Buffers: How we fell in love with them, and why we need a divorce

Hot Interconnects, Stanford 2004

Nick McKeown
High Performance Networking Group
Stanford University

nickm@stanford.edu
http://www.stanford.edu/~nickm
Which would you choose?

**DSL Router 1**
- $50
- 4 x 10/100 Ethernet
- 1.5Mb/s DSL connection
- 1Mbit of packet buffer

**DSL Router 2**
- $55
- 4 x 10/100 Ethernet
- 1.5Mb/s DSL connection
- 4Mbit of packet buffer
Network religion

Bigger buffers are better
Outline

- How we fell in love with buffers
- 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
  - We’d often be better off with smaller buffers
- Some examples

- How small could we make the buffers?
What we learn in school

- 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 buffers are cheap
Statistical Multiplexing

WARNING!!!
If you see this text, it means you need to have the Shockwave Flash files in the same directory as the powerpoint file. The Flash files are available, with this talk, at: http://www.stanford.edu/~nickm/talks

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. Can pick buffer size for a given loss rate.
2. Loss rate falls fast with increasing buffer size.
3. Bigger is better.
What we learn in school

2. Large Deviations

\[ P[X \geq k] \leq e^{-\delta k} \]

Observations
1. Arrival and service rates determine \( \delta \).
2. Can find bound on loss rate.
3. Bigger is better.
3. Networks of queues

**Observations**
1. Can find bound on loss rate.
2. Bigger is better.

**Queueing Theory:** Jackson networks, BCMP networks, …

**Large Deviations:** Additivity of effective bandwidths, decoupling bandwidth, …
Moore’s Law: Memory is plentiful and halves in price every 18 months.
- 1Gbit memory holds 500k packets and costs $25.

Conclusion:
- Make buffers big.
- Choose the $55 DSL router.
Why bigger isn’t 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
- We’d often be better off with smaller ones
Example

- 10Gb/s linecard
  - Rule-of-thumb: 250ms of buffering
  - Requires 300Mbytes of buffering.
  - Read and write 40 byte packet every 32ns.

- Memory technologies
  - DRAM: require 4 devices, but too slow.

- Problem gets harder at 40Gb/s
Sizing buffers

Packets are generated by a *closed-loop* feedback system

- 95% of traffic is TCP: End-to-end window-based flow control
- Queues with closed-loop source behave very differently
- TCP requires packet loss. Loss is not bad.
- Throughput is a better metric.
Review: TCP Congestion Control

Only $W$ packets may be outstanding

Rule for adjusting $W$

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

Window size

$W_{\text{max}}$

$W_{\text{max}} \div 2$

$t$
Review: TCP Congestion Control

Rule for adjusting $W$

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

WARNING!!!
If you see this text, it means you need to have the Shockwave Flash files in the same directory as the powerpoint file. The Flash files are available, with this talk, at:
http://www.stanford.edu/~nickm/talks
Some Examples

Example 1:
Make backbone router buffers 99% smaller!

Example 2:
Make access router buffers much smaller too!

Example 3:
Heck, things aren’t so bad with no buffers at all.
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

- Mandated in backbone and edge routers.
- Appears in RFPs and IETF architectural guidelines.
- Already known by inventors of TCP [Van Jacobson, 1988]
- Has major consequences for router design
Backbone router buffers

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

- Where does the rule of thumb come from?
  (Answer: TCP)
Single TCP Flow
Router with large enough buffers for full link utilization

WARNING!!!
If you see this text, it means you need to have the Shockwave Flash files in the same directory as the powerpoint file.
The Flash files are available, with this talk, at:
http://www.stanford.edu/~nickm/talks
Observations

- Sending rate is constant
- If buffer doesn’t go empty when window size halves, then we have 100% throughput.

It follows that

\[ B = 2T \times C \]
Buffer = rule of thumb

Time evolution of a single TCP flow through a router, Buffer is 2T*C

- Congestion Window [Pkts]
- Estimated RTT [ms]
- Bottleneck Link Utilization
- Buffer Occupancy [Pkts]
Over-buffered Link

Time evolution of a single TCP flow through a router, Buffer is 2T°C
Under-buffered Link

Time evolution of a single TCP flow through a router, Buffer is 2T°C

- Congestion Window [Pkts]
- Estimated RTT [ms]
- Bottleneck Link Utilization
- Buffer Occupancy [Pkts]
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.
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

\[ \sum W \]

Buffer Size

Probability Distribution

Gaussian with Mean 7729.1 Packets, StdDev 252.3
Quantitative Model

- Model congestion window as random variable

\[ W_i(t) \]

\[ E[W_i] = \mu_w \quad \text{var}[W_i] = \sigma^2_w \]

- If congestion windows are independent, central limit theorem tells us

\[
\sum_{n} W_i(t) \to \mu_{n=1} + \frac{1}{\sqrt{n}} \sigma_{n=1} N(0,1)
\]

- Thus as \( n \) increases, buffer size should decrease

\[ B \to \frac{B_{n=1}}{\sqrt{n}} \]
Required buffer size

Minimum Required Buffer to Achieve 95% Goodput

\[ 2T \times C / \sqrt{n} \]

Simulation
Short Flows

So far we were assuming a congested router with long flows in congestion avoidance mode.

- What about flows in slow start?
- Do buffer requirements differ?

Answer: Yes, however:

- Required buffer in such cases is independent of line speed and RTT (same for 1Mbit/s or 40 Gbit/s)
- In mixes of flows, long flows drive buffer requirements
  - Short flow result relevant for uncongested routers
Average Queue length

Average queue length for a router serving flows of a fixed length

\[
\frac{\rho^2 E[S^2]}{2(1-\rho)E[S]}
\]

- capacity: \( C = 40 \text{ Mbit/s} \)
- load: \( \rho = 0.8 \)
Queue Distribution

- Large-deviation estimate of queue distribution

\[ P(Q > b) = e^{-bk} \]

\[ \kappa = \frac{2(1-\rho)}{\rho} \frac{E[S]}{E[S^2]} \]

\[ P(Q > b) \]

Buffer B
Packet Loss

\[ b \]
Buffer requirements for short flows

- Independent of line speed and RTT
- Only depends on load and burst size distribution
- Example - for bursts of up to size 16 at load 0.8
  - 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

In mixes of flows, long flows dominate buffer requirements
Experimental Evaluation

- Simulation with ns2
  - Over 10,000 simulations that cover range of settings
    - Simulation time 30s to 5 minutes
    - Bandwidth 10 Mb/s - 1 Gb/s
    - Latency 20ms -250 ms,

- Physical router
  - Cisco GSR with OC3 line card
  - In collaboration with University of Wisconsin

- Operational Networks
  - Stanford University
  - Internet 2
Long Flows - Utilization
Small Buffers are sufficient - OC3 Line, ~100ms RTT

\[ 98.0\% \text{ Utilization} \]
\[ 99.5\% \text{ Utilization} \]
\[ 99.9\% \text{ Utilization} \]

\[ 2 \times \frac{2T \times C}{\sqrt{n}} \]

Number of long-lived flows
## Long Flows – Utilization

Model vs. ns2 vs. Physical Router

GSR 12000, OC3 Line Card

<table>
<thead>
<tr>
<th>TCP Flows</th>
<th>Router Buffer</th>
<th>Link Utilization</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\frac{2T \times C}{\sqrt{n}}$</td>
<td>Pkts</td>
</tr>
<tr>
<td>100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.5 x</td>
<td>32</td>
<td>64</td>
</tr>
<tr>
<td>1 x</td>
<td>64</td>
<td>129</td>
</tr>
<tr>
<td>2 x</td>
<td>258</td>
<td>258</td>
</tr>
<tr>
<td>3 x</td>
<td>387</td>
<td>387</td>
</tr>
<tr>
<td>400</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.5 x</td>
<td>32</td>
<td>32</td>
</tr>
<tr>
<td>1 x</td>
<td>64</td>
<td>64</td>
</tr>
<tr>
<td>2 x</td>
<td>128</td>
<td>128</td>
</tr>
<tr>
<td>3 x</td>
<td>192</td>
<td>192</td>
</tr>
</tbody>
</table>
**Operational Networks**

1. Stanford University Dorm Traffic 20Mb/s

<table>
<thead>
<tr>
<th>TCP Flows</th>
<th>Router Buffer $\frac{2T \times C}{\sqrt{n}}$</th>
<th>Pkts</th>
<th>Link Utilization</th>
</tr>
</thead>
<tbody>
<tr>
<td>333-1800</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.8 x</td>
<td>46</td>
<td></td>
<td>98.0%</td>
</tr>
<tr>
<td>1.2 x</td>
<td>65</td>
<td></td>
<td>100%</td>
</tr>
<tr>
<td>1.5 x</td>
<td>85</td>
<td></td>
<td>100%</td>
</tr>
<tr>
<td>&gt;&gt;2 x</td>
<td>500</td>
<td></td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>97.4%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>97.6%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>98.5%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>99.9%</td>
</tr>
</tbody>
</table>

2. Internet2 10Gb/s link, Indianapolis → Kansas City
   - Cut buffers from 1 second to 5ms (99.5%)
   - Measured loss: < $10^{-7}$
   - Even for 6Gb/s transfers from CERN to SLAC.
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
    - Completion time halved for short-flows

- **40Gb/s linecard with 40,000 x 1Mb/s flows**
  - Rule-of-thumb: Buffer = 10Gbits
  - Becomes: Buffer = 50Mbits
Some Examples

Example 1:
Make backbone router buffers 99% smaller!

Example 2:
Make access router buffers much smaller too!

Example 3:
Heck, things aren’t so bad with no buffers at all.
Access routers have too much buffering

Ever noticed the following?

- Even with the long file transfer, shouldn’t the quick download take as long as it would over a 150kb/s link?
- Why does it always seem to take much longer?
Access routers have too much buffering

Quick download
(14 packets)

Long file transfer

Client

DSL router

300kb/s

0.1s

< 0.5 seconds

Pkts sent

syn 2 4 8

1

2

1


2

8
Access routers have too much buffering.

About 5 seconds! 10x increase.

Quick download (14 packets)

Long file transfer

Buffer Size

\[ W_{\text{max}} = 64\text{kbbytes for Linux} > 1.5 \text{ seconds} \]
Access routers have too much buffering

Observations

1. With less buffering:
   - Downloads would complete faster
   - Long transfers would be unaffected

2. Because the access router is the bottleneck, packets aren’t buffered in the core.
   - We could reduce or remove the core buffers
Some Examples

Example 1:
Make backbone router buffers 99% smaller!

Example 2:
Make access router buffers much smaller too!

Example 3:
Heck, things aren’t so bad with no buffers at all.
TCP with no buffers

WARNING!!!
If you see this text, it means you need to have the Shockwave Flash files in the same directory as the powerpoint file. The Flash files are available, with this talk, at:
http://www.stanford.edu/~nickm/talks

Utilization of bottleneck link = 75%
Summary

- Buffering and queueing delay is everything
- We don’t really understand buffer sizing in the Internet…

- …We need more research.

Many thanks to Guido Appenzeller at Stanford. New Flash expert. For more details, see our Sigcomm 2004 paper available at: http://www.stanford.edu/~nickm/papers
Imagine you want to build an all-optical router for a backbone network.

It’s very hard to build optical buffers...maybe 5-10 packets in delay lines?
LASOR Project

- DARPA funded since 2004
- **Partners:** UCSB, Cisco, JDSU, Calient, Agility, Stanford
- **Program:** Building high capacity all-optical router with 40Gb/s interfaces
- **Problem:**
  - If you can design the routing mechanism, packet scheduling, and congestion control, then how small can we make the buffers?
What if...?

Theory (benign conditions)

\[ \rho \xrightarrow{\text{M/M/1}} X \]

\[ EX = \frac{\rho}{1 - \rho} \]

\[ P[X \geq k] = \rho^k \]

\[ \rho = 50\%, \ EX = 1 \text{ packet} \]
\[ \rho = 75\%, \ EX = 3 \text{ packets} \]

Practice

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

\[ \rho = 50\%, \ P[X > 10] < 10^{-3} \]
\[ \rho = 75\%, \ P[X > 10] < 0.06 \]

What if we randomize launch times and paths, so traffic looks Poisson...?