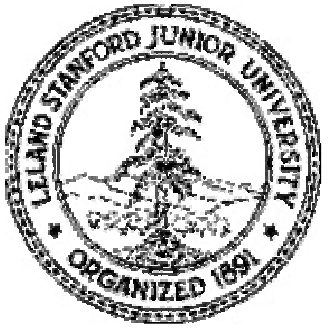
Sizing Router Buffers

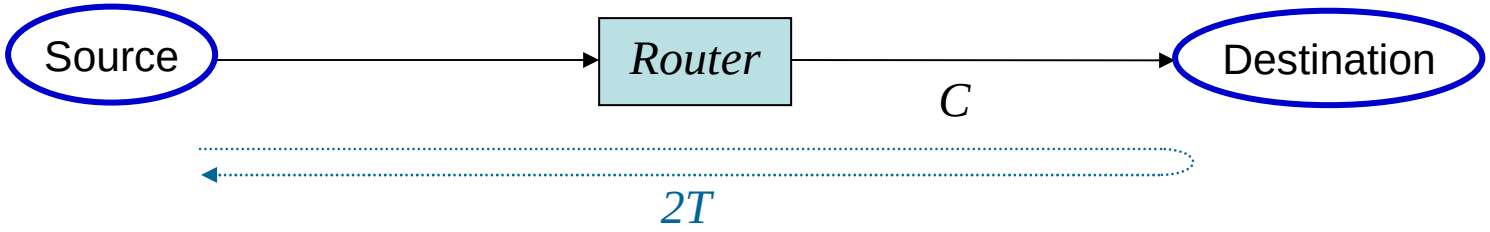


Guido Appenzeller
Isaac Keslassy
Nick McKeown

Stanford University

http://yuba.stanford.edu/~appenz

How much Buffer does a Router need?



- Universally applied rule-of-thumb:
  - A router needs a buffer size: $B = 2T \times C$
    - 2T is the two-way propagation delay (or just 250ms)
    - C is capacity of bottleneck link
- Context
  - Mandated in backbone and edge routers.
  - Appears in RFPs and IETF architectural guidelines.
  - Usually referenced to Villamizar and Song: “High Performance TCP in ANSNET”, CCR, 1994.
  - Already known by inventors of TCP [Van Jacobson, 1988]
  - Has major consequences for router design

Example

- 10Gb/s linecard
  - Requires 300Mbytes of buffering.
  - Read and write 40 byte packet every 32ns.
- Memory technologies
  - DRAM: require 4 devices, but too slow.
  - SRAM: require 80 devices, 1kW, $2000.
- Problem gets harder at 40Gb/s
  - Hence RLDRAM, FCRAM, etc.

Outline of this Talk

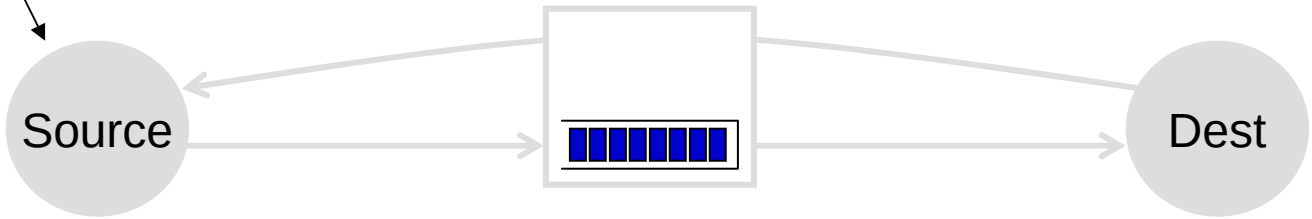
- The “Rule-of-Thumb” on Buffer Sizing is incorrect
  - Where the rule of thumb comes from
  - Why it is incorrect for a core router in the Internet today
- Real Buffer Requirements in case of Congestion
- Real Buffer Requirements without Congestion
- Experimental results from real Networks

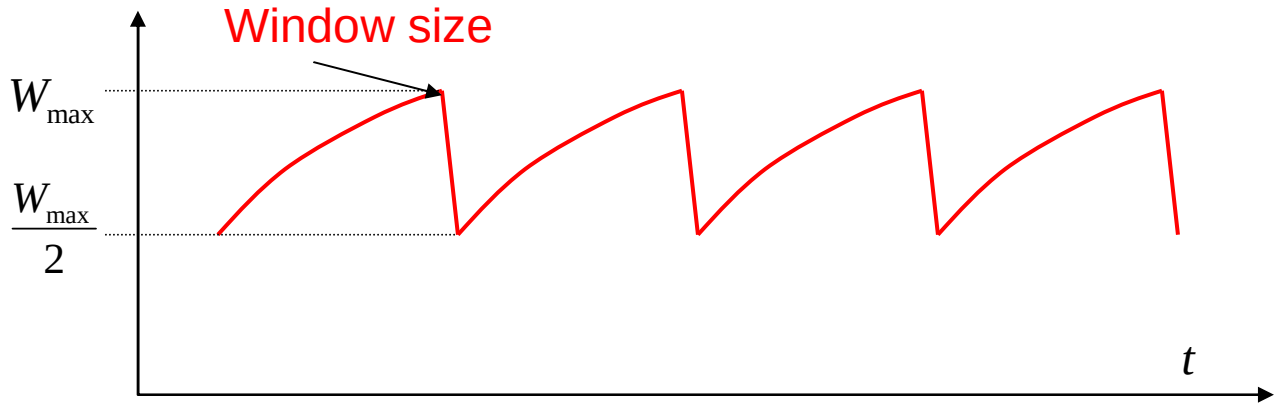
TCP Congestion Control

Rule for adjusting W

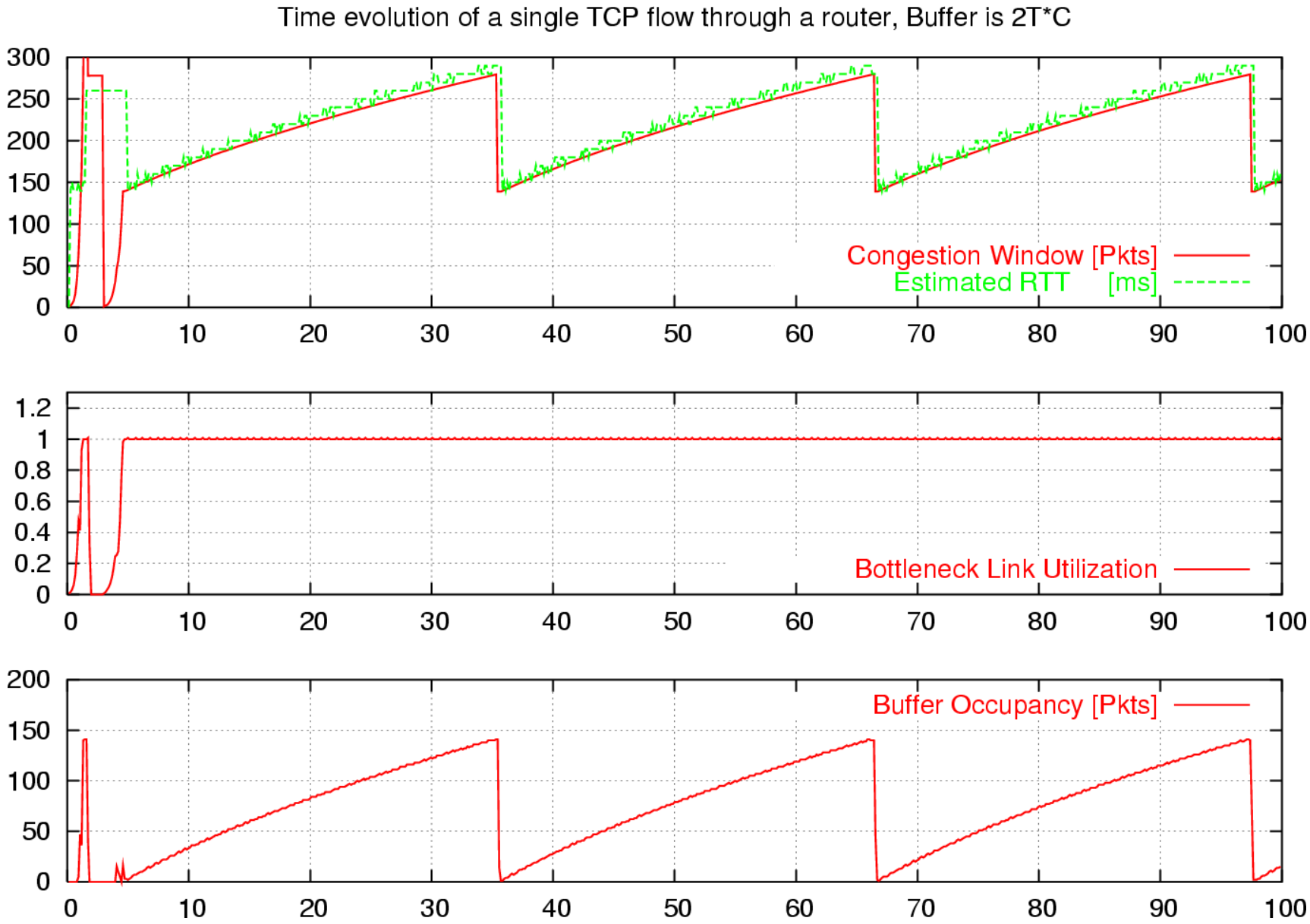
- If an ACK is received: W ← W+1/W
- If a packet is lost: W ← W/2

Only W packets may be outstanding

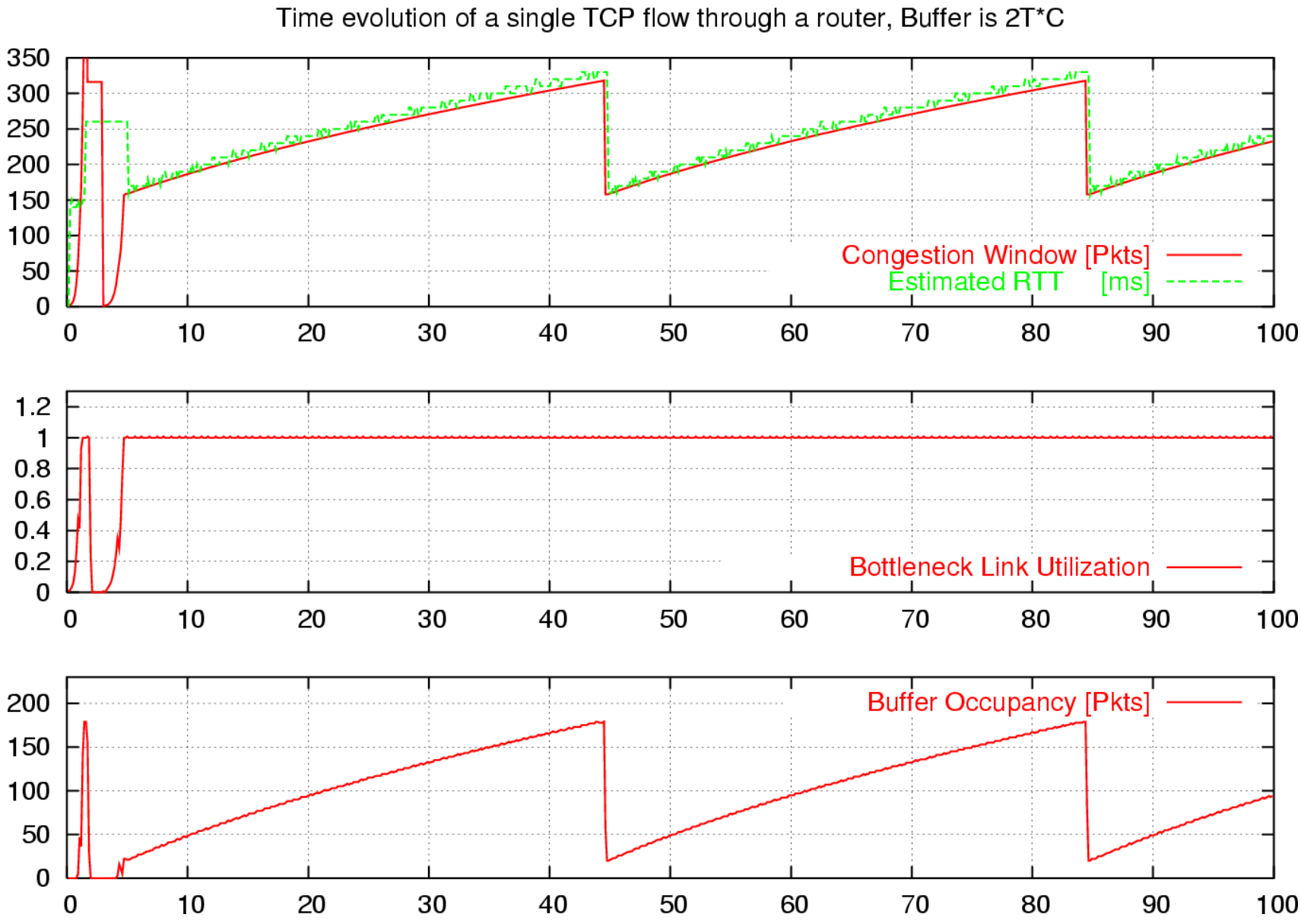




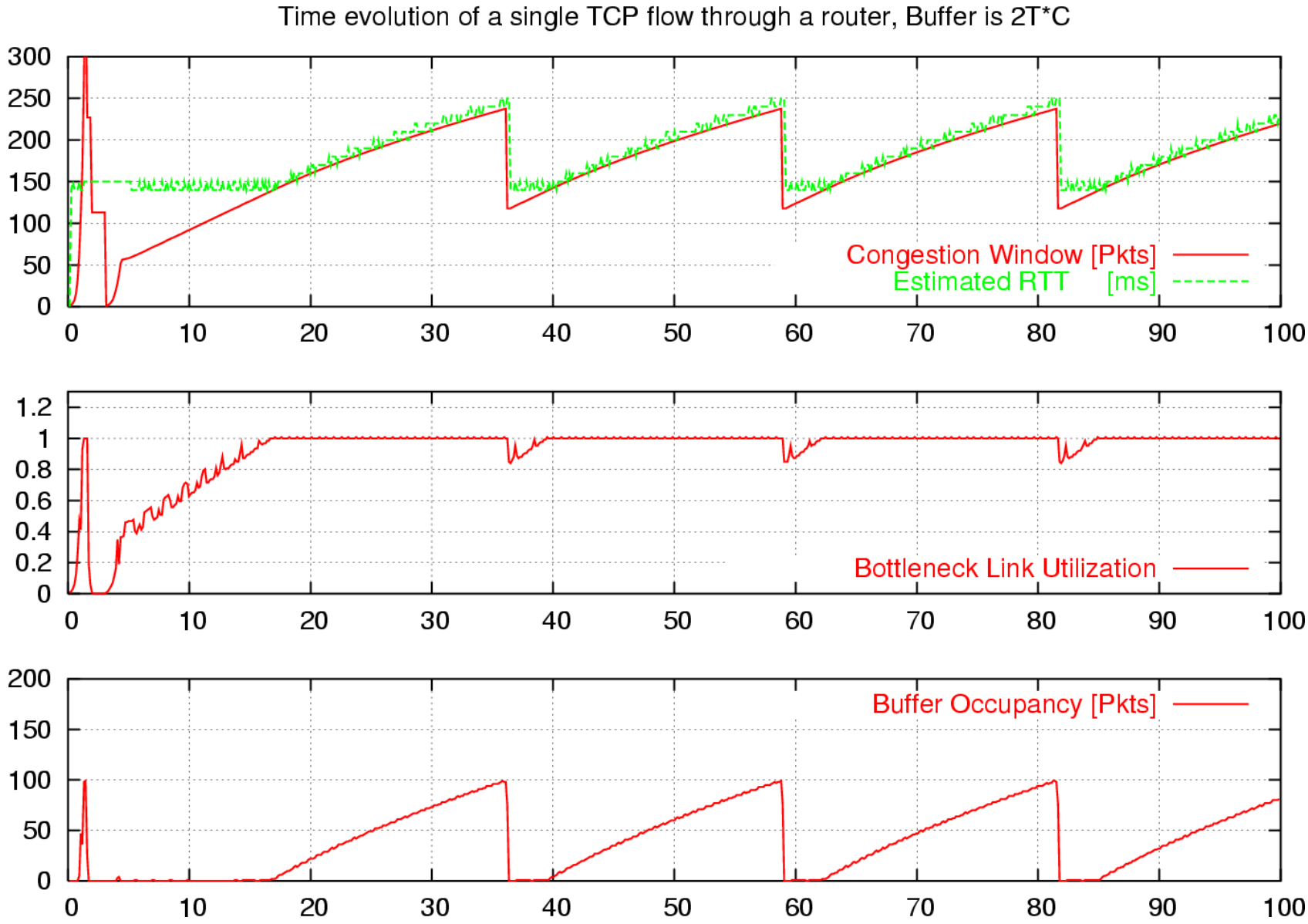
Buffer = rule of thumb



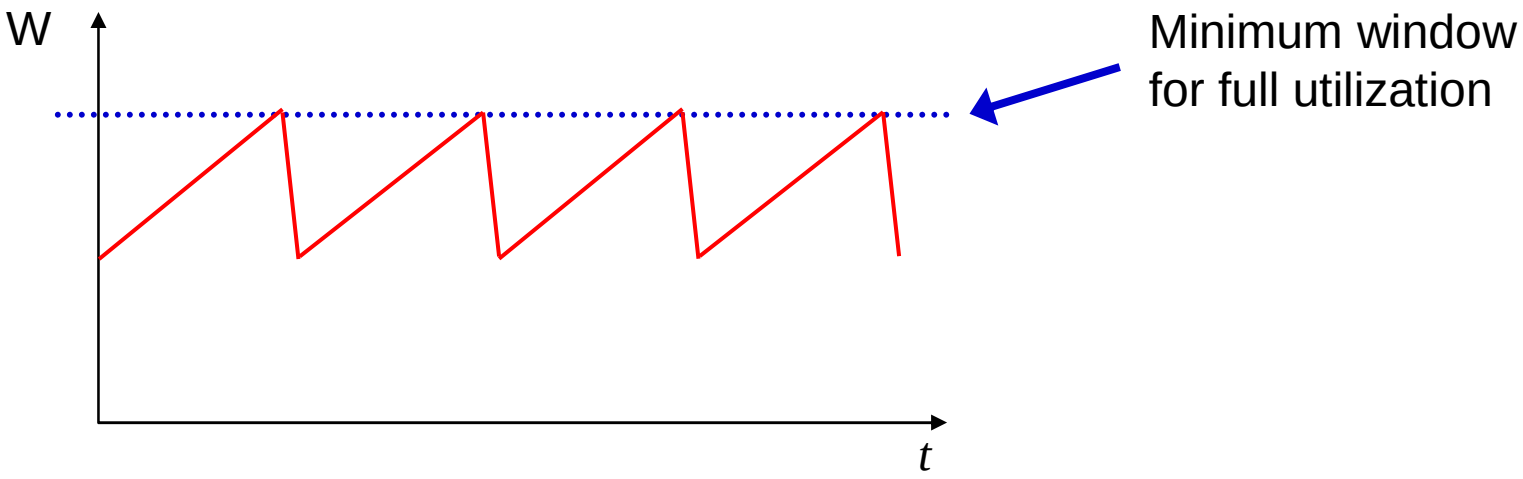
Over-buffered Link



Under-buffered Link

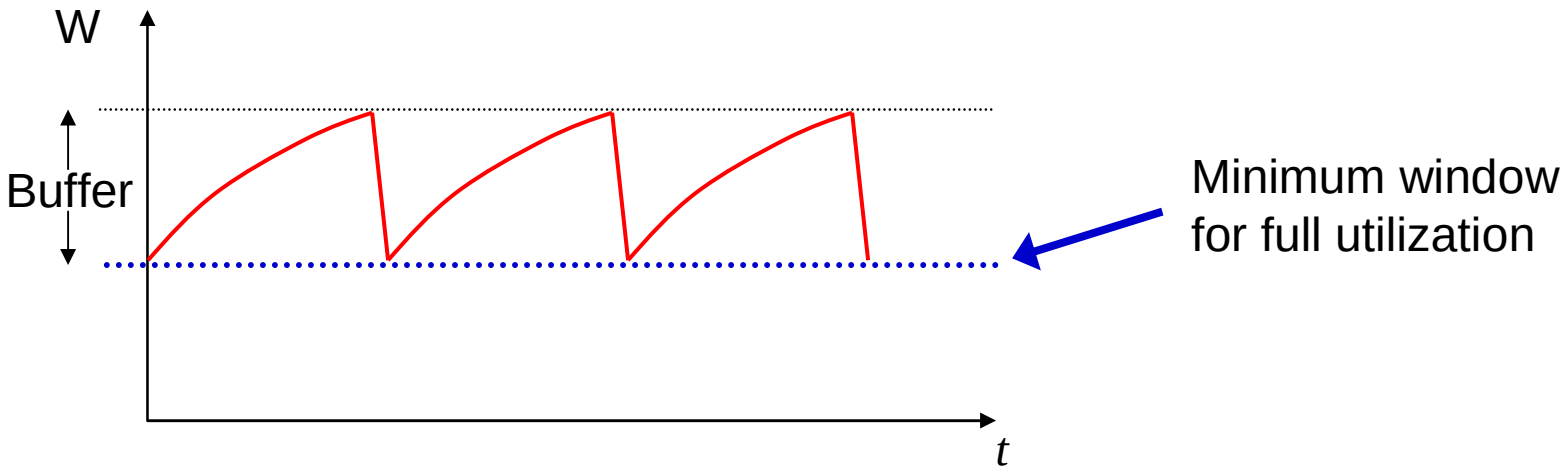


Summary Unbuffered Link



- The router can't fully utilize the link
  - If the window is too small, link is not full
  - If the link is full, next window increase causes drop
  - With no buffer it still achieves 75% utilization

Summary Buffered Link



- With sufficient buffering we achieve full link utilization
  - The window is always above the critical threshold
  - Buffer absorbs changes in window size
    - Buffer Size = Height of TCP Sawtooth
    - Minimum buffer size needed is 2T*C
  - This is the origin of the rule-of-thumb

Rule-of-thumb

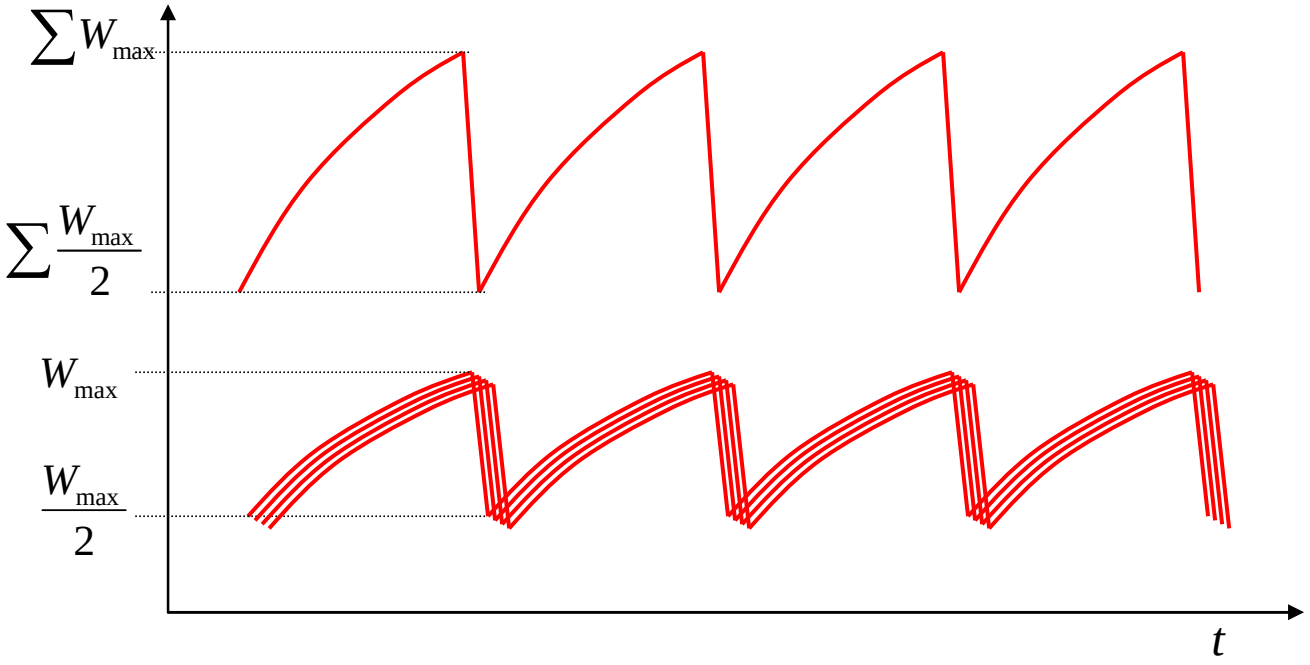
- Rule-of-thumb makes sense for one flow
- Typical backbone link has > 20,000 flows
- Does the rule-of-thumb still hold?

- Answer:
  - If flows are perfectly synchronized, then Yes.
  - If flows are desynchronized then No.

Outline of this Talk

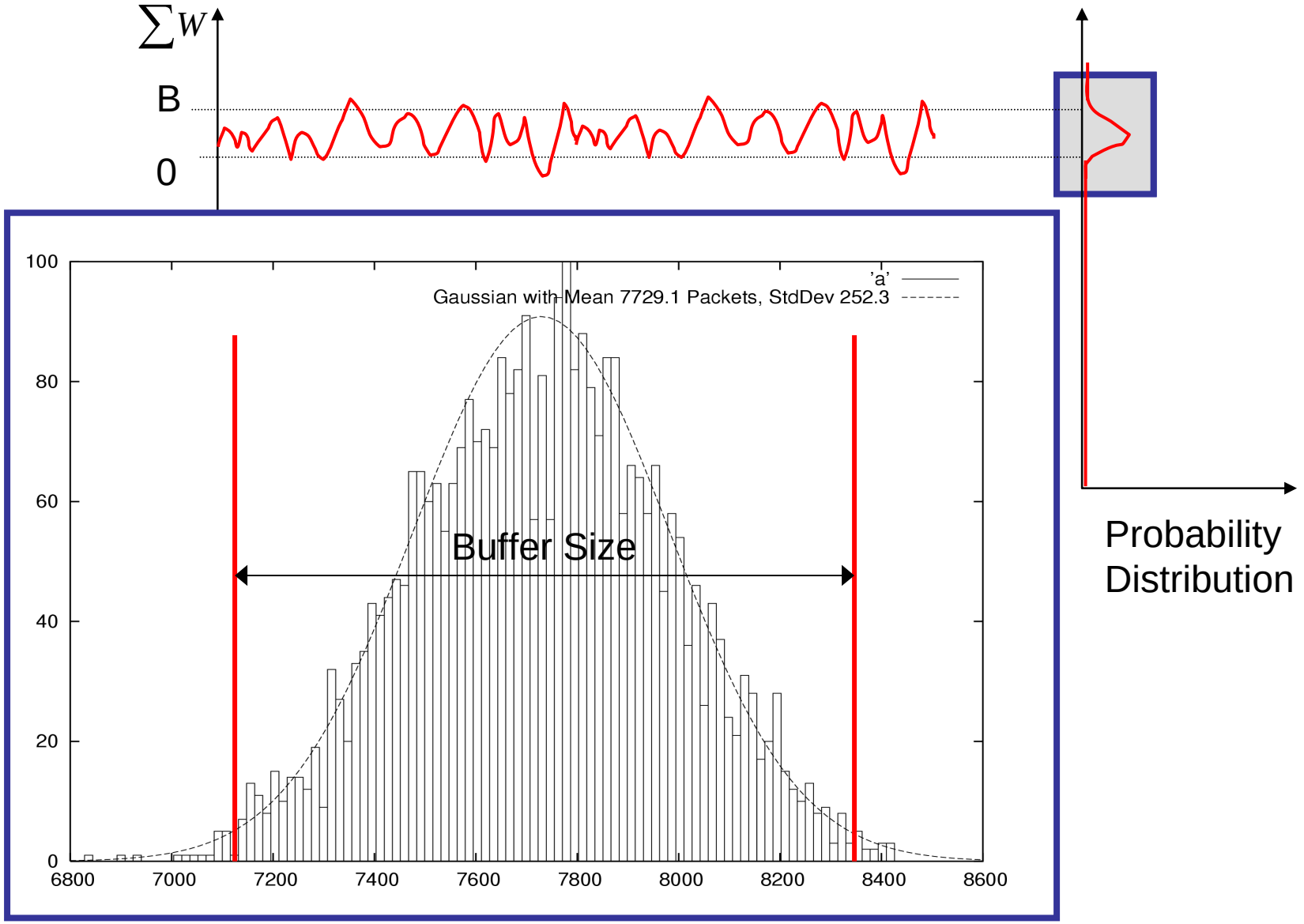
- The “Rule-of-Thumb” on Buffer Sizing is incorrect
- Real Buffer Requirements in case of Congestion
  - Correct buffer requirements for a congested router
  - Result: $B = 2T \times C / \sqrt{n}$
- Real Buffer Requirements without Congestion
- Experimental results from real Networks

If flows are synchronized



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

If flows are not synchronized

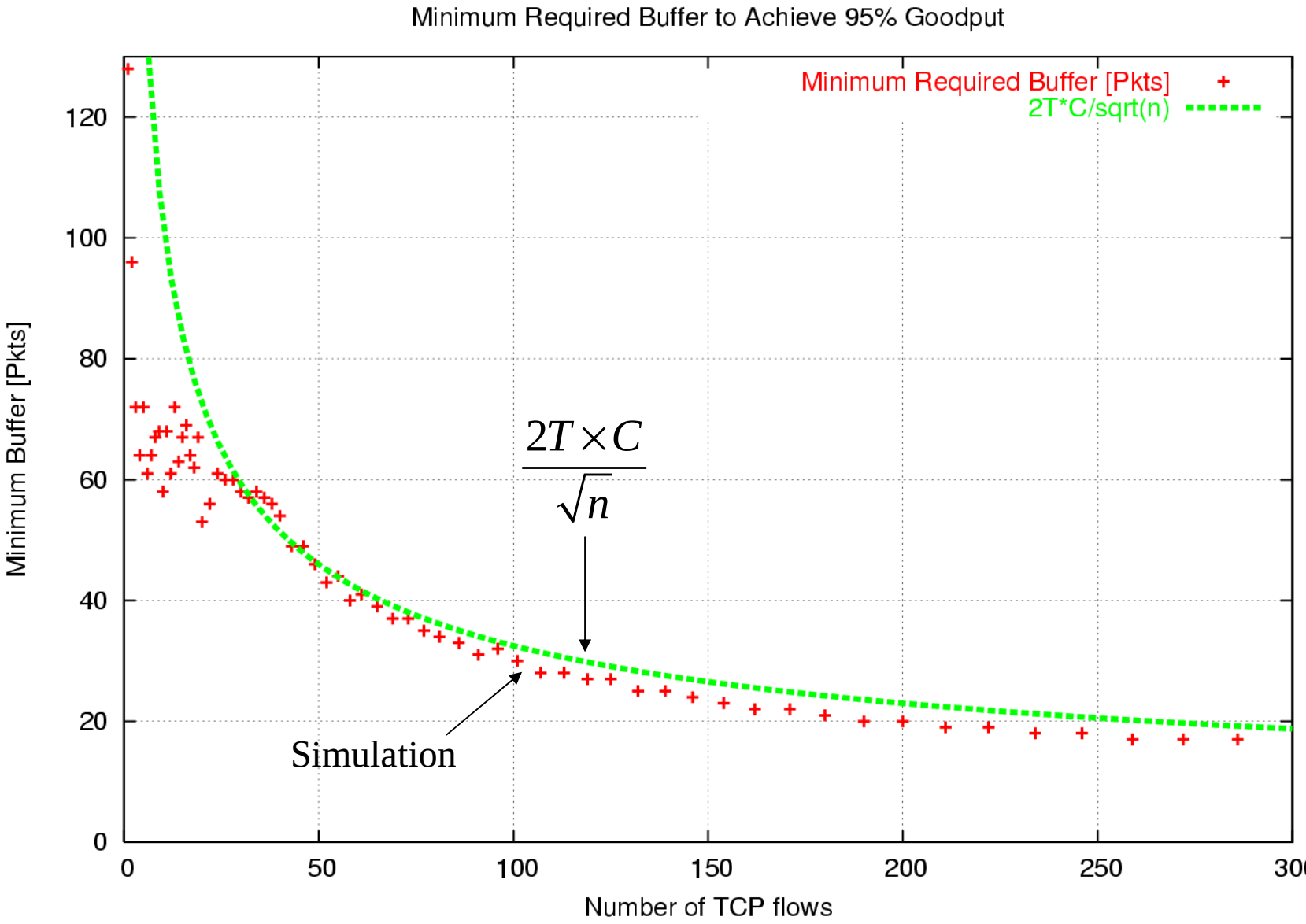


Central Limit Theorem

- CLT tells us that the more variables (Congestion Windows of Flows) we have, the narrower the Gaussian (Fluctuation of sum of windows)
  - Width of Gaussian decreases with $\frac{1}{\sqrt{n}}$
  - Buffer size should also decreases with $\frac{1}{\sqrt{n}}$

$$B \rightarrow \frac{B_{n=1}}{\sqrt{n}} = \frac{2T \times C}{\sqrt{n}}$$

Required buffer size



Summary Congested Router

- Flows in the core are desynchronized
  - Substantial experimental evidence
  - Supported by ns2 simulations

- For desynchronized, long-lived flows you need only buffers of

$$B = \frac{2T \times C}{\sqrt{n}}$$

Outline of this Talk

- The “Rule-of-Thumb” on Buffer Sizing is incorrect
- Real Buffer Requirements in case of Congestion
- Real Buffer Requirements without Congestion
  - Correct buffer requirements for an over-provisioned network
  - Result: Even smaller buffers
- Experimental results from real Networks

Uncongested Router

- So far we were assuming long flows in congestion avoidance mode.
  - What about flows in slow-start?
  - Do buffer requirements differ?

- Answer: Yes, you need even fewer buffers

  Caveat:

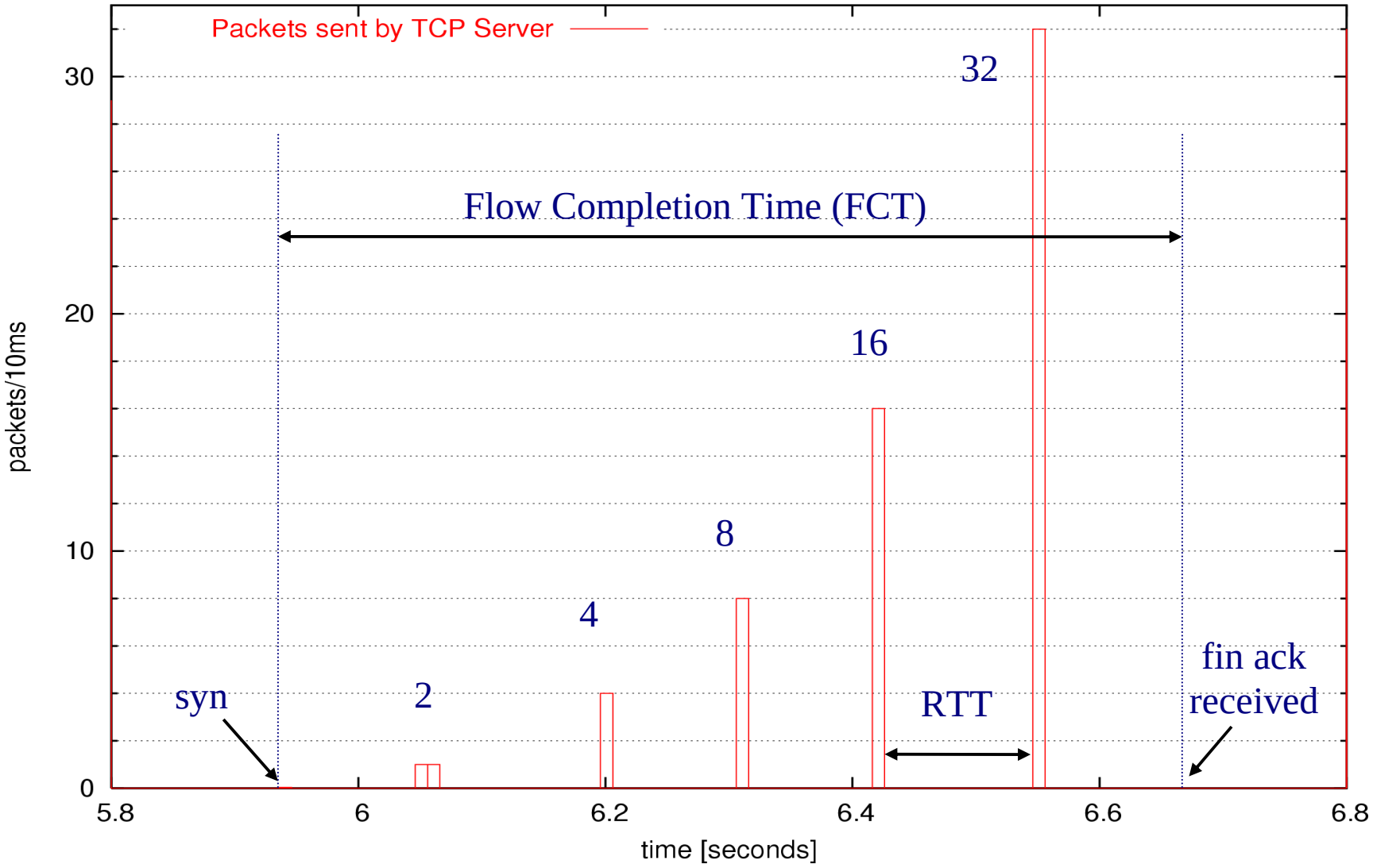
  In mixes of long and short flows, long flow effects dominate

  Therefore:

  - Short flow effects are only of interest on uncongested routers
  - Only useful if you have an overprovisioned network and that never is congested

A single, short-lived TCP flow

Flow length 62 packets, RTT ~140 ms



Short Flow Model

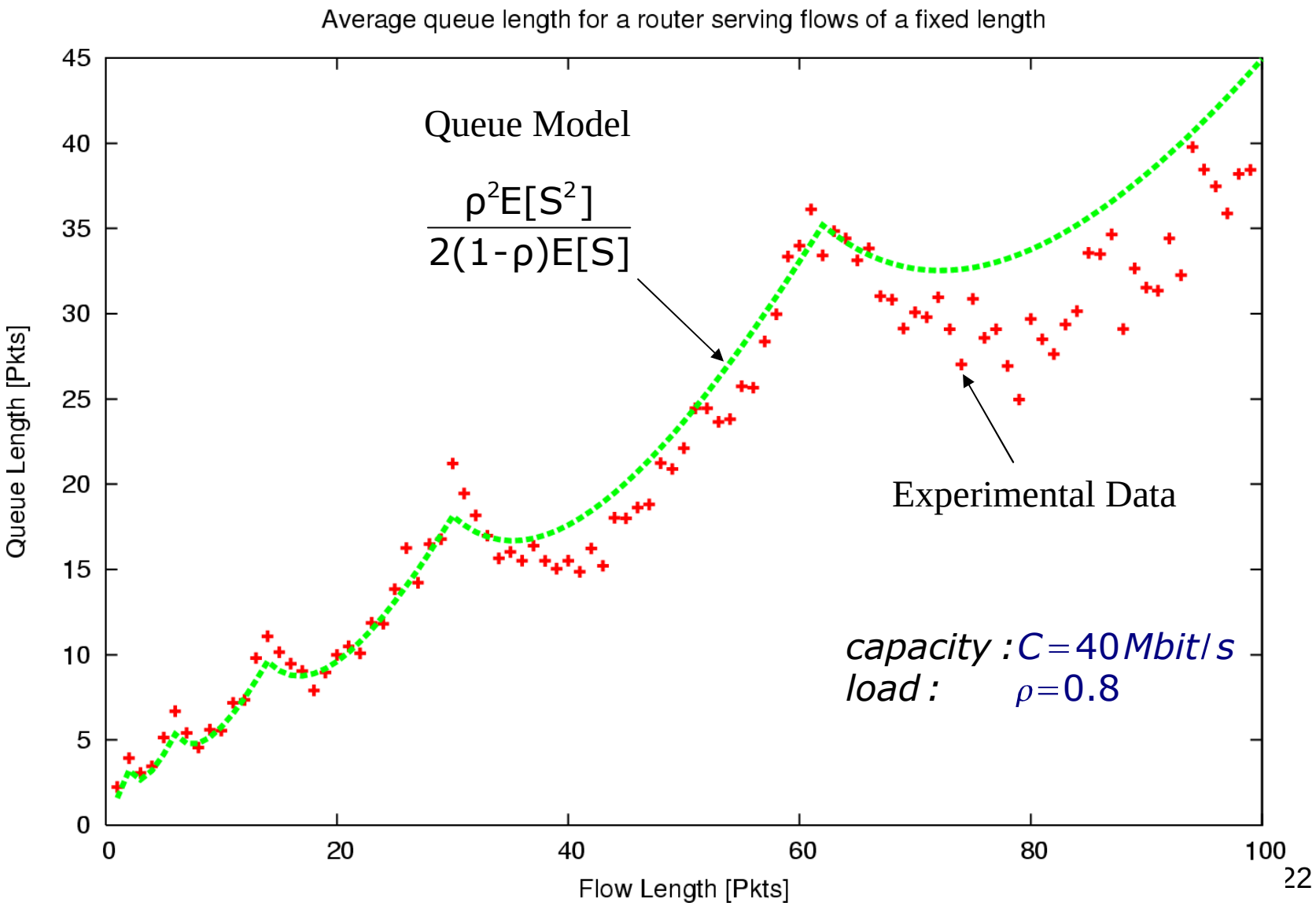
- Basic Idea: Model queue distribution

  Allows to calculate:

  - loss rate
  - average queue length
  - flow completion time

- Complete model is in the paper

Average Queue length



Buffers for Short Flows on an uncongested Router

- Results from Short-Flow Model
  - Buffer required only depends on lengths of bursts and load
  - Example - for bursts of up to size 16 at load 80%
    - For 1% loss probability B = 115 Packets
    - For 0.01% loss probability B = 230 packets etc.
    - Bursts of size 12 is maximum for Windows XP

  - This is independent of line speed and RTT
    - Same for a 1 Mb/s router and a 40 Gb/s router!

- In mixes of flows, long flow effects dominate
  - Also holds for length distributions, e.g. Pareto

Outline of this Talk

- The “Rule-of-Thumb” on Buffer Sizing is incorrect
- Real Buffer Requirements in case of Congestion
- Real Buffer Requirements without Congestion
- Results from real Networks
  - Lab results with a physical router
  - Experiments on production networks with real traffic

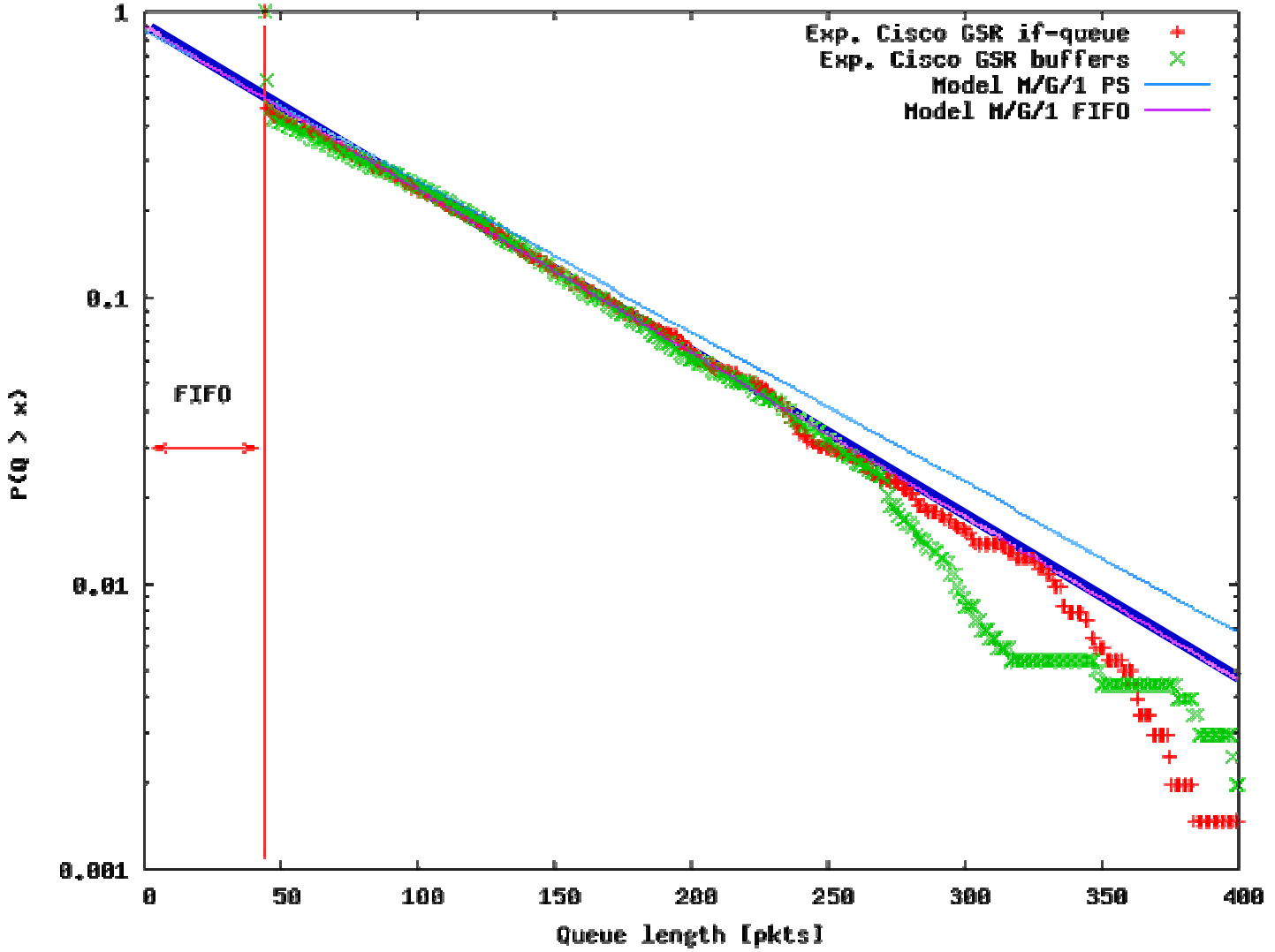
Long Flows – Utilization
Model vs. ns2 vs. Physical Router
GSR 12000, OC3 Line Card

TCP Flows	Router Buffer			Link Utilization		
	$\frac{2T \times C}{\sqrt{n}}$	Pkts	RAM	Model	Sim	Exp
100	0.5 x	64	1Mb	96.9%	94.7%	94.9%
	1 x	129	2Mb	99.9%	99.3%	98.1%
	2 x	258	4Mb	100%	99.9%	99.8%
	3 x	387	8Mb	100%	99.8%	99.7%
400	0.5 x	32	512kb	99.7%	99.2%	99.5%
	1 x	64	1Mb	100%	99.8%	100%
	2 x	128	2Mb	100%	100%	100%
	3 x	192	4Mb	100%	100%	99.9%

Thanks to Joel Sommers and Paul Barford of University of Wisconsin-Madison

Short Flows – Queue Distribution

Model vs. Physical Router, OC3 Line Card



Experiments with live traffic (I)

- Stanford University Gateway
  - Link from internet to student dormitories
  - Estimated 400 concurrent flows, 25 Mb/s
  - 7200 VXR (shared memory router)

TCP Flows	Router Buffer		Link Utilization	
	$\frac{2T \times C}{\sqrt{n}}$	Pkts	Model	Exp
400	0.8 x	46	95.9%	97.4%
	1.2 x	65	99.5%	97.6%
	1.5 x	85	99.9%	98.5%
	>>2 x	500	100%	99.9%

Thanks to Sunia Yang, Wayne Sung and the Stanford Backbone Team

Experiment with live traffic (II)

Internet2 link Indianapolis to Kansas City

- Link Setup
  - 10Gb/s link, T640
  - Default Buffer: ~1000 ms
  - Flows of 1 Gb/s
  - Loss requirement < 10^{-8}

- Experiment
  - Reduced buffer to 10 ms (1%) - nothing happened
  - Reduced buffer to 5 ms (0.5%) - nothing happened
  - Next: buffer of 2ms (0.2%)
  - Experiment ongoing…

Thanks to Stanislav Shalunov of Internet2 and Guy Almes (now at NSF)

How much buffer does a router need?

The old "Rule-of-Thumb"

Scenario	Buffer	Comments
▪ Single flow saturates router ▪ Few synchronized flows with congestion	$2T \times C$	Still applicable in a few select cases (e.g. I2 speed records)

The new "Rule-of-Thumb"

Scenario	Buffer	Comments
▪ Many flows ▪ Congestion	$\frac{2T \times C}{\sqrt{n}}$	Applicable for the core and edge of the internet today.
▪ One or many flows ▪ Not Congested, ρ << 1	100's of pkts	Works if there is ***never*** any congestion (optimists only)

Impact on Router Design

- 10Gb/s linecard with 200,000 x 56kb/s flows
  - Rule-of-thumb: Buffer = 2.5Gbits
    - Requires external, slow DRAM
  - Becomes: Buffer = 6Mbits
    - Can use on-chip, fast SRAM
- 40Gb/s linecard with 40,000 x 1Mb/s flows
  - Rule-of-thumb: Buffer = 10Gbits
  - Becomes: Buffer = 50Mbits

Thanks!