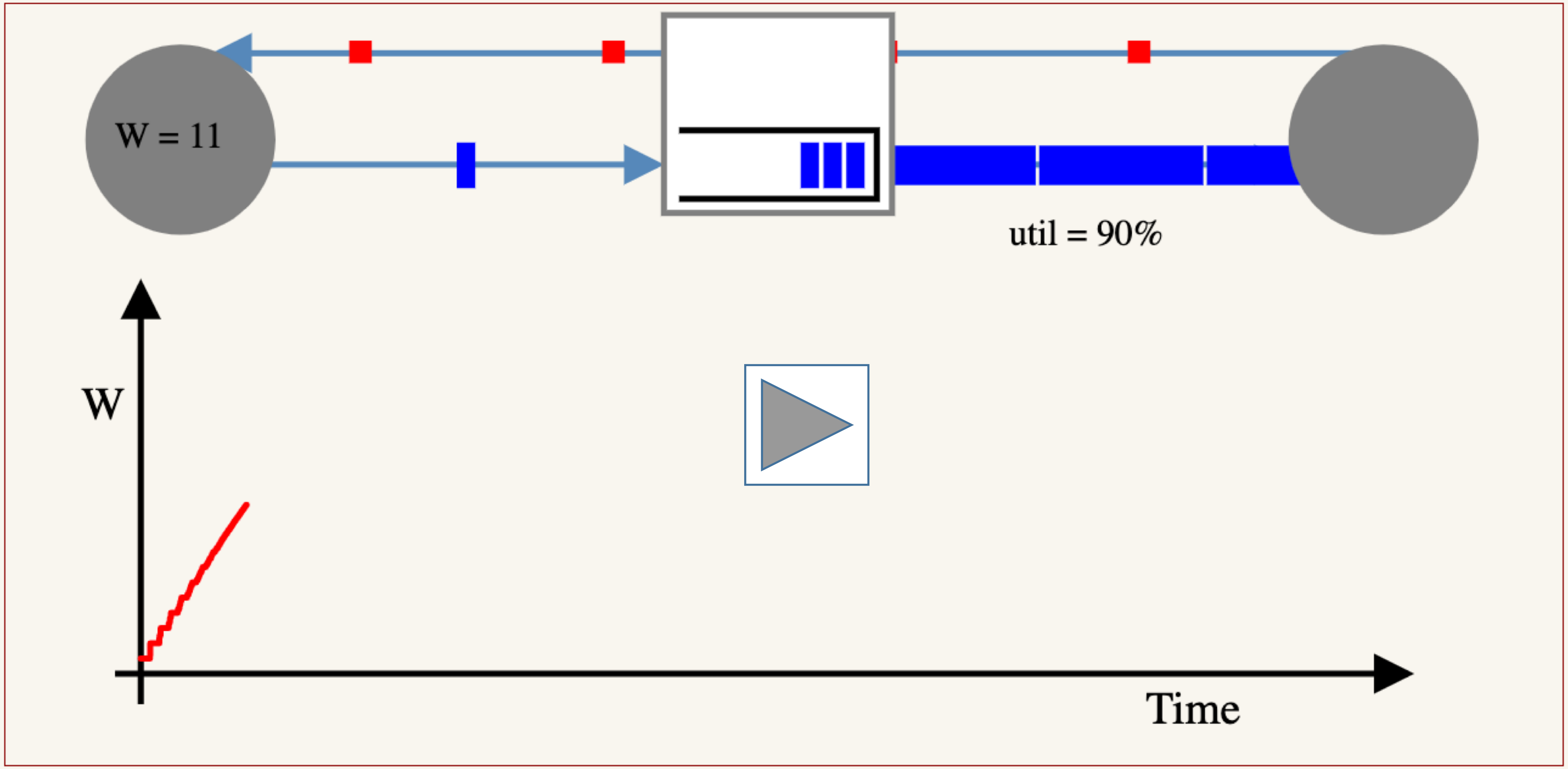

W = 11
util = 90%
W
Time

Why this meeting?

Background

- Research papers on experiments and theories of buffer size.
- Yet, no universal agreement on how big router buffers should be, and why.
- Personal confession: I have no idea what the general answer is
  - Incast
  - Data centers
  - For specific environments, like financial networks, SLAs, HPC, …

Our goal

- A workshop in October/November 2019: *“How Big Should Buffers be in Switches and Routers?”*
- Measurements: Invite operators of large networks to perform experiments in their networks.
- Theory: Invite researchers to develop theory explaining, supporting (challenging?) measurements.
- Report results publicly at workshop.
- Compare notes and write a report together, sharing our results to the world.

Organizers

1. Neda Beheshti
2. Christophe Diot
3. Tom Edsall
4. Nasser El-Aawar
5. Yashar Ganjali
6. Nick McKeown
7. Bruce Spang

Local logistics:
Andi Villanueva

Who we are

- Speakers from 14 companies and 2 universities
  - Network operators, cloud companies, router vendors, chip vendors
- Attendees from 22 companies and 2 universities
- Let's introduce ourselves…

Schedule for the day

10.30am – 1.30pm

Session 1: Network Operators

- Neda Beheshti Facebook
- Lincoln Dale Google
- TY Huang Netflix
- Honqqiang Liu Alibaba
- Ken Duell AT&T
- Joel Jaeggli Fastly

[12.00 – 12.30 Lunch]

- Simon Leinen Switch
- Bob Briscoe CableLabs
- Chuanxiong Guo Bytedance
- Igor Gashinsky Oath

1.45pm – 2.45pm

Session 2: Technology Providers

- Parvin Taheri Cisco
- Francois Labonte Arista
- Golan Schzukin Dune/BCM
- Chang Kim Barefoot

3.00pm – 4.00pm

Session 3: Discussion

- Conclusions Neda, Bruce, Nasser
- Actions and Next Steps Yashar, Nick

A brief history of buffer size

1988

Congestion Avoidance and Control

VJ & MK

1994

High Performance TCP in ANSNET

CV & CS

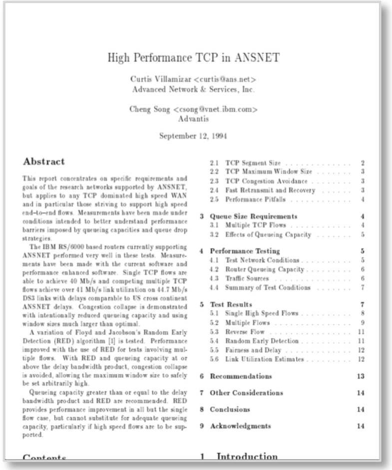
2019


Congestion Avoidance and Control*

Van Jacobson†
Lawrence Berkeley Laboratory
Michael J. Karels‡
University of California at Berkeley

November, 1988

Introduction


High Performance TCP in ANSNET

Curtis Villamizar <curtis@ans.net>
Advanced Network & Services, Inc.

Cheng Song <csong@vnet.ibm.com>
Advantis

September 12, 1994

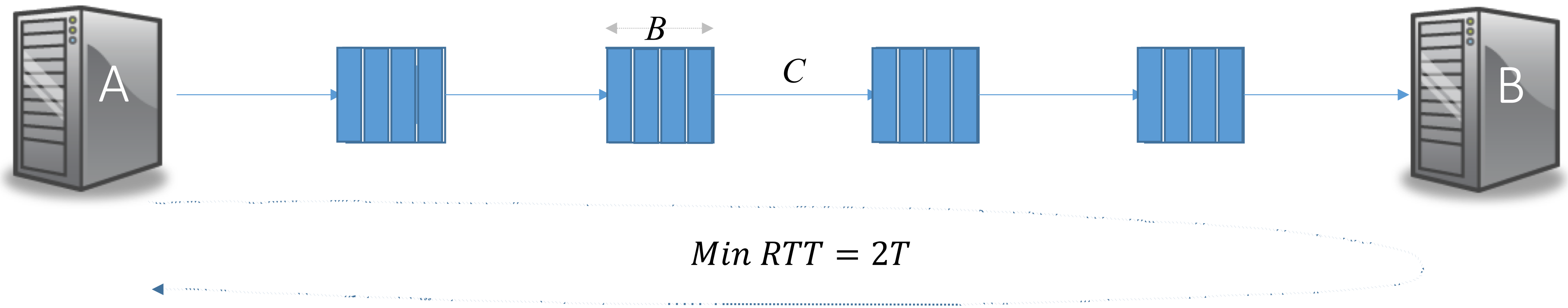
Abstract

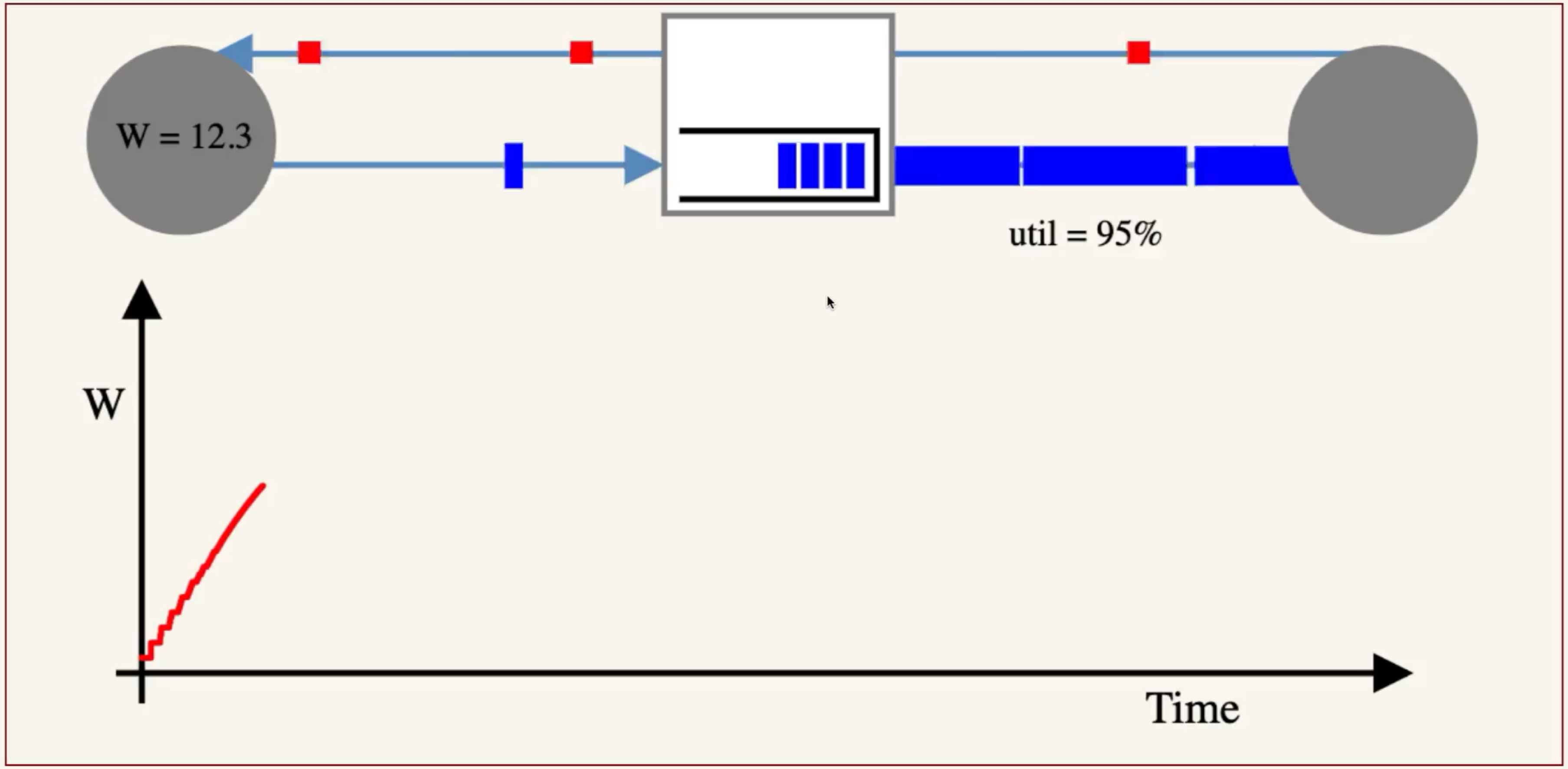
$$B = 2T \times C$$

"Buffer size should equal the bandwidth delay product"

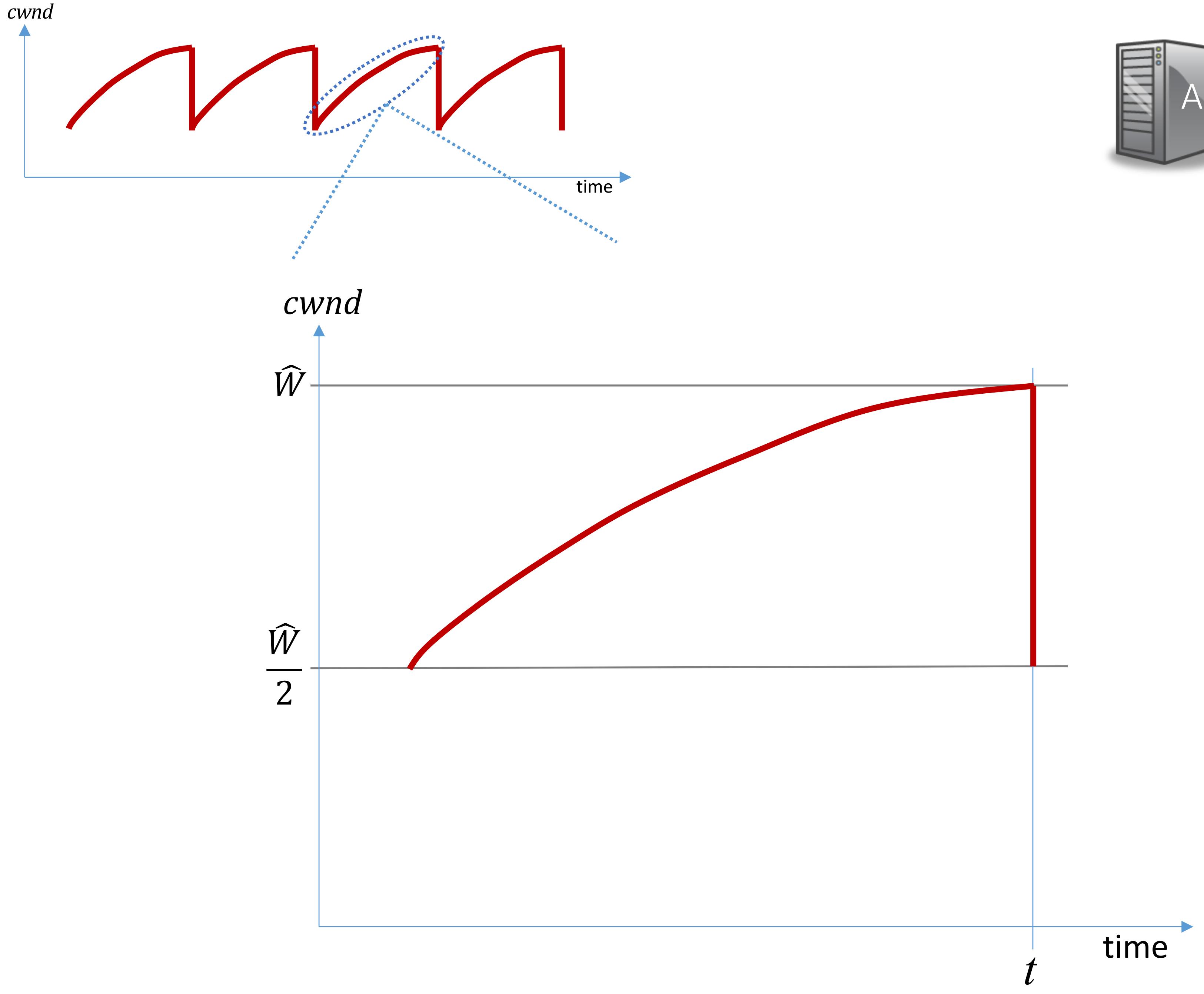
~~$Max\ RTT = 2T + B/C = 4T$~~

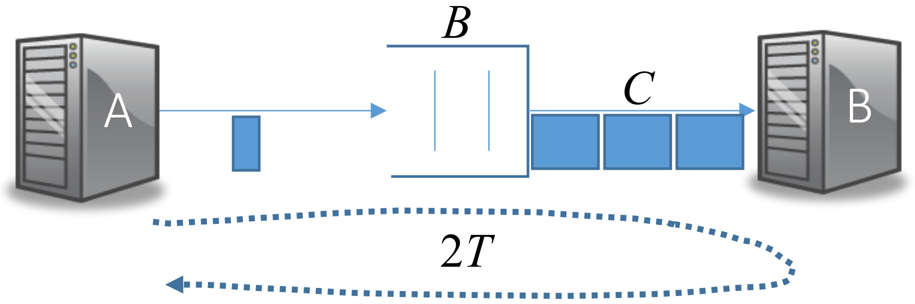
$$Max\ RTT = 2T + 4B/C = 10T$$

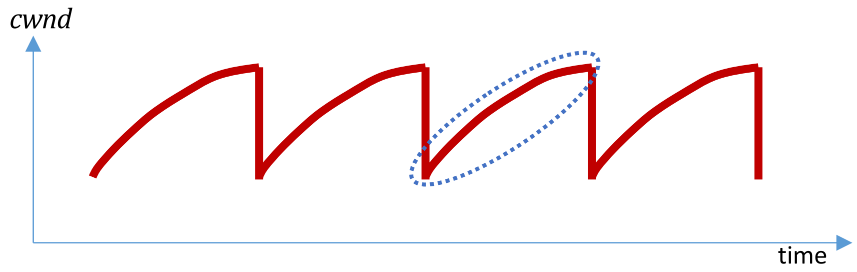


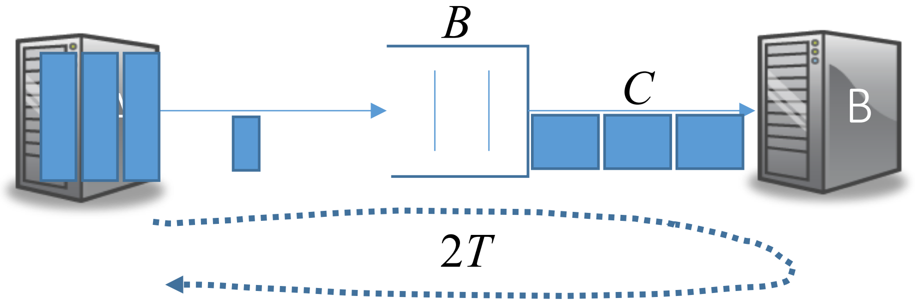


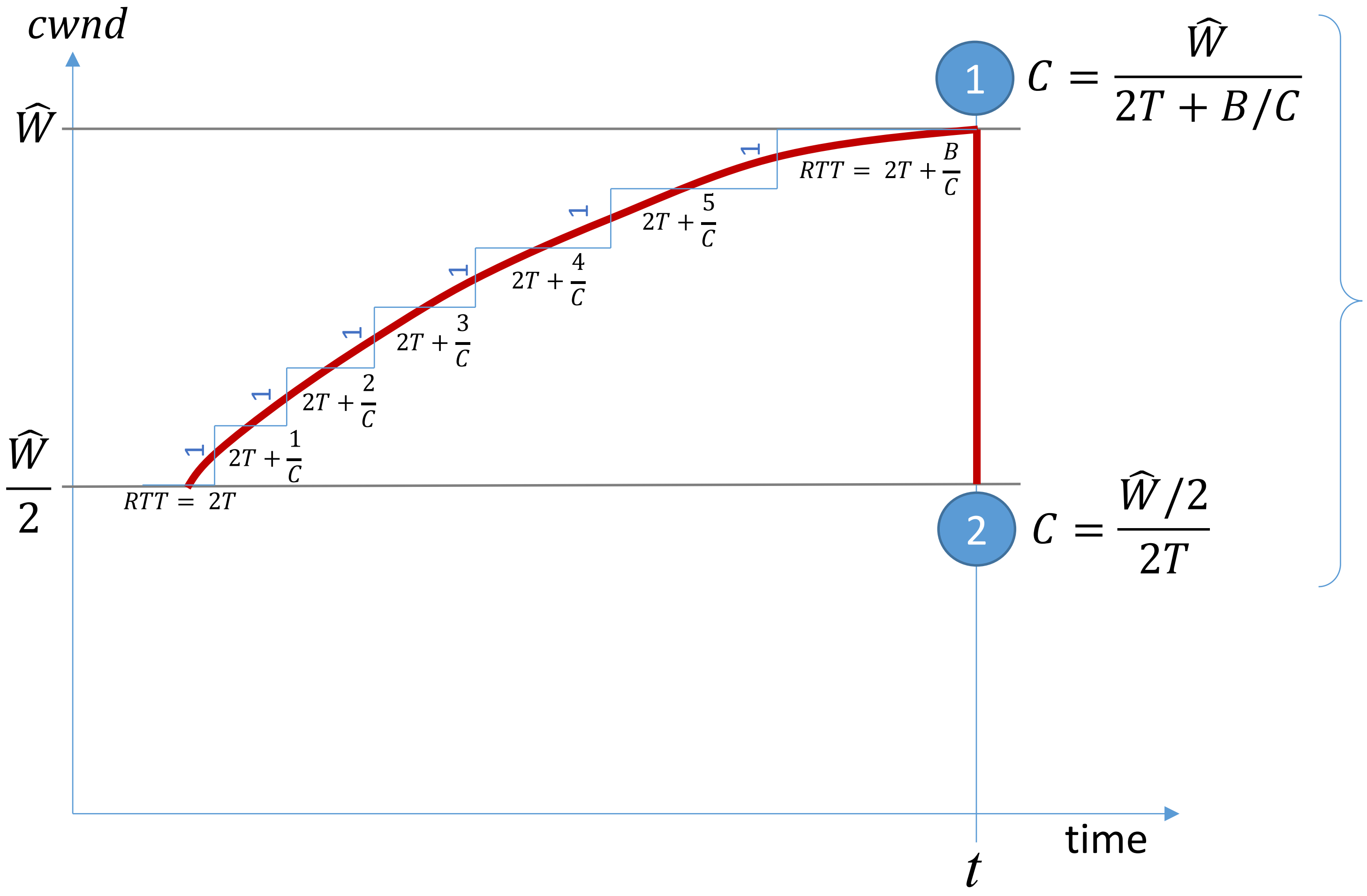
Controls: Left Arrow = Slow, Down Arrow = Medium, Right Arrow = Fast


cwnd
time
cwnd
$\widehat{W}$
$\frac{\widehat{W}}{2}$
t
time


A
B
C
B
2T



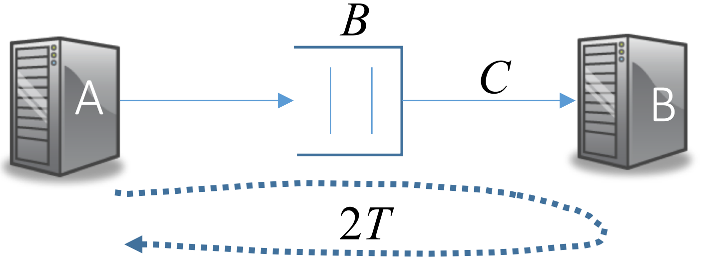




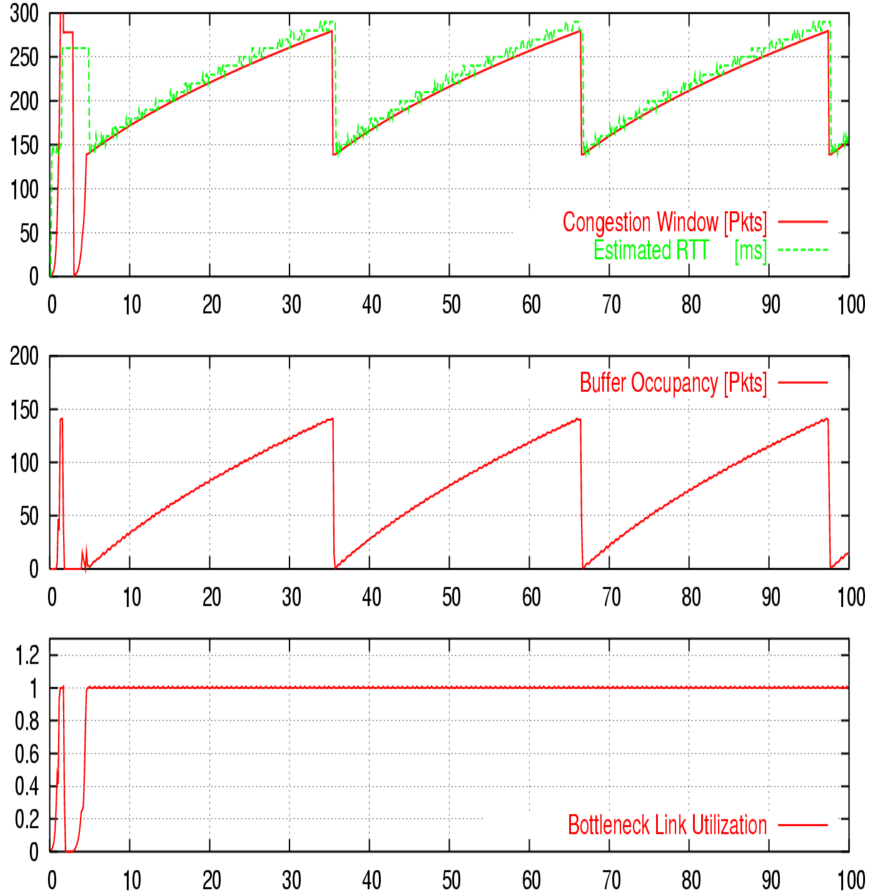
$$C = \frac{\widehat{W}}{2T + B/C} = \frac{\widehat{W}/2}{2T}$$

$$B = 2T \times C$$

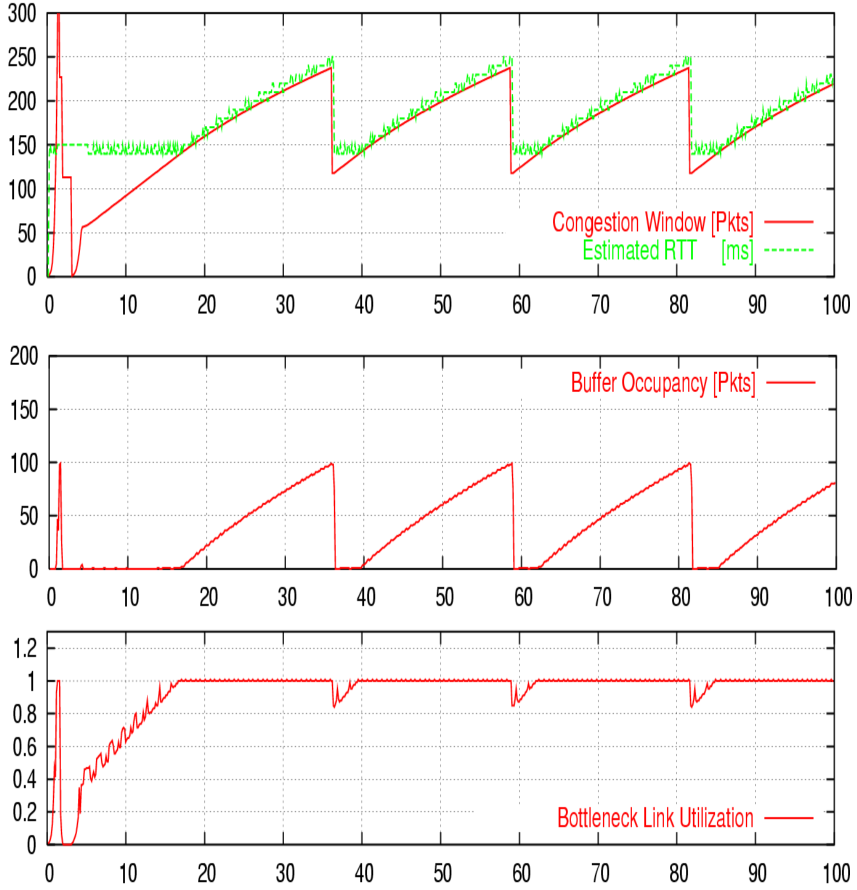
Time Evolution of a Single TCP Flow



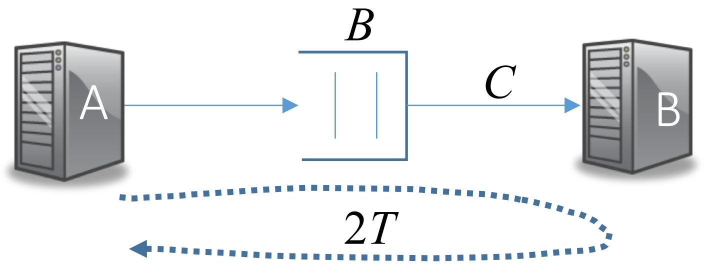
$$B = 2T \times C$$



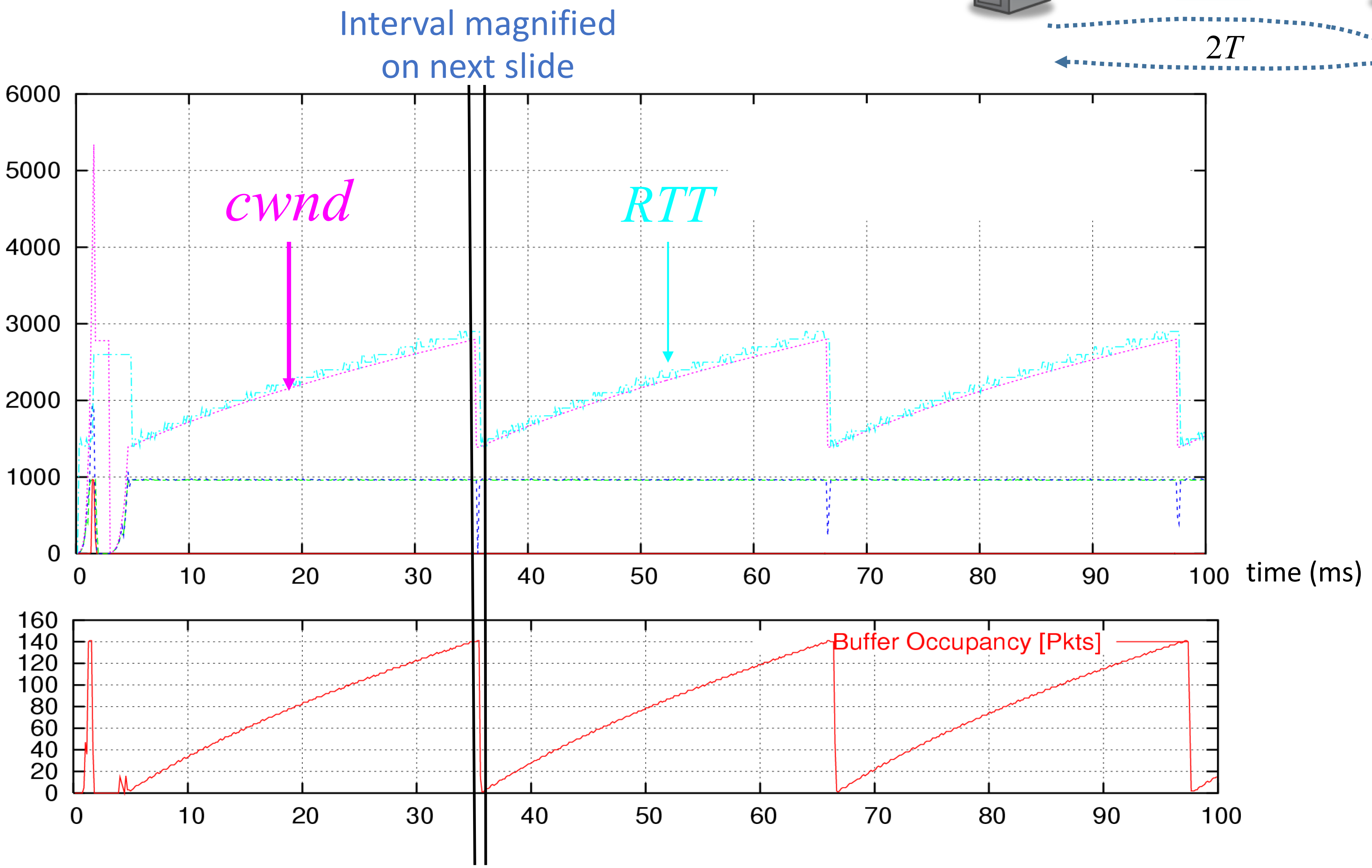
$$B < 2T \times C$$



$$B = 2T \times C$$



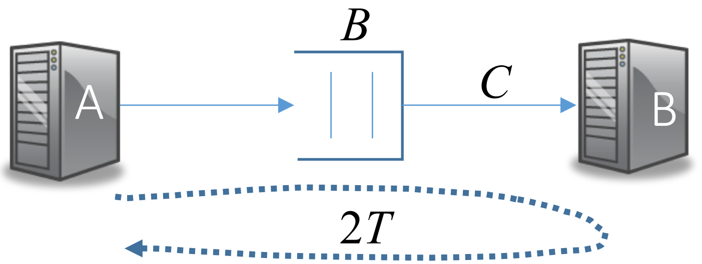
B
C
A
B
2T

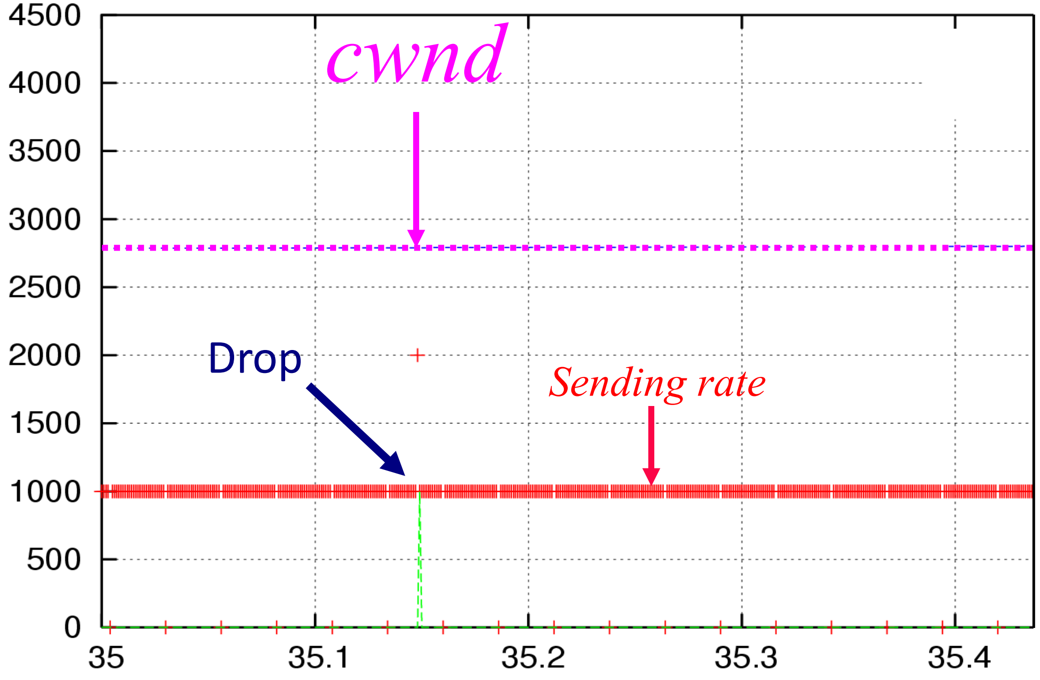


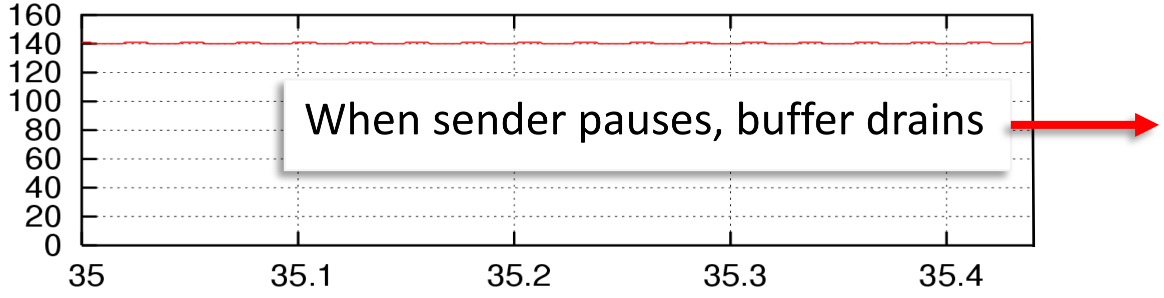
Interval magnified
on next slide
6000
5000
4000
3000
2000
1000
0
cwnd
RTT
0
10
20
30
40
50
60
70
80
90
100
time (ms)
160
140
120
100
80
60
40
20
0
Buffer Occupancy [Pkts]
0
10
20
30
40
50
60
70
80
90
100

$$B = 2T \times C$$

Zoom View







Single AIMD flow: 100% Throughput

1. If $\hat{W} \rightarrow \frac{\hat{W}}{2}$ then $B \geq 2T \times C$

Example: $2T = 100ms, C = 10Gb/s$
$B \geq 1Gbit$

2. If $\hat{W} \rightarrow \frac{\hat{W}}{k}$ then $B \geq 2T(k-1) \times C$

Example: $k = 1.5$
$B \geq 500Mbits$

Example: $k = 1.14$
$B \geq 140Mbits$

3. If $k = 1 + \frac{a}{2T}$ then $B \geq aC$

   i.e. if end host knows 2T, buffer size is independent of RTT

Example: a $= \frac{1}{100}$
$B \geq 50Mbits$

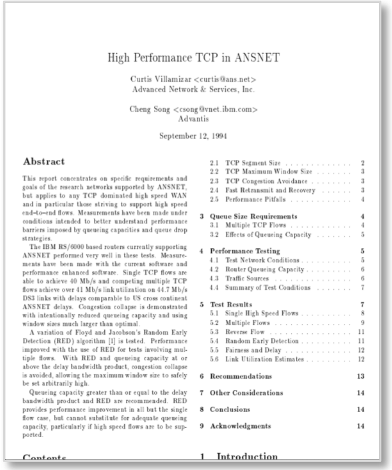
1988	1994	2004	2020
Congestion Avoidance and Control VJ & MK	High Performance TCP in ANSNET CV & CS	Sizing Router Buffers GA, IK, NM	


Congestion Avoidance and Control*

Van Jacobson†
Lawrence Berkeley Laboratory
Michael J. Karels‡
University of California at Berkeley

November, 1988

Introduction

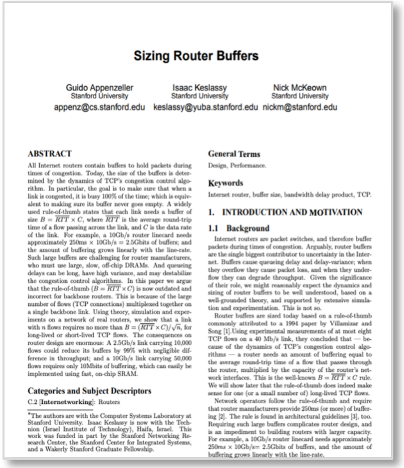

High Performance TCP in ANSNET

Curtis Villamizar <curtis@ans.net>
Advanced Network & Services, Inc.

Cheng Song <csong@vnet.ibm.com>
Advantis

September 12, 1994

Abstract

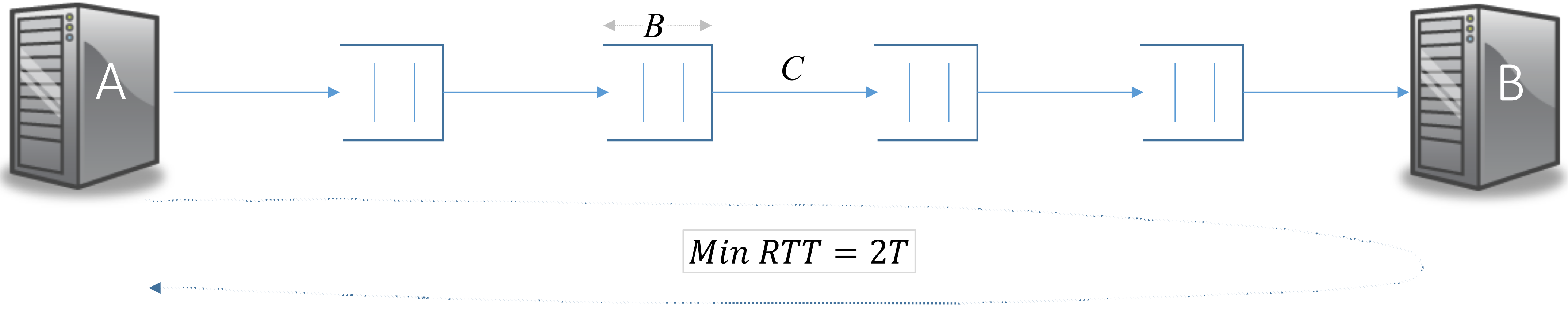

Sizing Router Buffers

Guido Appenzeller, Stanford University, appenz@cs.stanford.edu
Isaac Keslassy, Stanford University, keslassy@yuba.stanford.edu
Nick McKeown, Stanford University, nickm@stanford.edu

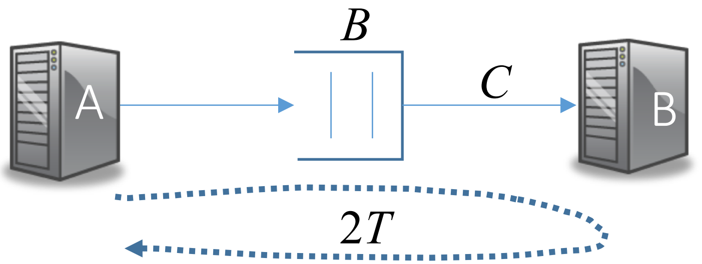
ABSTRACT

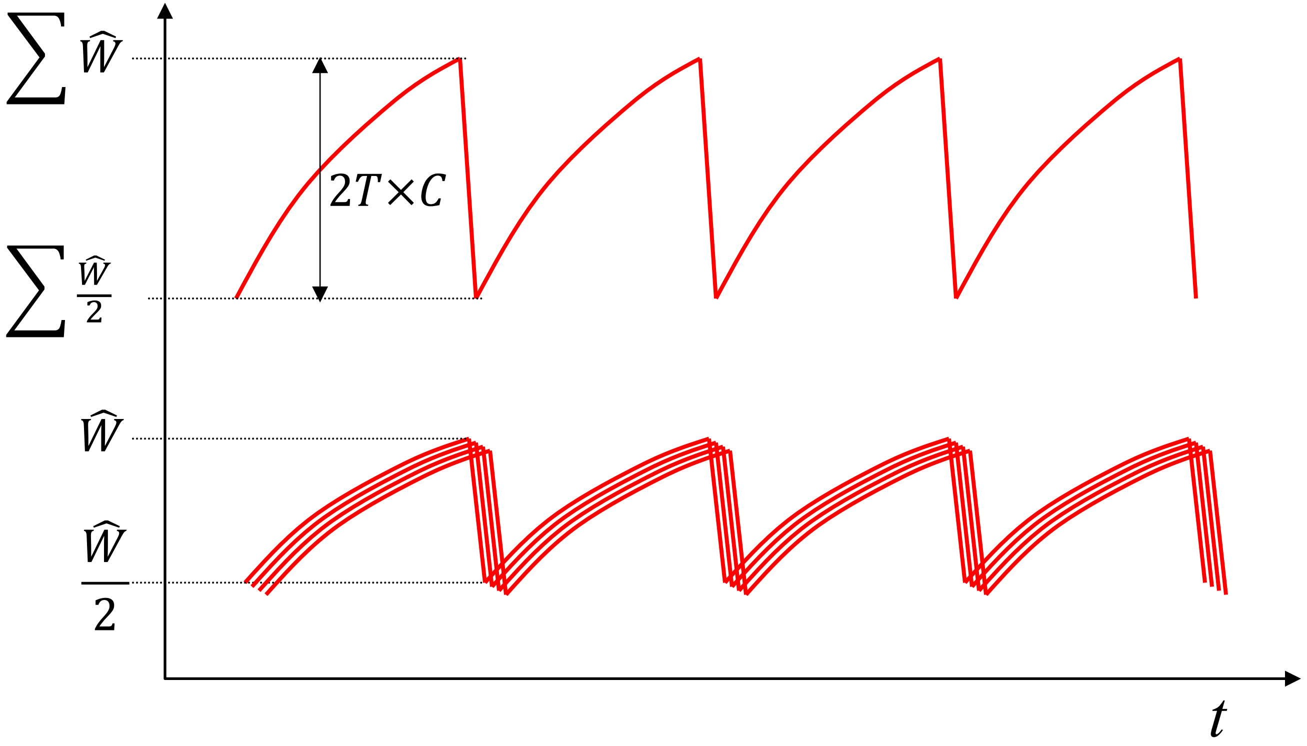
1. INTRODUCTION AND MOTIVATION

$$B \geq \frac{2T \times C}{\sqrt{N}}$$ where N is the number of long-lived flows



Synchronized Flows



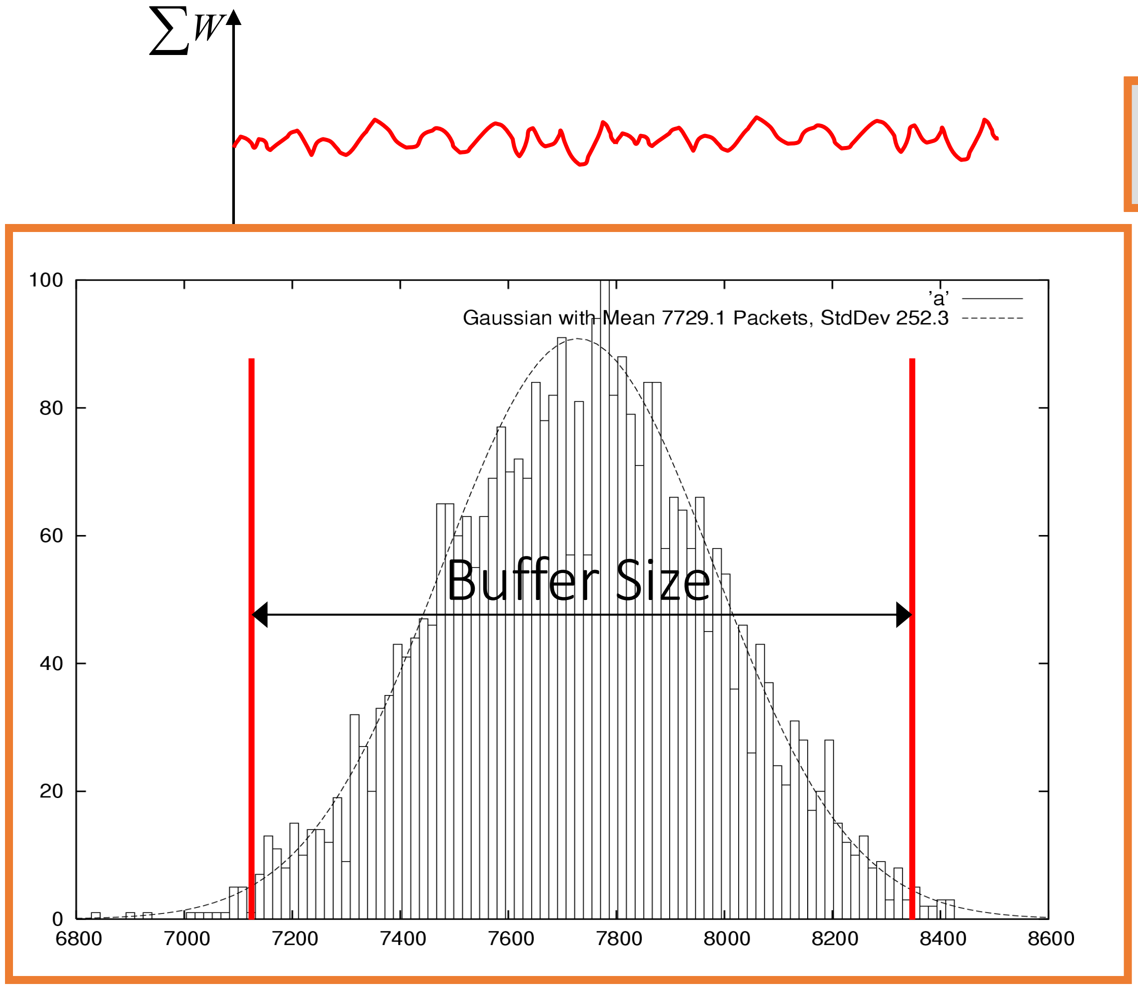


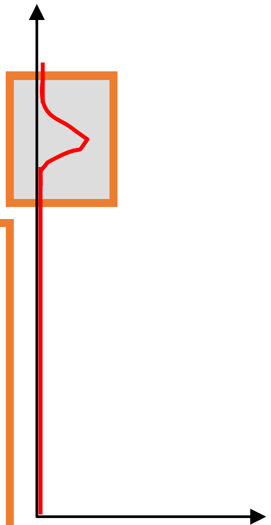
Aggregate window of all the flows has same dynamics

Therefore buffer occupancy has same dynamics

Rule-of-thumb $B \geq 2T \times C$ *still holds.*

Desynchronized TCP Flows





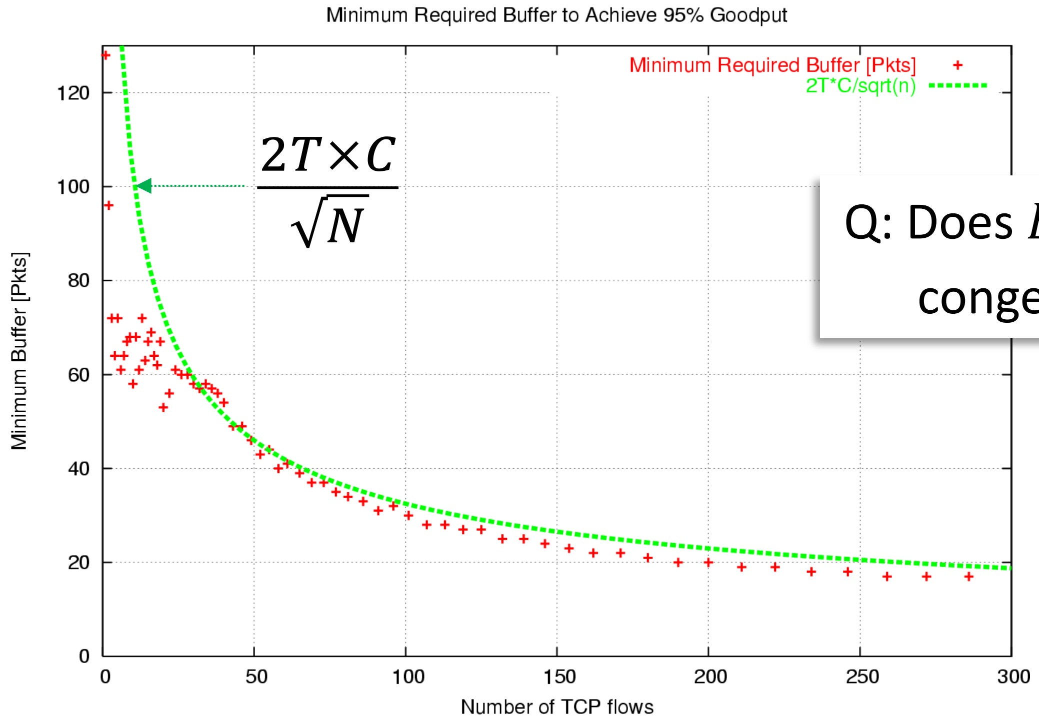
Probability Distribution

Many AIMD flows: 100% Throughput

$$B \geq \frac{2T \times C}{\sqrt{N}}$$

Example: $2T = 100ms, C = 10Gb/s, N = 1$
$B \geq 1Gbit$

Example: $2T = 100ms, C = 10Gb/s, N = 10{,}000$
$B \geq 10Mbit$



Q: Does $B \geq \frac{2T \times C}{\sqrt{N}}$ hold for all popular congestion control schemes…?

1988 — Congestion Avoidance and Control
VJ & MK

1994 — High Performance TCP in ANSNET
CV & CS

2004 — Sizing Router Buffers
GA, IK, NM

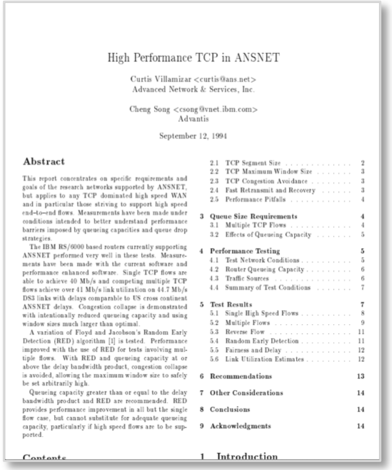
2006 — Routers with Very Small Buffers
ME, YG, AG, NM, TR

2020

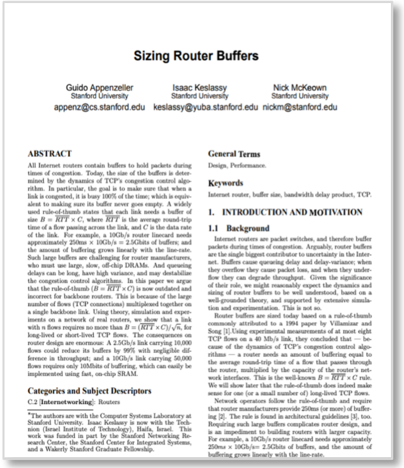
$$B = O(\log W)$$


Congestion Avoidance and Control*

Van Jacobson†
Lawrence Berkeley Laboratory
Michael J. Karels‡
University of California at Berkeley
November, 1988


High Performance TCP in ANSNET

Curtis Villamizar
Advanced Network & Services, Inc.
Cheng Song
Advantis
September 12, 1994


Sizing Router Buffers

Guido Appenzeller, Stanford University
Isaac Keslassy, Stanford University
Nick McKeown, Stanford University

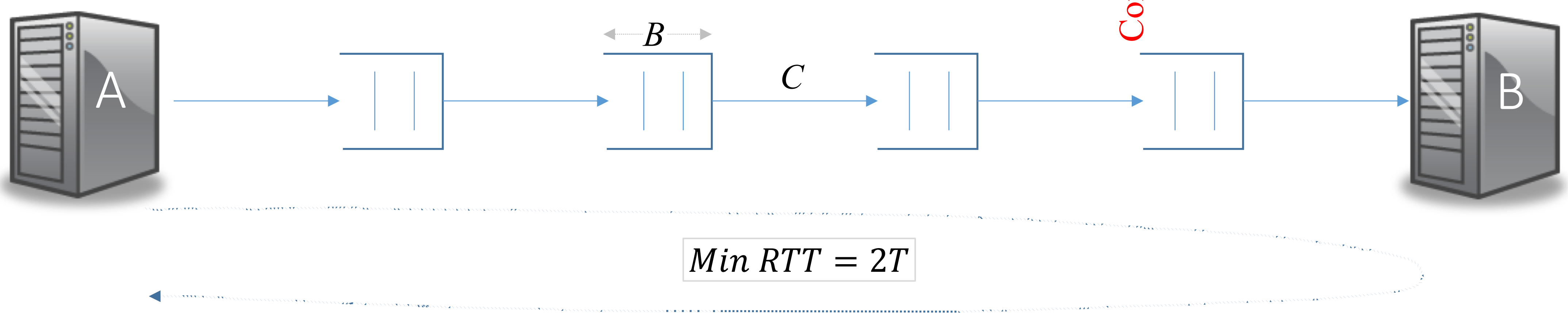

Routers with Very Small Buffers

Assumptions

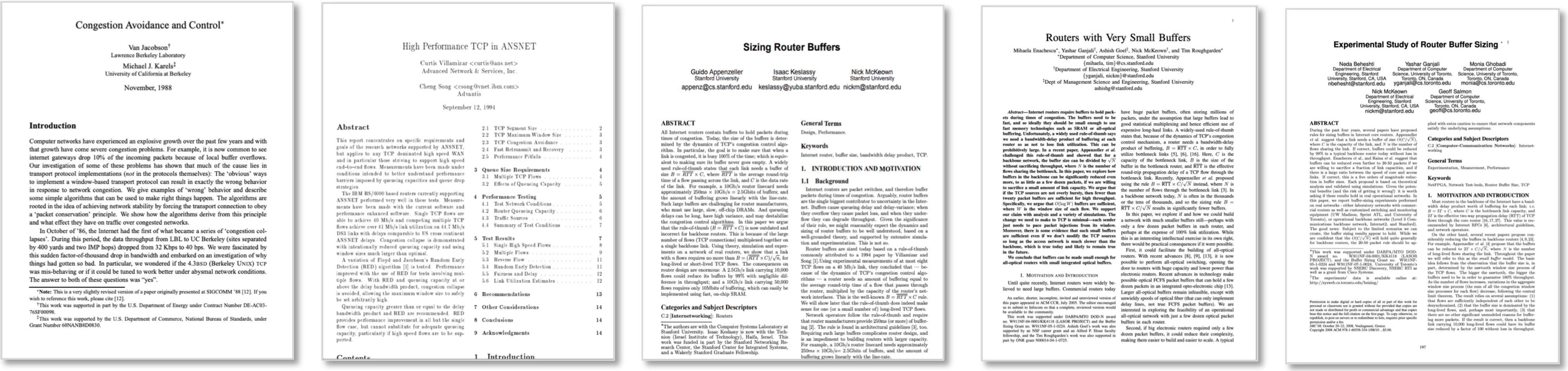
1. Paced Traffic
2. Link utilization < 80%

Consequences

20-50 packet buffers, all-optical routers.



1988	1994	2004	2006	2008	2020
Congestion Avoidance and Control	High Performance TCP in ANSNET	Sizing Router Buffers	Routers with Very Small Buffers	Experimental Study of Router Buffers	
VJ & MK	CV & CS	GA, IK, NM	ME, YG, AG, NM, TR	NB, YG, MG, NM, GS	


Congestion Avoidance and Control*

Van Jacobson†

Lawrence Berkeley Laboratory

Michael J. Karels‡

University of California at Berkeley

November, 1988

High Performance TCP in ANSNET

Curtis Villamizar

Advanced Network & Services, Inc.

Cheng Song

Advantis

September 12, 1994

Sizing Router Buffers

Guido Appenzeller, Stanford University, appenz@cs.stanford.edu

Isaac Keslassy, Stanford University, keslassy@yuba.stanford.edu

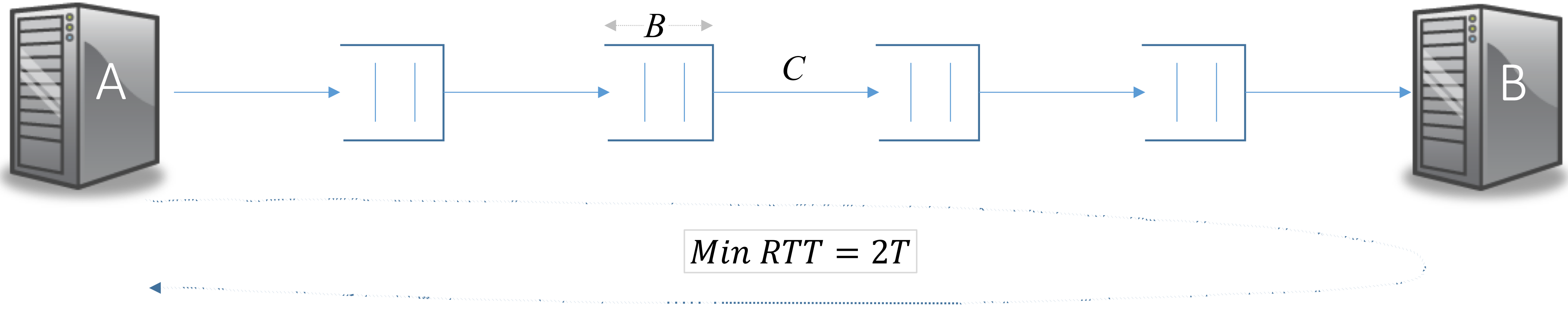
Nick McKeown, Stanford University, nickm@stanford.edu

Routers with Very Small Buffers

Mihaela Enachescu, Yashar Ganjali, Ashish Goel, Nick McKeown, and Tim Roughgarden

Experimental Study of Router Buffer Sizing

Neda Beheshti, Yashar Ganjali, Monia Ghobadi, Nick McKeown, Geoff Salmon



Buffer Sizing Experiments Are Challenging

Testbed experiments:

- Generate realistic traffic with high accuracy
- Explore a very large space (load, traffic shape, …)

Real network experiments:

- Packet drops *may* violate SLAs
- Adjusting buffers not straight forward (device limitations)

Both:

- Accurate measurement of performance metrics not straight forward

Buffer Sizing Experiments

Small Buffers

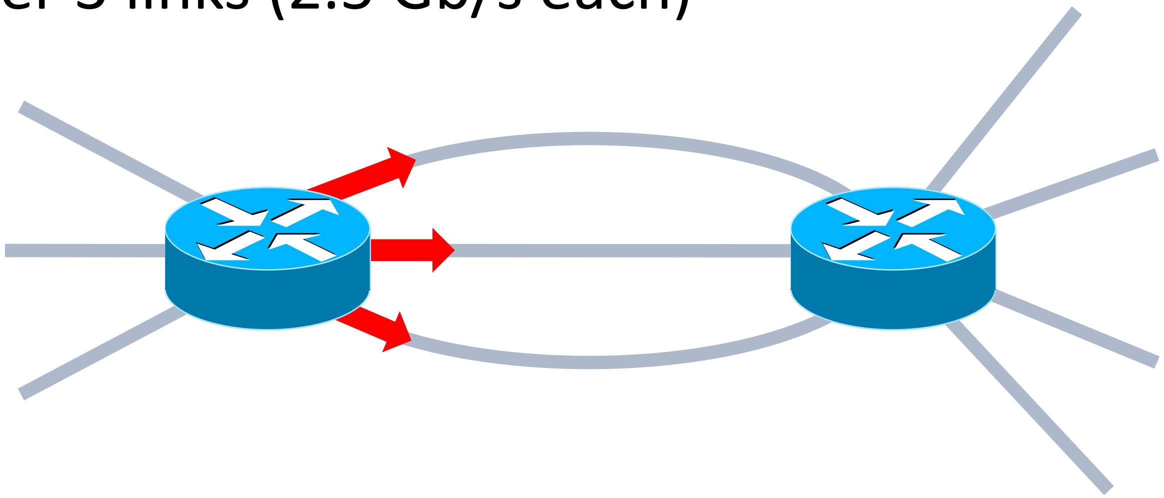
- Stanford University dorm network
- University of Wisconsin
- Internet2
- Level 3 Communications

Tiny Buffers

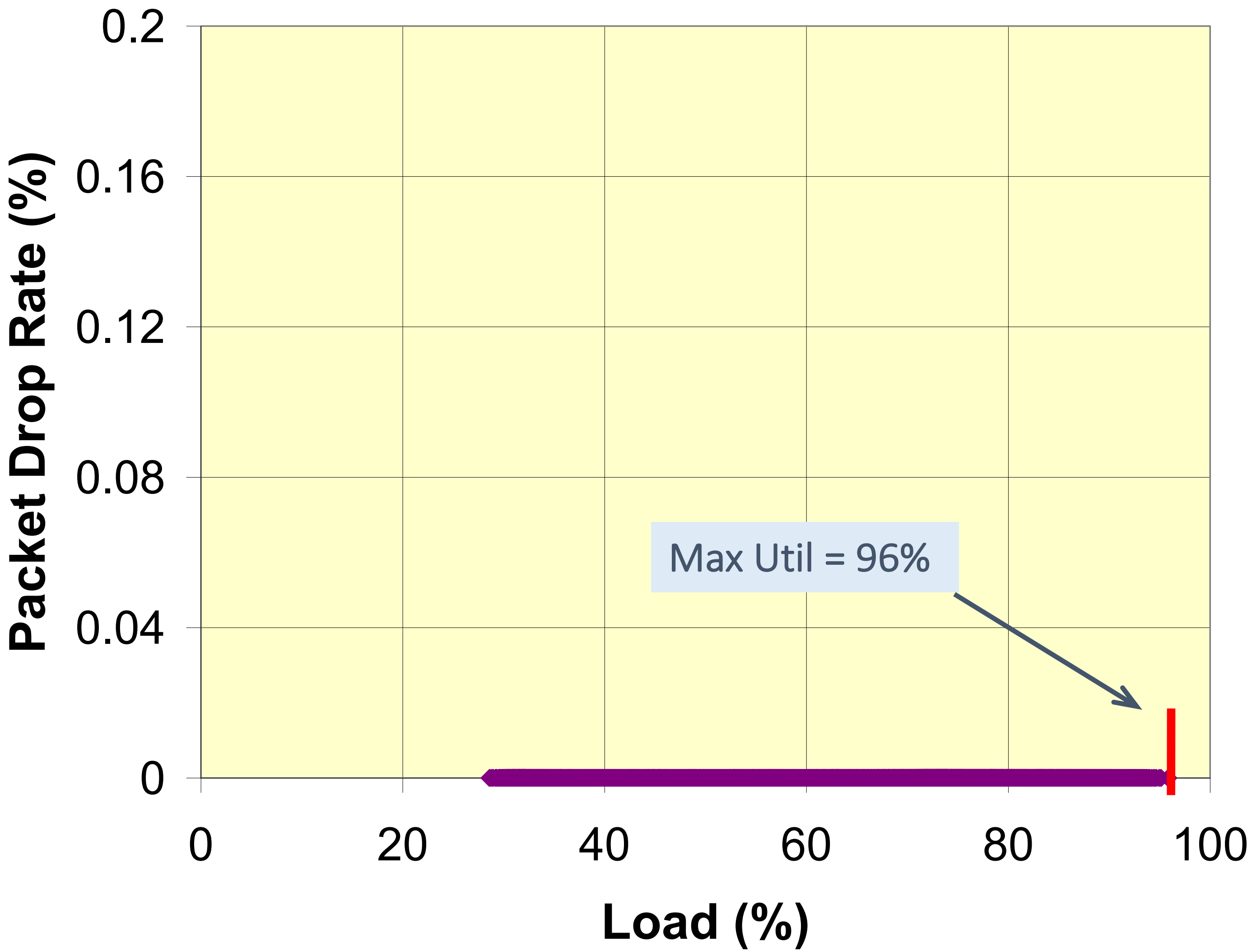
- Internet2
- Sprint Advanced Technology Lab
- University of Toronto

Level 3 Communications Experiments

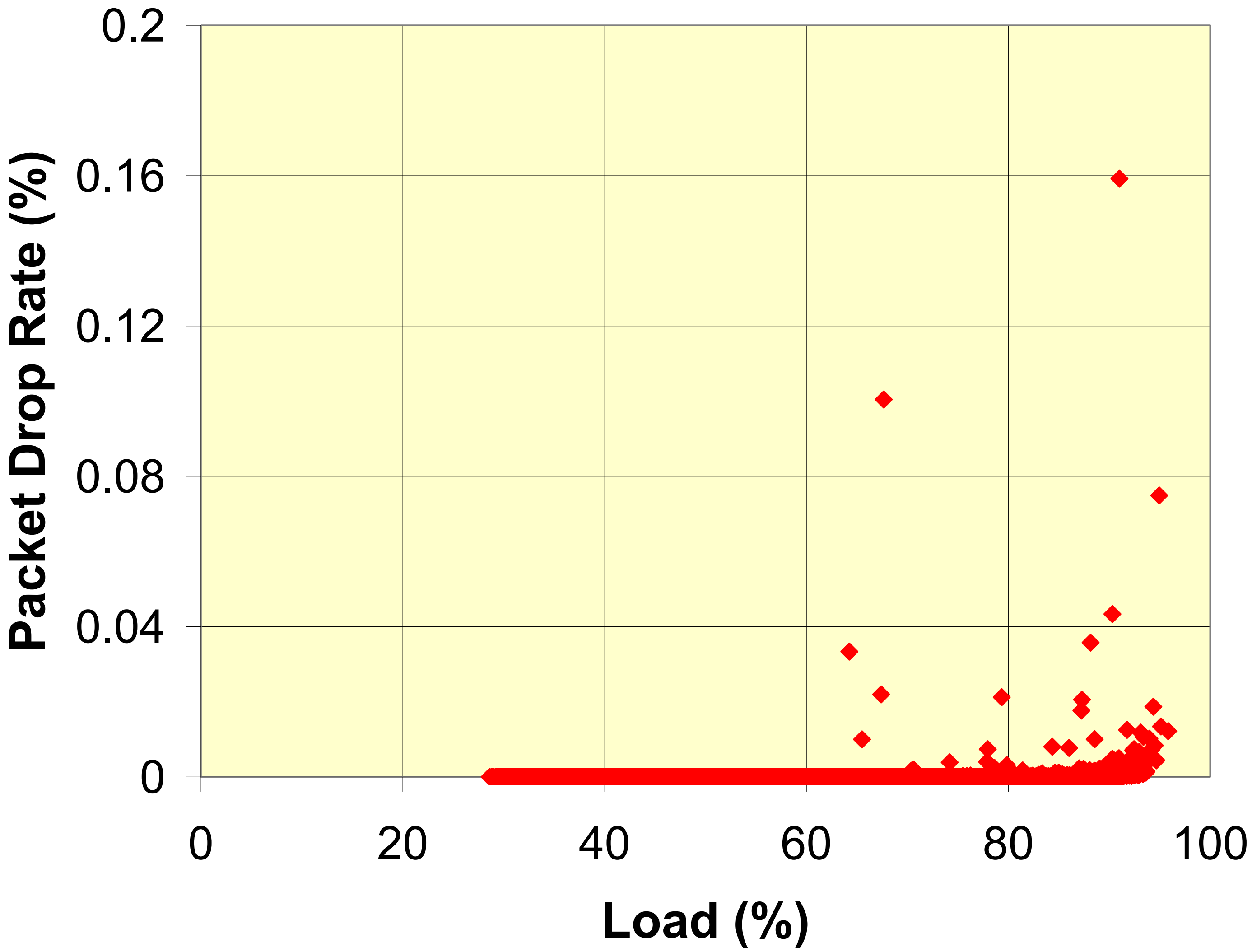
- High link utilization
- Long duration (about two weeks)
- Buffer sizes 190ms (250K packets), 10ms (10K packets), 2.5ms (2500 packets), 1ms (1000 packets)
- Load balancing over 3 links (2.5 Gb/s each)



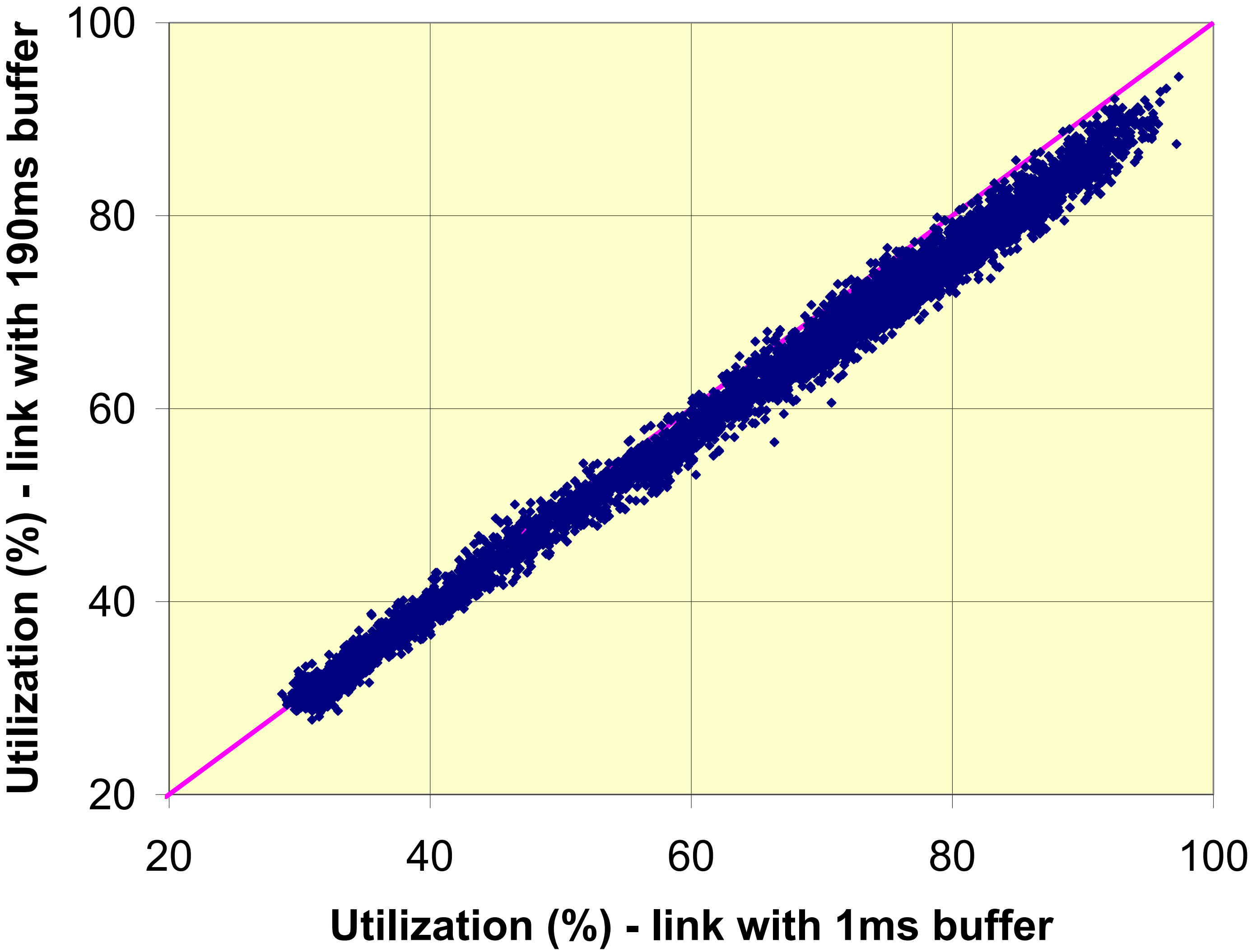
Drop vs. Load, Buffer = 190ms, 10ms



Drop vs. Load, Buffer = 1ms



Relative Link Utilization



Buffer Sizing Experiments

Small Buffers

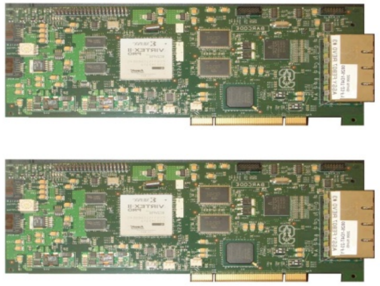
- Stanford University dorm network
- University of Wisconsin
- Internet2
- Level 3 Communications

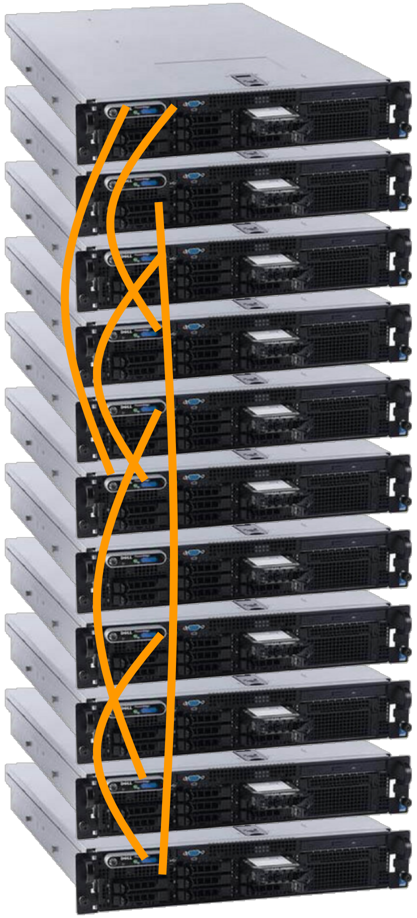
Tiny Buffers

- Internet2
- Sprint Advanced Technology Lab
- University of Toronto

Tiny Buffers Experiments

- Network of NetFPGA-based switches (20-100 machines)
  - 4 GigE interfaces
  - Programmable
- Accurate packet injections
- Complete TCP stack
- Accurate buffer size control
- No hidden buffers
- Added feature to measure queue occupancy time series





Experiment Results

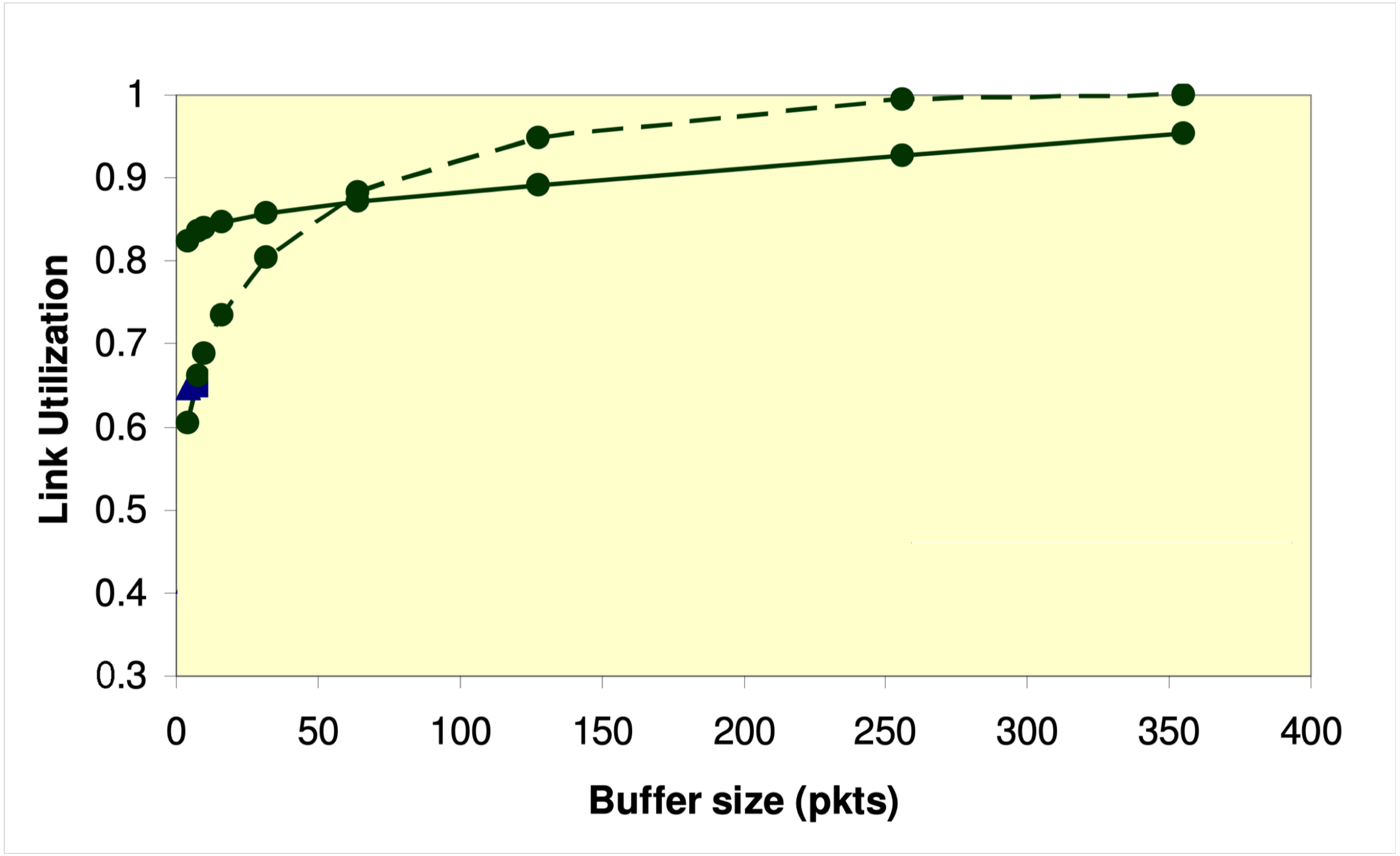
We measured:

- Throughput
- Flow completion times
- Packet drop rates
- ...

For various combinations of:

- Input traffic
- Delays
- Buffer sizes
- ...

Results: Pacing and Buffer Size



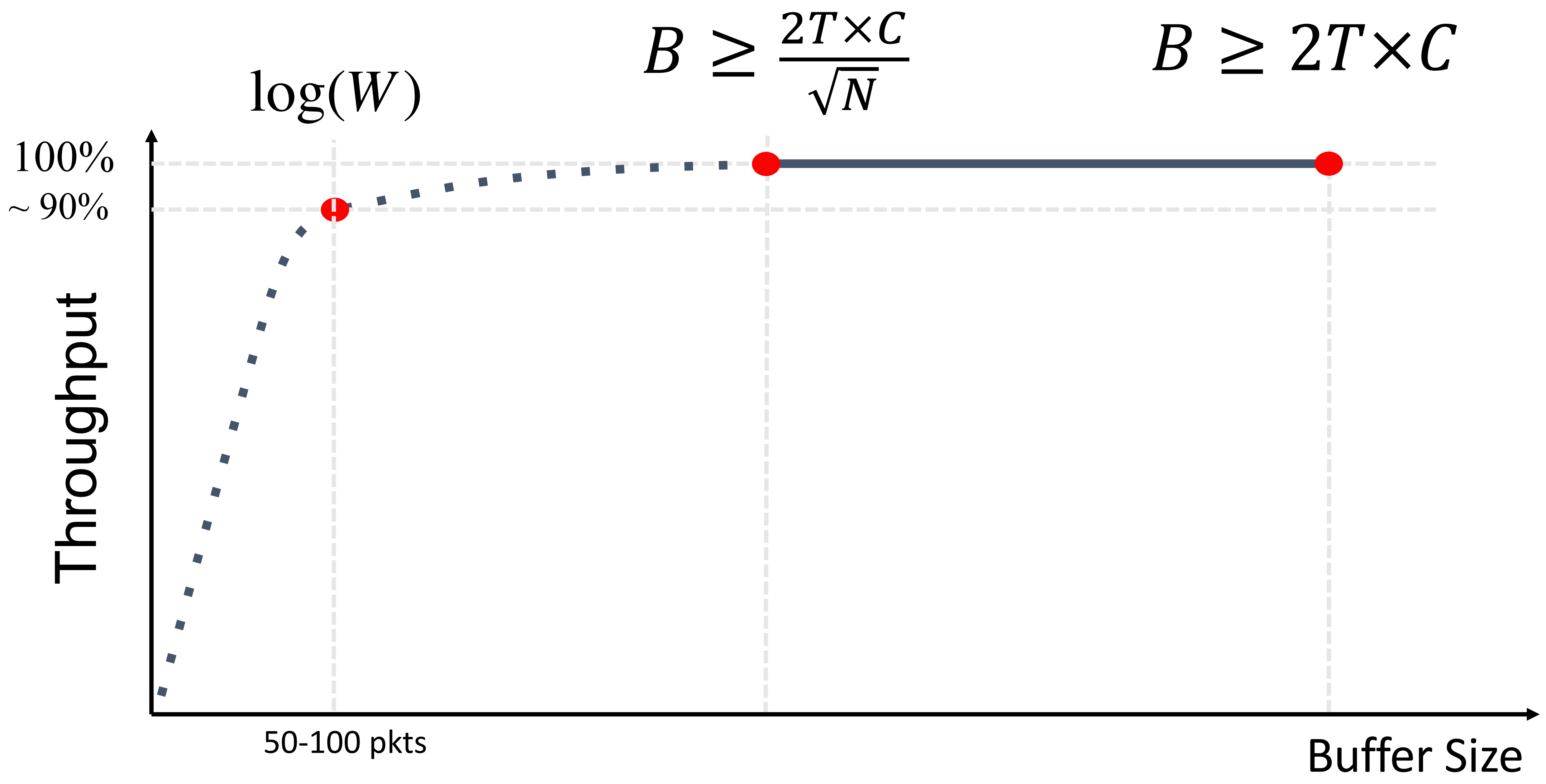
Experiment Conclusions

- Small and tiny buffer experiments inline with theoretical predictions

- Small buffers: no change needed

- Tiny buffers: assumptions are extremely important
  - Necessary to guarantee them all over the network
  - We need support from network components (both software and hardware)

Summary



$B \geq \frac{2T \times C}{\sqrt{N}}$
$B \geq 2T \times C$
$\log(W)$
100%
~ 90%
Throughput
50-100 pkts
Buffer Size

Some ground rules for the day

- 40 different experiences, 40 preconceived notions. Me too.
- Let’s check preconceptions at the door: None of us know for sure.

- Speakers: Please keep you to 15 minutes, including Q&A

- This afternoon, two discussion sessions:
  1. Conclusions: What do we take away from today?
  2. Actions: What are the next steps?

Schedule for the day

10.30am – 1.30pm

Session 1: Network Operators

- Neda Beheshti Facebook
- Lincoln Dale Google
- TY Huang Netflix
- Honqqiang Liu Alibaba
- Ken Duell AT&T
- Joel Jaeggli Fastly

[12.00 – 12.30 Lunch]

- Simon Leinen Switch
- Bob Briscoe CableLabs
- Chuanxiong Guo Bytedance
- Igor Gashinsky Oath

1.45pm – 2.45pm

Session 2: Technology Providers

- Parvin Taheri Cisco
- Francois Labonte Arista
- Golan Schzukin Dune/BCM
- Chang Kim Barefoot

3.00pm – 4.00pm

Session 3: Discussion

- Conclusions Neda, Bruce, Nasser
- Actions and Next Steps Yashar, Nick