Bufferbloat

Dark Buffers in the Internet

Jim Gettys

Bell Labs

March 24, 2011

james.gettys@alcatel-lucent.com, jg@freedesktop.org

If I have seen a little further, it is because I am on the shoulder of giants
- Isaac Newton

This is a personal history – but I've provided only a few pieces of the puzzle, and assembled the puzzles. It's not a pretty picture.

Many *more pieces, and much more important pieces*, come to me from many other people, including Dave Clark, Dave Reed, Van Jacobson, Nick Weaver, Vern Paxson, and many others some of whom I do not know... My apologies if I have overlooked your contributions to solving the puzzle.

Alcatel·Lucent

TCP's design assumption is that a congested network will generate *timely* notification of congestion, by packet loss or ECN (later). To do otherwise ***destroys TCP and other congestion avoiding protocol's control loops***

Some *timely* packet loss is not only normal, it is essential to their correct operation!

We judge other Internet protocols by whether they are "*at least as good*" as TCP at avoiding congestion

What happens if TCP's timeliness assumption is ***violated by a lot***?

What happens if packet loss is avoided by buffering?

Alcatel·Lucent

"The Internet is slow today, Daddy!"

You've heard this for years. I've heard this for years. I even tried to debug my network many times. Every time I would go look, the network would stop mis-behaving.

April, 2010 – simple bandwidth/latency test clearly demonstrated a problem: poor latency during continuous data transfers:
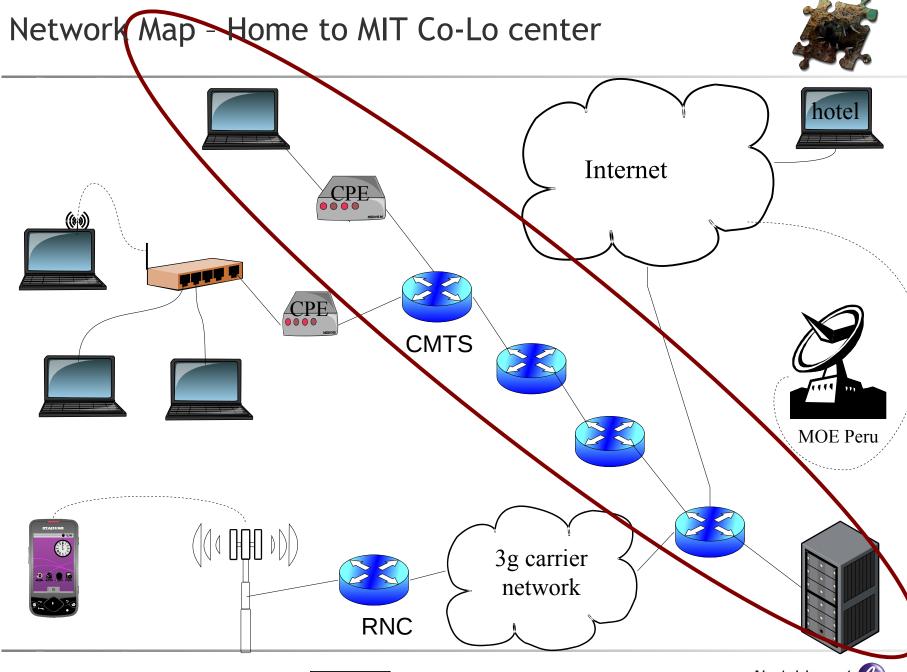*1-2 seconds latency, with very rapidly varying 1-2 seconds jitter*

- In following up, I thought Comcast Powerboost might be the problem

Alcatel·Lucent

hotel

Internet

CPE

CPE

CMTS

MOE Peru

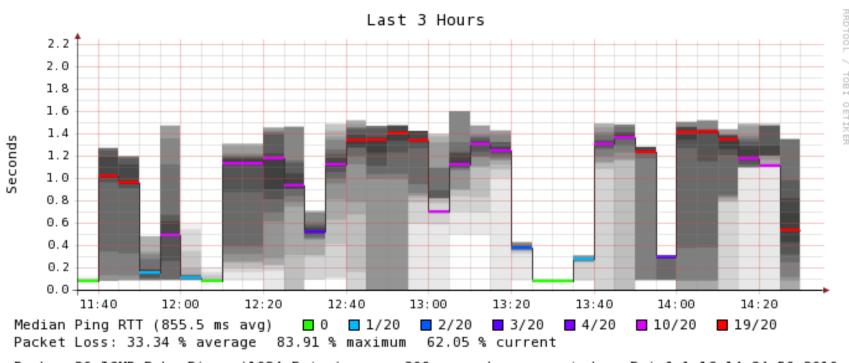3g carrier network

RNC

Alcatel·Lucent

Lunch with Comcast provided many, many, puzzle pieces that fell into place later.

- Suggests the "big buffers" problem, which he been chasing on suggestion of Dave Clark for over two years, but had had no proof of the problem

  - I could drop back to DOCSIS 2, and the differences between DOCSIS 2 and DOCSIS 3 could be used to rule out TurboBoost

- RED is often not enabled in major networks. It is distrusted due to the need to tune RED, which has made many network managers averse to it. Some network managers run RED or other AQM, some do not.

- ECN is blocked in some networks – Dave Oran later explained to me that ECN packets crash many (old) home routers

- Pointer to ICSI netalyzr.

Alcatel·Lucent

Last 3 Hours

Median Ping RTT (855.5 ms avg)   ■ 0   ■ 1/20   ■ 2/20   ■ 3/20   ■ 4/20   ■ 10/20   ■ 19/20
Packet Loss: 33.34 % average   83.91 % maximum   62.05 % current
Probe: 20 ICMP Echo Pings (1024 Bytes) every 300 seconds   created on Fri Jul 16 14:34:50 2010
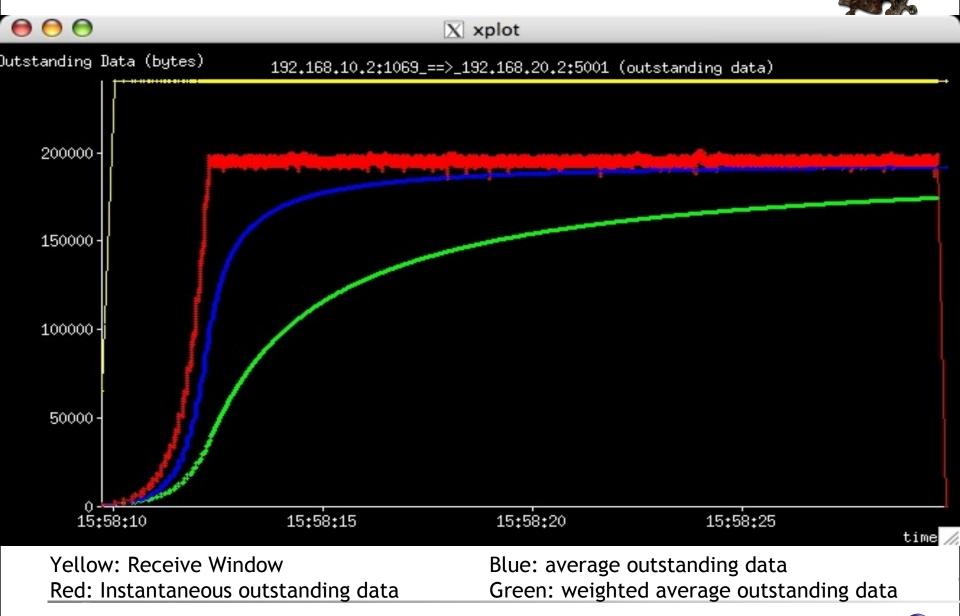
RTT of this path is less than 10ms!

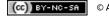Scp of X Consortium archives from my house to expo.x.org.

The periods of "good behavior" are when I suspended the copy to get work done.

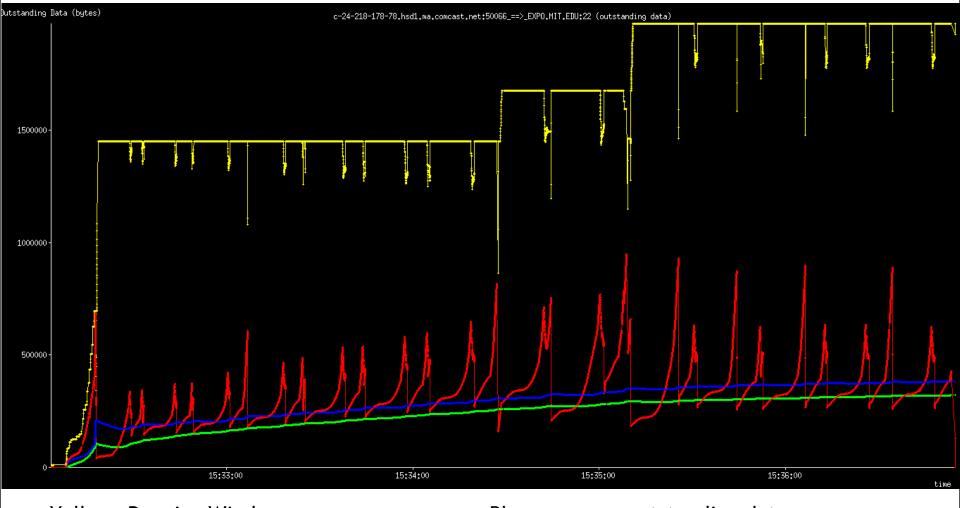Alcatel·Lucent

"Typical" tcptrace plot of a perfect TCP session

Yellow: Receive Window
Red: Instantaneous outstanding data

Blue: average outstanding data
Green: weighted average outstanding data

Alcatel·Lucent

HUH? …. Half a megabyte in flight over a 10ms path???   Spikes???

Yellow: Receive Window

Blue: average outstanding data

Red: Instantaneous outstanding data

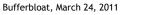Green: weighted average outstanding data

Alcatel·Lucent

Bursts of duplicate acks; bursts of retransmits; lots of SACK's – on long timescales

Cable move at my house due to lightning means that technician tests JG's cable at home end, and Comcast has technicians check my cable at CMTS end.
Nominal cable (once my interior TV wiring was removed, anyway); as good Cable service ever gets outside of a lab

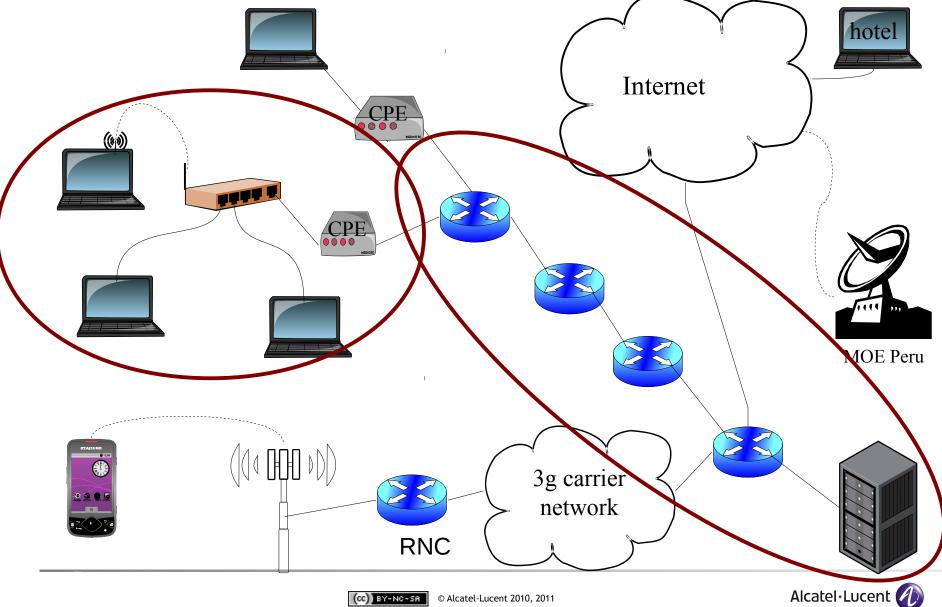Others reproduce the behavior: Partha, at Georgia Tech, Nick Weaver at ICSI
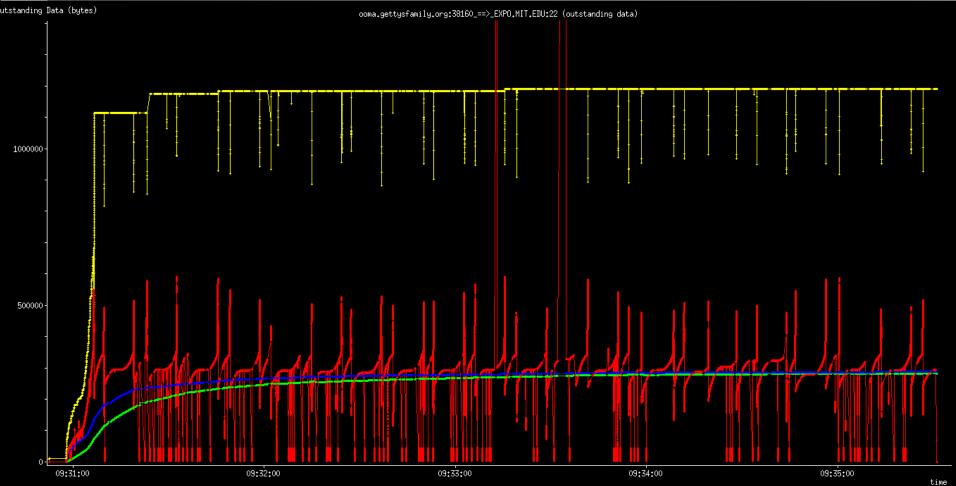
Alcatel·Lucent

hotel

Internet

CPE

CPE

MOE Peru

3g carrier network

RNC

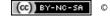Alcatel·Lucent

The Plot Thickens....... My Inlaw's FIOS service.

My inlaws wired FIOS service in Summit, New Jersey, 25/25 service, 7/30/2010
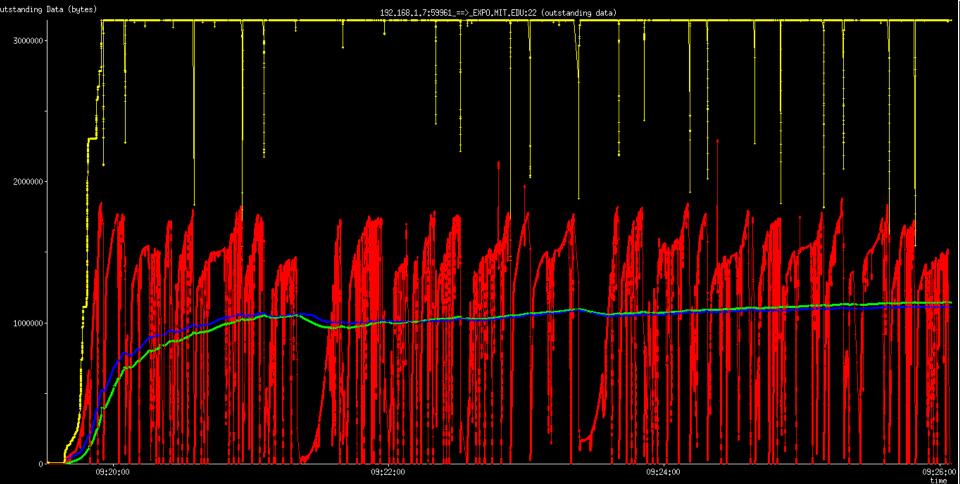
1/4Mbyte outstanding data on a 20ms path? 200Ms latency?

Alcatel·Lucent

And thickens.... Wireless side of the FIOS home router

Wireless side of inlaw's FIOS service, Verizon provided router, > 400ms, 9% loss

Alcatel·Lucent

I called the fire chiefs/consulting detectives for help

Some of you are real TCP experts; I'm not...

- Dave Clark, Dave Reed, Vern Paxon and Dick Sites have all looked this over, and agree with the conclusions

- Van Jacobson says there are timestamps in my data which proves the case for bufferbloat, since both ends were recent enough Linux systems

- What is more, the bloated buffers are *defeating congestion avoidance* since the buffers do not allow **timely** notification

Traffic classification (QOS) can not help you. These are stupid devices. Even if they did classify, that would only move where and when the pain occurs.

I wish I were wrong about Bufferbloat...

Alcatel·Lucent

Uploads and downloads – uploads may often be worse than downloads due to details of hardware and asymmetric provisioning, but some examples include…

- YouTube uploads, Crash dump uploads

- Email with large attachments

- Bittorrent

- File copies/backups

- Downloads of movies to disk

- Video teleconferencing

- Web browsing image heavy sites like YouTube, Google Images, etc.
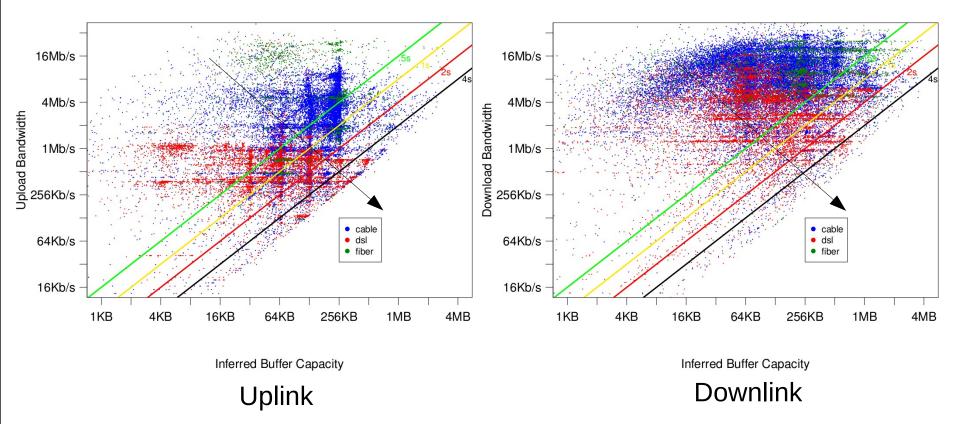
Alcatel·Lucent

Once a network/link exhibits high latency and bad packet loss, other critical, statistically insignificant but mission critical packets can't do their jobs

- DNS – adding 100's ms of latency to lookups kills web browser feel and losses cause lookup failures

- ARP - relies on timely resolution to find other devices on your network

- DHCP - if these packets are lost or excessively delayed, machines can't get on the network

- RA and ND - essential for Ipv6 functioning

- VOIP - needs about a single packet per 10ms flow in order to be good, and less than 30ms jitter.

- Gamers - will get fragged a lot more often with latencies above their twitch factor

- Responsiveness of all network applications, web or otherwise, suffers

Alcatel·Lucent

Uplink

Downlink

Arrow direction is increasing latency

This data is a *lower* bound on the severity of the broadband bufferbloat problem; there was a bug in netalyzr causing it to sometimes fail to fill buffers and therefore detect buffer sizes

This data mixes wireless and wired traffic, so may be contaminated with home router bufferbloat

Alcatel·Lucent

ICSI has already proven the broadband edge is broken!

But...
I think bufferbloat is (almost)
everywhere...

Alcatel·Lucent

hotel

Internet

CPE

CPE

MOE Peru

3g carrier network

RNC

Alcatel·Lucent

During my tests, I thought I caught my home router doing terrible things!

8 second latency over a path of less than 10ms! Much worse than my broadband only test

scp of files from my house to expo.x.org, while running speedtest.net

Recent commercial home router, 50/10 Comcast service, 9/10/2010

Alcatel·Lucent

Eight second latencies? Insanity... What is going on?

- I replicated this problem on several different modern home routers

Given the experience with broadband, and the ICSI data, I had become seriously paranoid

Eventually, I installed OpenWRT on a router, so I could understand better

- I had realized that Linux's transmit queue might be a problem

- I set the txqueuelen knob to zero on the router, but nothing happened...

- More hair lost, and then I realized: the queue is on my *laptop* in the upload direction! And on the *home router* in the download direction.

Any time broadband's bandwidth exceeds wireless goodput at home, the bottleneck becomes the device/home router hop

- this is an increasing problem, that I already suffer frequently from due to chimneys in my house + high broadband speed

Alcatel·Lucent

Since home routers are usually using general purpose operating systems under the covers (e.g. Linux), the problem is on both sides of your wireless link

Buffers hide in multiple places in modern OS's + hardware (*Linux, MacIntosh, Windows alike*) – more about this later

Let's do a simple calculation, presuming 10Mbps:

- 256 packets is of order 3,000,000 bits. So here's 1/3 of a second (one way)

What happens at a busy conference, where your "fair share" might be 100Kbps? 30 seconds: applications (and users) timeout entirely....

And your excessive packet loss rate induced by bufferbloat is increasing failures further
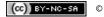
Alcatel·Lucent

You **DO** see bufferbloat trivially on Mac OSX and Linux on 100Mbps ethernet:
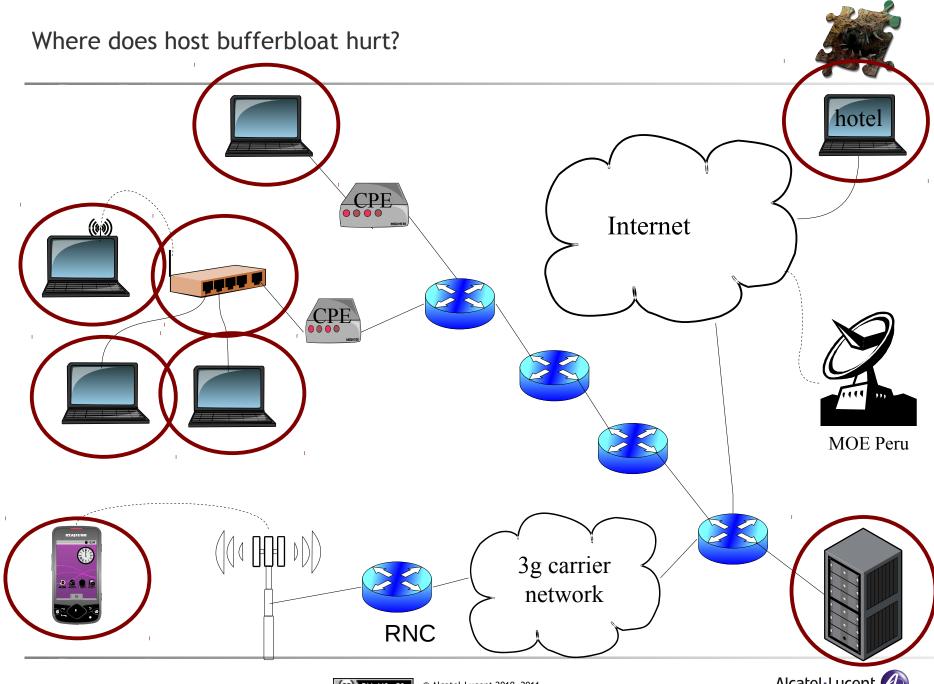
- Use a test program capable of saturating the connection....

- The network is slower than your 1Gbps NIC, so buffers build on your machine

- The driver ring buffers are/appear to be about 256 packets in size on modern hardware: you get ~10ms of latency with simple file copies

Windows is "interesting"... you get good latency, relative to Mac OSX and Linux

- The fastest any version of Windows will run (by default) is about 85Mbps; so the wire is faster than the OS, and therefore the OS buffers never fill

- A Microsoft tech note explains that to get better multimedia experience, Windows bandwidth shapes outgoing traffic

  - I think Microsoft noticed ethernet bufferbloat, but did not fully understand it, and implemented a pragmatic band-aid

Alcatel·Lucent

hotel

Internet

CPE

CPE

MOE Peru

3g carrier network

RNC

Alcatel·Lucent

Aggregate network behavior – Back to the Future

What happens if a network has buffers "all over"? Such as any wired network without AQM enabled in its routers... Classic congestion of the 1990's, yet again

Wireless: first seen in satellite links in the '80s.

- Base station protocol adaptation, to cover error correction in high error environments
- Static buffers used to cover radio bandwidth variation
- Wireless protocols themselves (802.11, 3g)

Buffers start working in an aggregate fashion

Latency will go up when loaded if there are buffers all over; but you won't observe much packet drop - Expect a diurnal (daily) pattern: people timeout before packets do

- This is exactly what Dave Reed reported in 2009 on end-to-end interest in 3G networks
- I've observed up to 6 second latency: Dave Reed has observed up to 30 seconds!
- Known to exist in some (all?) RNC's, and observed in back haul networks
- Back-haul networks are also failing to run RED or other AQM when they should

Alcatel·Lucent

Over ten years ago, Kathie Nichols walked into Van's office one afternoon, and showed him that RED has two bugs

- Van failed to get the paper with RED's problems and fixes published (twice).

Van says RED has several problems; these problems mean that RED requires tuning, and the 100 or more papers about RED tuning in the last decade confirm this. Ergo, network operator's reluctance to enable RED is understandable, even if their fears are (usually, but not completely) excessive: RED must be used carefully!
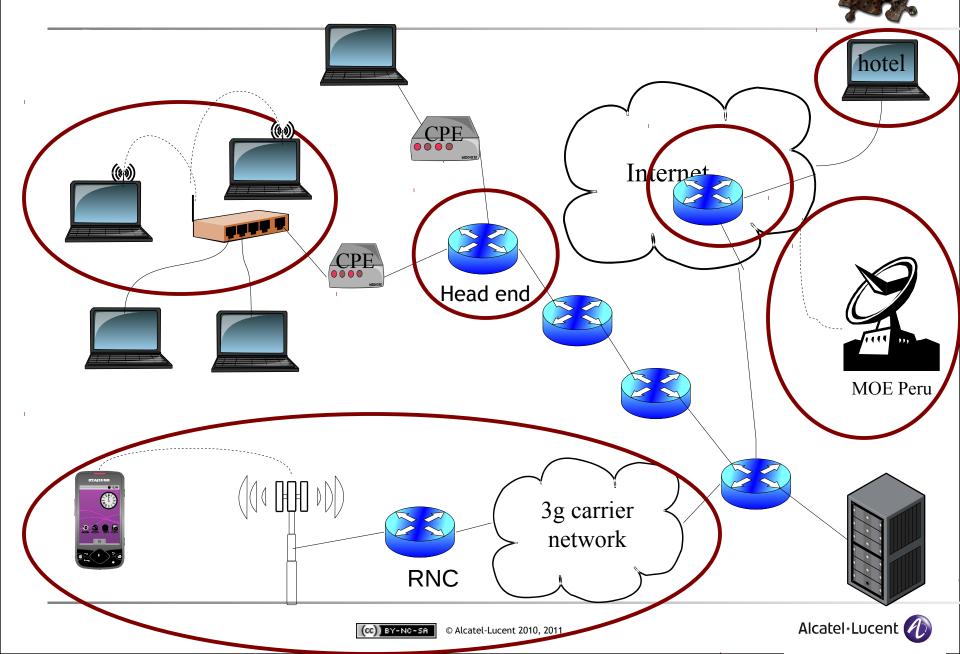
A 1999 draft of *RED in a Different Light* draft did escape; a finished paper should be available soon (I hope). Warning: old draft nRED algorithm there still has a bug

Van warns that time based behavior in Internet gear could also cause congestion and instability. Note that a number of recent broadband technologies clearly have this property. Van wishes one could pace packet transmission but hardware support is lacking - Bunching of initially well paced packets has also been observed already. Synchronization and global resonance does occur!

Alcatel·Lucent

hotel

CPE

Internet

Head end

CPE

MOE Peru

RNC

3g carrier network

Alcatel·Lucent

Concrete example, 25 802.11 nodes, each with a single packet buffer, trying to transmit to an access point.

- Some nodes are far away, and the AP adapts down to, say 2Mbps?

- You have 25 * 1500 bytes of buffering; this is > .15 seconds excluding any overhead, if **everything** goes well.

What happens if:

- You buffer 20 packets instead of 1 on each node? 200? or 1000?

- You keep trying to retransmit packets in the name of "reliability"?
  (some MAC's are known to try to transmit up to 255 times; 8 times in common)

- And, in the name of "reliability", any inherently unreliable multicast/broadcast traffic drops the radio bandwidth to minimum?

- You then try to run WDS or 802.11s, which both forward packets and/or respond to any multicast (e.g. ARP) with routing messages?

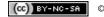Alcatel·Lucent

Conferences using 802.11

Schools

Hotels

Some network operator's networks

Painful personal example/failure: OLPC's network melted under load

- Do you know what happens in Mongolia, where the teachers have the children all open their laptops at the same minute in the morning?

- Net result: no packets would get through

- Well before absolute meltdown, our applications failed due to timeouts

- We had several problems including mesh routing problems; but we missed bufferbloat **entirely** in our failure analysis

Alcatel·Lucent

Buffers are **only** detectable when they are **next** to a saturated bottleneck – at all other times they are "dark"

Your hosts, in your applications, and in socket buffers and network layers

- Your MAC itself may have packet buffers internally;

- Network device drivers themselves

- Your network interface's ring buffer potentially buffers large numbers of packets, often put there to hide x86 SMM (system management mode) behavior and for marketing

- And the VM system your OS may be running on top of may add yet more layers

Your wireless access, at both ends

- 3g has buffers for fragmentation reassembly (how big?); I don't know about LTE

- 802.11 has similar issues: long packet delays destroy timely notification

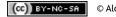Your switch fabric (8 ms/switch at 1GBPS): how many hops, how congested?

Your home router – potentially megabytes

Your CPE/cable modem/FIOS box – potentially megabytes

The head-ends of those connections (e.g. DSLAM, CMTS, etc.)

Each and every router in your path, and the line cards in those routers

Speed of light in glass and vacuum

Alcatel·Lucent

Corporate networks

- Smokeping made me suspect my corporate network wasn't right: latency spikes were present

- ALU corporate network example: sophisticated classification is present, but no AQM enabled; when we convert to Windows 7 and other systems, this becomes an immediate problem, as all hosts will be able to saturate all network links

Satellite Links

- I have observed > 20 second RTT's on links to the MOE in Peru, for example

Example Middlebox
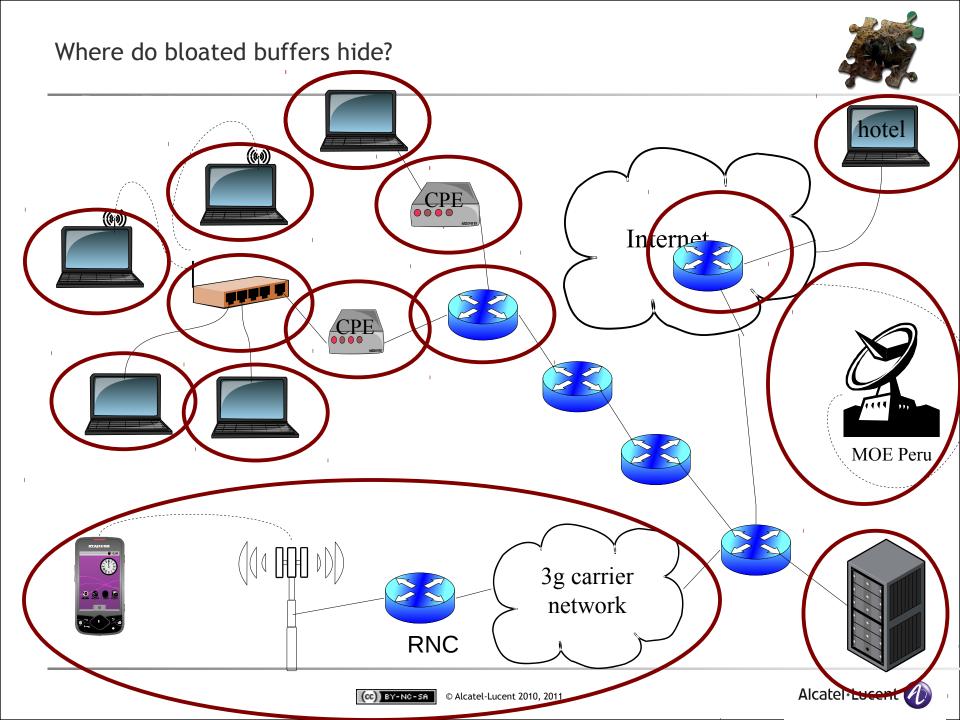
- Tunnel devices

  - Detected in my company's IPSEC infrastructure, at the firewall complex where the tunnels land

- Firewall relays?

  - Haven't looked; but I expect so; if not in the OS, then in the relay applications

Elsewhere?

- Encryption buffers?

Alcatel·Lucent

Where do bloated buffers hide?

CPE

CPE

hotel

Internet

MOE Peru

RNC

3g carrier network

Alcatel·Lucent

HTTP/1.1 spec prohibited more than 2 connections to the same path: chrome opens up to 6 connections to the same path. Initial inflight packets are therefore 12 data packets in current browsers such as chrome: already a problem... Firefox is 15 connections, therefore 30....

Proposal is currently before the IETF to increase TCP's initial congestion window

Microsoft has apparently turned off the initial congestion window entirely; Google has upped its initial window to 8 packets

The potential "impulse" into the network when visiting a web page with many images is therefore 120 packets (or more!)

This is up to 1,500,000 bits that may be in fight, to arrive *SPLAT* at your home bottleneck, which may be anywhere between 1 and 50 megabits in bandwidth. I've observed >150ms latency @ my 20Mbps bottleneck

No classification or fair queuing is deployed at the bottlenecks where these impulses arrive!

Is this a good idea? I think – **NOT!** I'd like to be able to use my Internet connection for applications other than surfing the web, without complaints from my family, strangely...

**Right solution: Replace current HTTP with something better.
Stop papering over the sins of a bad protocol!**

Alcatel·Lucent

Network Neutrality Implications of Bufferbloat

I believe (part of) Bittorrent's problems were incorrectly diagnosed. Consumer ISP's may believe control of applications is an existential issue

- Buffers were already grossly excessive then, and uplinks were much slower

- Windows XP does not enable TCP Window scaling, so it was not obvious what was happening

Carrier's telephony currently has a major quality advantage over SIP VOIP or Skype

- I do not believe it intentional: they get the service calls due to bufferbloat and everyone who builds equipment has been making the same mistake

Innovation is at risk

- *Reliable* low latency applications (games, hosted desktops, immersive teleconferencing) have become impossible to deploy

Fixing bufferbloat is essential to innovation in the Internet

Alcatel·Lucent

Update RFC 2309 – "The RED Manifesto", given the history and knowledge

BCP on router buffering – research shows conventional wisdom that more is better in routers is almost exactly backwards! (e.g. Nick McKeown)

Education – break down layered thinking!

Standardization & testing of AQM algorithms – particularly for the edge & hosts

Advice on applicability of AQM algorithms – needs to be a living document(s) for both for guidance of network operators and those building equipment and systems

Start mitigating bufferbloat while we work on actual solutions that "just work"

How to get ECN actually debugged and turned on in the net?

The HTTP group need to talk to transport to sort out the issues around # of simultaneous TCP connections with the initial congestion window change proposal, in light of bufferbloat and the desire to run **other** protocols simultaneously over our broken Internet of the present

Replace HTTP (I said this over 10 years ago!) What to do about the failure of HTTP pipelining due to lack of sequence numbers in HTTP to enable out of order delivery and deal with buggy implementations?

Alcatel·Lucent

Remember, we are all in this bloat together!

My Blog – http://gettys.wordpress.com

Bufferbloat.net – http://bufferbloat.net
Please come help!

This talk – http://mirrors.bufferbloat.net/Talks/Prague

Full length version of this talk:
http://mirrors.bufferbloat.net/Talks/BellLabs01192011/

Alcatel·Lucent