CS244

Advanced Topics in Networking

Lecture 10: Buffer Sizing

Nick McKeown

"Sizing Router Buffers"
[Appenzeller, et al. 2004]

Spring 2020

Context

Guido Appenzeller

- At the time: CS PhD student
- Founded Big Switch Networks
- CTO at VMware for networking
- Sigcomm Test of Time Award, 2015

At the time of writing

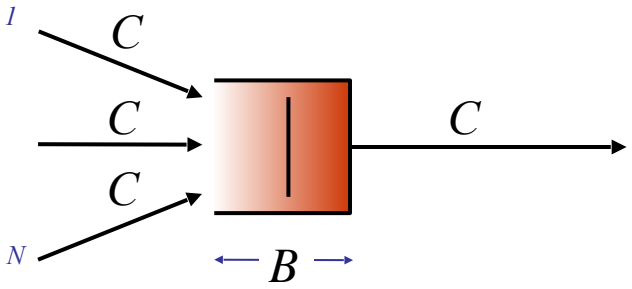
- Challenging to build ISP routers with big buffers
- 80% of world's SRAMs used for router and switch buffers/counters
- ISP routers sold with ~90% profit margin

Why should we care?

Background

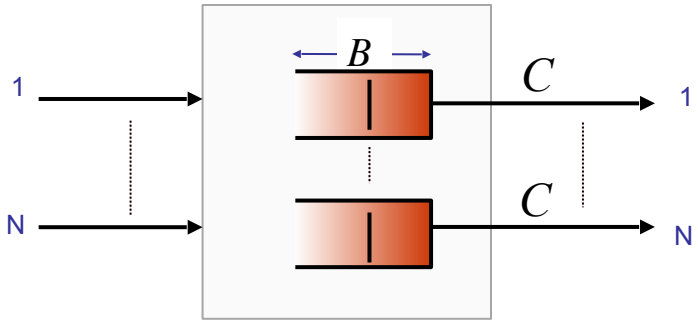
- No universal agreement on how big a router buffer should be
- Or why.
- Yet buffers are a major cause of variation in packet delay.
- Big buffers require large, slow DRAM memories…
- …which complicate the design of large routers.
- It would be nice if we could use single chip switches and routers

Simple model of FCFS router buffer

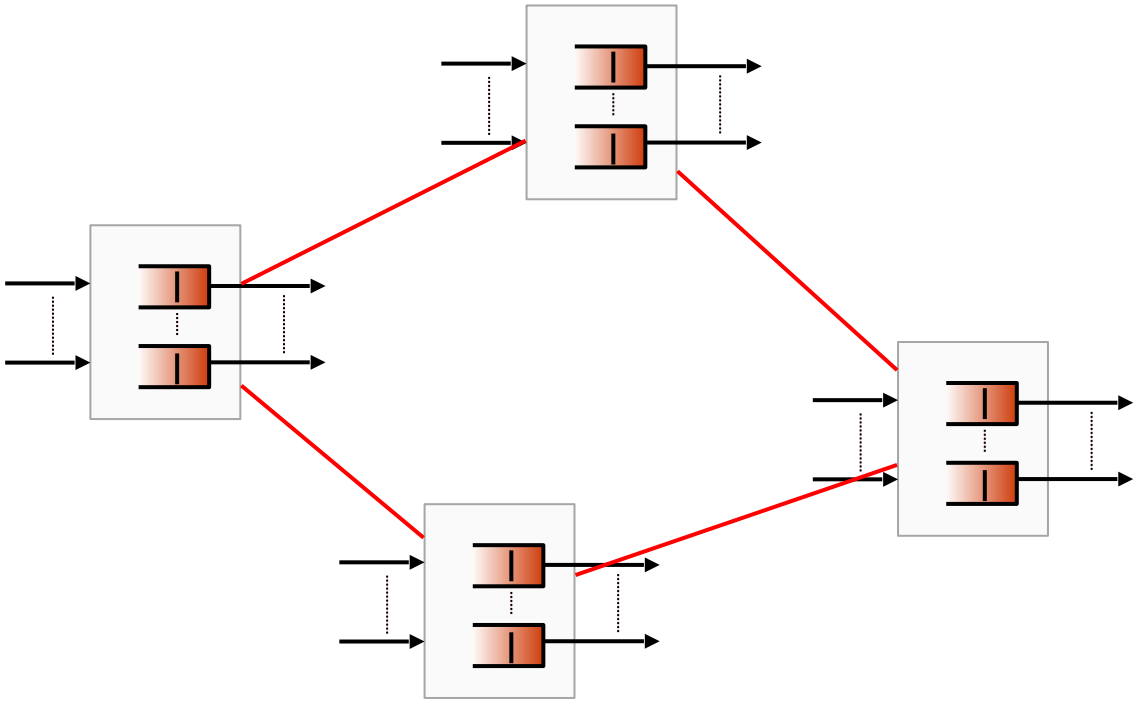


Q: Why does the router have a buffer (or queue)?
Q: What factors determine the buffer's size?

Simple model of a router



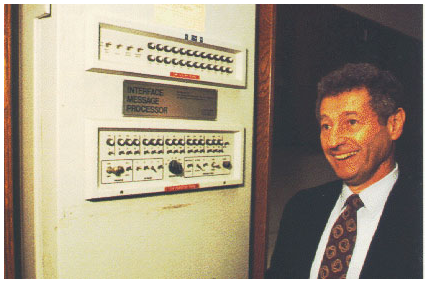
Simple Internet queueing model



Early Internet Models

Leonard Kleinrock
Professor at UCLA

1963: 1st theoretical study of packet switching using queueing theory

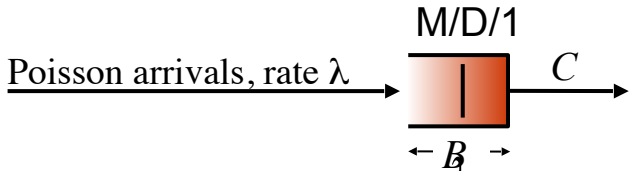


1969: 1st message sent over Internet.

UCLA IMP (Interface Message Processor)

Example model of single packet queue

Poisson Traffic, fixed size packets



Define: Load, $\rho = \frac{\lambda}{C}$

Then: drop rate $< \rho^B$

$e.g. \quad \rho = 0.8, \; B = 20pkts \rightarrow$ loss < 1%

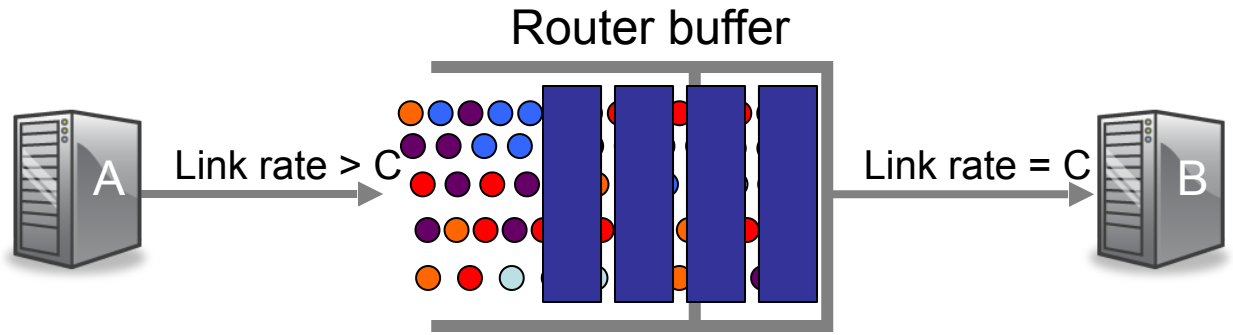
Observations:

- Packet drop rate is small
- Independent of C, RTT, number of flows, etc.

Q: How well does this model fit today's Internet?

The model assumes traffic is generated "open loop" by the source. Today's Internet carries mostly TCP traffic, which uses a closed loop congestion control algorithm.

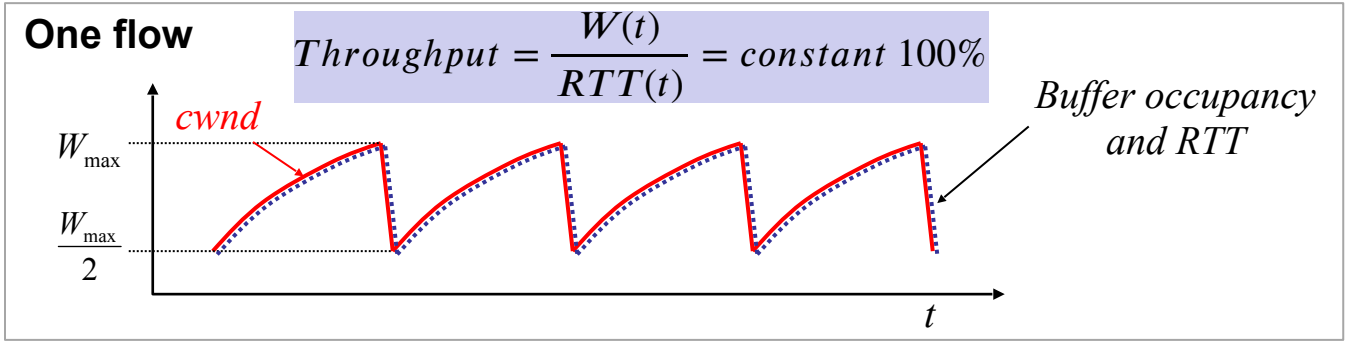
With multiple flows, RTT is less variable

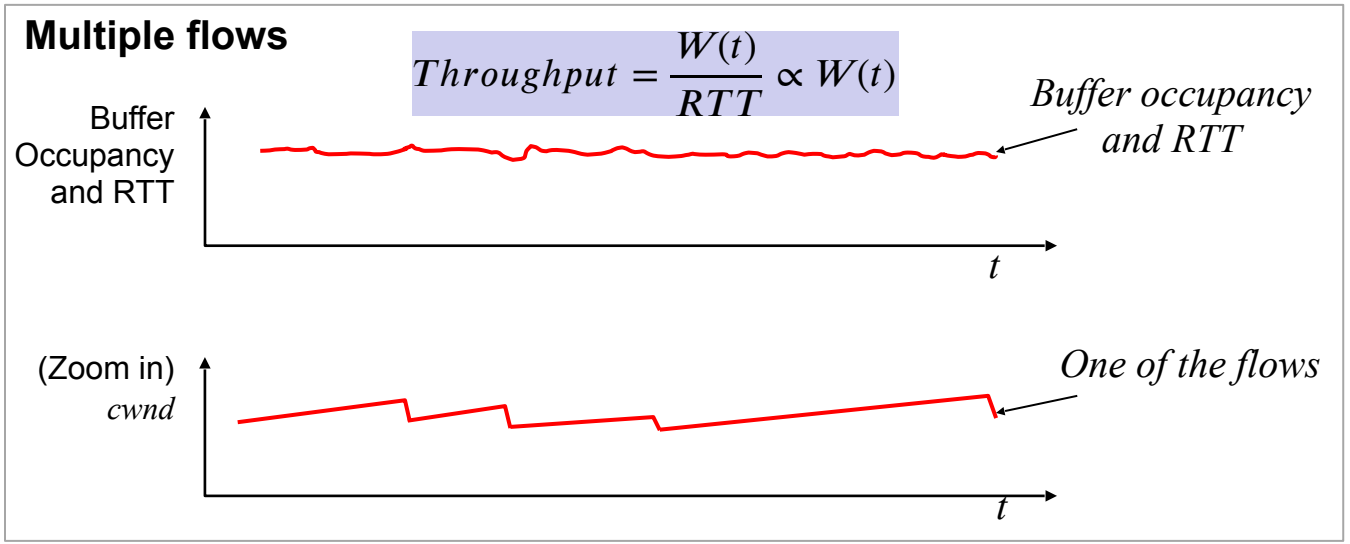


Observations:

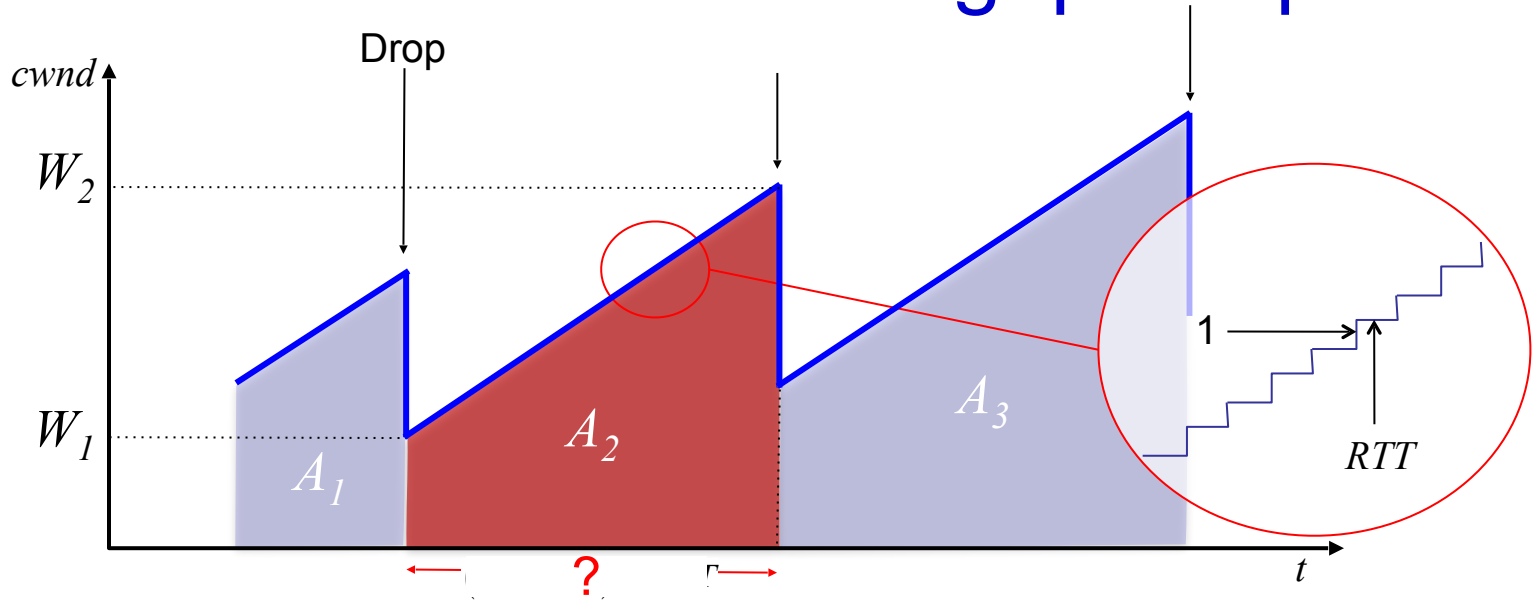
- An arriving packet sees a usually-full queue
- RTT is pretty much constant (very different from single flow case)
- Drops are random events

One flow vs multiple flows





Geometric intuition for throughput equation



Note: To convert from packets/second to bits/second, multiply throughput by the packet size (e.g. MSS)

Expected Area, $E[A]$

Area $A_2 = (W_2 - W_1)W_1 + \frac{1}{2}(W_2 -$

$= \frac{1}{2}(W_2 - W_1)^2$

But $E[W_1] = \frac{1}{2}E[W_2]$

Throughput, $E[T]$

Segments sent during A_2: $T_2 = \frac{A_2}{(W_2 - W_1) \bullet$

$= \frac{\frac{1}{2}(W_2 - W_1)^2}{(W_2 - W_1) \bullet RTT}$

Throughput: $E[T] = \frac{1}{4 \bullet RTT} E[W]$

Combining

Drop rate, $p \approx \frac{1}{E[A]} = \frac{1}{\frac{3}{8}E[W]^2}$

Combining…

$E[T] = \frac{k}{\sqrt{p} \bullet RTT}$ segments/se

Interpreting the throughput equation

$$E[T] = \frac{k}{\sqrt{p} \bullet RTT} \text{ segments/sec}$$

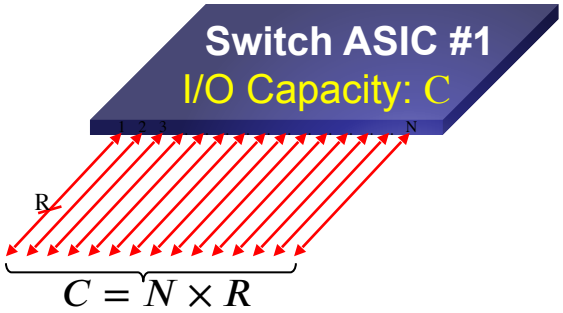
- If RTT doubles, throughput halves.
- If packet drop rate increases from 1% to 4%, throughput halves.

One reason to care about buffer size

With on-chip buffers we can build
higher capacity switch ASICs

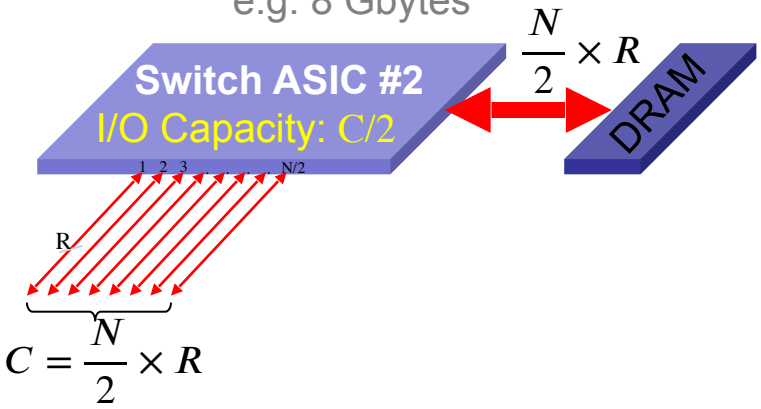
Switch Chips are Limited by Serial I/O Capacity to the outside world

Single chip switch ASIC
Small on-chip buffering
e.g. 64 Mbytes

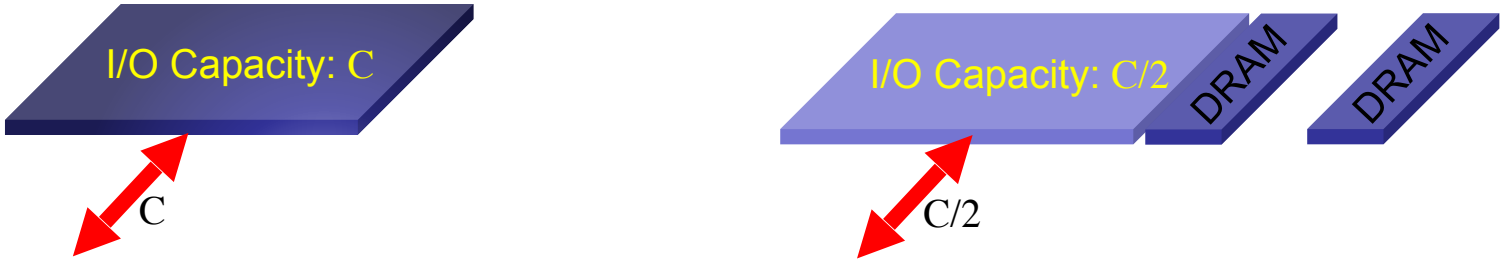


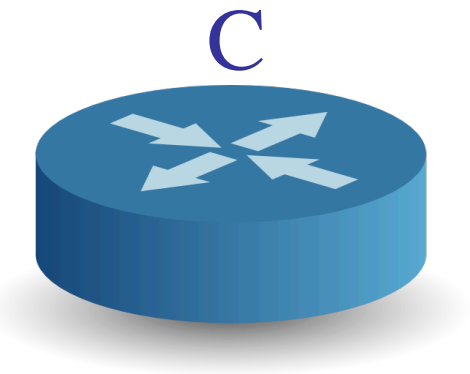
e.g. 12.8 Tb/s = 128 x 100Gb/s

Switch ASIC with external memory
Large off-chip buffering
e.g. 8 Gbytes



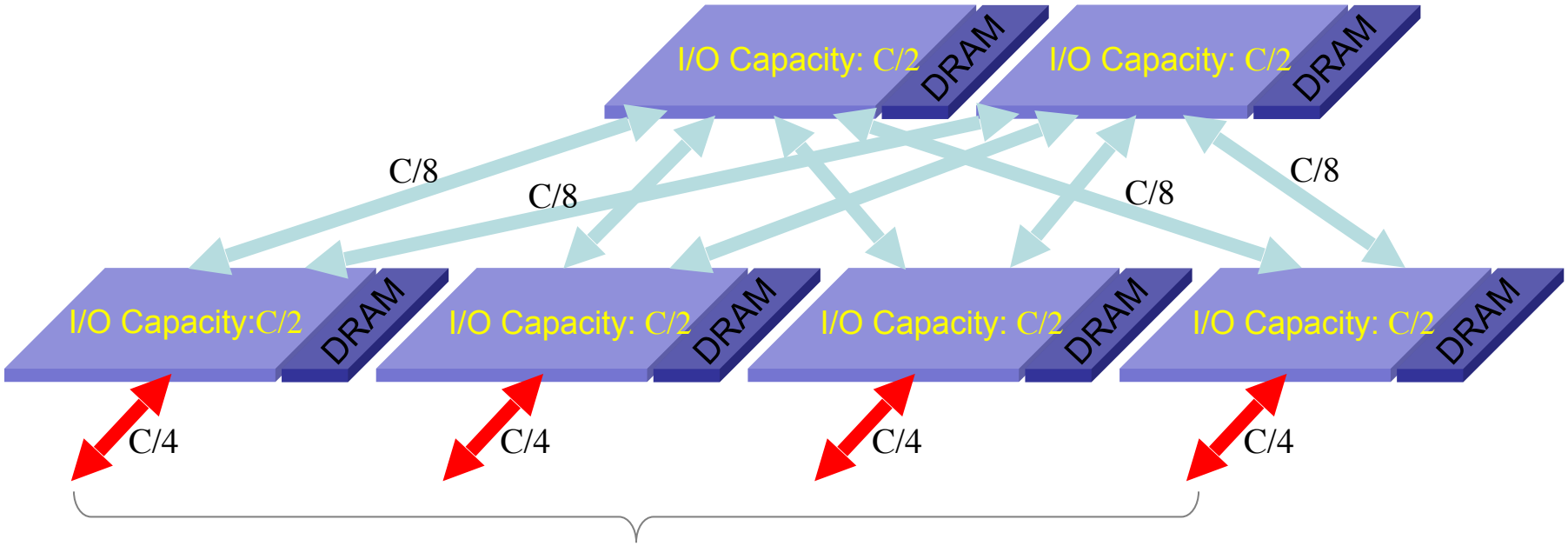
Switch Chips are Limited by Serial I/O Capacity





How many switch chips with capacity *C/2* do we need to make a router with capacity *C*?

We need *6* ASICs with capacity C/2



Total Router Capacity: $C = N \times R$

It is worth understanding
where and when
small on-chip buffers suffice

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

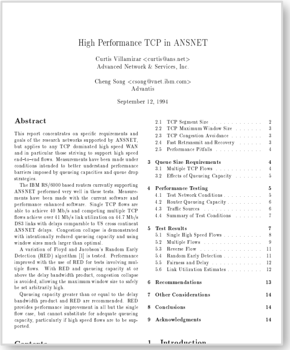
November, 1988

Introduction

Computer networks have experienced an explosive growth over the past few years and with that growth have come severe congestion problems. For example, it is now common to see internet gateways drop 10% of the incoming packets because of local buffer overflows. Our investigation of some of these problems has shown that much of the cause lies in transport protocol implementations (*not* in the protocols themselves): The ‘obvious’ ways to implement a window-based transport protocol can result in exactly the wrong behavior in response to network congestion. We give examples of ‘wrong’ behavior and describe some simple algorithms that can be used to make right things happen. The algorithms are rooted in the idea of achieving network stability by forcing the transport connection to obey a ‘packet conservation’ principle. We show how the algorithms derive from this principle and what effect they have on traffic over congested networks.

In October of ’86, the Internet had the first of what became a series of ‘congestion collapses’. During this period, the data throughput from LBL to UC Berkeley (sites separated by 400 yards and two IMP hops) dropped from 32 Kbps to 40 bps. We were fascinated by this sudden factor-of-thousand drop in bandwidth and embarked on an investigation of why things had gotten so bad. In particular, we wondered if the 4.3BSD (Berkeley UNIX) TCP was mis-behaving or if it could be tuned to work better under abysmal network conditions. The answer to both of these questions was “yes”.

*Note: This is a very slightly revised version of a paper originally presented at SIGCOMM ’88 [12]. If you wish to reference this work, please cite [12].
†This work was supported in part by the U.S. Department of Energy under Contract Number DE-AC03-76SF00098.
‡This work was supported by the U.S. Department of Commerce, National Bureau of Standards, under Grant Number 60NANB8D0830.


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

This report concentrates on specific requirements and goals of the research networks supported by ANSNET, but applies to any TCP dominated high speed WAN and in particular those striving to support high speed end–to–end flows. Measurements have been made under conditions intended to better understand performance barriers imposed by queueing capacities and queue drop strategies.

The IBM RS/6000 based routers currently supporting ANSNET performed very well in these tests. Measurements have been made with the current software and performance enhanced software. Single TCP flows are able to achieve 40 Mb/s and competing multiple TCP flows achieve over 41 Mb/s link utilization on 44.7 Mb/s DS3 links with delays comparable to US cross continent ANSNET delays. Congestion collapse is demonstrated with intentionally reduced queueing capacity and using window sizes much larger than optimal.

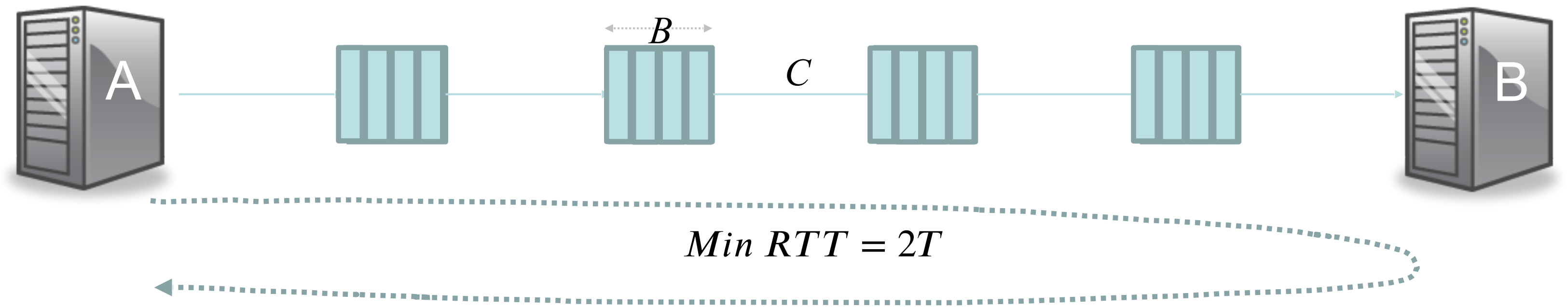
A variation of Floyd and Jacobson's Random Early Detection (RED) algorithm [1] is tested. Performance improved with the use of RED for tests involving multiple flows. With RED and queueing capacity at or above the delay bandwidth product, congestion collapse is avoided, allowing the maximum window size to safely be set arbitrarily high.

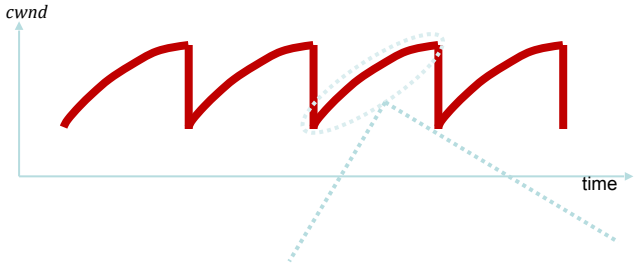
Queueing capacity greater than or equal to the delay bandwidth product and RED are recommended. RED provides performance improvement in all but the single flow case, but cannot substitute for adequate queueing capacity, particularly if high speed flows are to be supported.

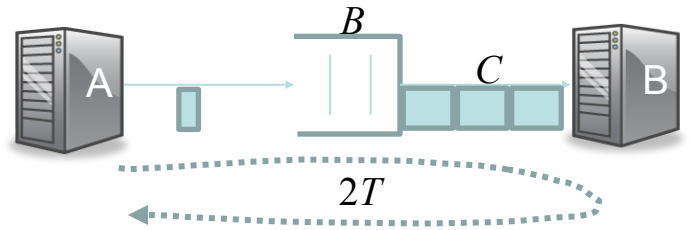
$$B = 2T \times C$$

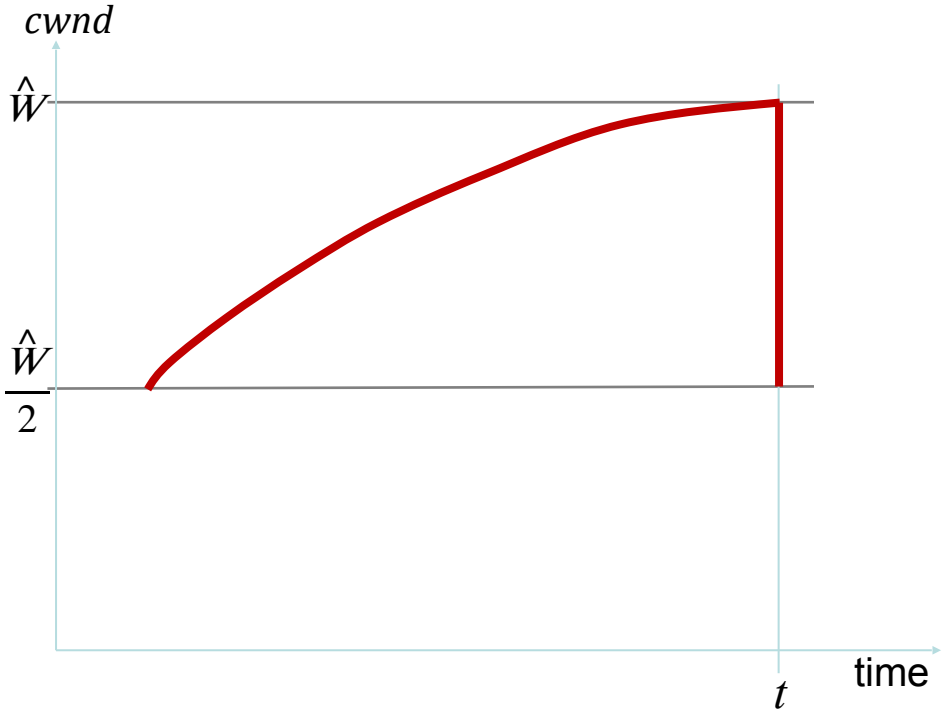
“Buffer size should equal the bandwidth delay product”

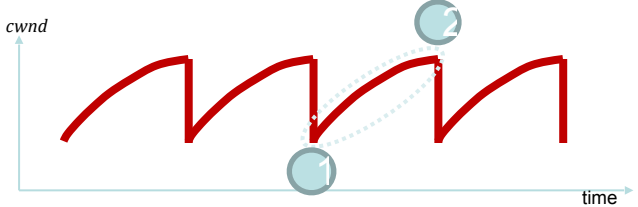
$$Max\ RTT = 2T + B/C = 4T$$

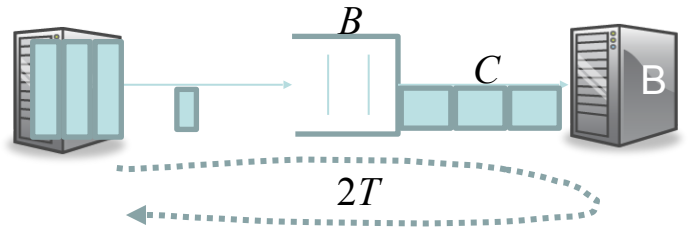


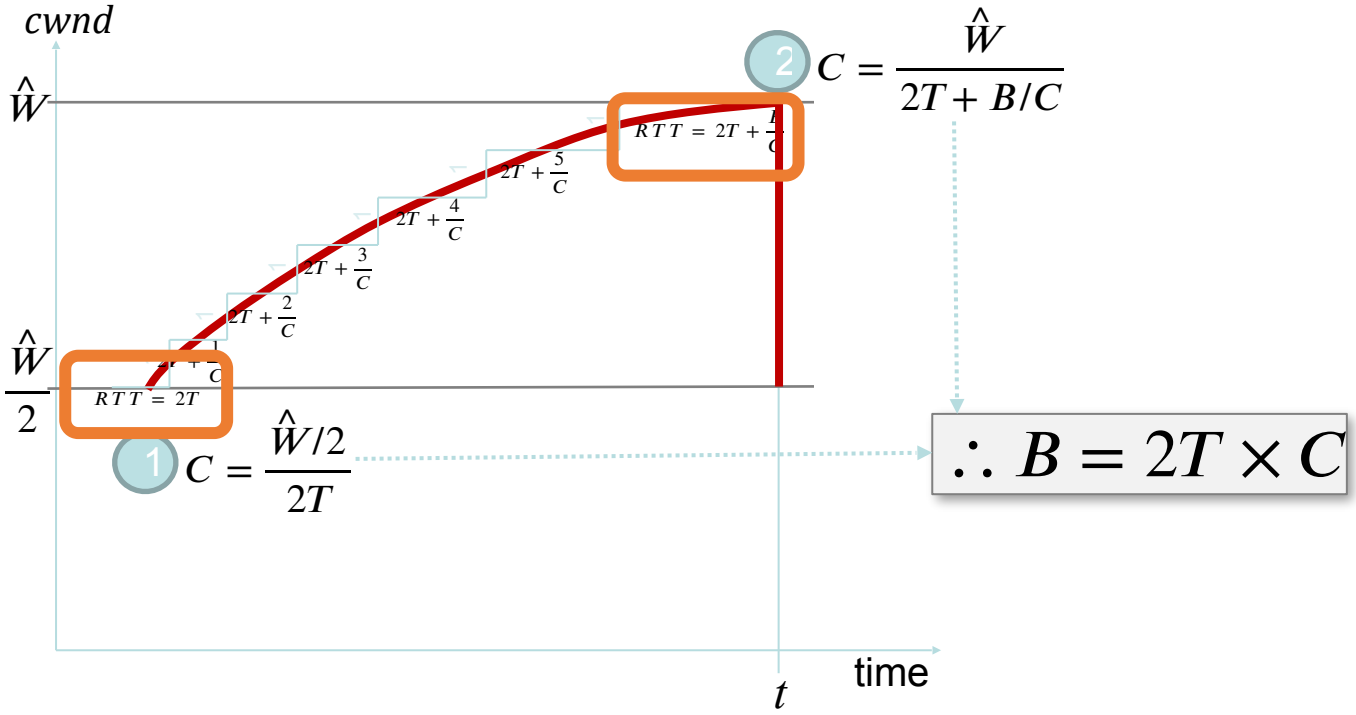

cwnd
time


B
C
A
B
2T

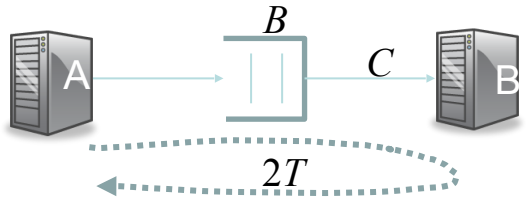

cwnd
$\hat{W}$
$\frac{\hat{W}}{2}$
t
time


cwnd
2
1
time

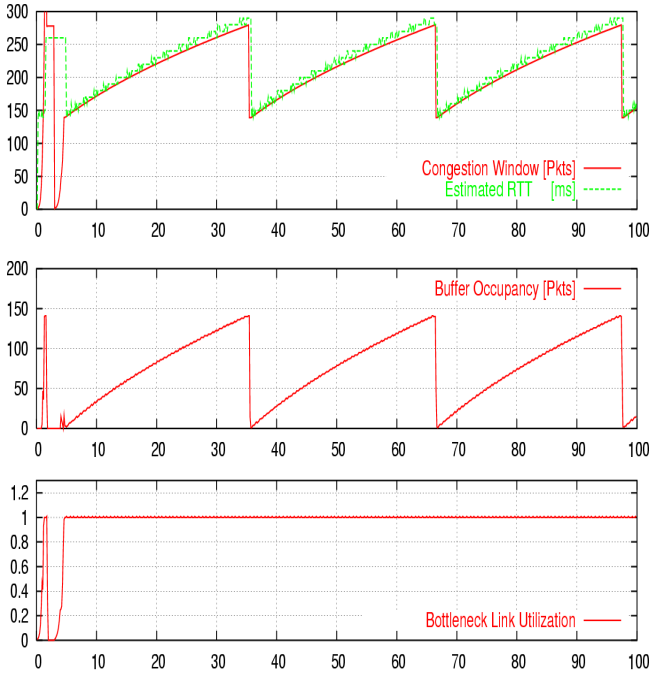

B
C
B
2T


cwnd
$\hat{W}$
$\frac{\hat{W}}{2}$
$RTT = 2T$
$2T + \frac{1}{C}$
$2T + \frac{2}{C}$
$2T + \frac{3}{C}$
$2T + \frac{4}{C}$
$2T + \frac{5}{C}$
$RTT = 2T + \frac{1}{C}$
2
$C = \frac{\hat{W}}{2T + B/C}$
1
$C = \frac{\hat{W}/2}{2T}$
$\therefore B = 2T \times C$
t
time

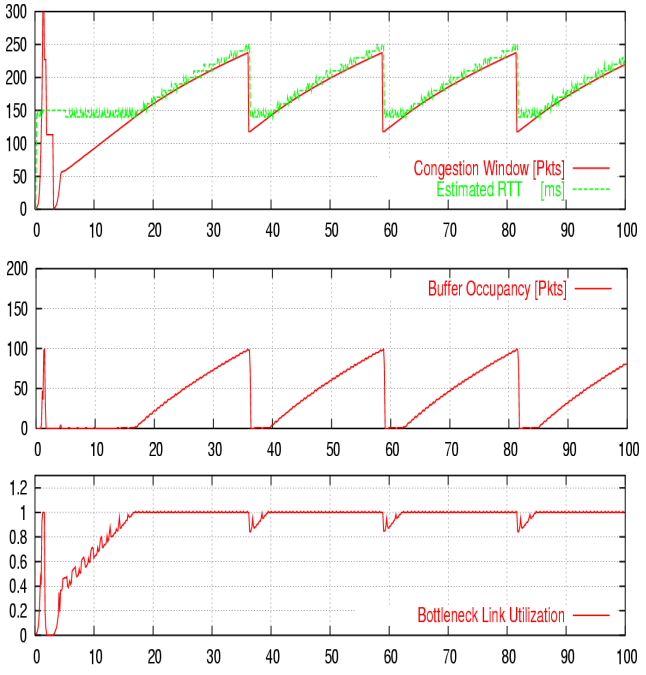
Time Evolution of a Single TCP Flow



$B = 2T \times C$



$B < 2T \times C$



Single TCP New Reno 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: $k = 1.14$
$B \geq 140Mbits$

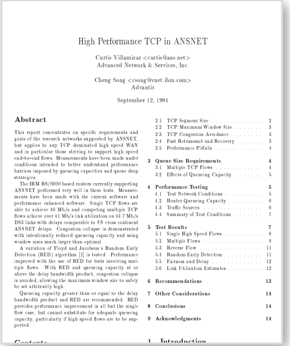
3. If $k = 1 + \frac{a}{2T}$ then $B \geq aC$

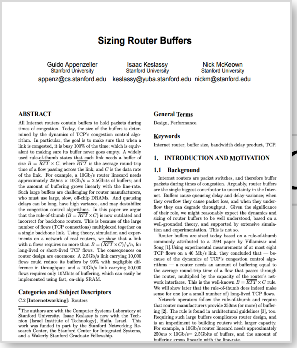
Example: $a = \frac{1}{100}$
$B \geq 50Mbits$

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

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

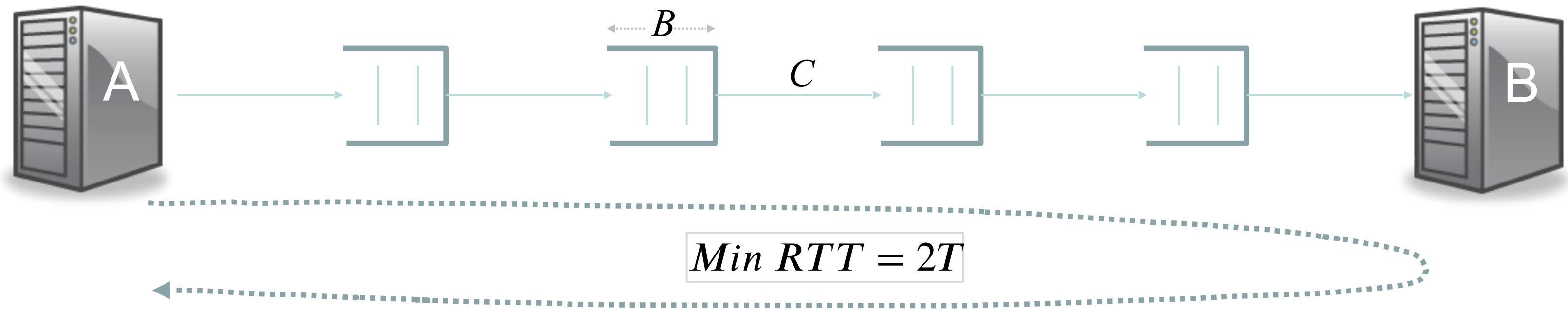

Congestion Avoidance and Control*
Van Jacobson†
Lawrence Berkeley Laboratory
Michael J. Karels‡
University of California at Berkeley
November, 1988


High Performance TCP in ANSNET
Curtis Villamizar <curtis@ans.net>
Advanced Network & Services, Inc.
Cheng Song <csong@vnet.ibm.com>
Advantis
September 12, 1994

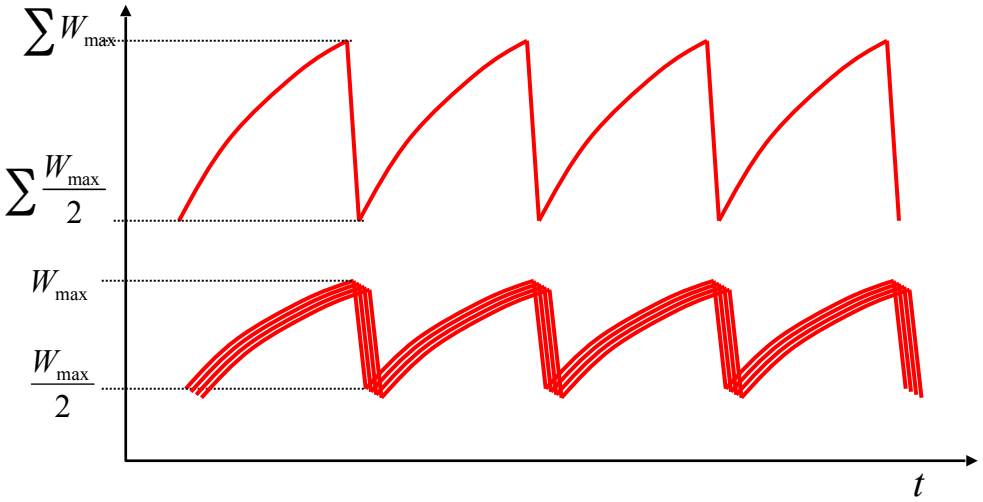

Sizing Router Buffers
Guido Appenzeller, Stanford University, appenz@cs.stanford.edu
Isaac Keslassy, Stanford University, keslassy@yuba.stanford.edu
Nick McKeown, Stanford University, nickm@stanford.edu

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

where N is the number of long-lived flows

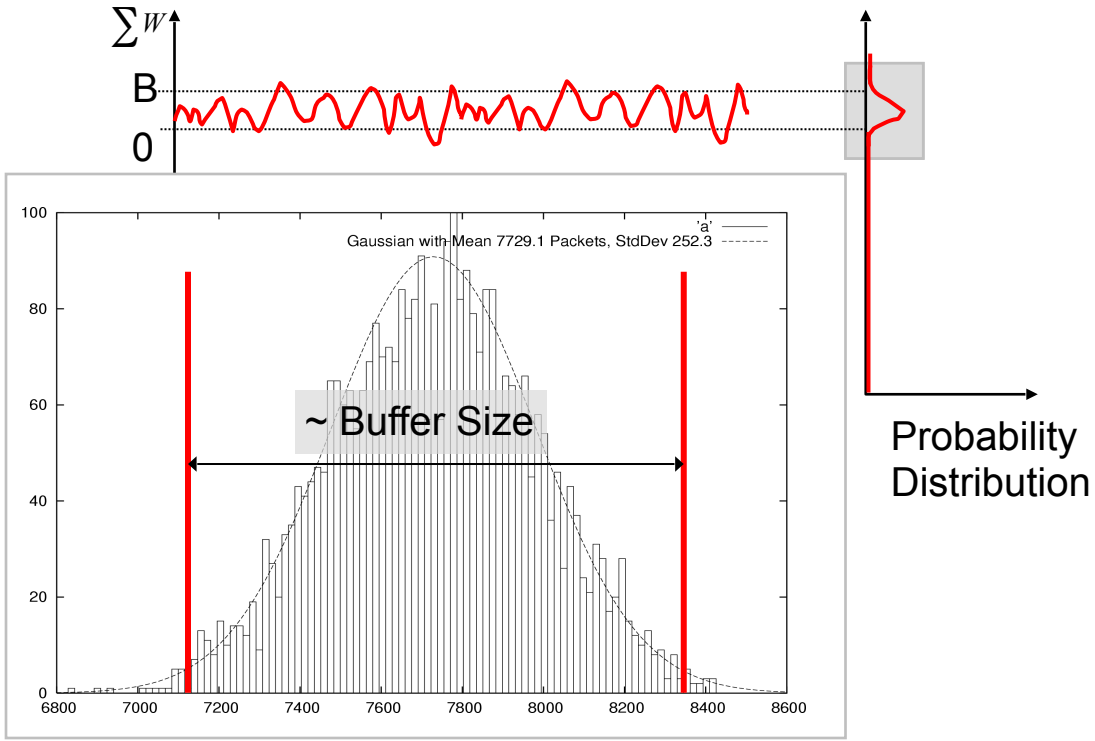


Synchronized Flows



- Aggregate window has same dynamics
- Therefore buffer occupancy has same dynamics
- Rule-of-thumb still holds.

Many TCP Flows



Many AIMD flows: 100% Throughput

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

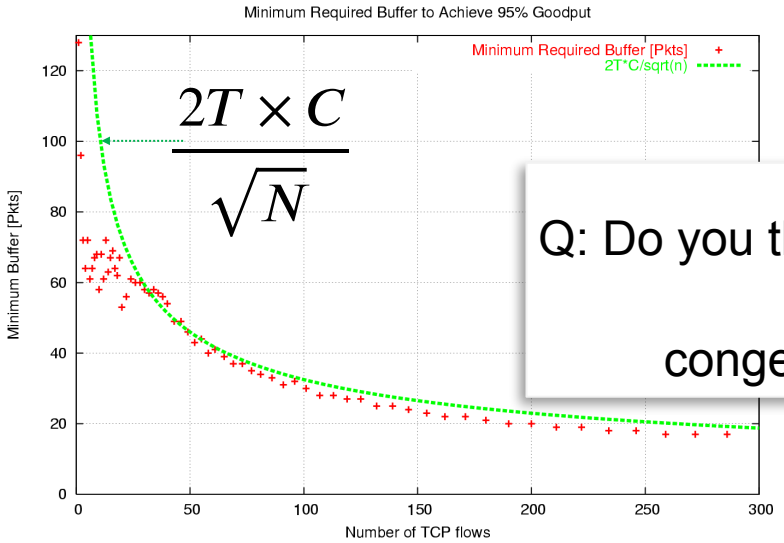
Example: $2T = 100ms,\ C = 10Gb/s,\ N = 1$

$B \geq 1Gbit$

Example:

$2T = 100ms,\ C = 10Gb/s,\ N = 10{,}000$

$B \geq 10Mbit$



Q: Do you think $B \propto \frac{1}{\sqrt{N}}$ will hold for other congestion control algorithms…?

You said

Nikhil Athreya

I think the assumptions made in the original paper are beginning to change. While they acknowledge that their work is mostly based off of TCP, they argue that single-packet sources (e.g. DNS) and constant-rate UDP sources (e.g. online games) can be modeled with short flows. However, a quick Google search tells me that (https://www.caida.org/research/traffic-analysis/tcpudpratio/) that the amount of UDP flows either has surpassed TCP flows or is at least comparable to it. I'm unsure how this would affect the results in this paper; perhaps because the results for short flows shows that average queue length is only dependent on link load and flow length, this shouldn't be a problem, but I'd be interested to hear more about it.

You said

Kevin Baichoo

It seems like from the backbone router vendor's will disavow this smaller buffer sizes, as otherwise they can't justify the prices for their equipment.

Esther Goldstein

Do buffer sizes today go by the rule-of-thumb equation, or has the equation been refined in a different way than what was proposed in the paper?

Isabel Victoria Papadimitriou

It seems that this is working under an assumption that TCP is TCP, and router manufacturers have to deal with that regarding buffer size. Why isn't it the other way around ("this is the buffer size, make congestion control work"), or is it possible to think about some sort of joint optimization of buffer size and paradigm?

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

$$B = O(logW)$$


Congestion Avoidance and Control*

Van Jacobson†
Lawrence Berkeley Laboratory

Michael J. Karels‡
University of California at Berkeley

November, 1988

Introduction

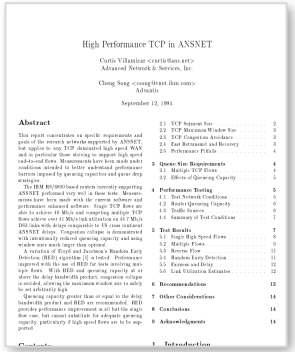
Computer networks have experienced an explosive growth over the past few years and with that growth have come severe congestion problems. For example, it is now common to see internet gateways drop 10% of the incoming packets because of local buffer overflows. Our investigation of some of these problems has shown that much of the cause lies in transport protocol implementations (not in the protocols themselves). The 'obvious' ways to implement a window-based transport protocol can result in exactly the wrong behavior in response to network congestion. We give examples of 'wrong' behavior and describe some simple algorithms that can be used to make right things happen. The algorithms are rooted in the idea of achieving network stability by forcing the transport connection to obey a 'packet conservation' principle. We show how the algorithms derive from this principle and what effect they have on traffic over congested networks.

In October of '86, the Internet had the first of what became a series of 'congestion collapses'. During this period, the data throughput from LBL to UC Berkeley (sites separated by 400 yards and two IMP hops) dropped from 32 Kbps to 40 bps. We were fascinated by this sudden factor-of-thousand drop in bandwidth and embarked on an investigation of why things had gotten so bad. In particular, we wondered if the 4.3BSD (Berkeley UNIX) TCP was mis-behaving or if it could be tuned to work better under abysmal network conditions. The answer to both of these questions was "yes".

*Note: This is a very slightly revised version of a paper originally presented at SIGCOMM '88 [12]. If you wish to reference this work, please cite [12].

†This work was supported in part by the U.S. Department of Energy under Contract Number DE-AC03-76SF00098.

‡This work was supported by the U.S. Department of Commerce, National Bureau of Standards, under Grant Number 60NANB8D0830.

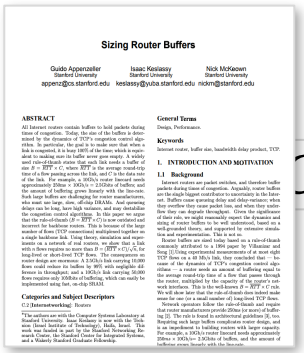

High Performance TCP in ANSNET

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

Cheng Song <cssong@ans.net>
ANSnet

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

General Terms

Keywords

1. INTRODUCTION AND MOTIVATION

1.1 Background


Routers with Very Small Buffers

Mihaela Enachescu*, Yashar Ganjali*, Ashish Goel*, Nick McKeown†, and Tim Roughgarden*
*Department of Computer Science, Stanford University
{mihaela, yganjali, ashishg}@cs.stanford.edu
†Department of Electrical Engineering, Stanford University
nickm@stanford.edu

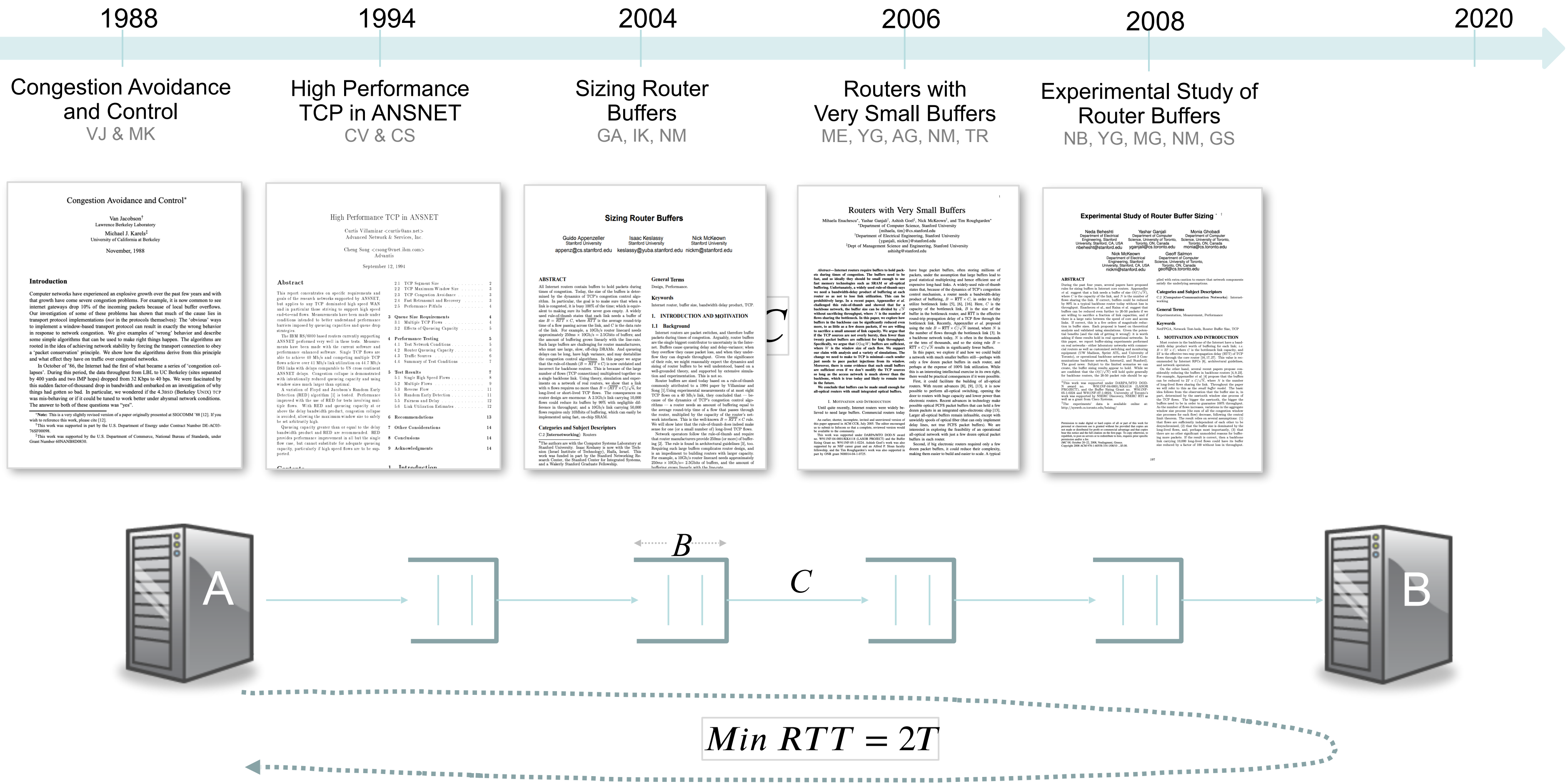
Assumptions

1. Paced Traffic
2. Link utilization < 80%

Consequences

Only 20-50 packet buffers.

$$Min\ RTT = 2T$$


1988
1994
2004
2006
2008
2020
Congestion Avoidance and Control
VJ & MK
High Performance TCP in ANSNET
CV & CS
Sizing Router Buffers
GA, IK, NM
Routers with Very Small Buffers
ME, YG, AG, NM, TR
Experimental Study of Router Buffers
NB, YG, MG, NM, GS
Congestion Avoidance and Control
High Performance TCP in ANSNET
Sizing Router Buffers
Routers with Very Small Buffers
Experimental Study of Router Buffer Sizing
A
B
C
B
Min RTT = 2T

Buffer Sizing Experiments

Small Buffers

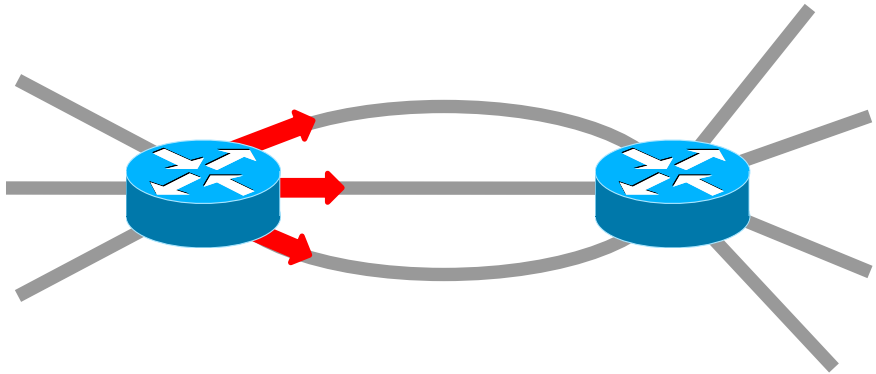
- Stanford University dorm network
- University of Wisconsin
- Internet2
- Level 3 Communications

Tiny Buffers

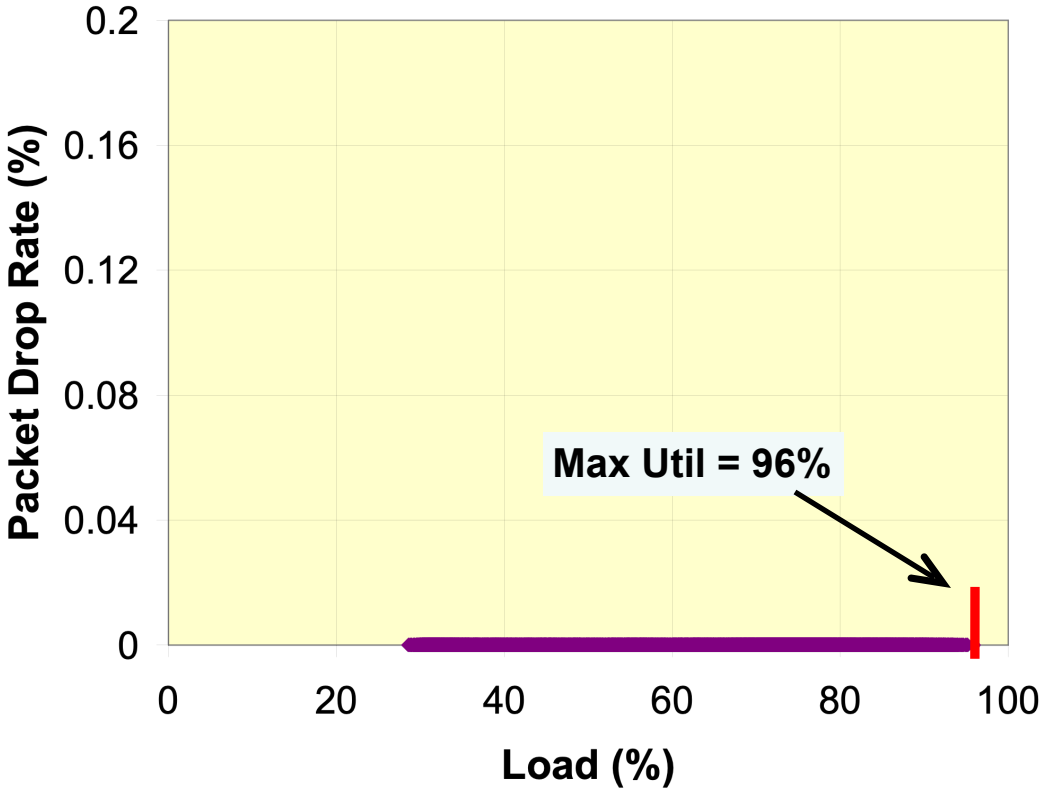
- Internet2
- Sprint Advanced Technology Lab
- University of Toronto

Level 3 Communications Experiments

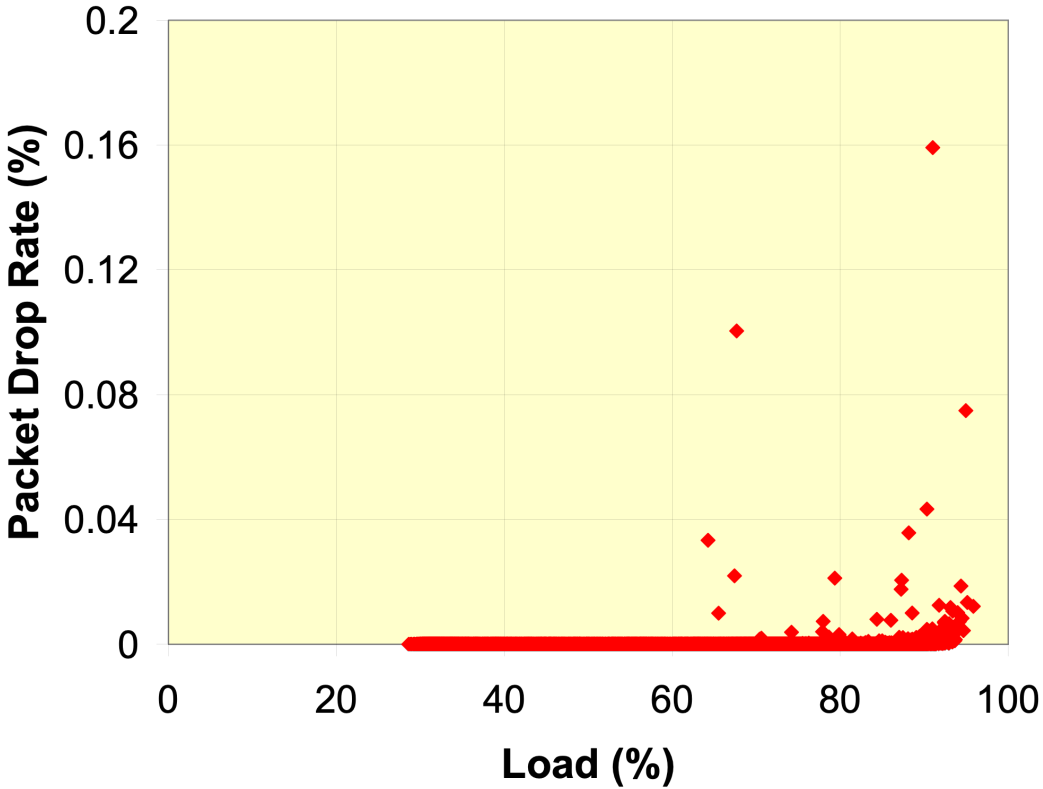
- High link utilization
- Long duration (about two weeks)
- Buffer sizes 190ms (250K packets), 10ms (10K packets), 2.5ms (2500 packets), 1ms (1000 packets)
- Load balancing over 3 links (2.5 Gb/s each)



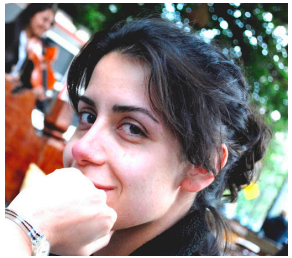
Drop vs. Load, Buffer = 190ms, 10ms



Drop vs. Load, Buffer = 1ms



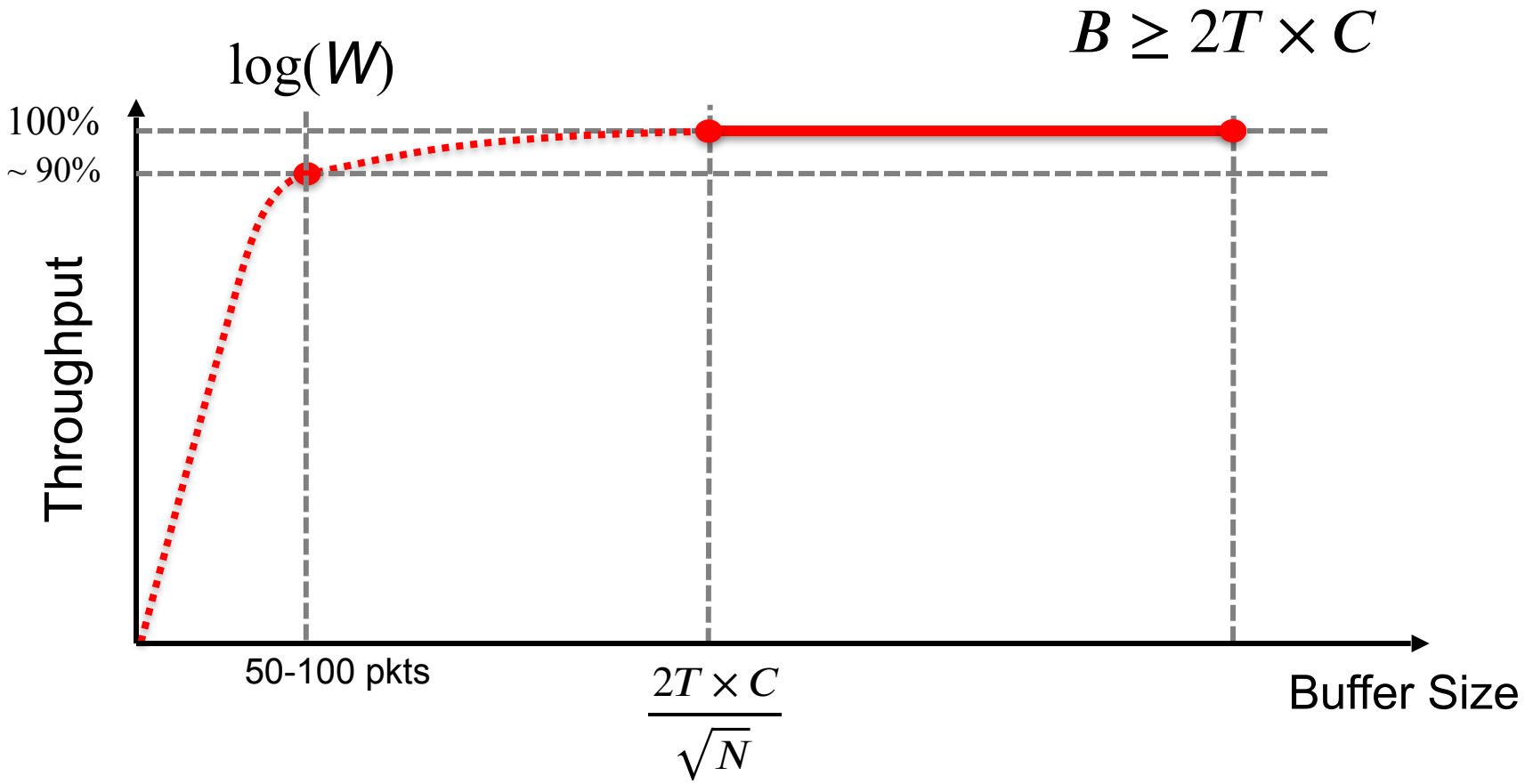
Internet2 Experiments



Neda Beheshti

2010

Summary of throughput and buffer size



End.