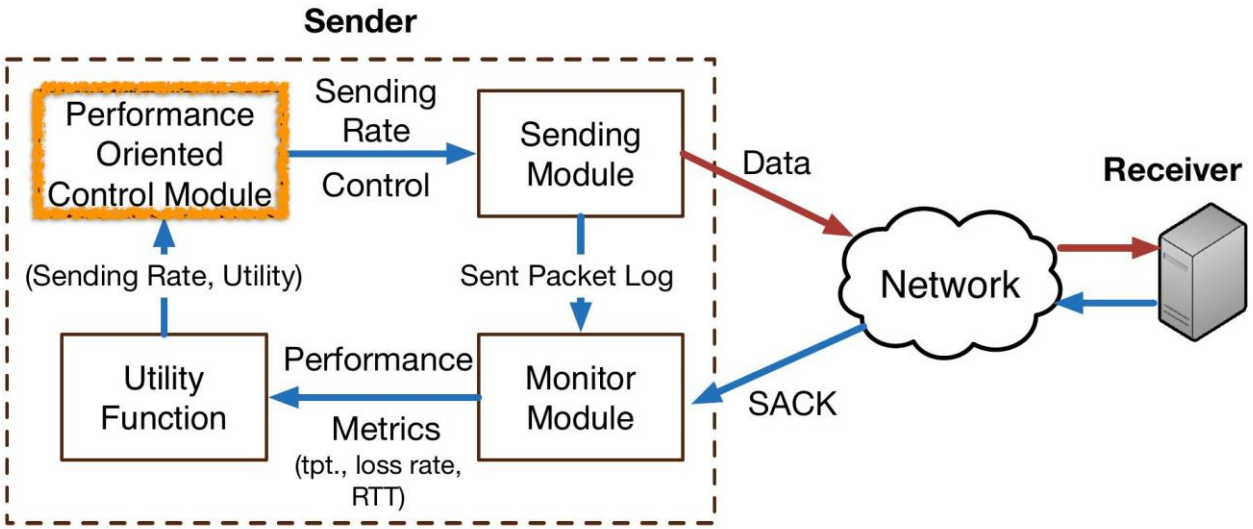
Close to Optimal
Rapidly Changing Networks

5X in median
Global Commercial
Internet

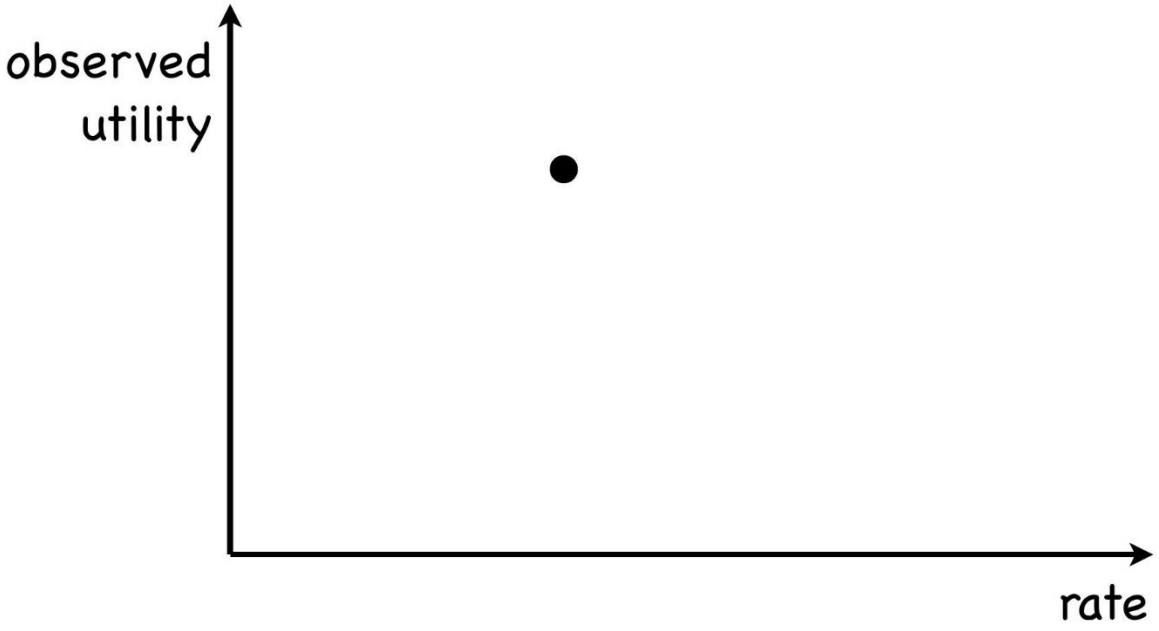
Software Components



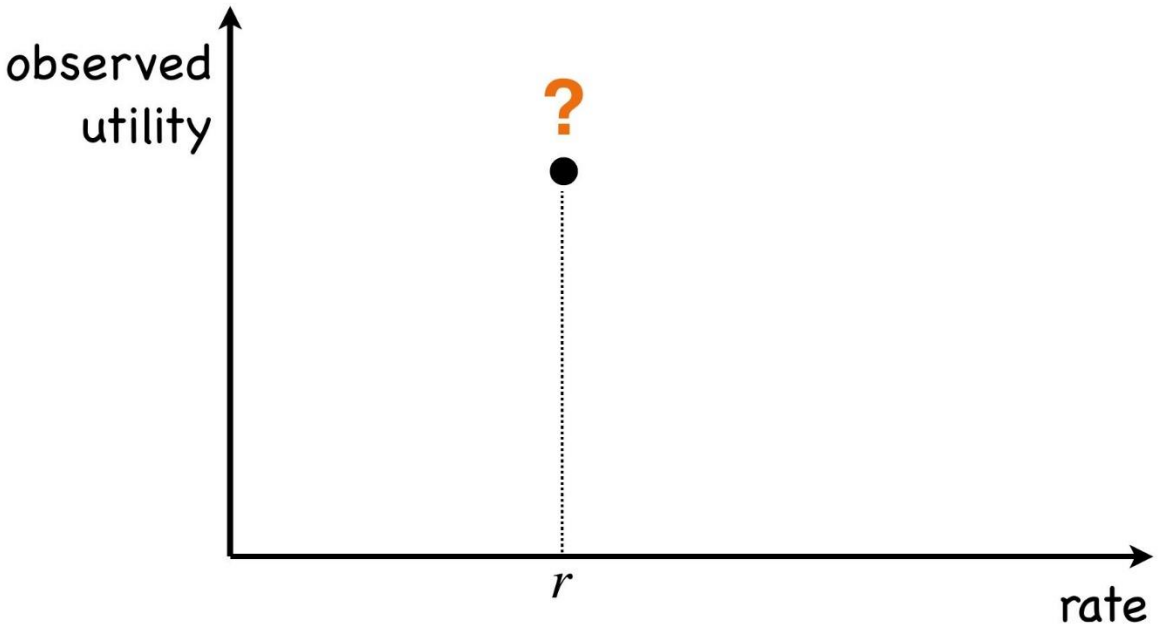
Software Components



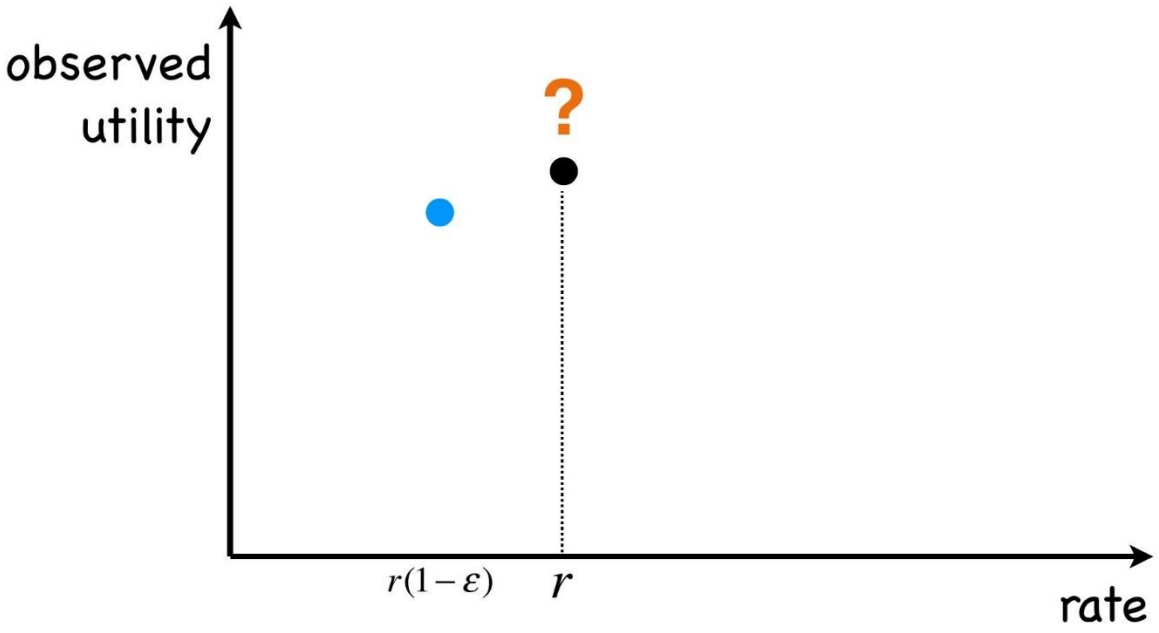
Performance Oriented Control



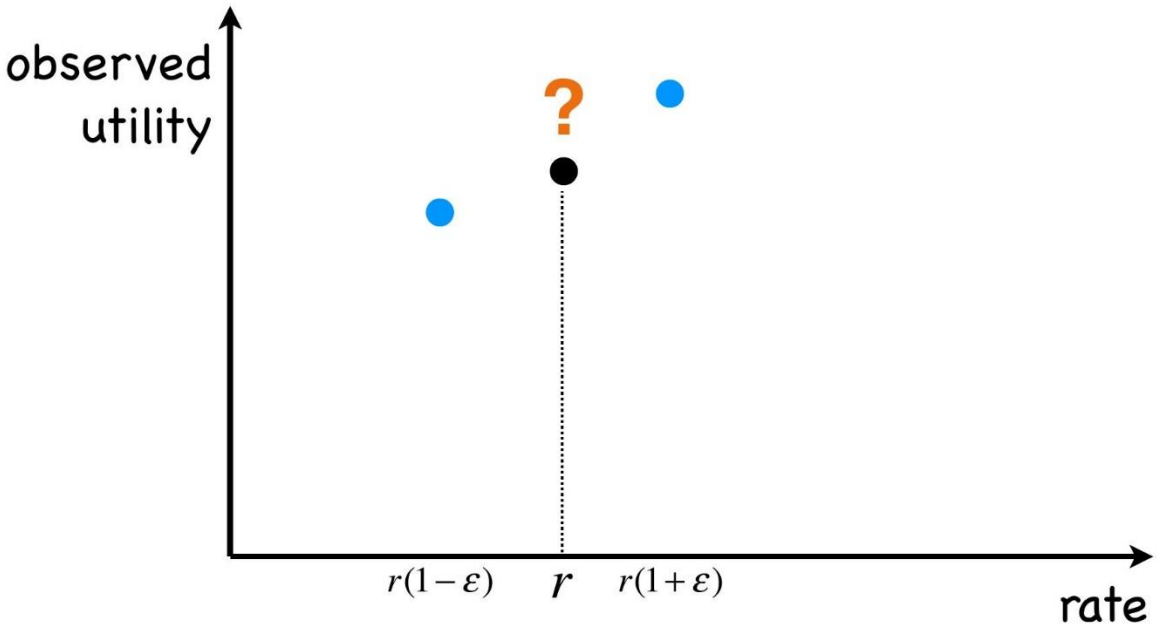
Performance Oriented Control



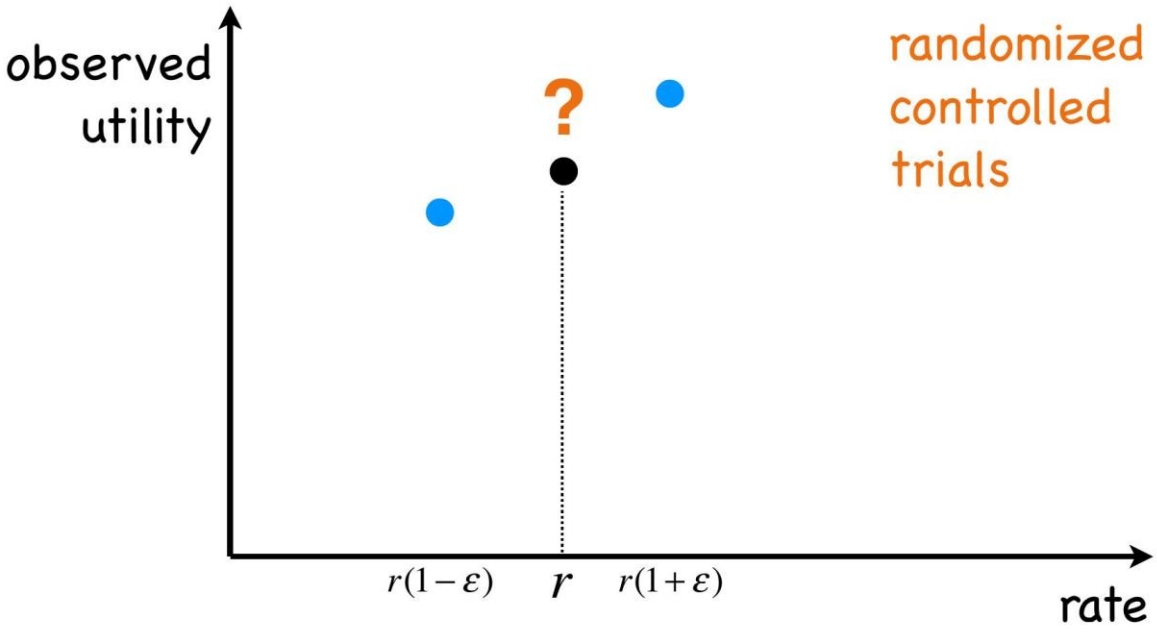
Performance Oriented Control



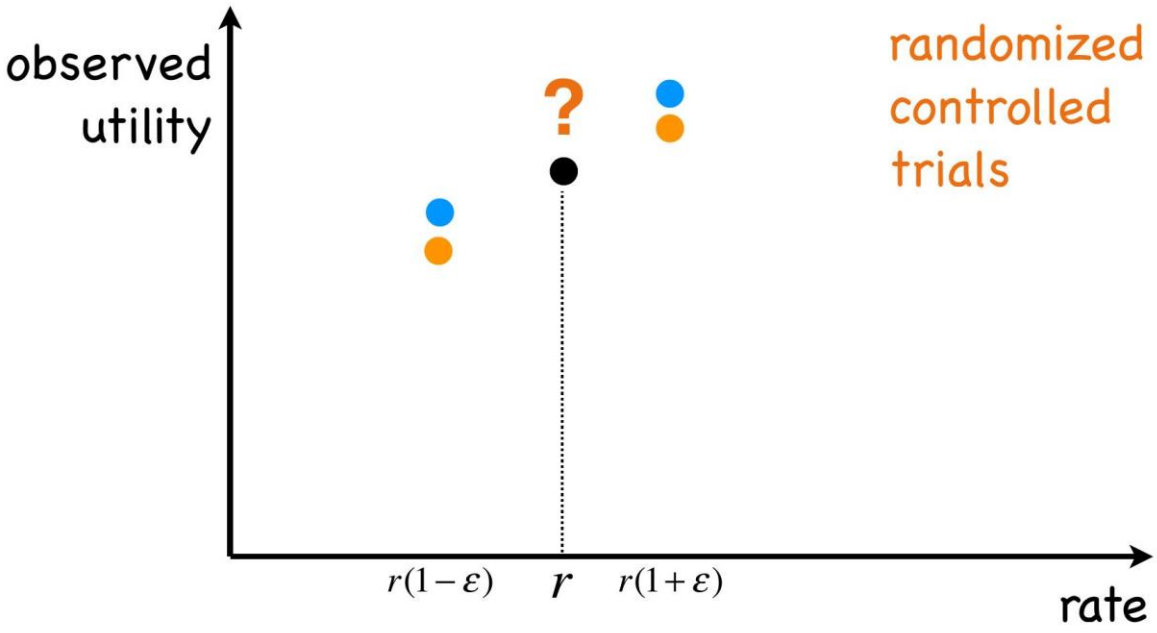
Performance Oriented Control



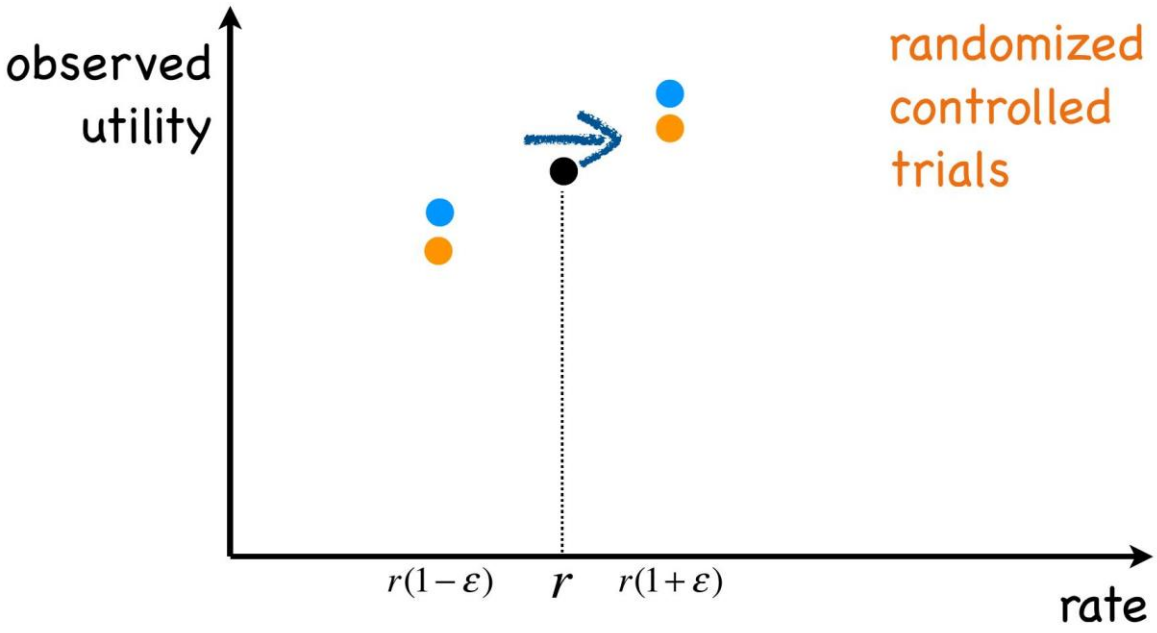
Performance Oriented Control



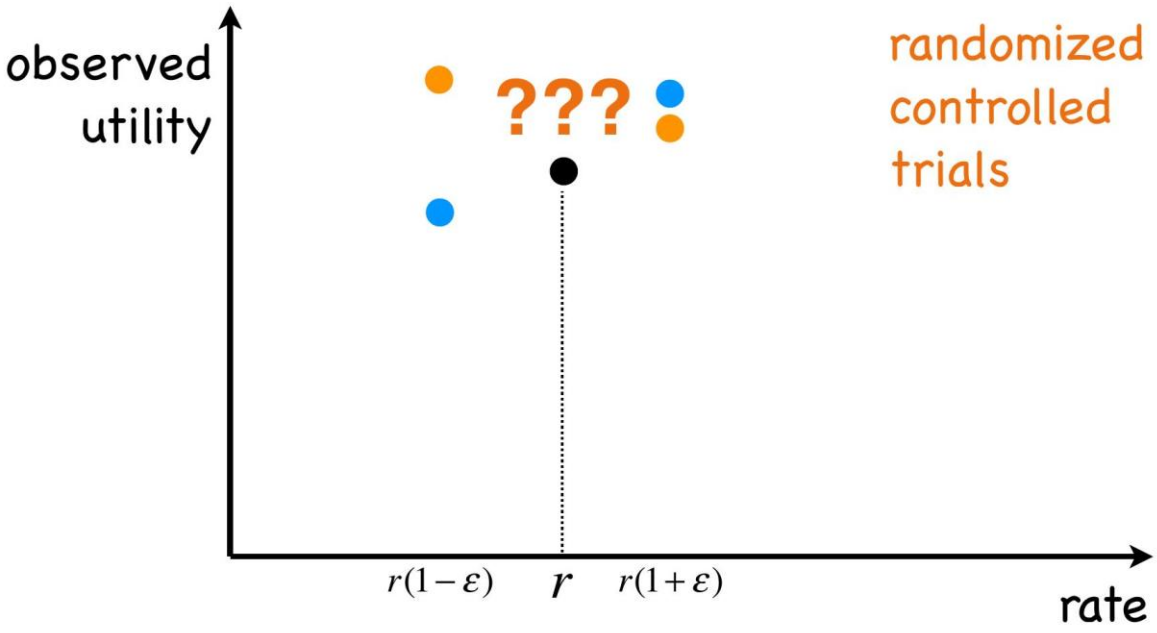
Performance Oriented Control



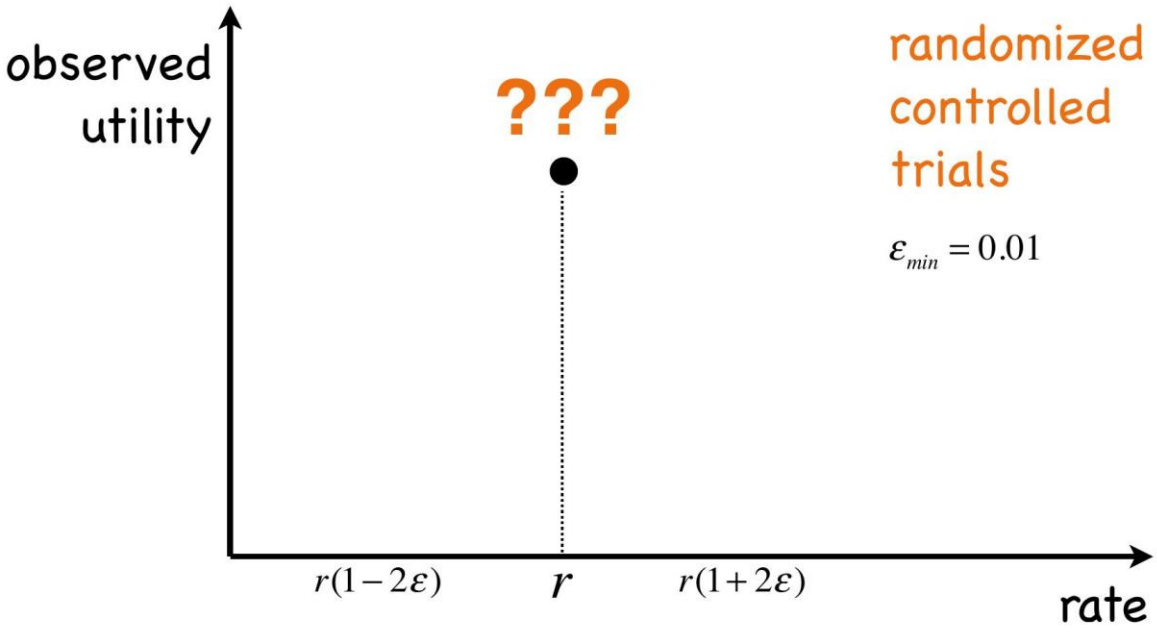
Performance Oriented Control



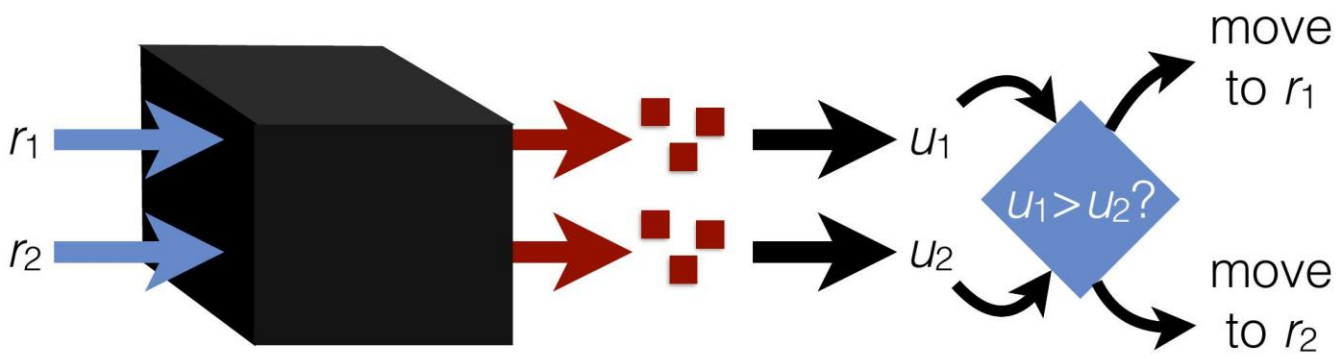
Performance Oriented Control



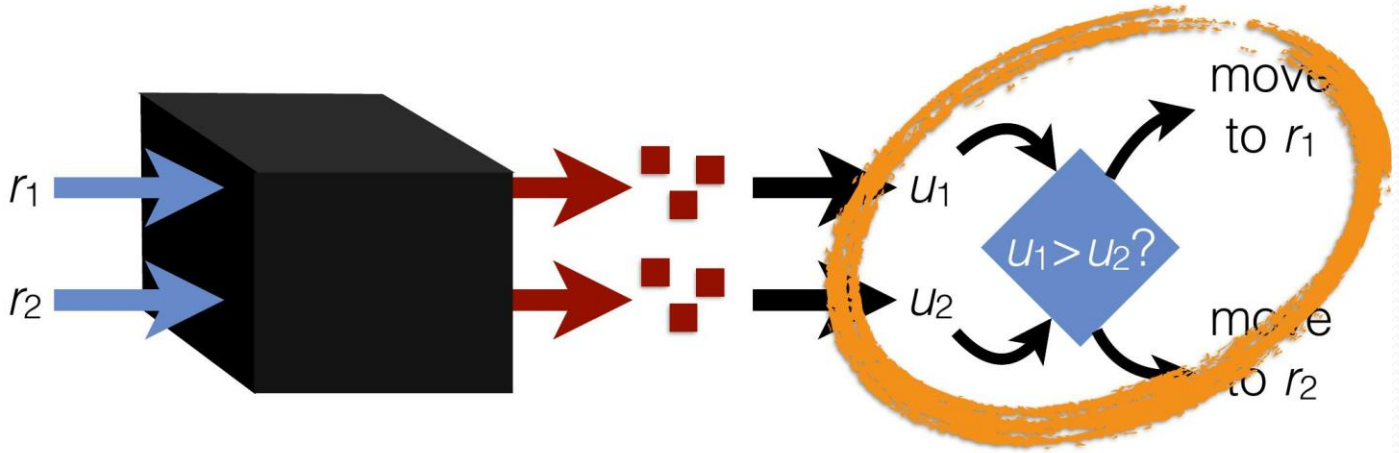
Performance Oriented Control



Where is Congestion Control?

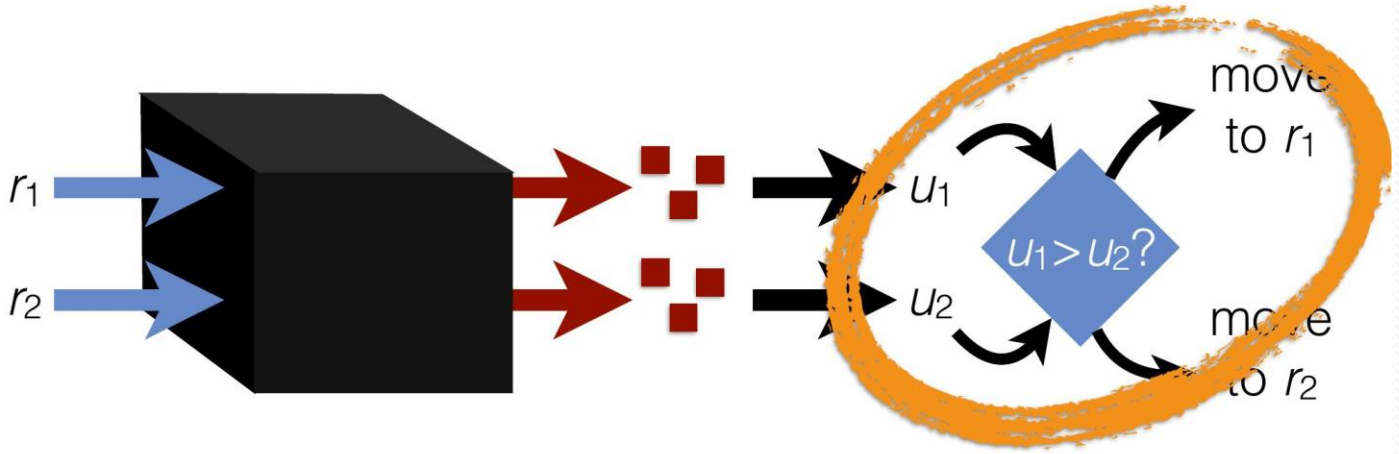


Where is Congestion Control?



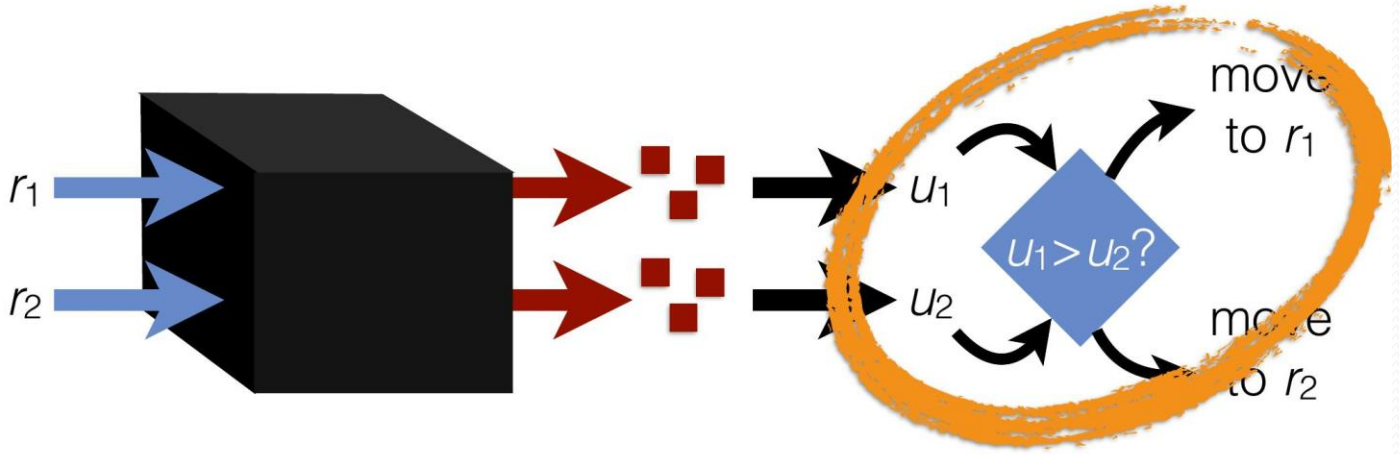
Selfishly maximizing utility

Where is Congestion Control?



Selfishly maximizing utility
=> non-cooperative game

Where is Congestion Control?



Selfishly maximizing utility
=> non-cooperative game

Do we converge to a fair Nash equilibrium?

A class of utility functions converge to a fair and efficient Nash Equilibrium

A class of utility functions converge to a fair and efficient Nash Equilibrium

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

A class of utility functions converge to a fair and efficient Nash Equilibrium

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

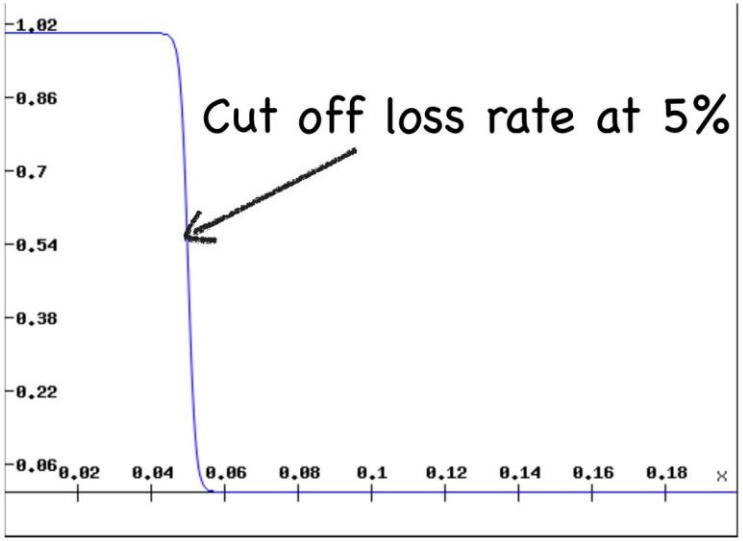
throughput

sending rate

observed loss rate

A class of utility functions converge to a fair and efficient Nash Equilibrium

$$u_i(x) = T_i * sigmoid(L_i - 0.05) - x_i * L_i$$



Loss Rate

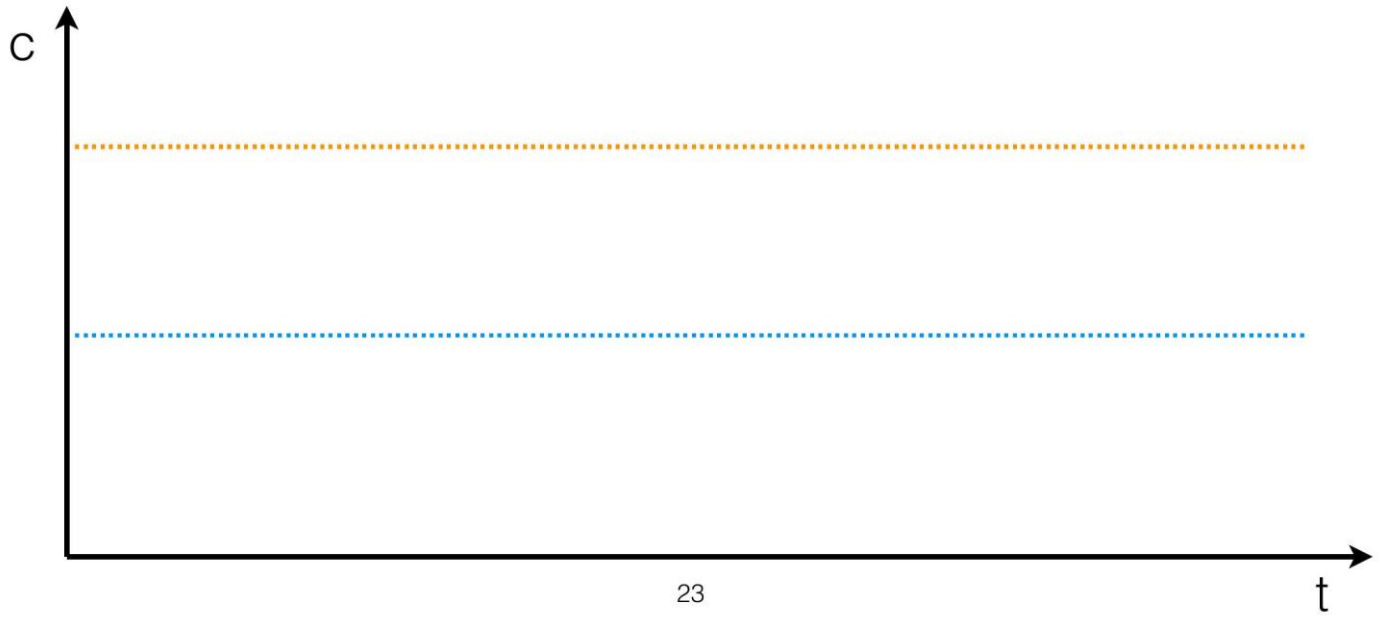
TCP Dynamics

TCP Dynamics

AIMD as a "hack" to asymptotic fairness

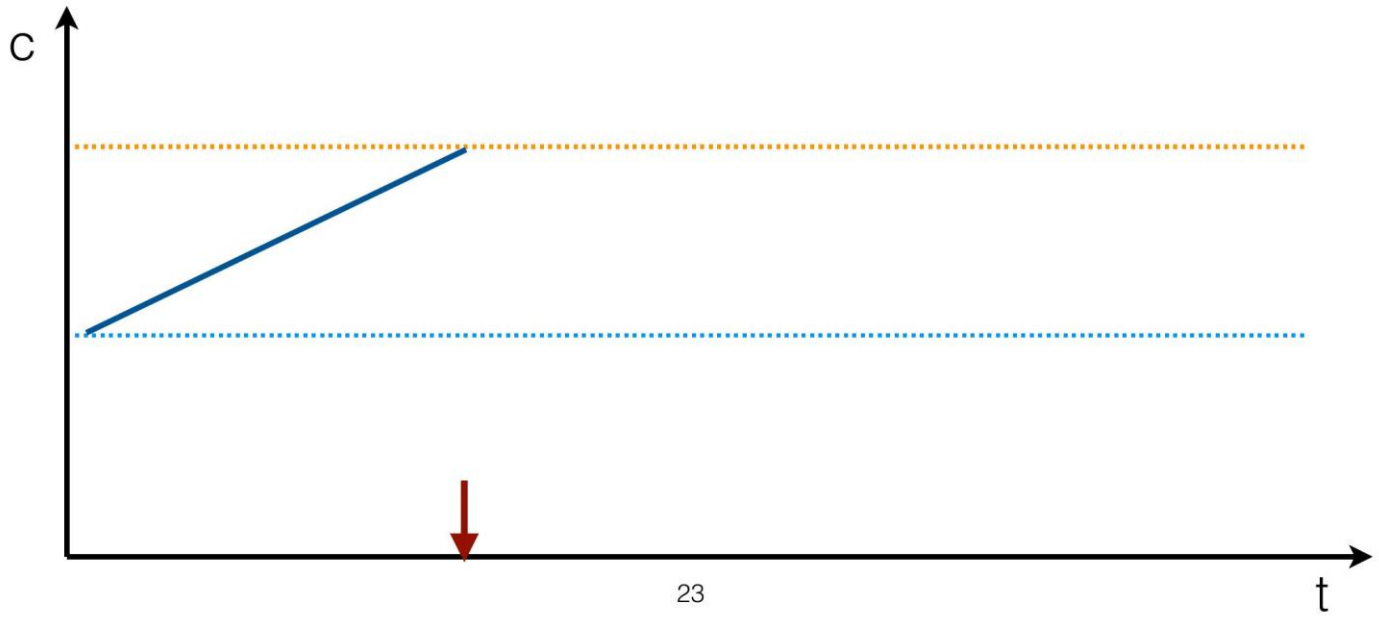
TCP Dynamics

AIMD as a "hack" to asymptotic fairness



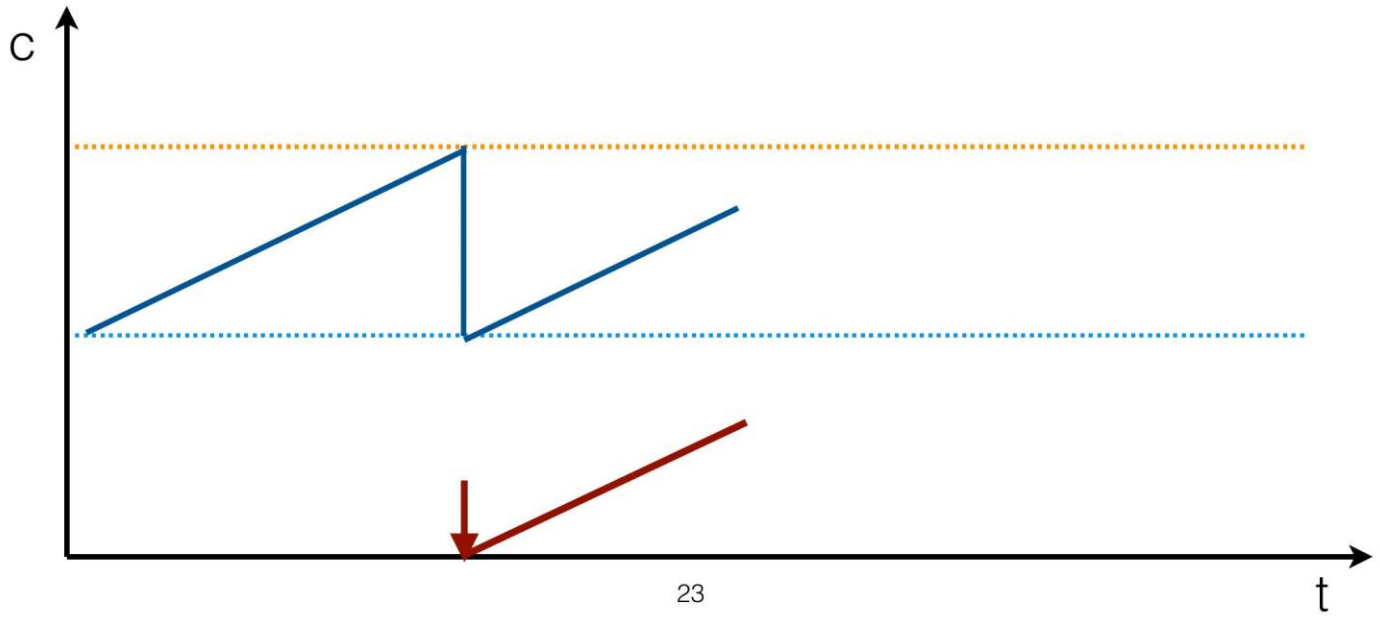
TCP Dynamics

AIMD as a "hack" to asymptotic fairness



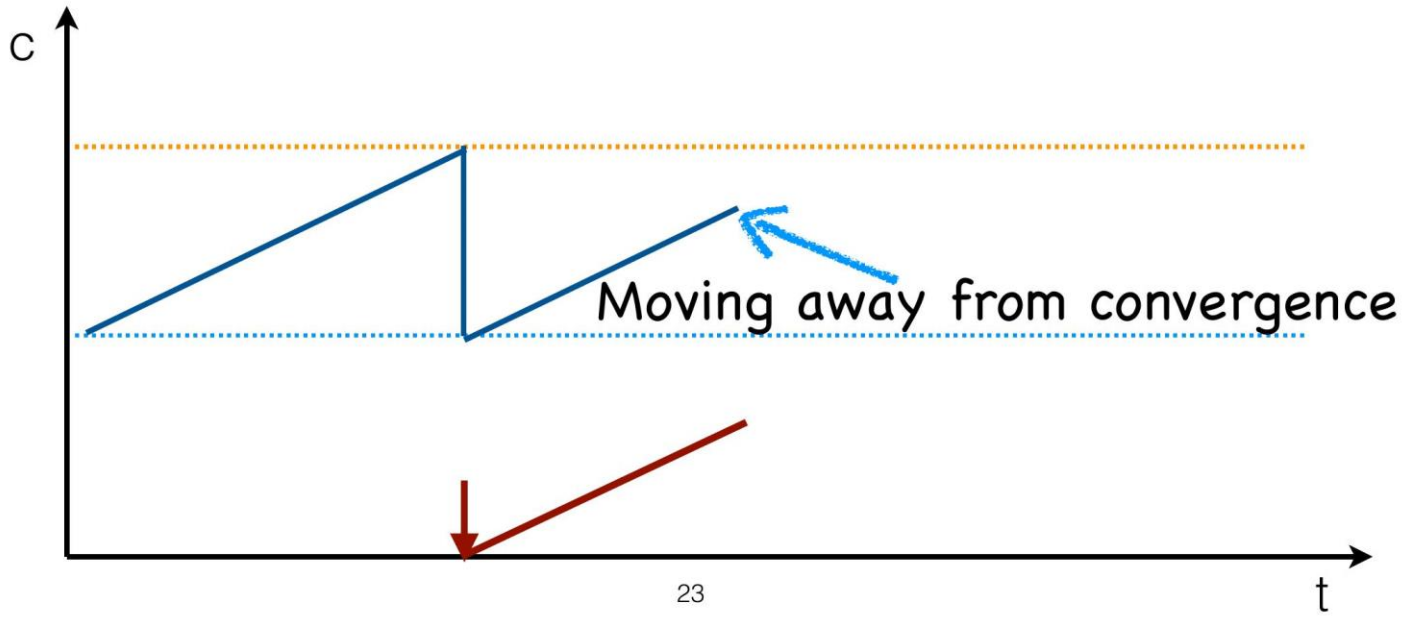
TCP Dynamics

AIMD as a "hack" to asymptotic fairness

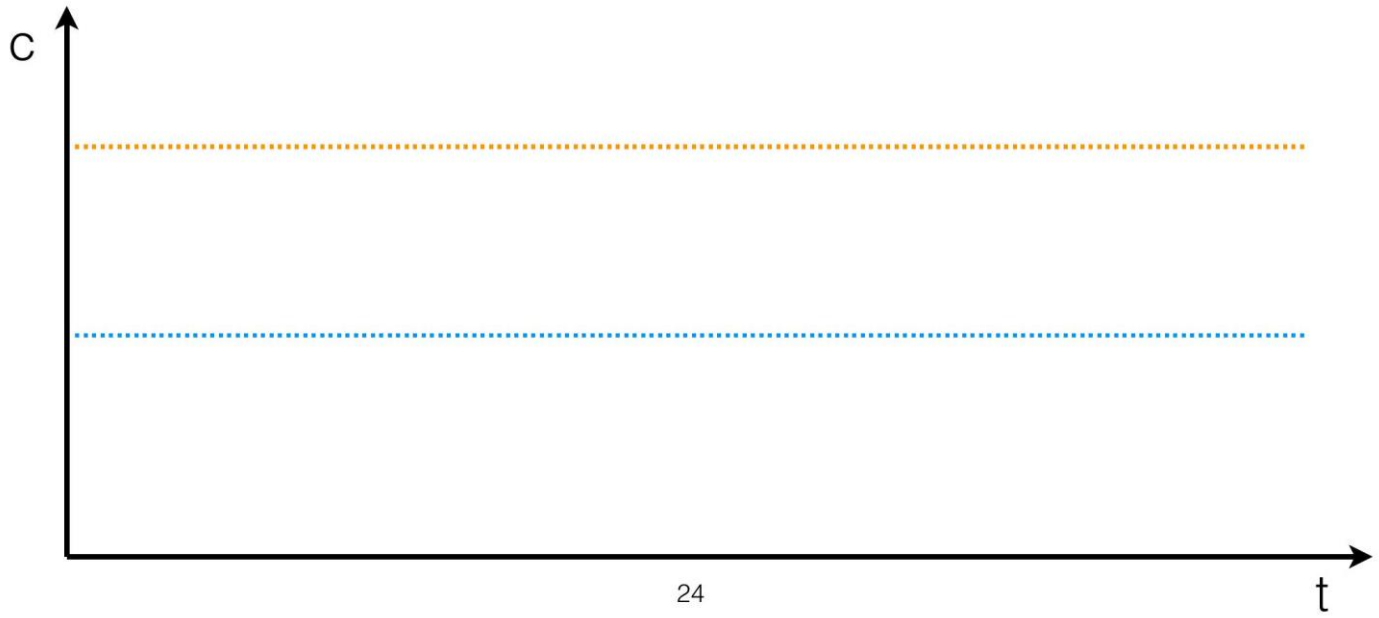


TCP Dynamics

AIMD as a "hack" to asymptotic fairness

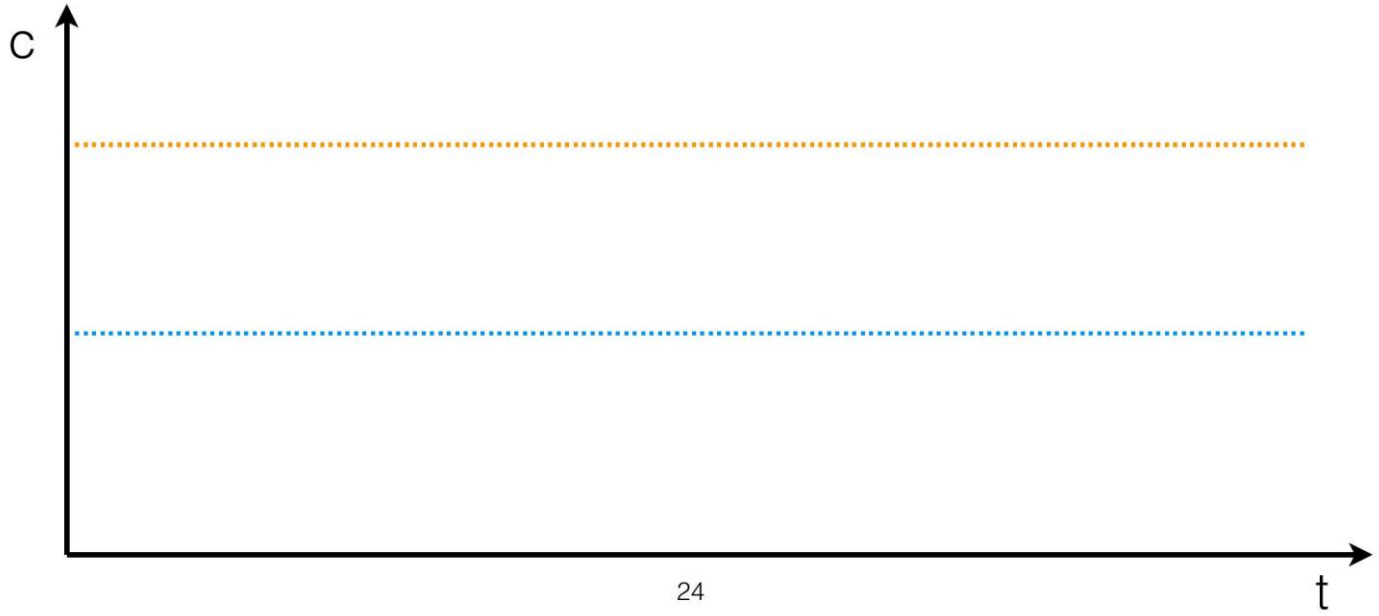


PCC Dynamics



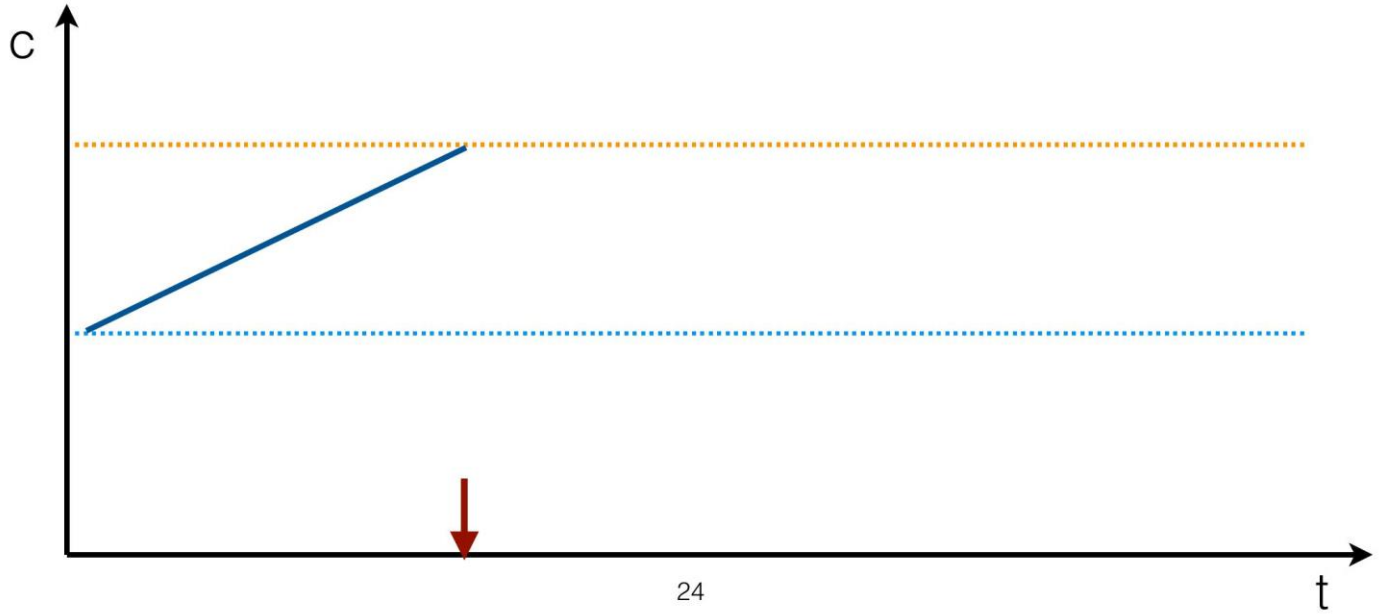
PCC Dynamics

PCC does not need AIMD because it looks at real performance



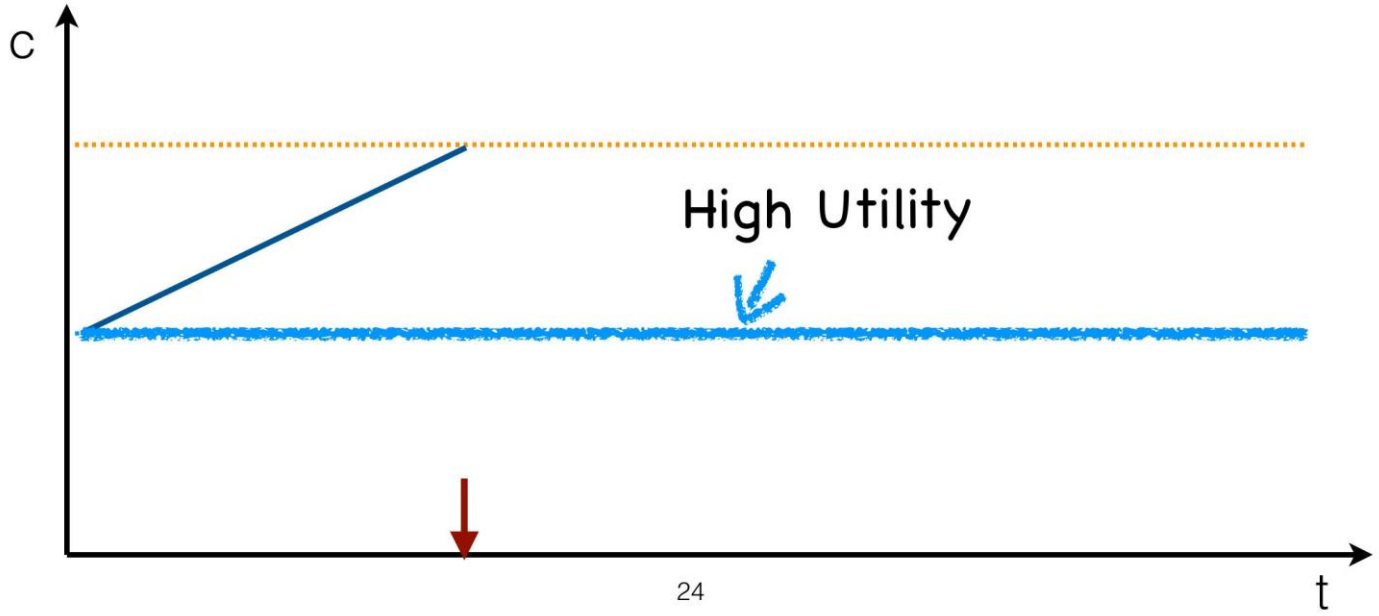
PCC Dynamics

PCC does not need AIMD because it looks at real performance



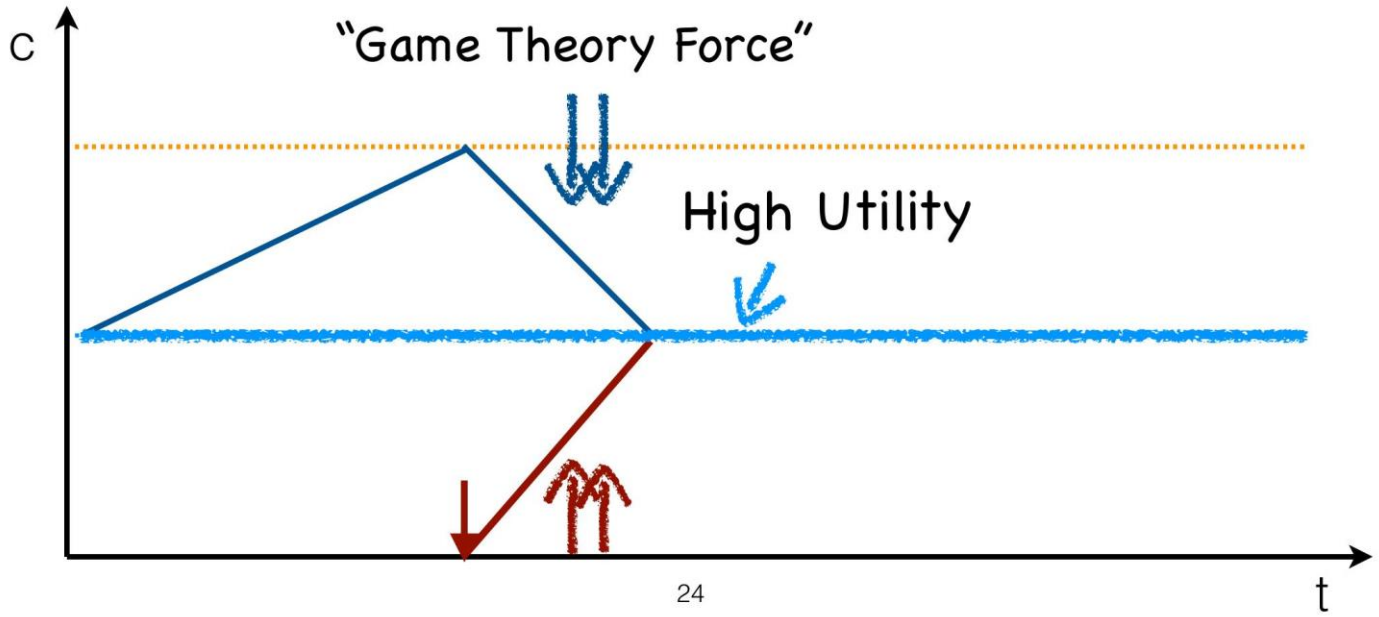
PCC Dynamics

PCC does not need AIMD because it looks at real performance



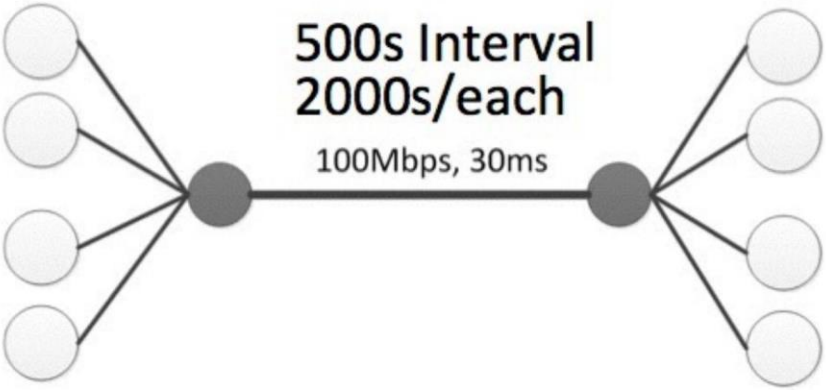
PCC Dynamics

PCC does not need AIMD because it looks at real performance

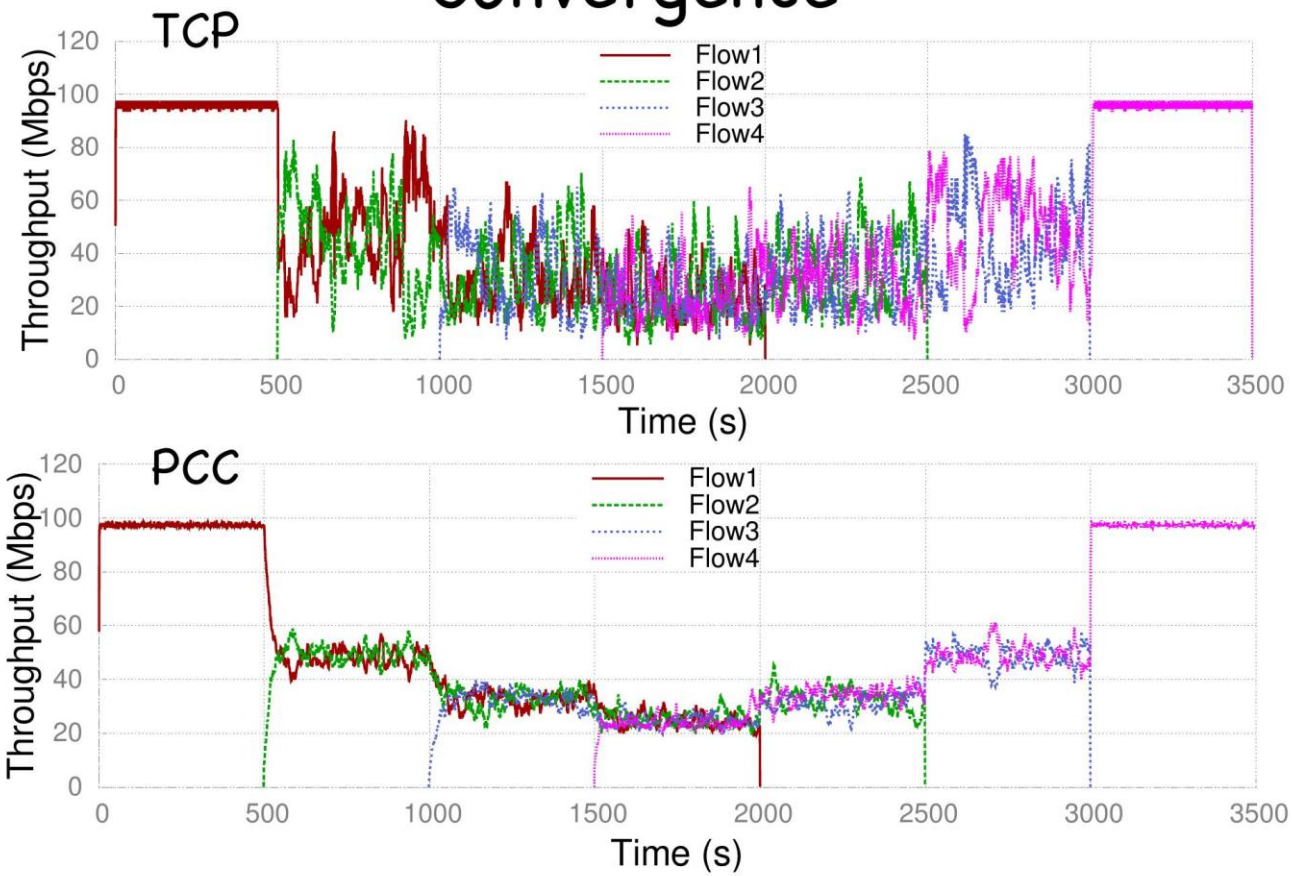


Convergence

Convergence



Convergence



Deployment

- No hardwired support, packet header, protocol change needed
- Where to deploy
  - CDN backbone, Inter-data center, dedicated scientific nw
  - "In the Wild"?

TCP Friendliness

TCP Friendliness

- Is PCC TCP-friendly?

TCP Friendliness

- Is PCC TCP-friendly?

TCP Friendliness

TCP Friendliness

- PCC's default utility function is not TCP Friendly

TCP Friendliness

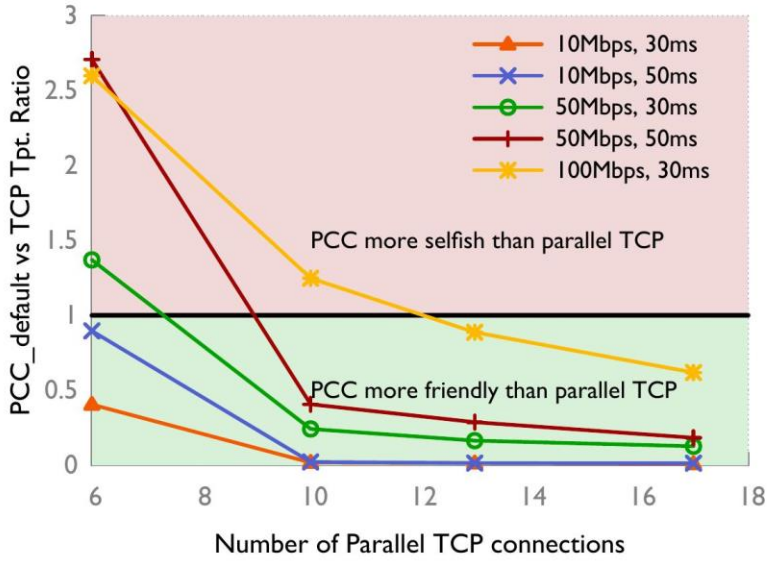
- PCC's default utility function is not TCP Friendly

But not **that** bad

TCP Friendliness

- PCC's default utility function is not TCP Friendly

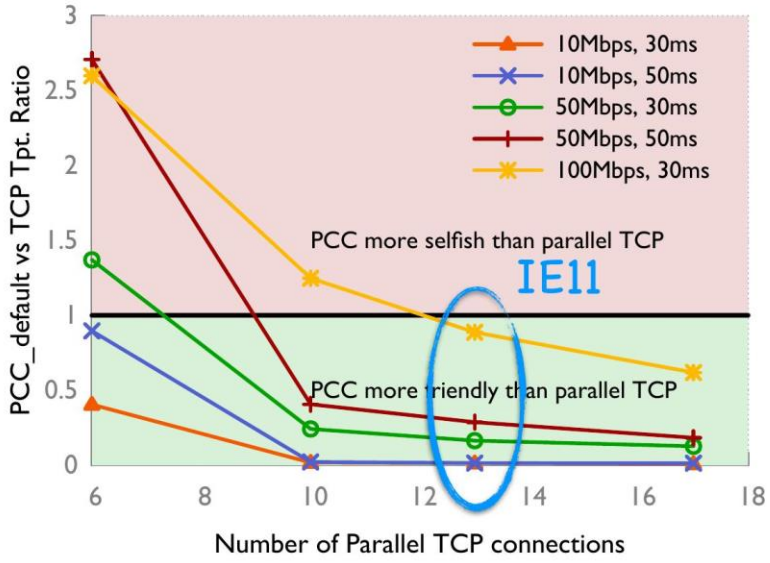
But not **that** bad



TCP Friendliness

- PCC's default utility function is not TCP Friendly

But not **that** bad



TCP Friendliness

- PCC's default utility function is not TCP Friendly
- Different utility functions can be a solution

TCP Friendliness

- PCC's default utility function is not TCP Friendly
- Different utility functions can be a solution

TCP vs TCP TCP vs PCC

TCP Friendliness

- PCC's default utility function is not TCP Friendly
- Different utility functions can be a solution

TCP Friendliness

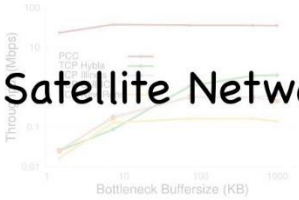
- PCC's default utility function is not TCP Friendly
- Different utility functions can be a solution

TCP vs TCP → TCP vs PCC

		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

Consistent High Performance

Consistent High Performance



Satellite Networks

Consistent High Performance

17X vs Hybla

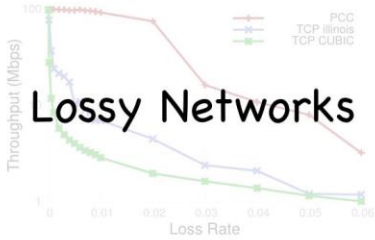
Satellite Networks

Consistent High Performance

17X vs Hybla

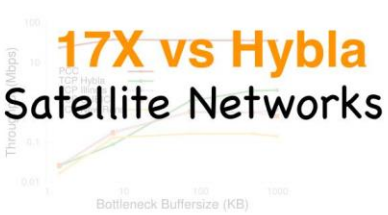
Satellite Networks

Lossy Networks

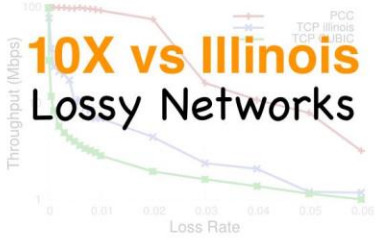


Consistent High Performance

17X vs Hybla
Satellite Networks

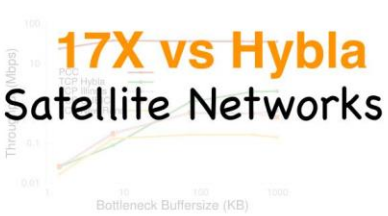


10X vs Illinois
Lossy Networks

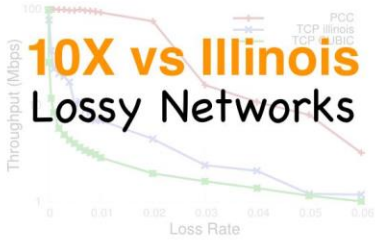


Consistent High Performance

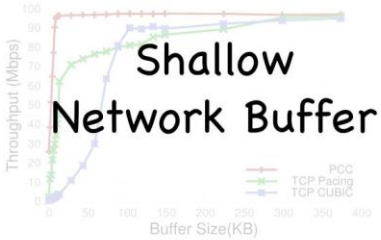
17X vs Hybla
Satellite Networks



10X vs Illinois
Lossy Networks

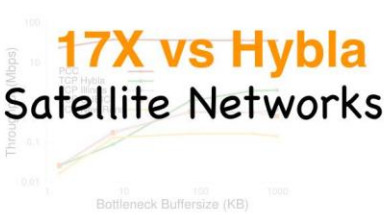


Shallow
Network Buffer

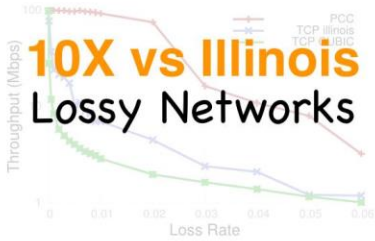


Consistent High Performance

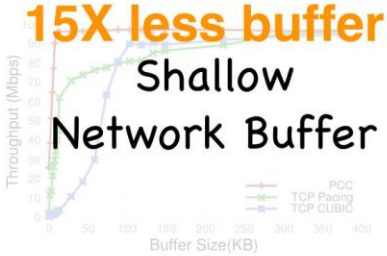
17X vs Hybla
Satellite Networks



10X vs Illinois
Lossy Networks



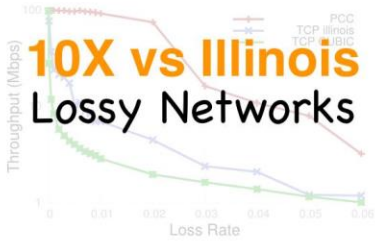
15X less buffer
Shallow
Network Buffer



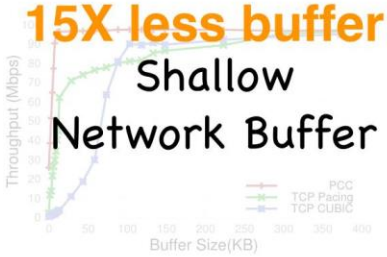
Consistent High Performance

17X vs Hybla
Satellite Networks

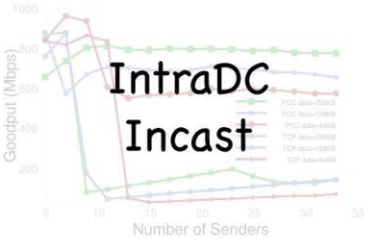
10X vs Illinois
Lossy Networks



15X less buffer
Shallow
Network Buffer



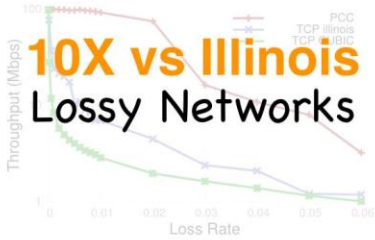
IntraDC
Incast



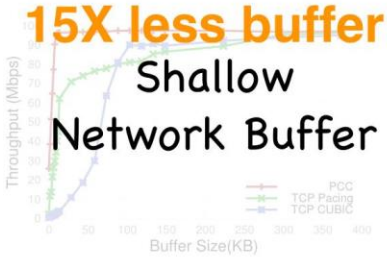
Consistent High Performance

17X vs Hybla
Satellite Networks

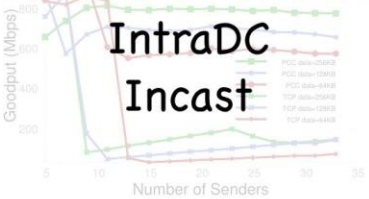
10X vs Illinois
Lossy Networks



15X less buffer
Shallow
Network Buffer



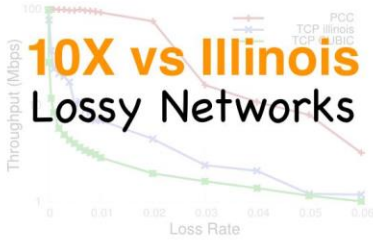
Similar to ICTCP
IntraDC
Incast



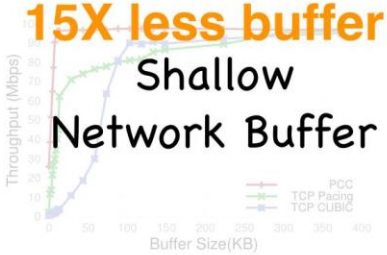
Consistent High Performance

17X vs Hybla
Satellite Networks

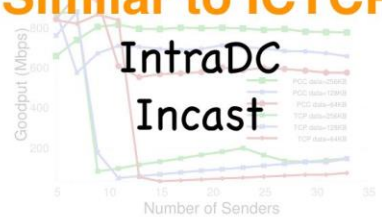
10X vs Illinois
Lossy Networks



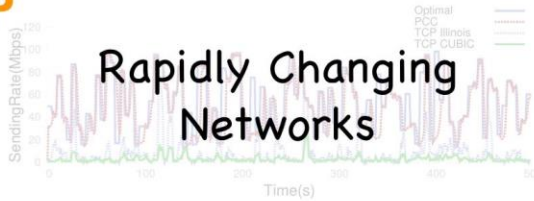
15X less buffer
Shallow
Network Buffer



Similar to ICTCP
IntraDC
Incast



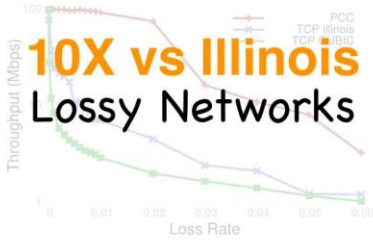
Rapidly Changing
Networks



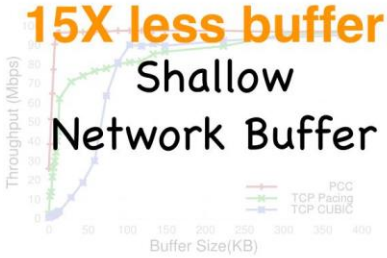
Consistent High Performance

17X vs Hybla
Satellite Networks

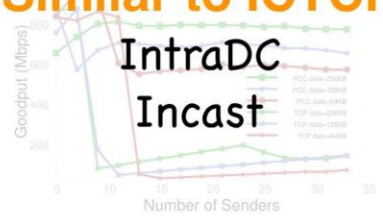
10X vs Illinois
Lossy Networks



15X less buffer
Shallow
Network Buffer



Similar to ICTCP
IntraDC
Incast

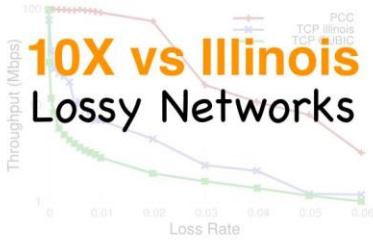


Close to Optimal
Rapidly Changing
Networks

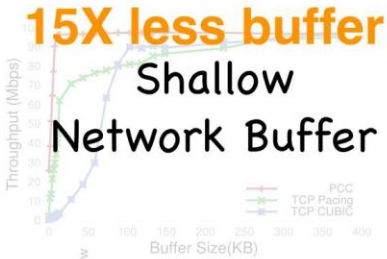
Consistent High Performance

17X vs Hybla
Satellite Networks

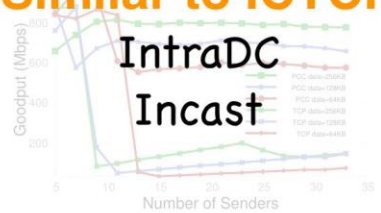
10X vs Illinois
Lossy Networks



15X less buffer
Shallow
Network Buffer

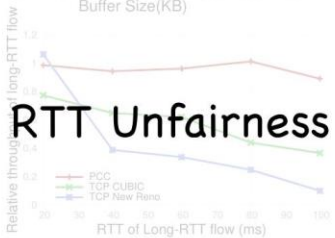


Similar to ICTCP
IntraDC
Incast



Close to Optimal
Rapidly Changing
Networks

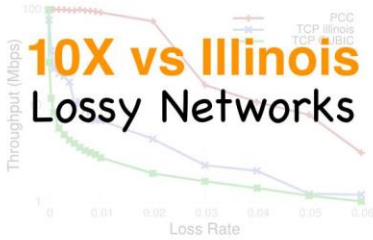
RTT Unfairness



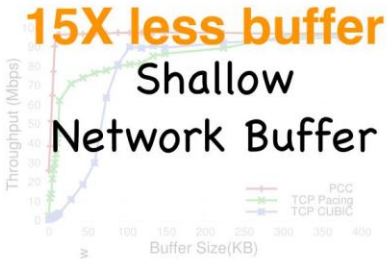
Consistent High Performance

17X vs Hybla
Satellite Networks

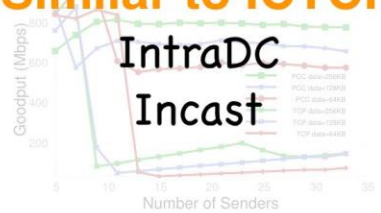
10X vs Illinois
Lossy Networks



15X less buffer
Shallow
Network Buffer



Similar to ICTCP
IntraDC
Incast



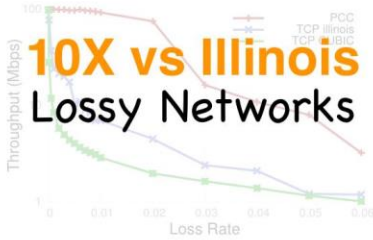
Close to Optimal
Rapidly Changing
Networks

Solves
RTT Unfairness

Consistent High Performance

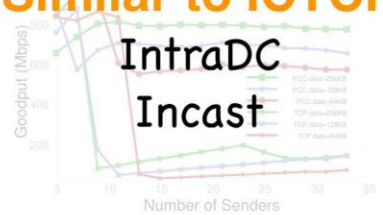
17X vs Hybla
Satellite Networks

10X vs Illinois
Lossy Networks



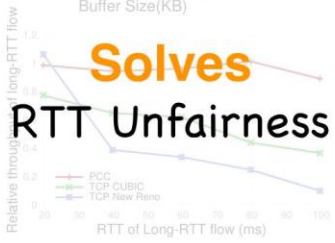
15X less buffer
Shallow
Network Buffer

Similar to ICTCP
IntraDC
Incast



Close to Optimal
Rapidly Changing
Networks

Solves
RTT Unfairness



InterDC

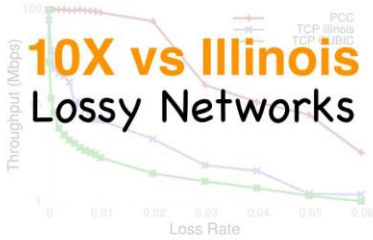
Table 1: PCC significantly outperforms TCP in inter-data center environments. RTT is in msec; throughput in Mbps.

Transmission Pair	RTT	PCC	SABUL	CUBIC	Illinois
GPO → NYSERNet	12.1	791	563	129	326
GPO → Missouri	47.4	808	664	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

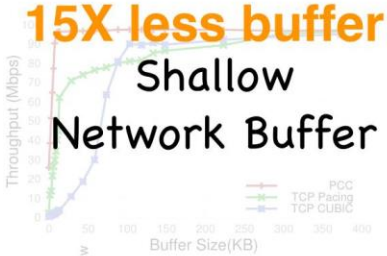
Consistent High Performance

17X vs Hybla
Satellite Networks

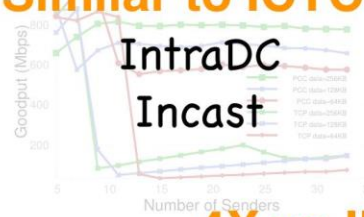
10X vs Illinois
Lossy Networks



15X less buffer
Shallow
Network Buffer

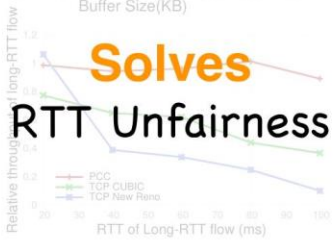


Similar to ICTCP
IntraDC
Incast



Close to Optimal
Rapidly Changing
Networks

Solves
RTT Unfairness



4X vs Illinois
1.23X vs UDT
InterDC

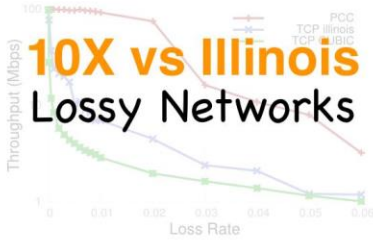
Table 1: PCC significantly outperforms TCP in inter-data center environments. RTT is in msec; throughput in Mbps.

GPO → NYSERNet	47.4	816	662	129	326
GPO → Missouri	9.01	801	700	80.7	90.1
GPO → Illinois	35.4	796	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	265
NYSERNet → Illinois	36.1	808	674	141	141

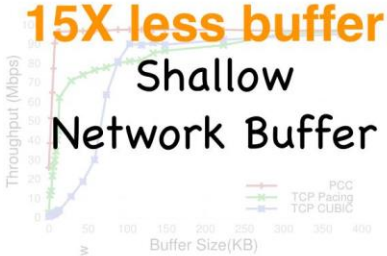
Consistent High Performance

17X vs Hybla
Satellite Networks

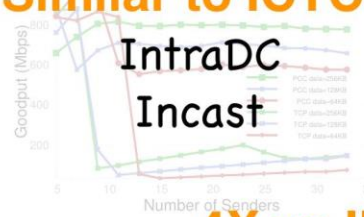
10X vs Illinois
Lossy Networks



15X less buffer
Shallow
Network Buffer

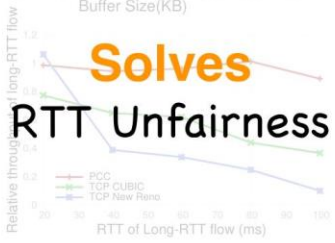


Similar to ICTCP
IntraDC
Incast



Close to Optimal
Rapidly Changing
Networks

Solves
RTT Unfairness



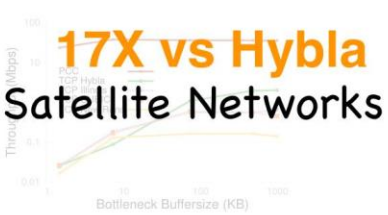
4X vs Illinois
1.23X vs UDT
InterDC


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

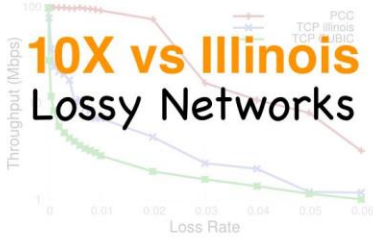
GPO → NYSERNet	47.4	816	662	129	326
GPO → Missouri	9.01	801	700	80.7	90.1
GPO → Illinois	35.4	796	664	84.5	102
NYSERNet → Missouri	47.4	816	662	108	109
Wisconsin → Illinois	9.01	801	700	547	562
GPO → Wisc.	38.0	783	487	79.3	120
NYSERNet → Wisc.	38.3	791	673	134	134
Missouri → Wisc.	20.9	807	698	259	265
NYSERNet → Illinois	36.1	808	674	141	141

Consistent High Performance

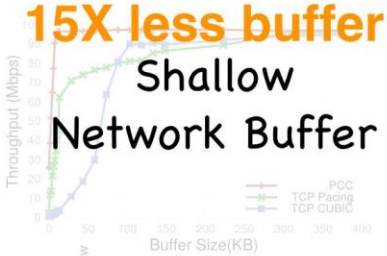
17X vs Hybla
Satellite Networks



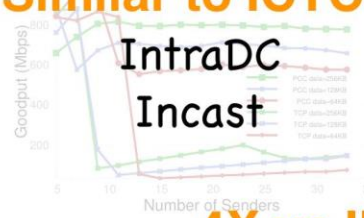
10X vs Illinois
Lossy Networks



15X less buffer
Shallow
Network Buffer

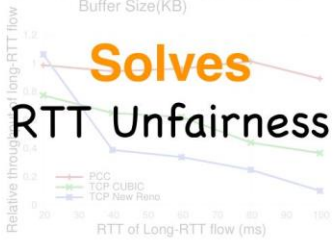


Similar to ICTCP
IntraDC
Incast



Close to Optimal
Rapidly Changing
Networks

Solves
RTT Unfairness

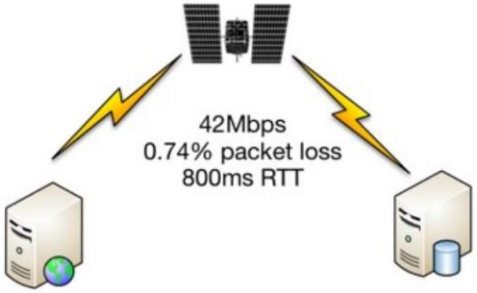


4X vs Illinois
1.23X vs UDT
InterDC

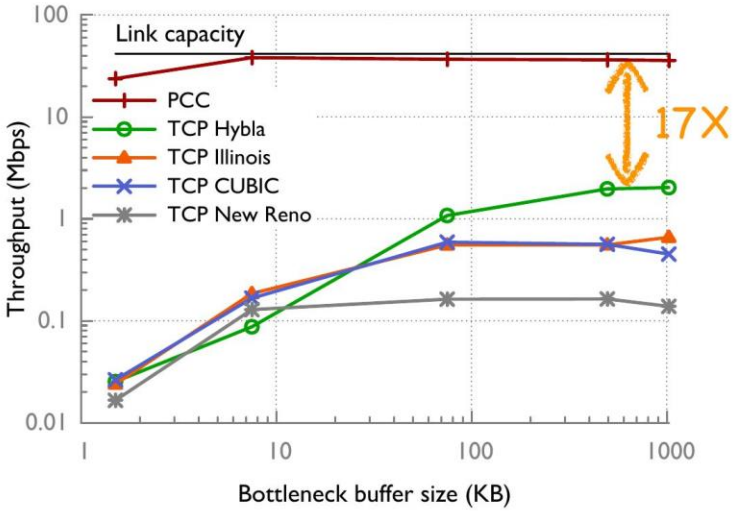
4X median vs CUBIC
1.48X vs UDT + 4X less loss
Global Commercial
Internet

Consistent High Performance

Satellite Network



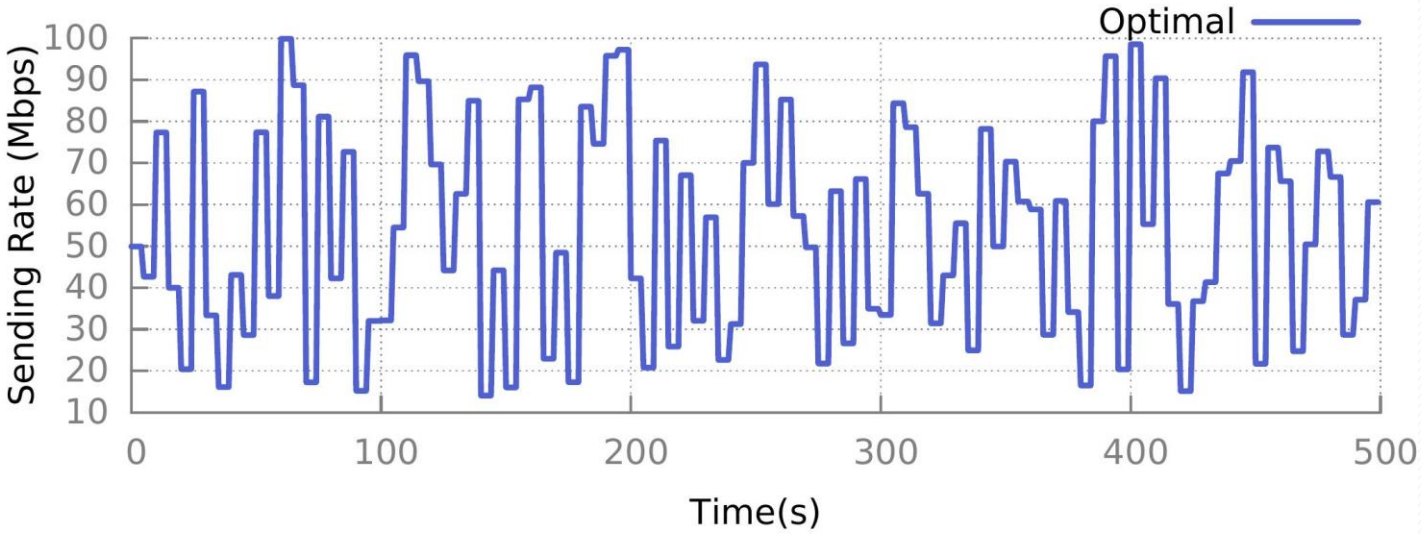
WINDS System



Consistent High Performance

Rapidly Changing Networks

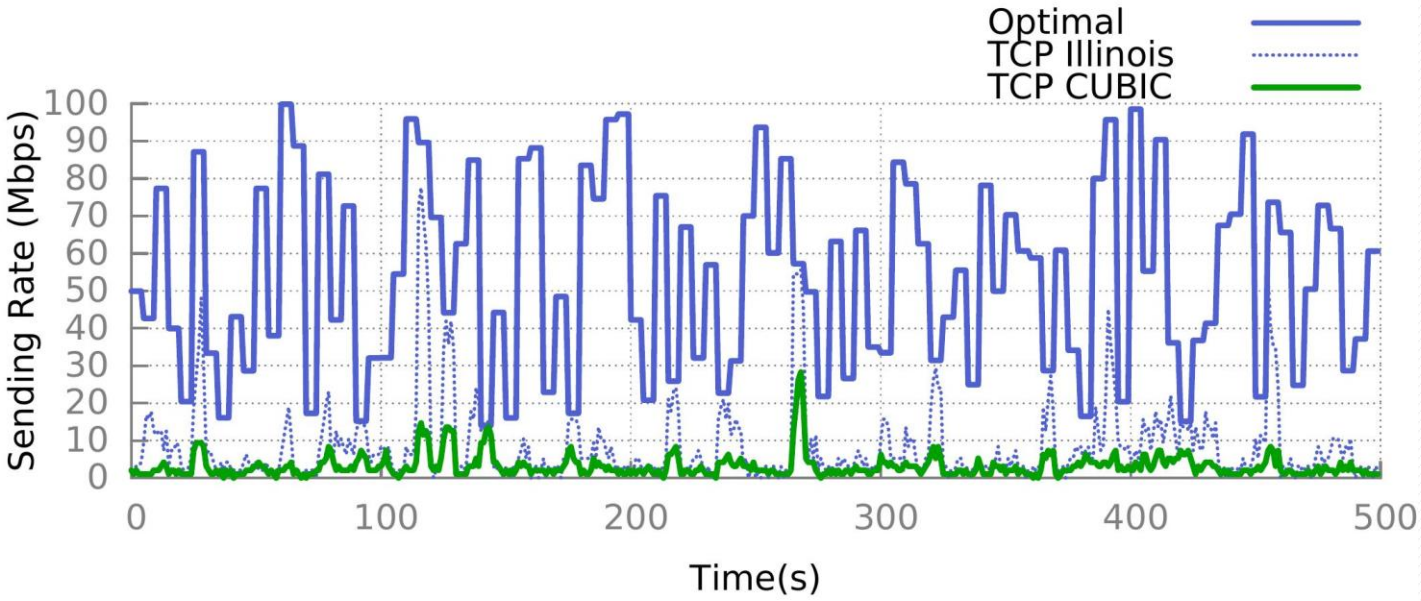
BW: 10-100Mbps; RTT: 10-100ms; Loss Rate: 0-1%
Change every 5 seconds



Consistent High Performance

Rapidly Changing Networks

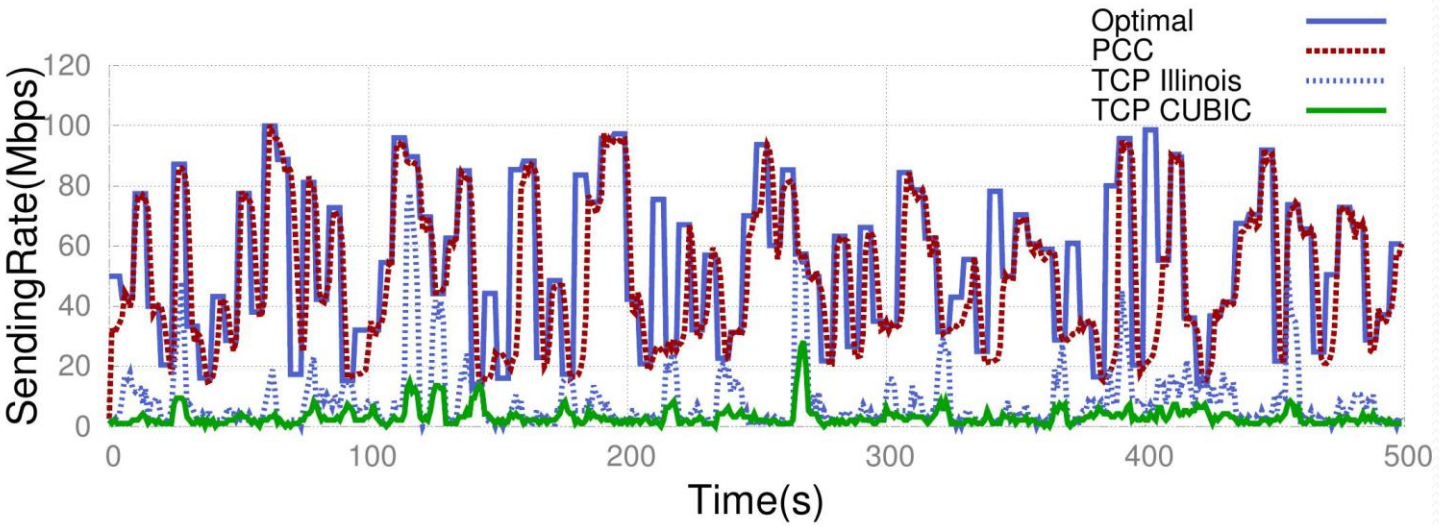
BW: 10-100Mbps; RTT: 10-100ms; Loss Rate: 0-1%
Change every 5 seconds

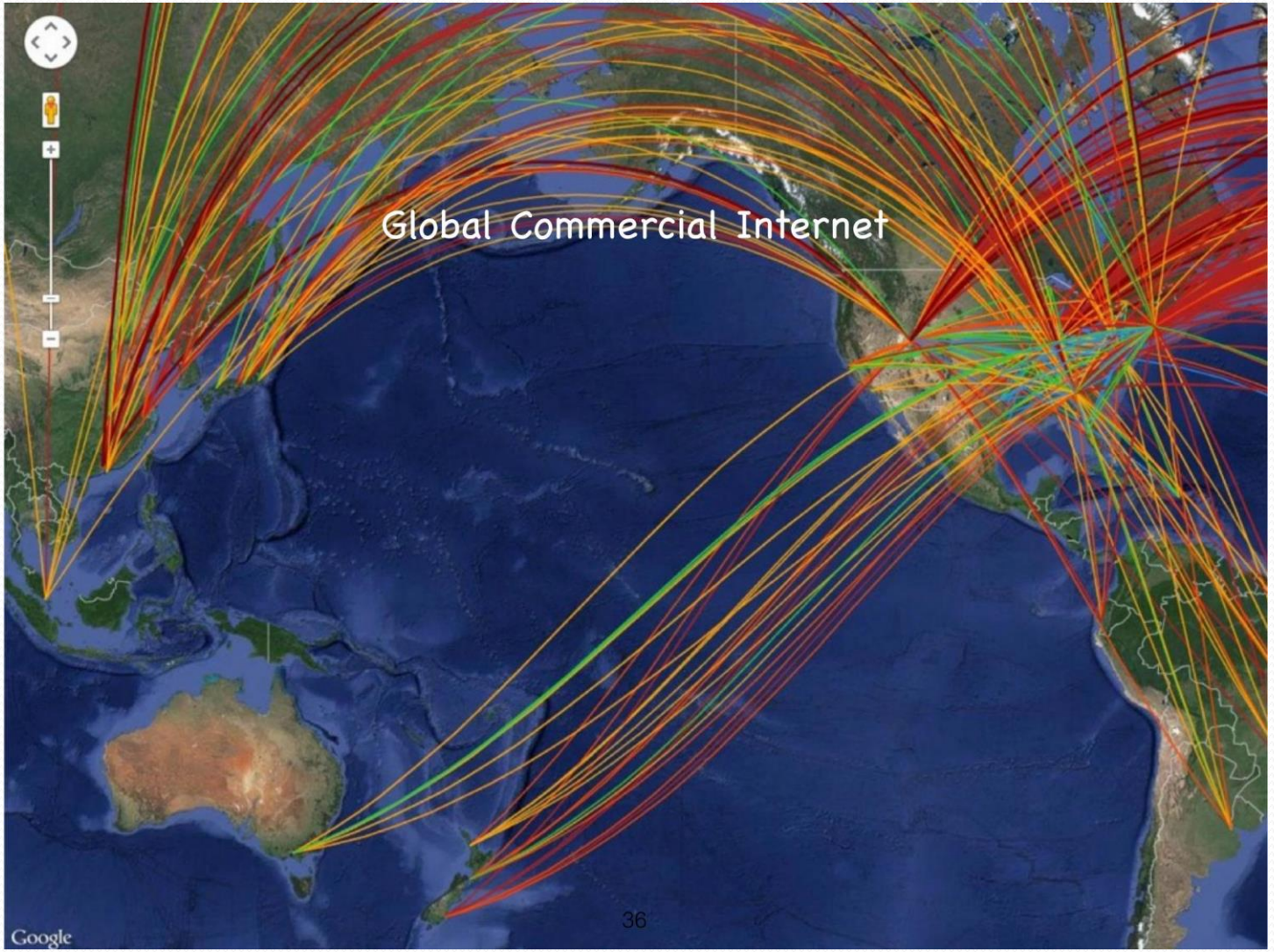


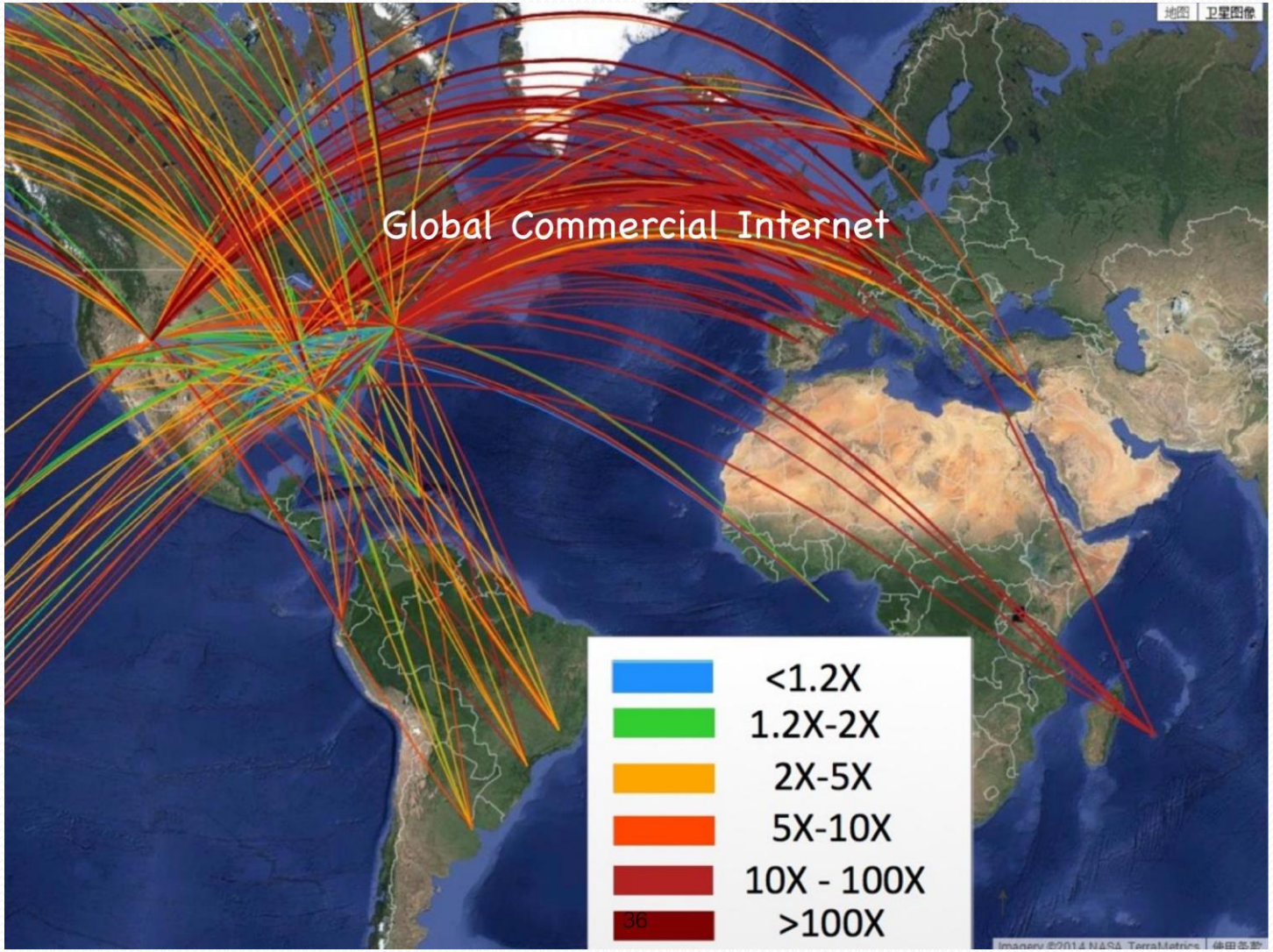
Consistent High Performance

Rapidly Changing Networks

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



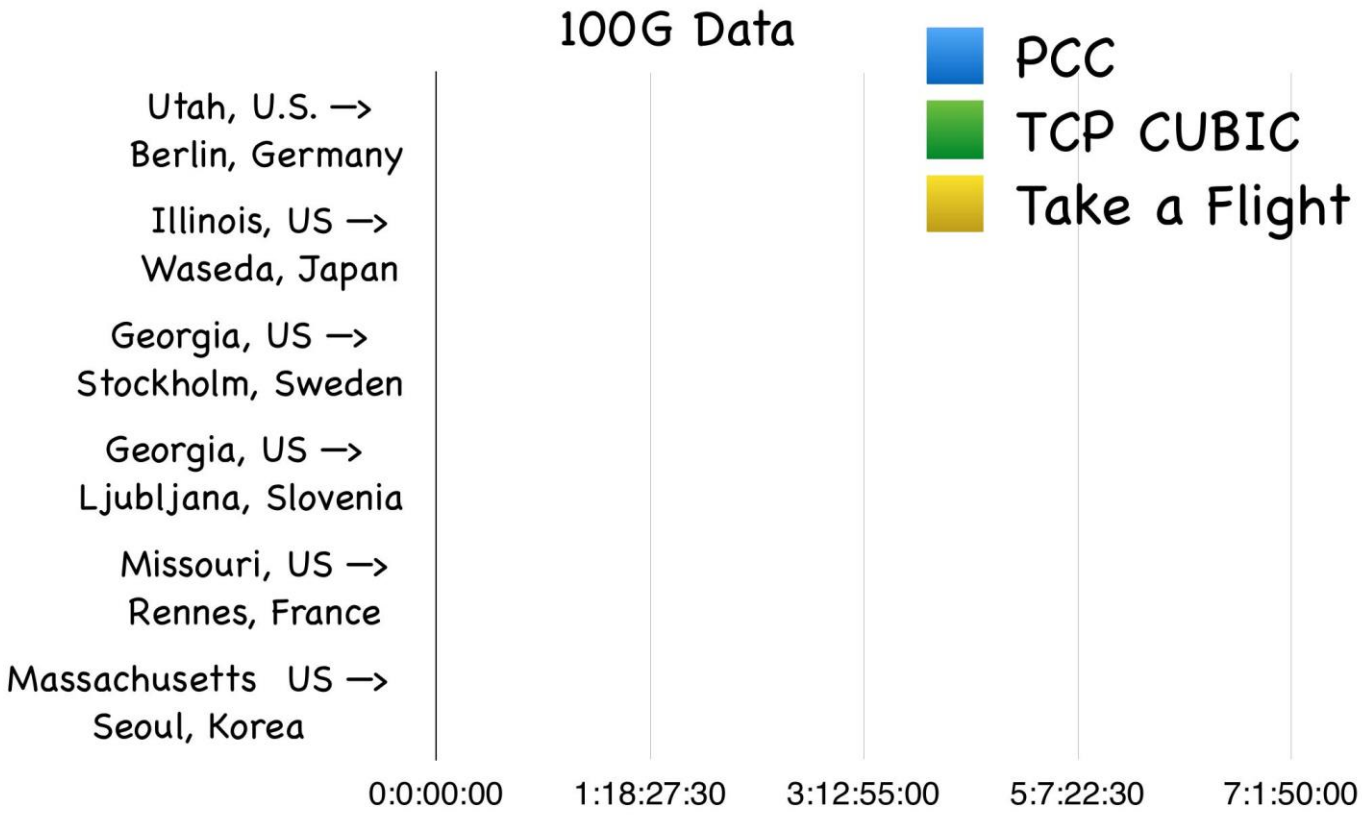

Global Commercial Internet
Google


Global Commercial Internet
<1.2X
1.2X-2X
2X-5X
5X-10X
10X - 100X
>100X

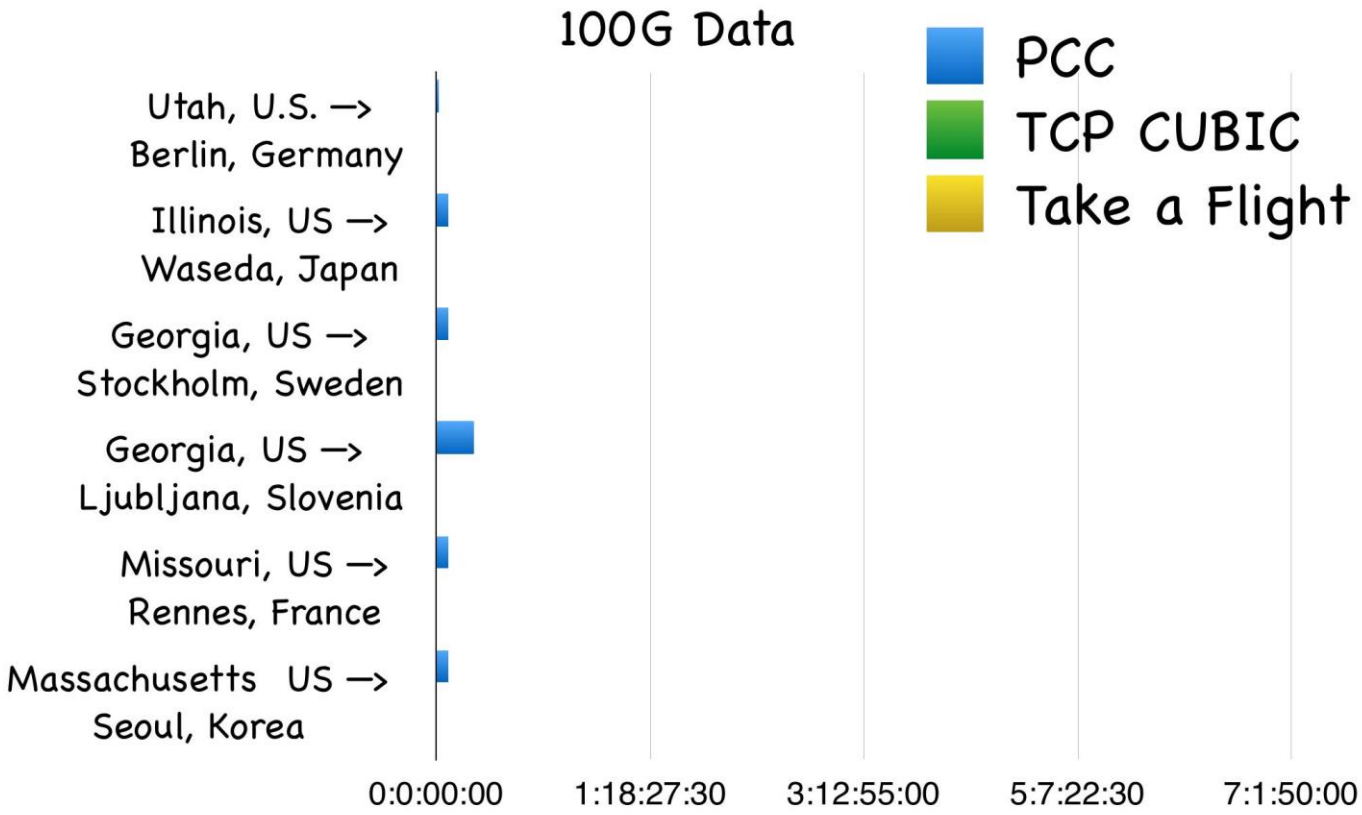
PCC vs TCP vs Take a Flight

100G Data

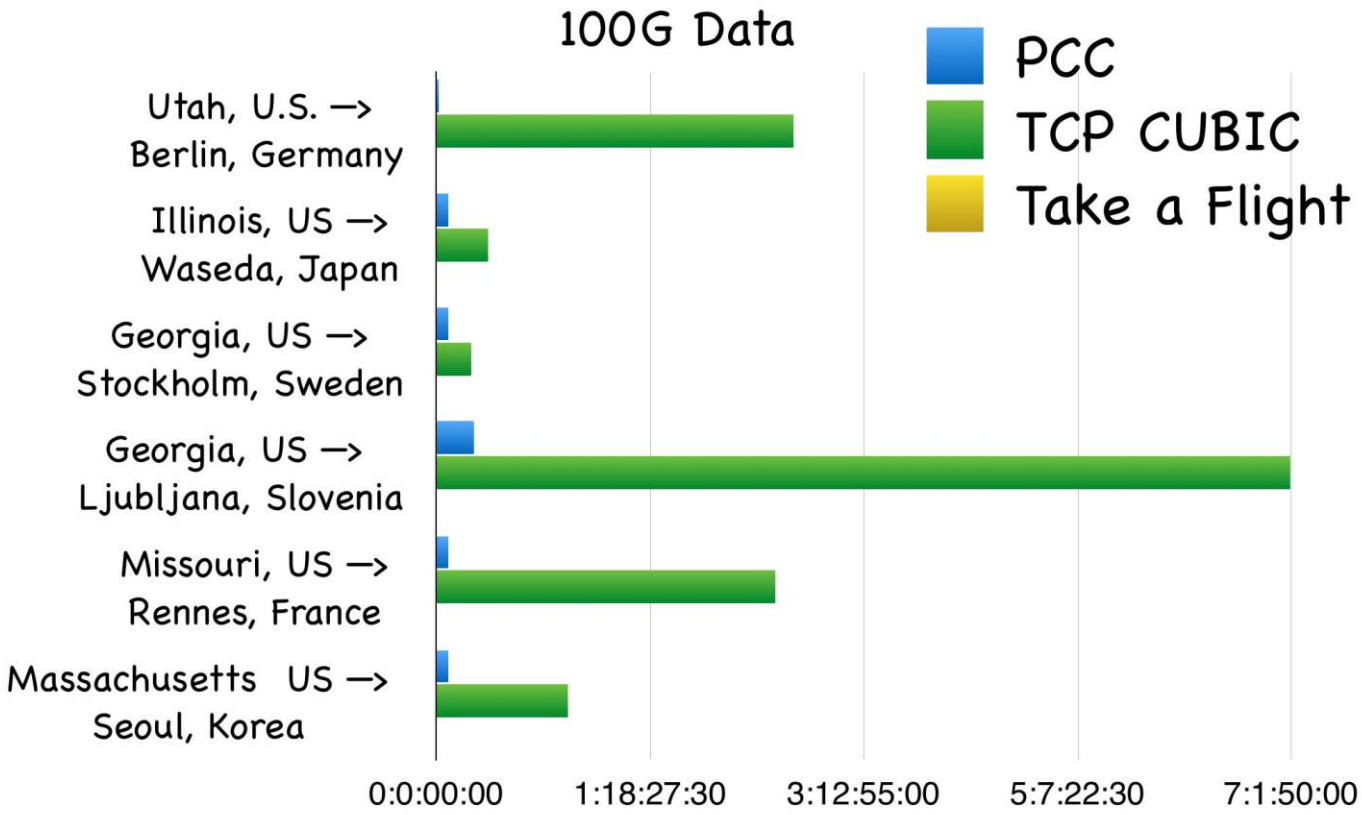
PCC vs TCP vs Take a Flight



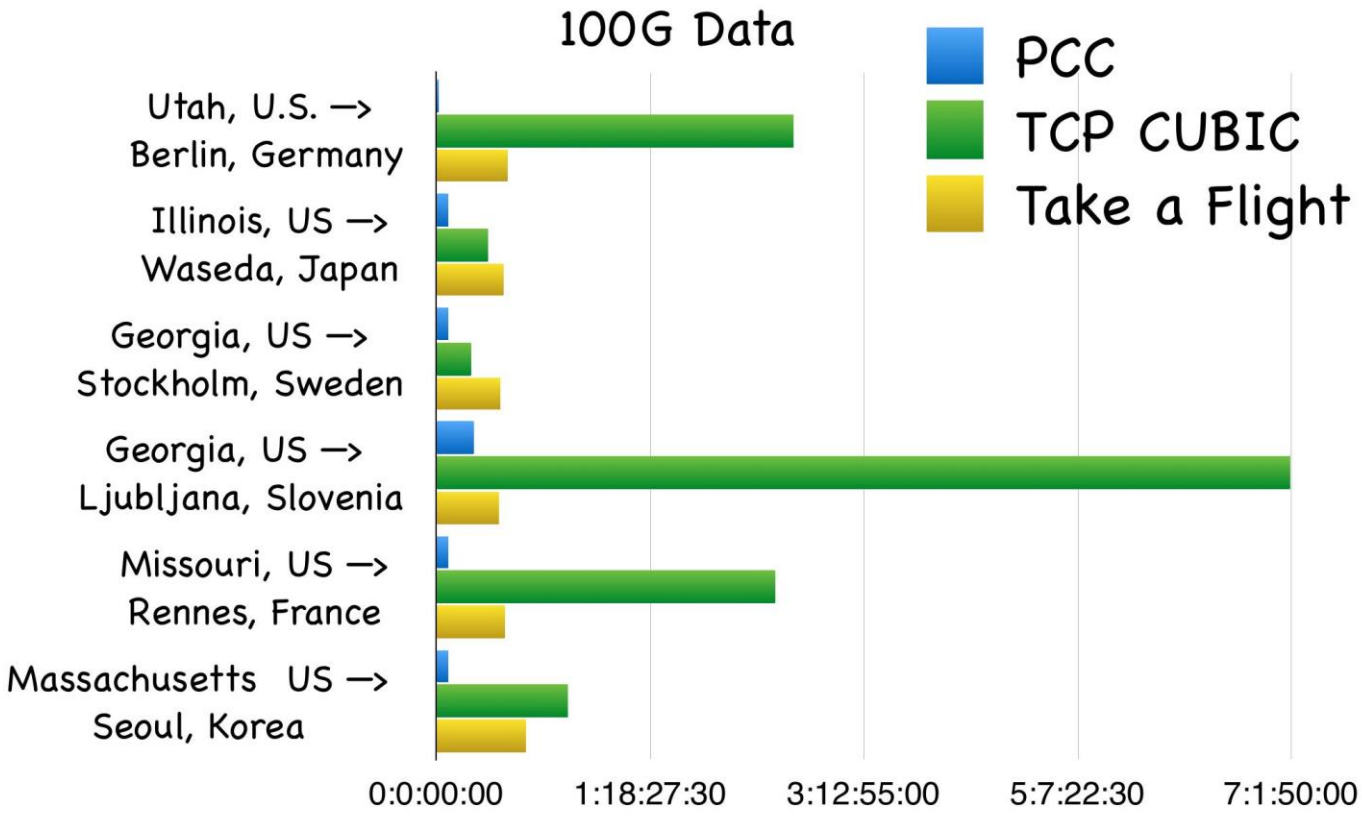
PCC vs TCP vs Take a Flight



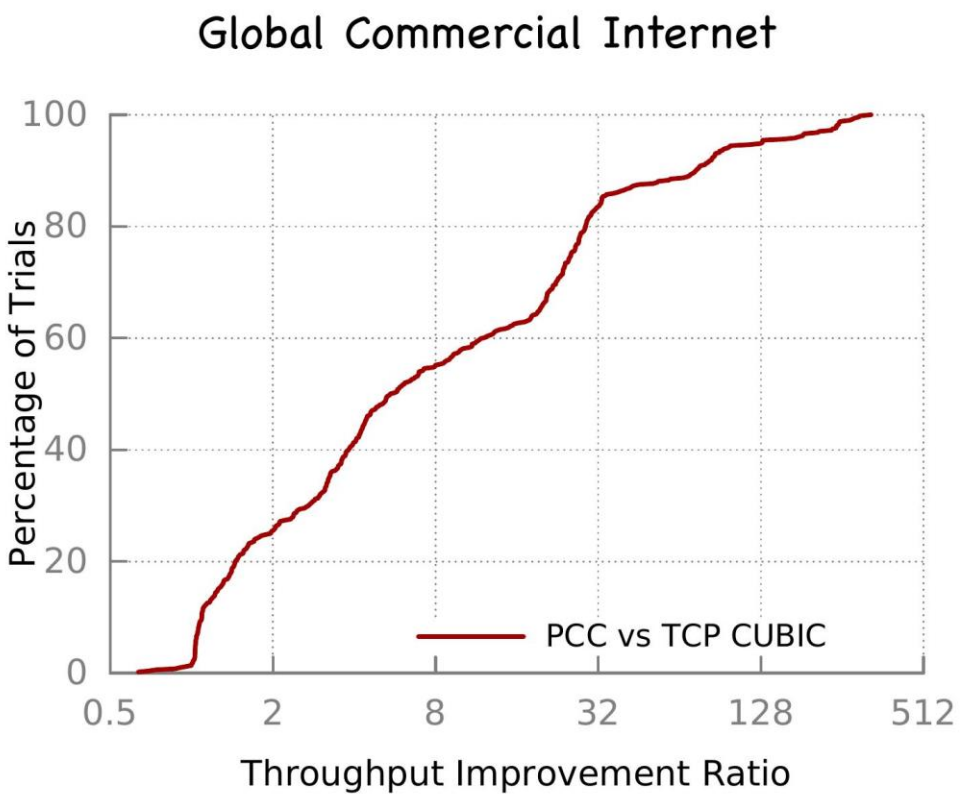
PCC vs TCP vs Take a Flight



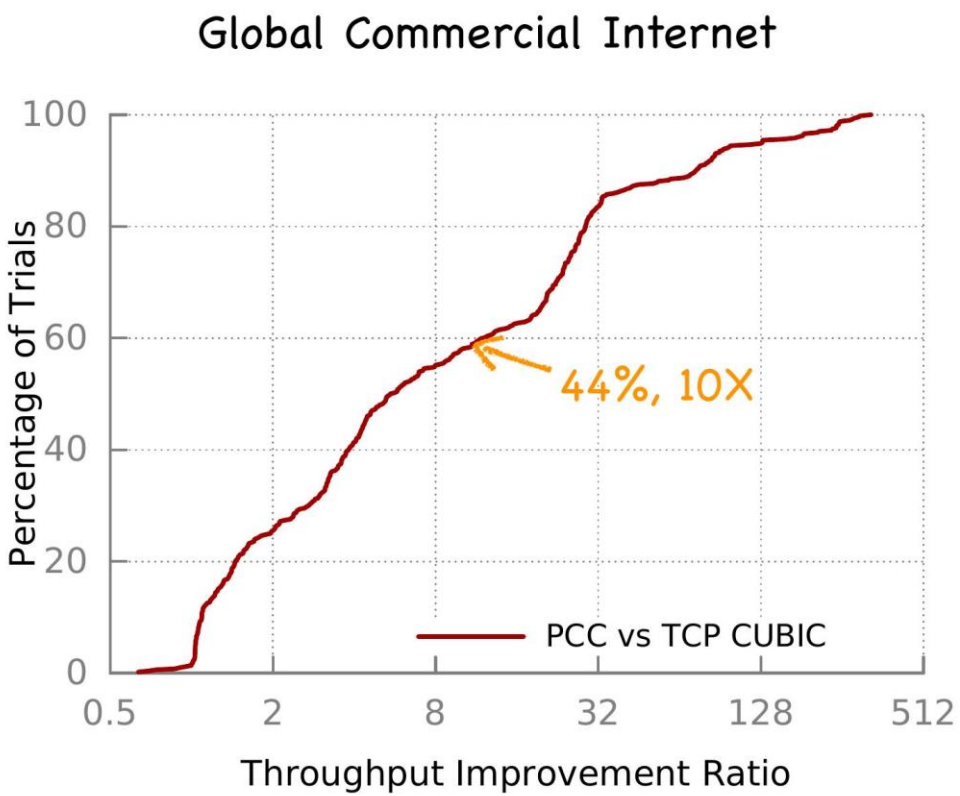
PCC vs TCP vs Take a Flight



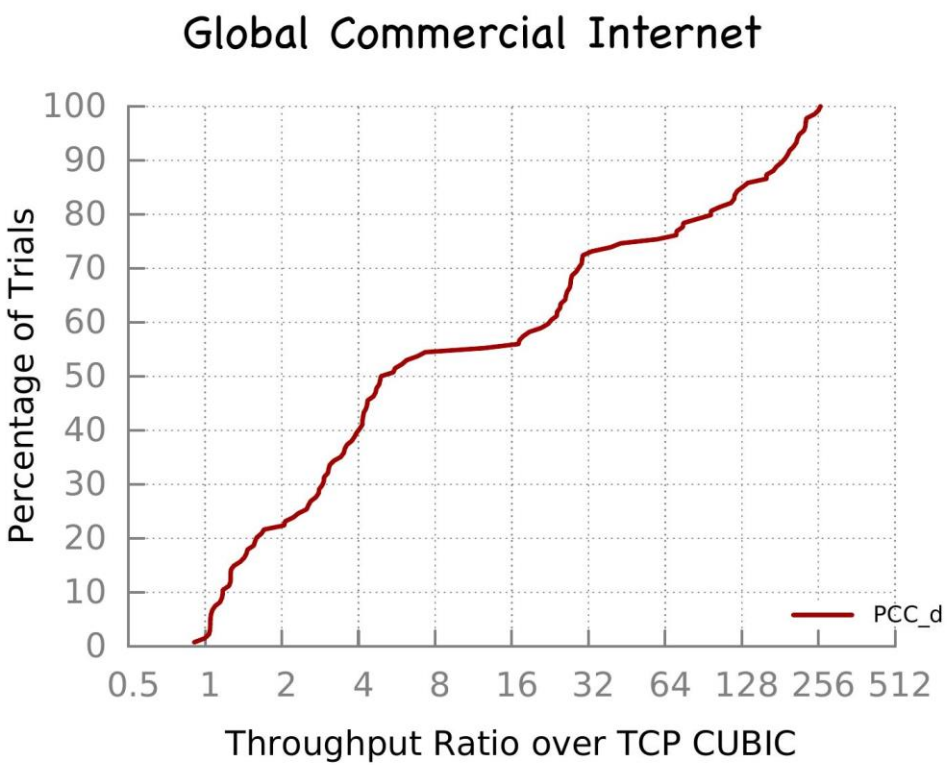
Consistent High Performance



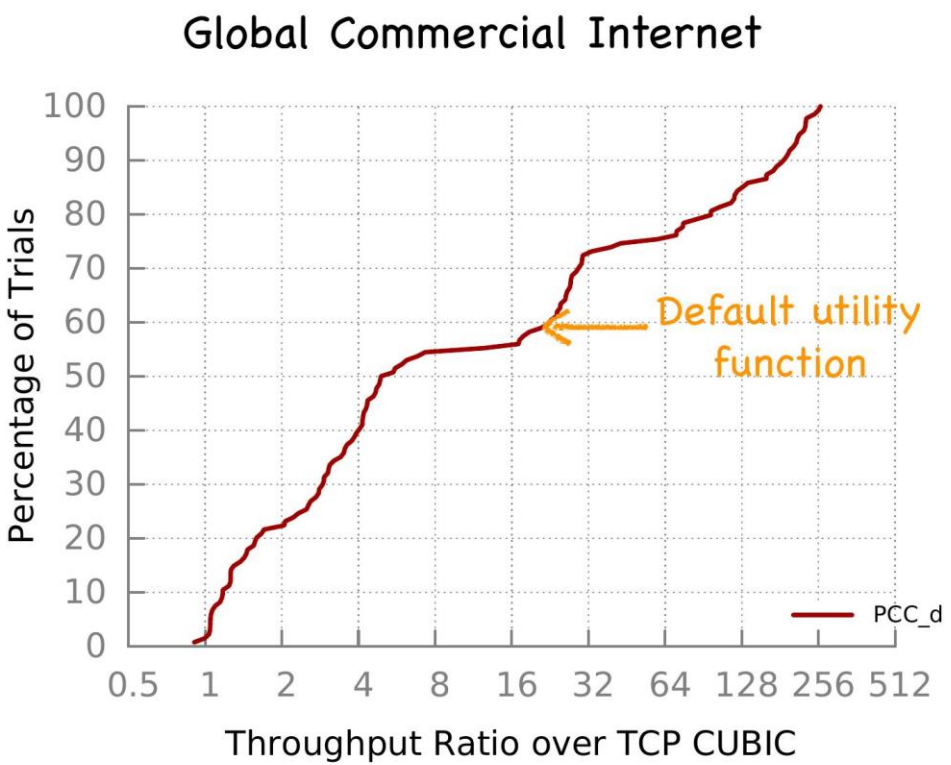
Consistent High Performance



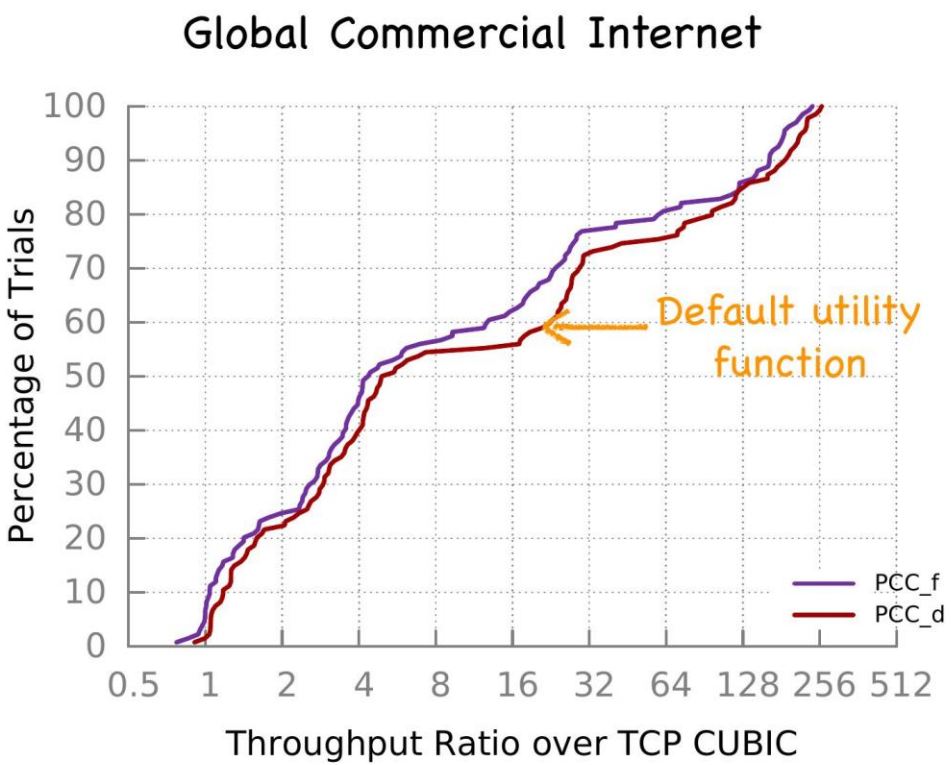
Consistent High Performance



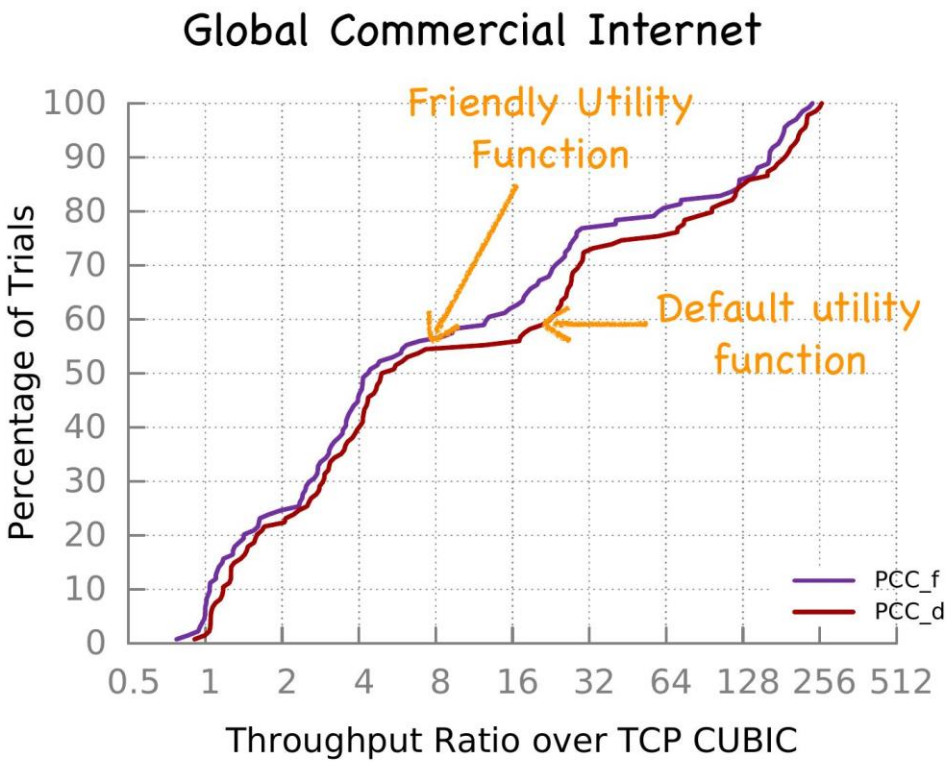
Consistent High Performance



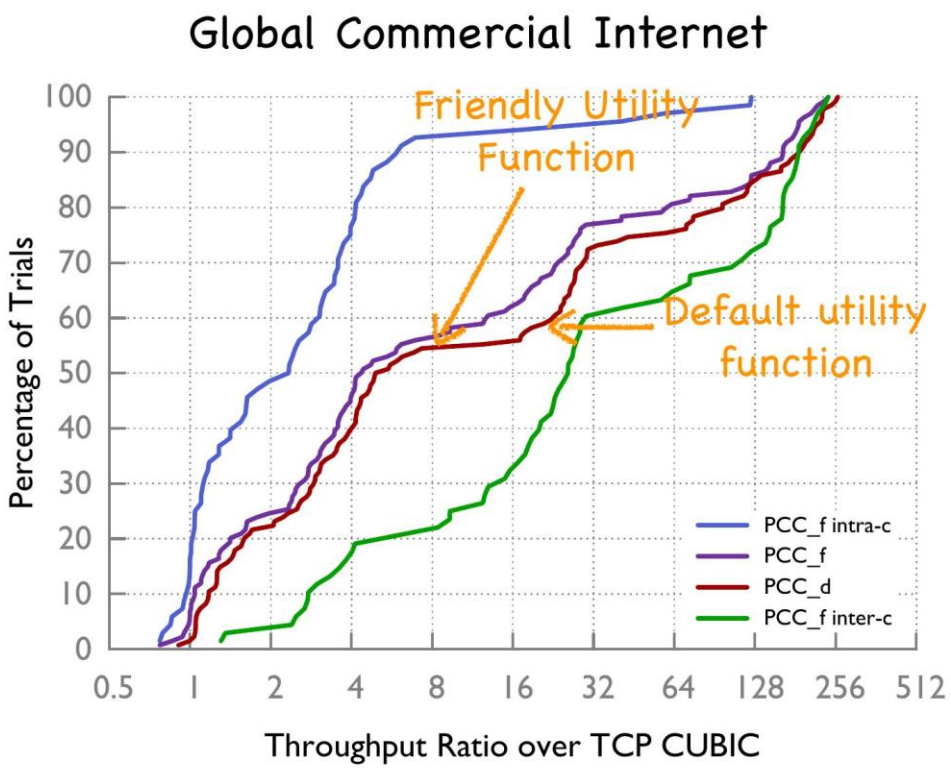
Consistent High Performance



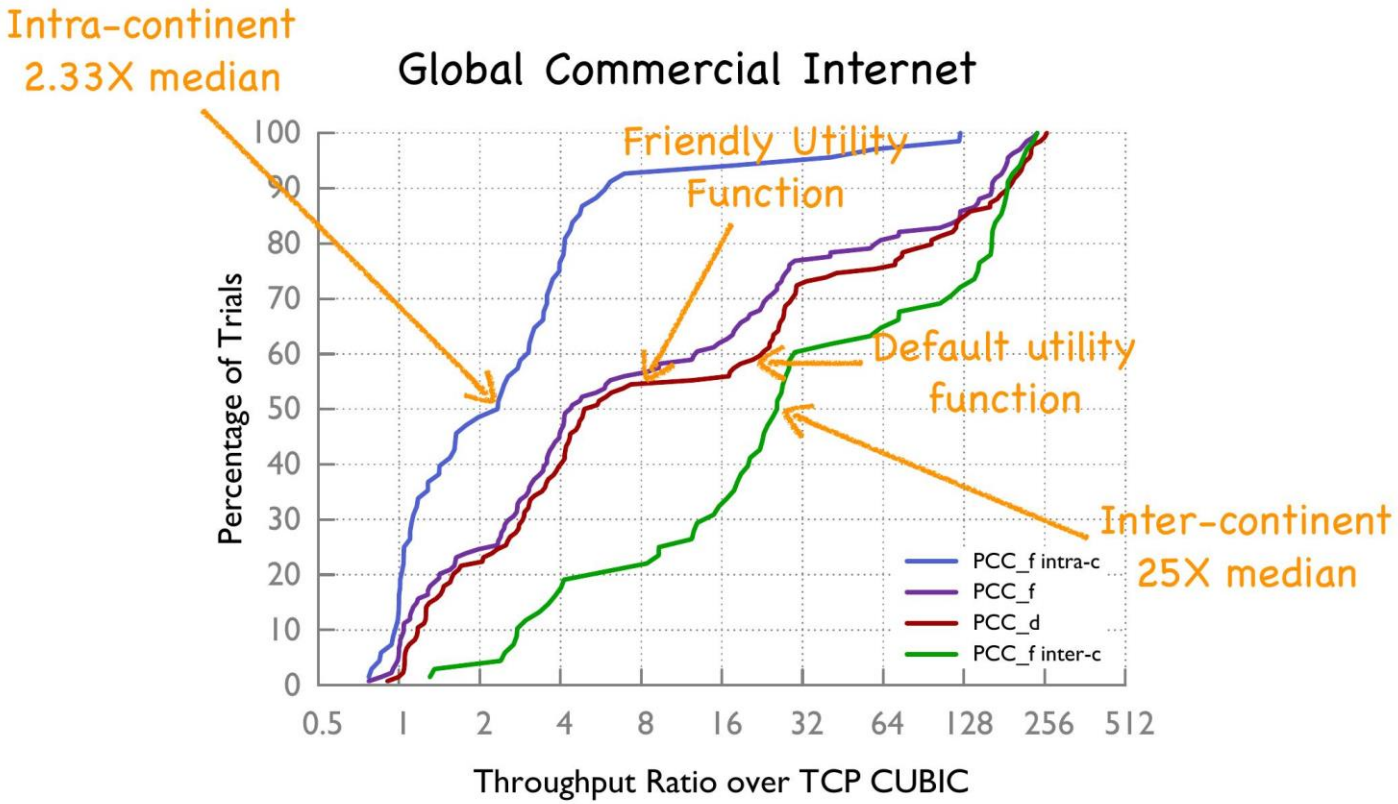
Consistent High Performance



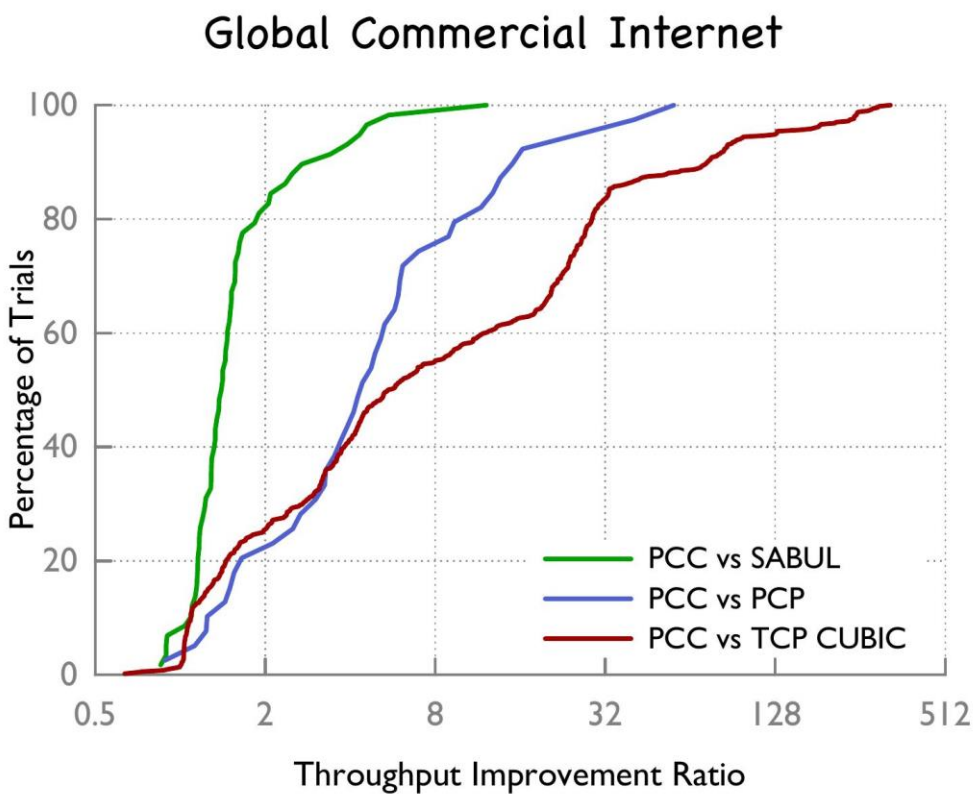
Consistent High Performance



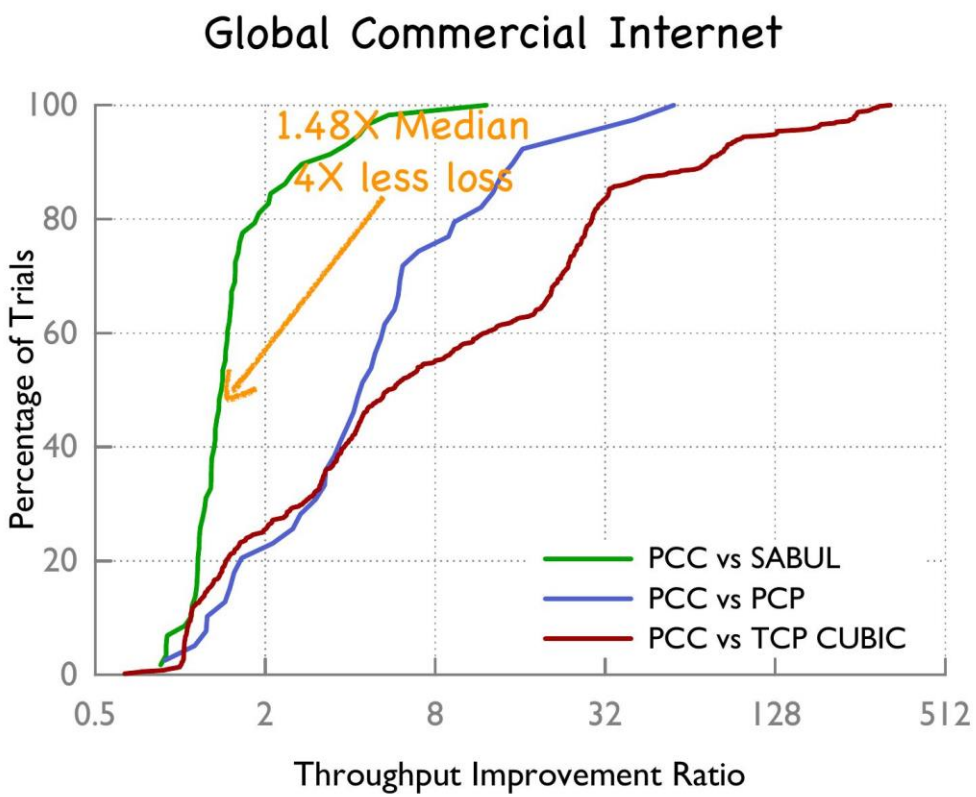
Consistent High Performance



Consistent High Performance



Consistent High Performance



Same Rate Control Algorithm

Same Rate Control Algorithm

+

Same Rate Control Algorithm

+

Different Utility Function

=

Flexibility

Limitation and Future Work

Limitation and Future Work

- Understanding of Utility Function and Convergence
- Better Control Algorithm
- Better Scalability