

# Introducing optical switching into the network



*ECOC 2005, Glasgow*

***Nick McKeown***

***High Performance Networking Group  
Stanford University***

***[nickm@stanford.edu](mailto:nickm@stanford.edu)***

***<http://www.stanford.edu/~nickm>***

# Network religion

## Bigger buffers are better

### Widely held assumptions

1. Big buffers (millions of packets) are necessary in Internet routers
2. You can't build big optical buffers
3. Therefore, you can't build an all-optical Internet router

So how can we use optical switching in the backbone.....?

# Optical switching in the backbone



## Approach 1: Don't use packet switching

- Dynamic Circuit Switching

## Approach 2: Packet switching with small buffers

- Challenge the assumption
  - What we learn in school
  - What we know about the Internet
- Conclusion: 20 packet buffers might be enough.

**We'll consider both...**

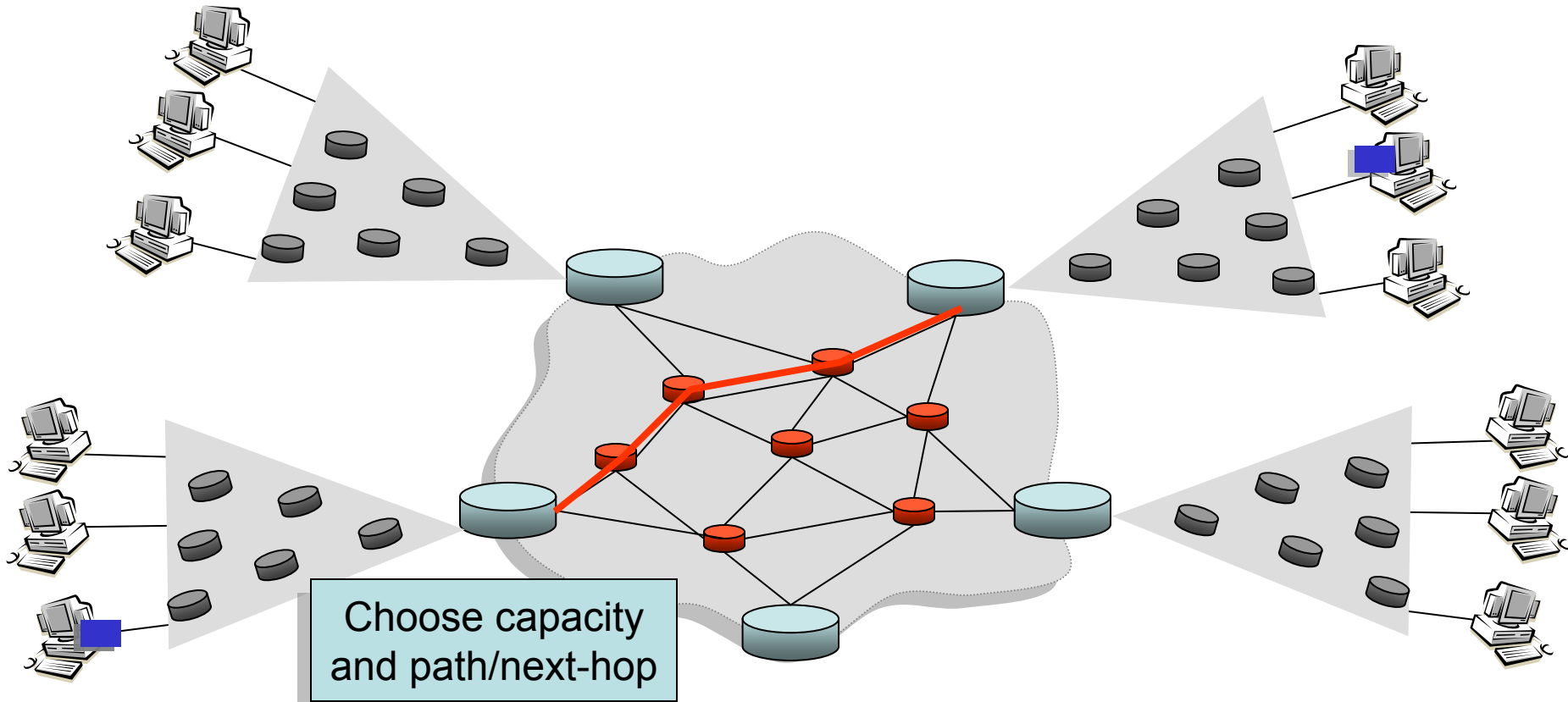
# Dynamic circuit switching in the backbone

- There seem to be some advantages
  - Well-suited to optics
  - Circuit switches are simple
    - “Start with a packet switch and throw most of it away”*
  - Higher capacity per unit volume
  - Higher capacity per watt
  - Lower cost per Gb/s
- There are some disadvantages
  - They are unfashionable

# Dynamic Circuit Switching (DCS)

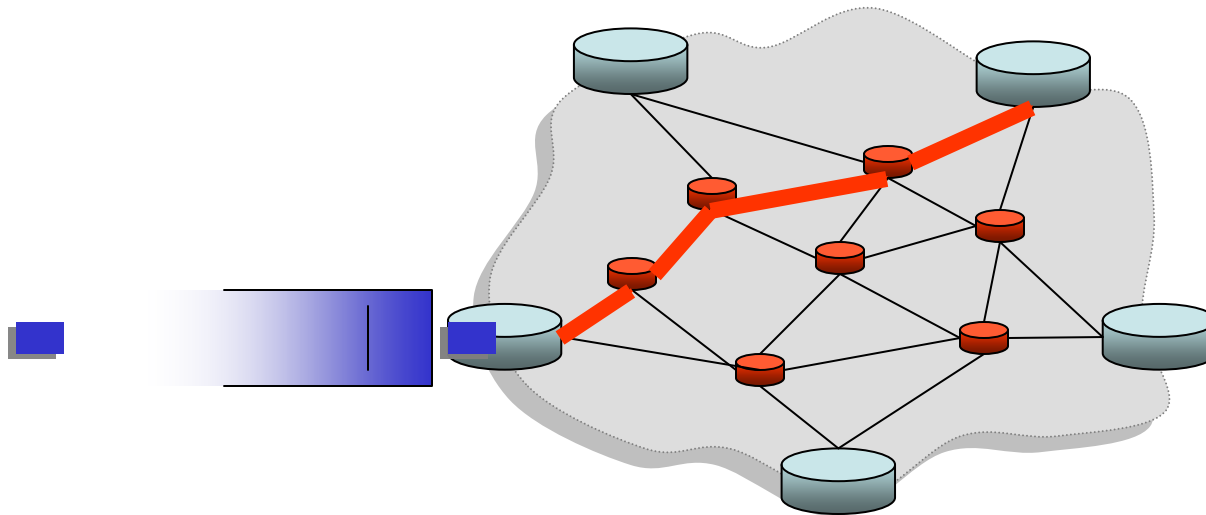
- Some “hurdles” need not be a problem:
  - Routing circuits
  - Signaling
  - Maintaining state
- Technical hurdles
  - When to create a new circuit?
  - How large to make the circuit?

# DCS: A circuit per flow *aka "TCP Switching"*



**Simple:** 10Gb/s boundary router prototyped on a desktop PC  
**Extreme granularity:** Demonstrates feasibility of DCS

# DCS: Capacity on demand between border routers



## Rule of thumb

Predict the need for capacity by monitoring how quickly new flows are created, rather than waiting for the buffer to fill

# My conclusion on dynamic circuit switching

- Compelling to operator: Cost, reliability, management, predictability
- Scalable with optical circuit switching
- Users can't tell the difference
  
- **Prediction:** The backbone will use DCS in 10 years time



# Optical switching in the backbone

## Approach 1: Don't use packet switching

- Dynamic Circuit Switching



## Approach 2: Packet switching with small buffers

- Challenge the assumption
  - What we learn in school
  - What we know about the Internet
- Conclusion: 20 packet buffers is probably enough.

**We'll consider both...**

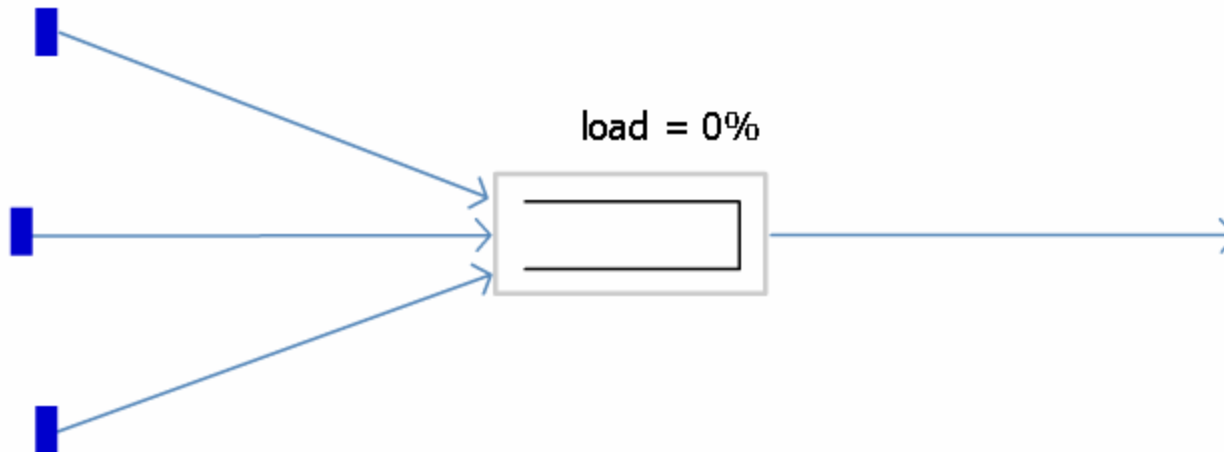
# Why bigger is not better

- Network users don't like buffers
- Network operators don't like buffers
- Router architects don't like buffers
  
- We don't need big buffers

# Where did the religion come from?

- Packet switching is good
  - Long haul links are expensive
  - Statistical multiplexing allows efficient sharing of long haul links
- Packet switching requires buffers
- Packet loss is bad
- Use big buffers
- Luckily, big electronic buffers are cheap

# Statistical Multiplexing

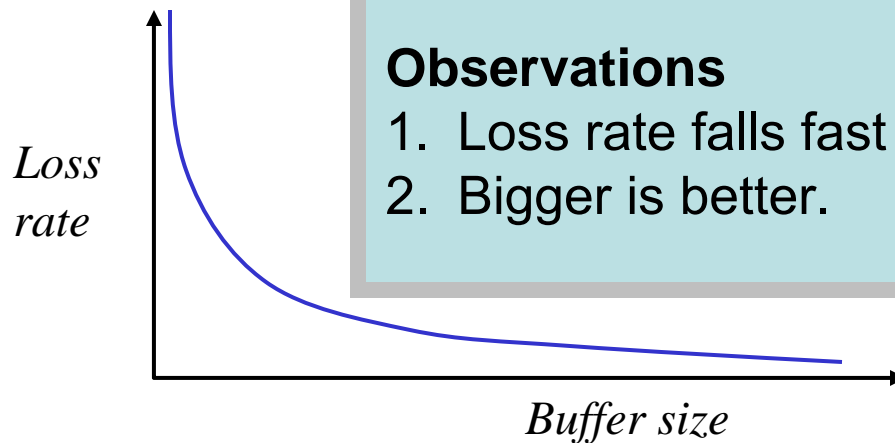
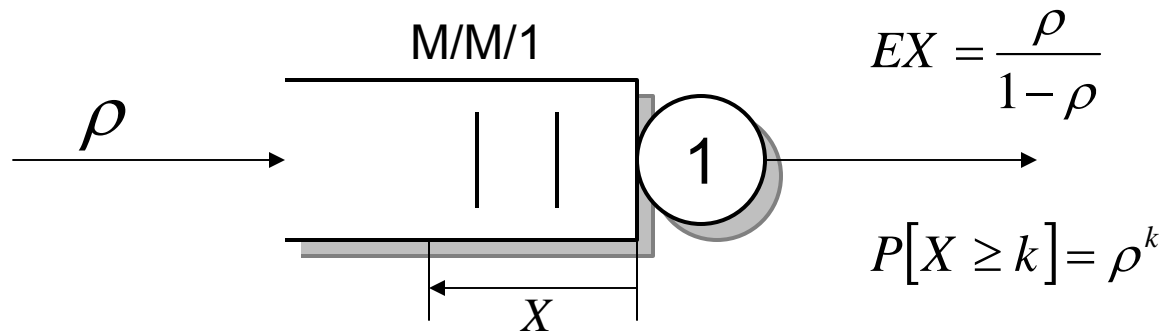


## Observations

1. The bigger the buffer, the lower the packet loss.
2. If the buffer never goes empty, the outgoing line is busy 100% of the time.

# What we learn in school

## 1. Queueing Theory



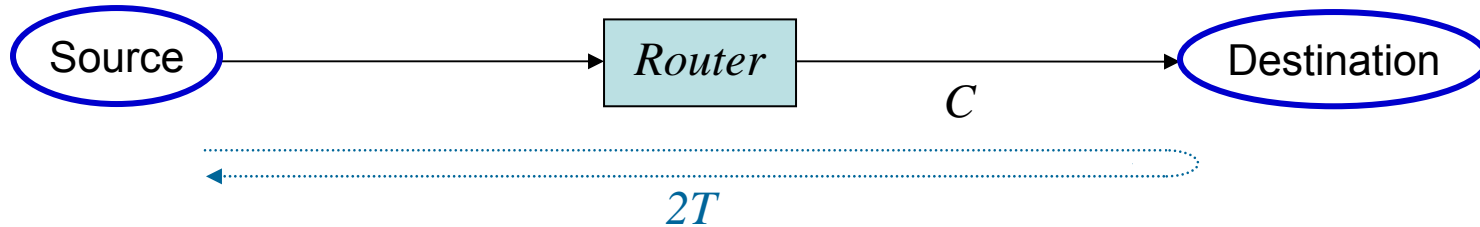
### Observations

1. Loss rate falls fast with increasing buffer size.
2. Bigger is better.

# What we learn in school

- **Moore's Law:** Memory is plentiful and halves in price every 18 months
  - 1Gbit memory: 500k packets for \$25
- **Conclusion**
  - Make buffers big

# Backbone router buffers



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

# Example

- 10Gb/s linecard
  - Rule-of-thumb: Buffer = 1 Million packets
- 40Gb/s linecard
  - Rule-of-thumb: Buffer = 4 Million packets
- Requires DRAM, but DRAM too slow
- Almost unthinkable at 160Gb/s



# Sizing buffers

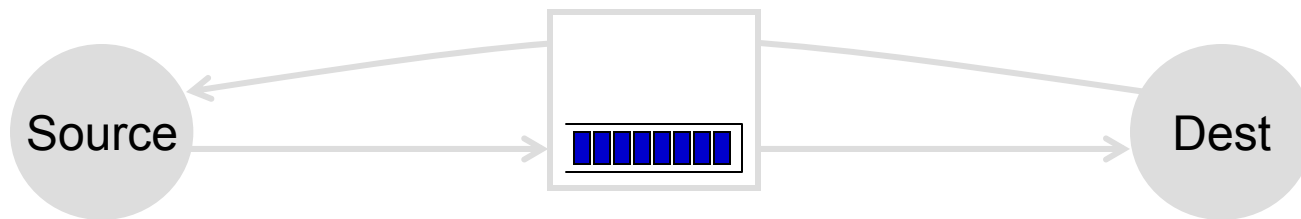
So where did the rule-of-thumb come from?

## **Answer**

With  $B = 2T \times C$ , a single TCP flow can keep the bottleneck link busy.

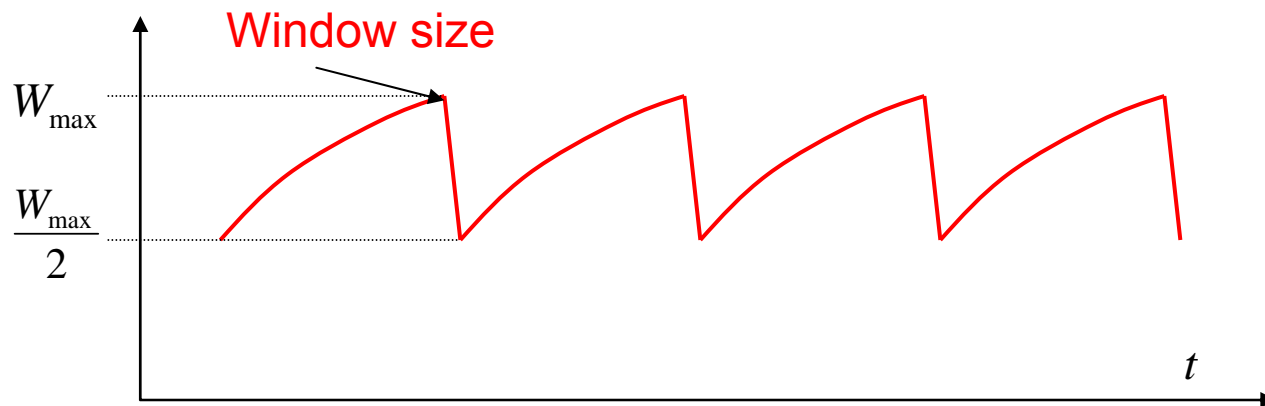
# TCP Congestion Control

Only  $W$  packets  
may be outstanding



## Rule for adjusting $W$

- If an ACK is received:  $W \leftarrow W + 1/W$
- If a packet is lost:  $W \leftarrow W/2$



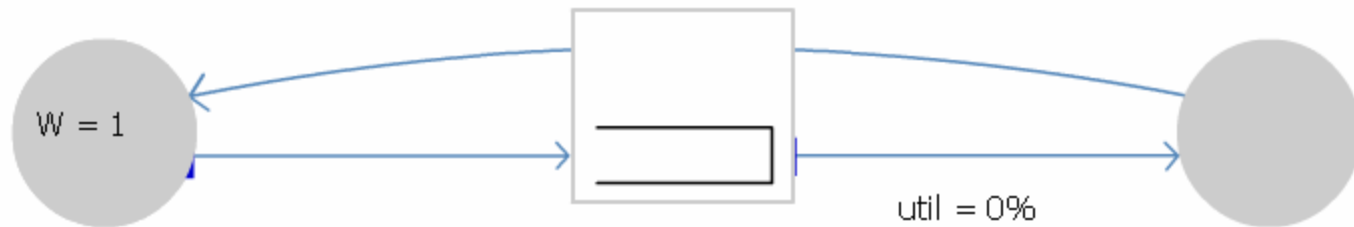
# TCP Congestion Control

Only  $W$  packets  
may be out

It follows that

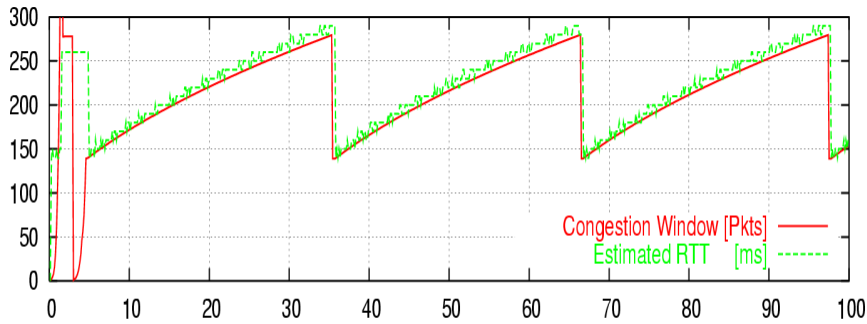
$$B = 2T \times C$$

$\leftarrow W+1/W$   
 $\leftarrow W/2$

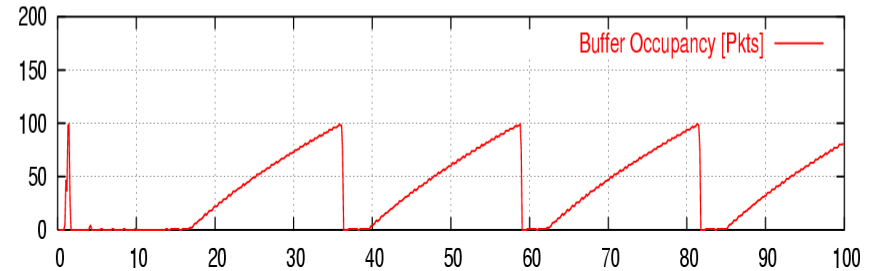
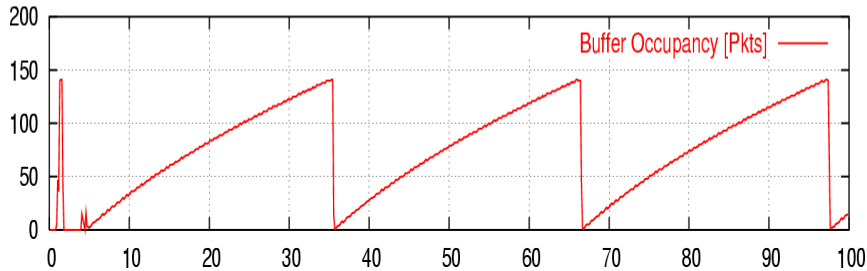
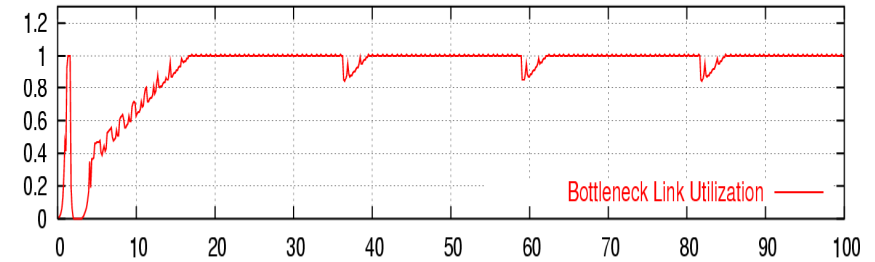
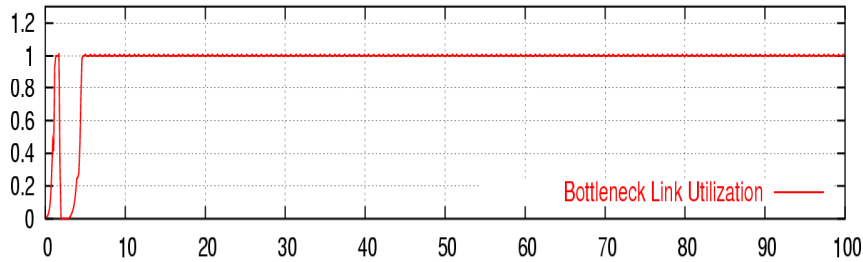
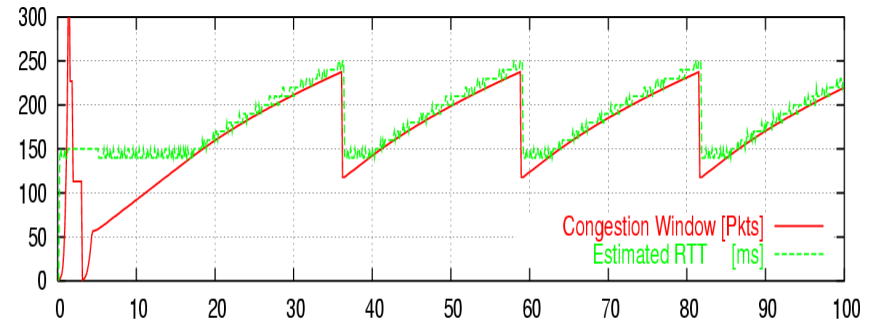


# Time evolution of a single TCP flow

Time evolution of a single TCP flow through a router. Buffer is  $2T \cdot C$



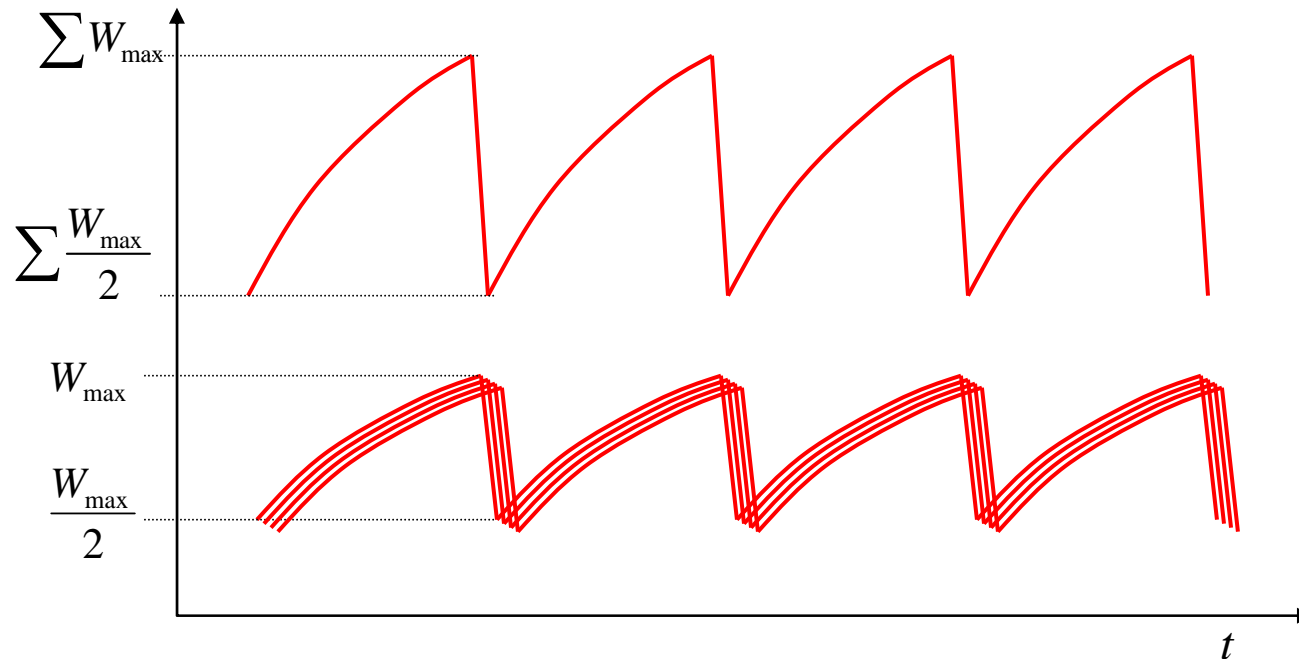
Time evolution of a single TCP flow through a router. Buffer is  $< 2T \cdot C$



# Rule-of-thumb

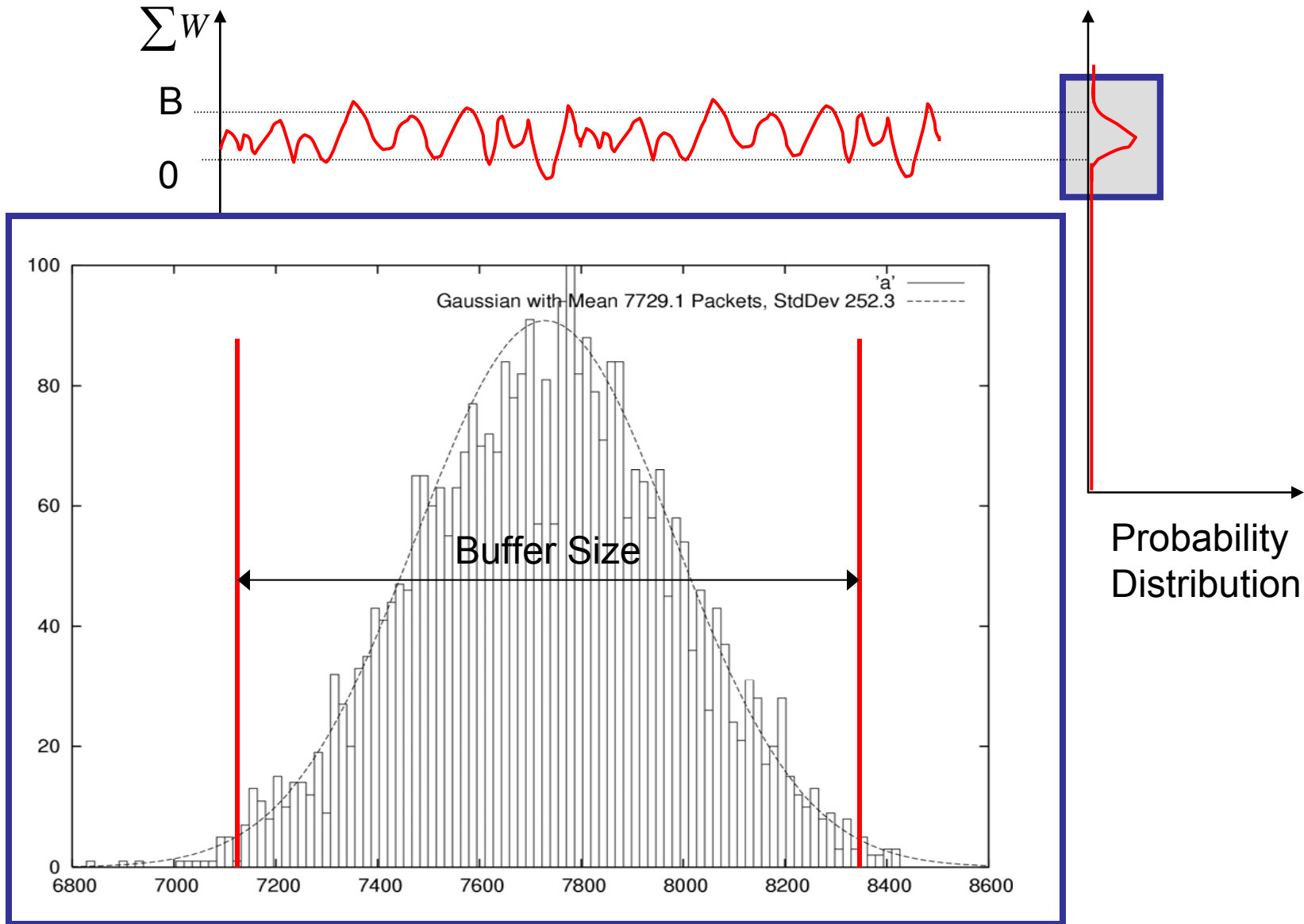
- Rule-of-thumb makes sense for one flow
- Typical backbone link has  $> 20,000$  flows
- Does the rule-of-thumb still hold?
  
- Answer: **No!**

# If flows were all synchronized



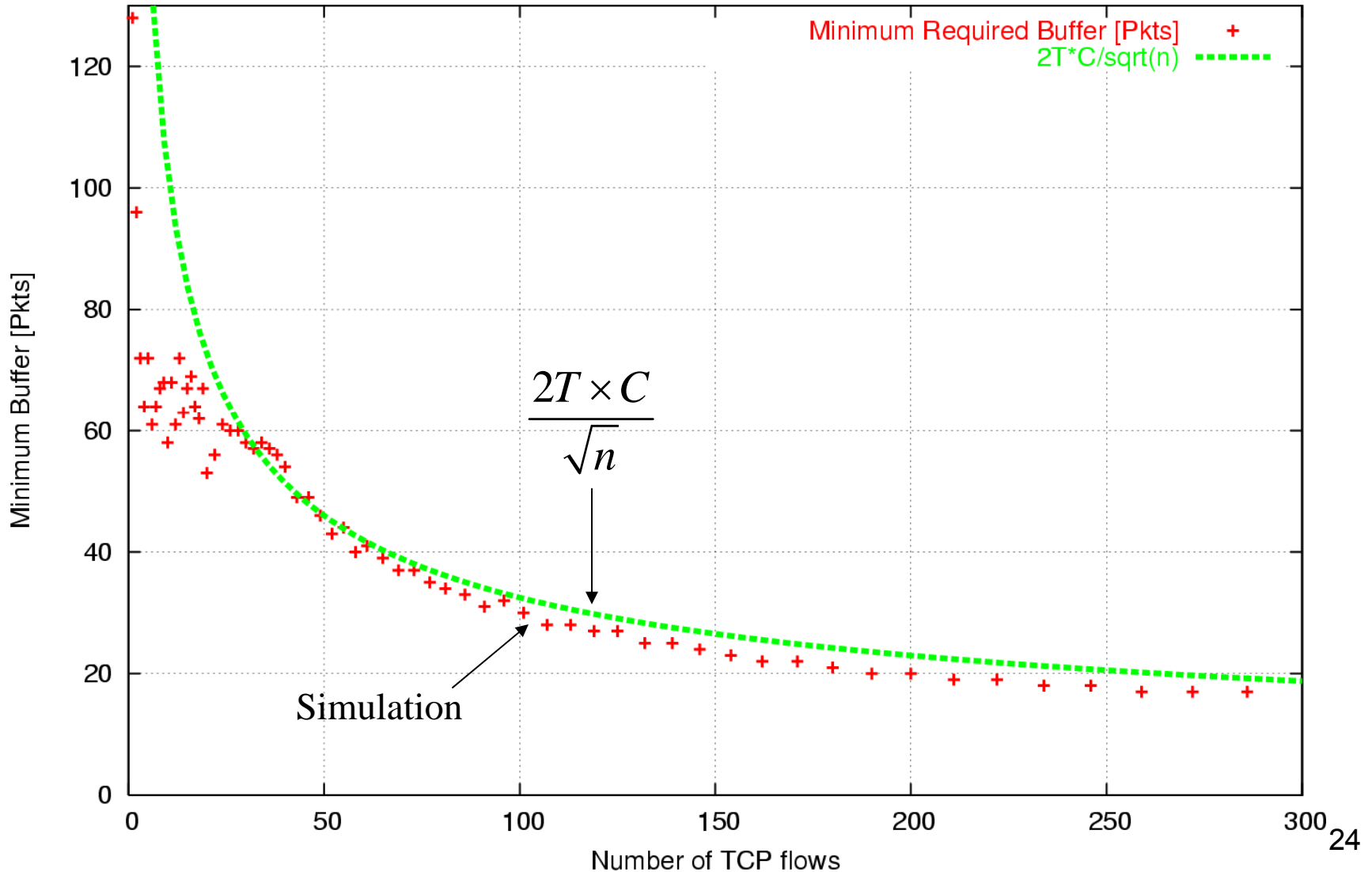
- Aggregate window has same dynamics
- Therefore buffer occupancy has same dynamics
- Rule-of-thumb still holds.

# If flows are not synchronized



# Required buffer size

Minimum Required Buffer to Achieve 95% Goodput





# Backbone router buffers

## ➤ Summary

- The rule of thumb is wrong for a core routers today
- Required buffer is  $\frac{2T \times C}{\sqrt{n}}$  instead of  $2T \times C$

## ➤ Validation

- Theory validated in two small networks (Stanford, University of Wisconsin), two operator networks, and over 10,000 simulations.
- But more work needed to understand worst-cases...

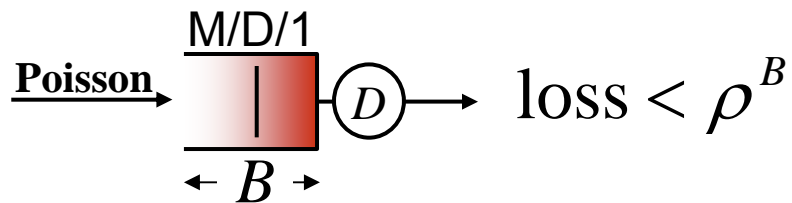
# Impact on Router Design

- 10Gb/s linecard with 200,000 x 56kb/s flows
  - Rule-of-thumb: Buffer = 1 Million packets
  - Becomes: Buffer = 3,000 packets
- 40Gb/s linecard with 40,000 x 1Mb/s flows
  - Rule-of-thumb: Buffer = 4 Million packets
  - Becomes: Buffer = 25,000 packets

Still way too many packets to store optically...

# What if...?

Theory (benign conditions)



*i.e.*  $\rho = 80\%$ ,  $B = 20\text{pkts} \Rightarrow \text{loss} < 1\%$

Loss independent of link rate,  
RTT, number of flows, etc.

Practice



Typical OC192 router linecard  
buffers over 1,000,000 packets

**5 orders of magnitude  
difference!**

Can we make traffic look “Poisson-enough”  
when it arrives to the routers...?

# Optical routers with very small buffers



## In collaboration with:

1. Stanford: Yashar Ganjali, Guido Appenzeller, Ashish Goel, Tim Roughgarden
2. LASOR team at DARPA, UCSB, Cisco, JDSU, Calient, Agility
3. Buffer Sizing teams: Frank Kelly (Cambridge), Damon Wischik (UCL), Don Towsley (UMass)



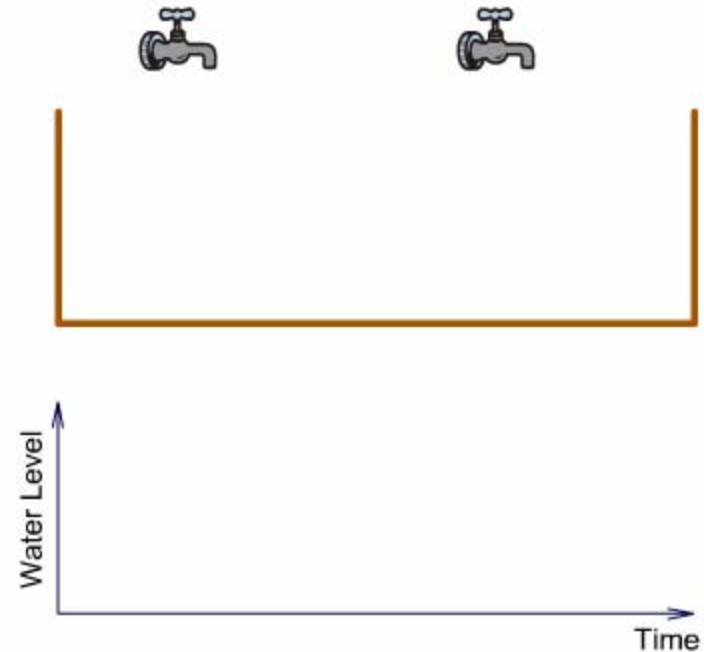
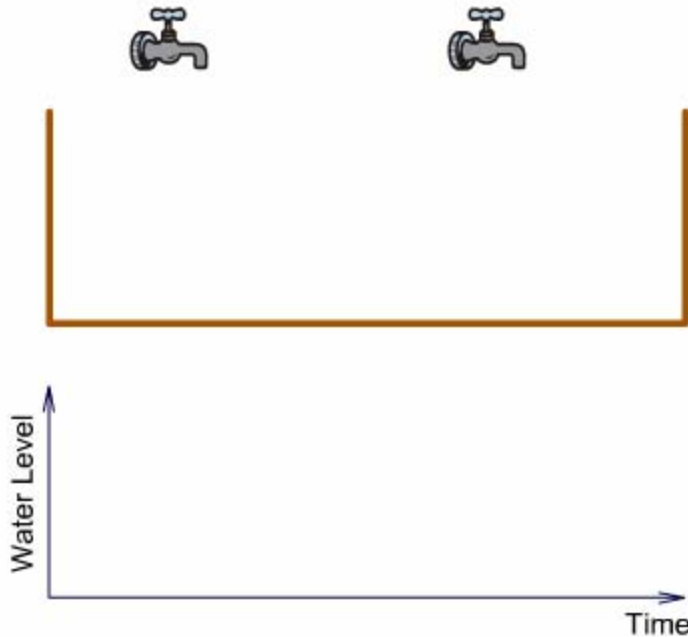
# Recent work to make buffers smaller

<b># packets at 10Gb/s</b>	<b>1,000,000</b>	<b>10,000</b>	<b>20</b>
$2T \times C \xrightarrow{(1)} \frac{2T \times C}{\sqrt{n}} \xrightarrow{(2)} O(\log W)$			
<b>Intuition &amp; Proofs</b>	<b>Sawtooth Peak-to-trough</b>	<b>Smoothing of many sawtooths</b>	<b>Non-bursty arrivals</b>
<b>Evidence</b>	<b>Simulated Single TCP Flow</b>	<b>Simulated Many TCP Flows</b> <b>Also: Real n/w experiments</b>	<b>Simulated Many TCP Flows</b> <b>Just starting: Real n/w experiments</b>

$W$  = window-size of a flow

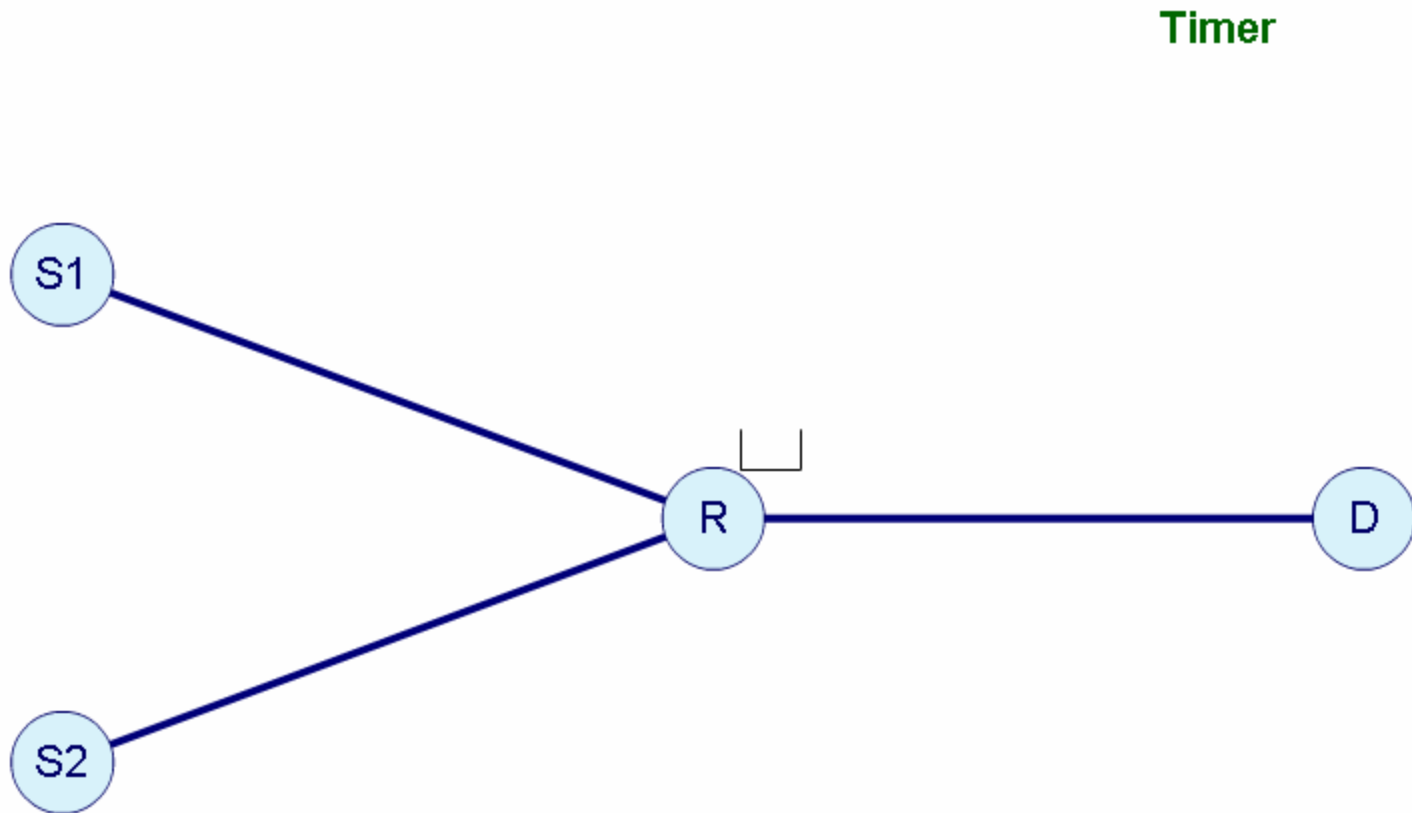
# Bursty TCP vs. Spacing packets

Bucket drains with a constant rate. Load is 90% for both cases.



# TCP Reno

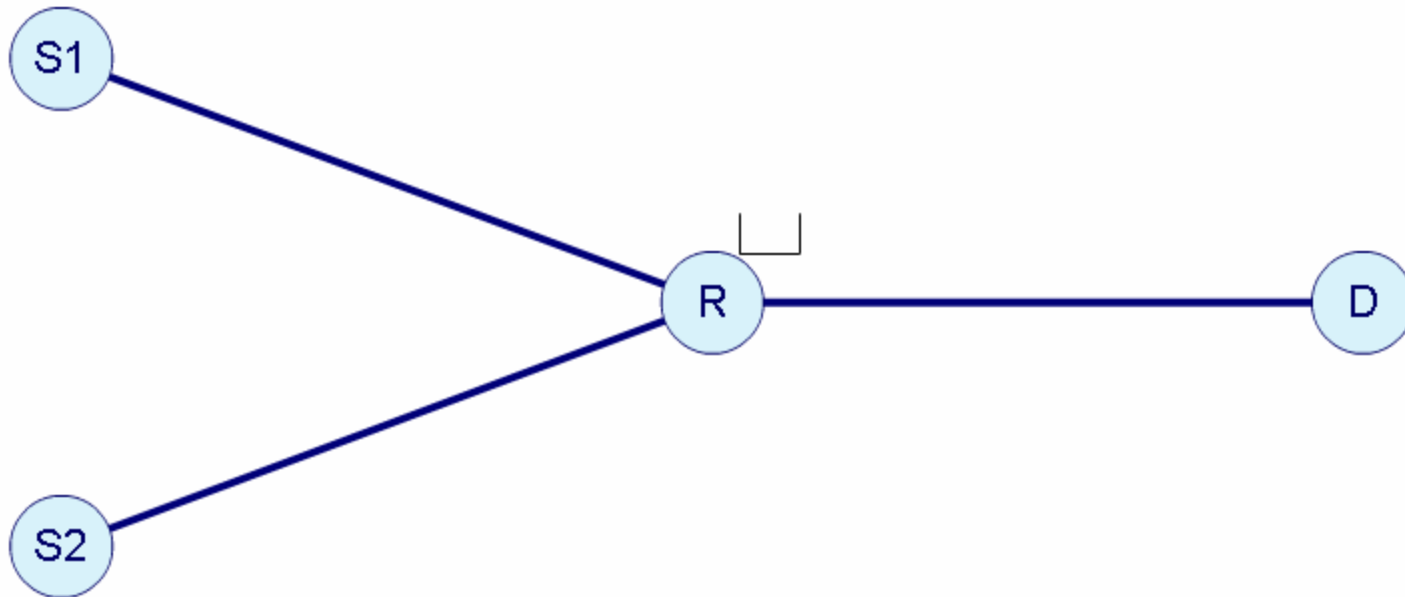
TCP Reno sends packets in bursts → High drop rate



# Spacing packets

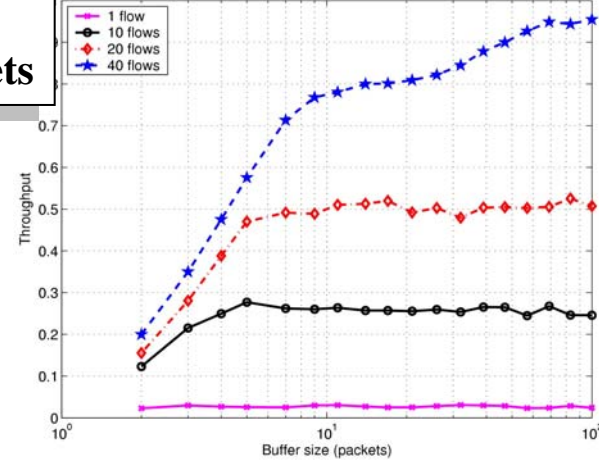
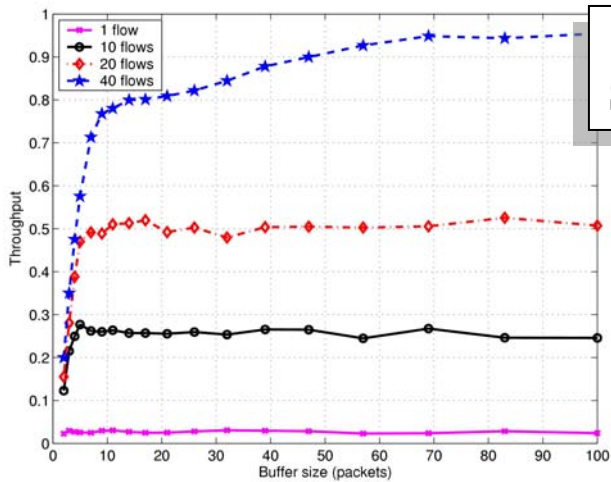
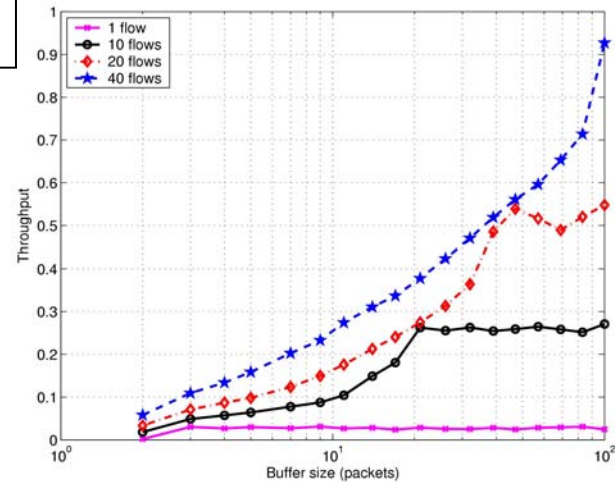
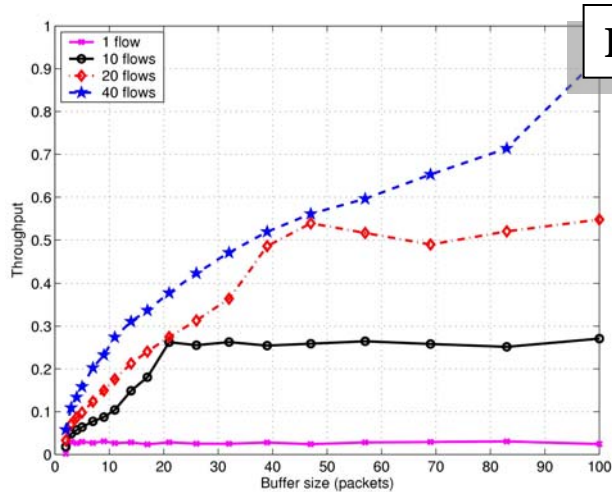
Spacing packets → Much lower drop rate

Timer



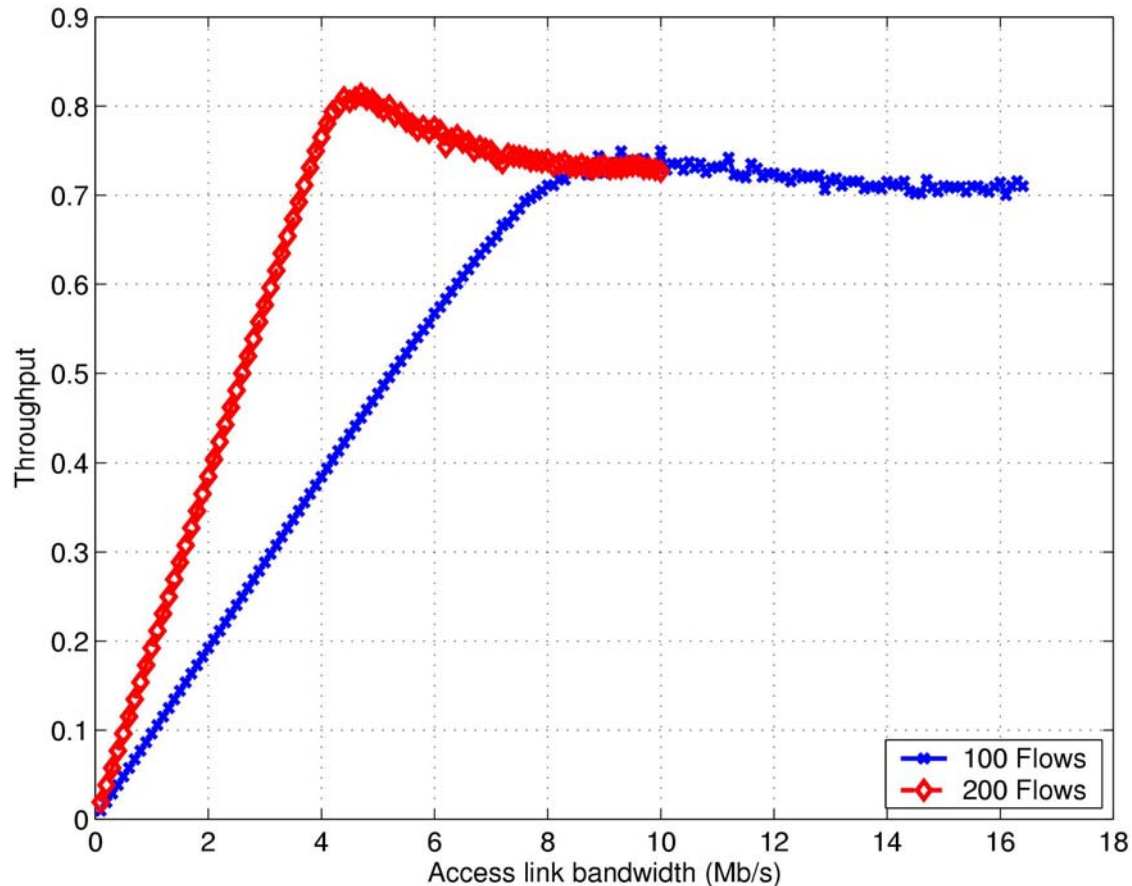


# $O(\log W)$ Buffers



# $O(\log W)$ Buffers

With a large ratio between core and access link bandwidth  
Bottleneck Bandwidth = 1Gb/s



# O(log W) Buffers

## Model

Assume: link over-provisioned by  $1/\rho C$ , and packets from each flow launched smoothly, or spaced out



$$B > \log_{1/\rho} \left( \frac{W_{\max}^2}{2(1-\theta)} \right) < 20$$

Similar results from Cambridge/UCL, UMass and Stanford  
See papers in: **ACM Computer Communications Review, July 2005**

# Backbone router buffers

## ➤ Summary

- If the backbone links are 100x faster than the access links, it seems 20 packet buffers will suffice
- Backbone links will lose 25-30% of their capacity. E.g. a 40Gb/s link will seem like a 28-30Gb/s link
- For faster flows or access links, use TCP Pacing to space packets

## ➤ Validation

- Theory and simulation so far
- Experiments underway...

# Conclusions

Will it be optical DCS or optical packet switching?

- Technically, both seem feasible
- Perhaps we shouldn't care
- Both are unfashionable

I'm hedging my bets...