IETF AQM and Packet Scheduling Working Group Jul 22, 2014

The Case for Comprehensive Queue Management

Dave Taht
bufferbloat.net

Are these Non-AQM/PS WG Problems?

- Layer 2
- Non-AQM but latency saving abstractions
- Software Rate shaping headache
- Ingress Policing
- Products of other working groups (Classification)
- Reproducible experiments, tools and benchmarks

The Layer 2 Dependency Problem

- Ethernet - Byte Queue Limits ”BQL” necessary to mediate between TX-Ring and AQM/FQ technologies
- DOCSIS-PIE: Tightly wound around layer 2 aggregation and packet scheduling
- CEROWRT-SQM: Multiple compensations for ATM and PPP-OE framing required for software rate limiting with HTB.
- WIFI: Packet aggregation and TXOP scheduling do not work well with AQM/FQ strictly layered above. Unification is needed.

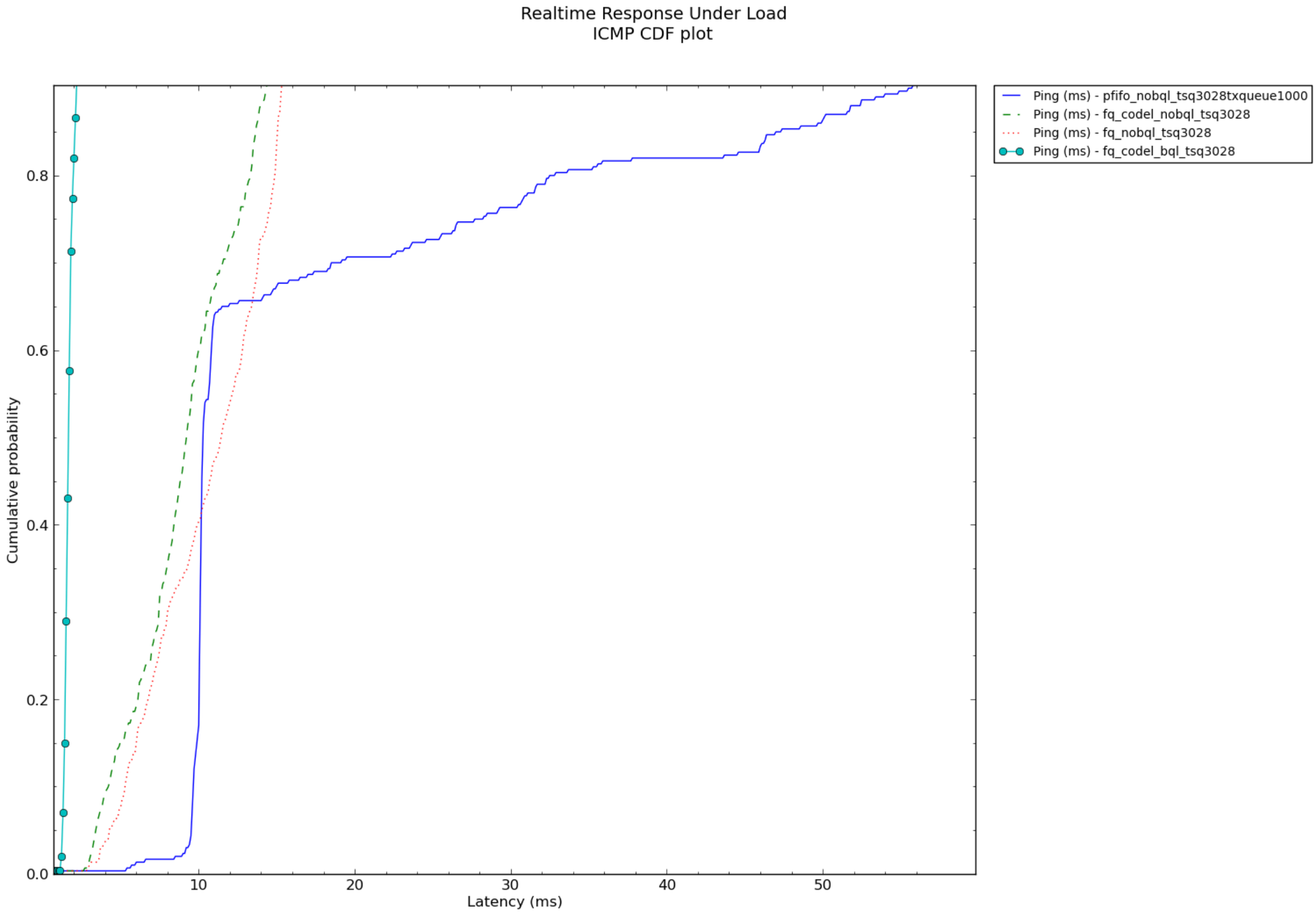
What other network types does AQM and packet scheduling apply to?

- Do we need "AQM over carrier pigeons with QOS"?
  - (updating http://tools.ietf.org/html/rfc2549)
- Do we have to reach out to
- IEEE?
- 3gpp?
- Wifi Alliance?
- ITU?
- UL?
- Elsewhere?

Useful: Byte Queue Limits

- Dynamically controls the hardware ring buffers by keeping enough bytes outstanding to keep the hardware busy, but no more. Typical tx ring: 1024 (up to) 64K packets.
- Typical BQL reductions on the ring: 10Mbit - 1500 bytes, 100Mbit, 3k, GigE - 2 TSO sized packets (with TSO), 20k (without TSO)
- Still is not unified with the overlying AQM/PS layer.
- Not ideal, but makes a radical improvement:

Host latency with a BeagleBone Black without BQL, With BQL, and with various qdiscs at 100Mbit



Ingress Policing

- It seems unlikely head end hardware makers will adopt these technologies anytime fast…
- Resellers of bandwidth often use dumb policers; conventional (byte based policing) doesn't work well
- Using an rate limiter with AQM/Packet Scheduler does work halfway decently on CPE.
- Do we do testing/make requirements to make for better policing?

Rate Limiting

- Used universally by ISPs and Virtual machine providers to sell bands of service.
- Widely used with AQM/Packet Scheduling
- Naively used, can lead to trouble
- Are things like HTB, HFSC, CBQ in scope?

Other WG activity with classification

- RMCAT
- DART
- ?
- Usually 4 tiers of service defined, with a dozen + code points defining drop behavior.
- No implementations that I know of.

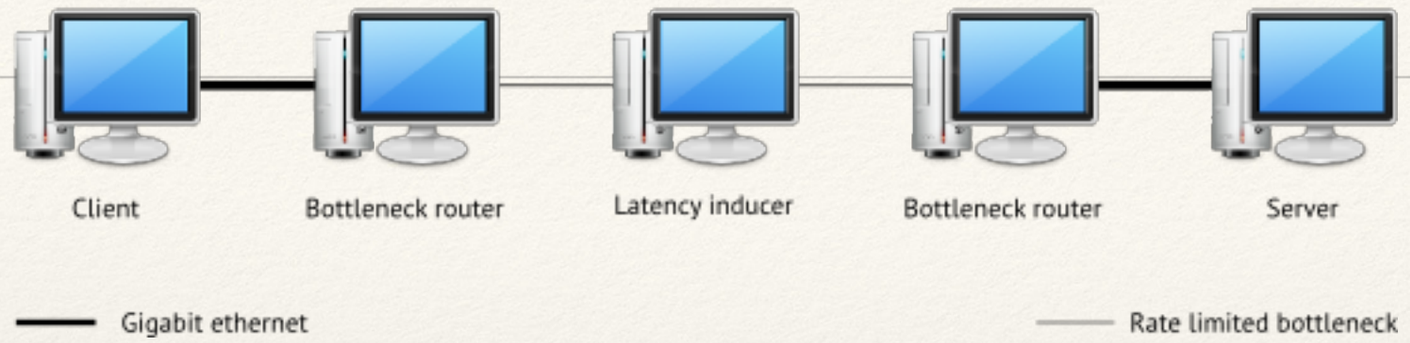
Some updates on my models

- ns-3 models for CoDel, FQ-CoDel, and SFQ-CoDel under development in a Google Summer of Code project for ns-3.
  - Includes asymmetric bandwidth and latency model
  - CoDel may make the ns-3.21 release (August); FQ-CoDel and SFQ-CoDel likely for ns-3.22 release (December)
- ns-2 models for CoDel, SFQ-CoDel, PIE, and DocsisLink developed by Kathie Nichols, CableLabs, and Cisco Systems
  - Available in ns-2 CVS tree, and scheduled for ns-2.36 (August) release
- Public repositories if you want to track the work

Netperf-wrapper update

- Client/server works on linux and OSX.
- Public servers: netperf-{east,west,eu}.bufferbloat.net (good to at least 200Mbit)
- Has support for tcp up/down/bidir/rrul/voip/web tests
- Duplicated several other tests people are using
- 20+ plot types, batch support for more complex repeatable test runs
- https://github.com/tohojo/netperf-wrapper

AQM/PS evaluation Testbed



- Two very large datasets now available:
- http://tohojo-pc.eki.kau.se/deployable-queueing/ (Extensive dataset comparing ared, codel, pie, fq_codel, fq_nocodel, sfq at 10mbit/10mbit, and 10/1)
- http://snapon.lab.bufferbloat.net/~d/residential-tests.tar.gz (subset of the above tests for 8/1, 5/1, 10/1, 22/5, 50/10, 100/10 asymmetric networks, fq_codel and pie byte mode (docsis-pie emulation) only)

The classic Bufferbloat Experiment

- Is: *1 TCP flow up, 1 TCP flow down, and a ping or other isochronous traffic, simultaneously on a network with asymmetric and limited bandwidth, measured against your other variables.*
- Despite documenting extensively how to do this, can't seem to get any experimenters to duplicate it… So…
  - ns3 model for it in progress, netperf-wrapper has multiple combinations of this test.
- Honestly: all you have to do is do one test like this somewhere in your paper or test suite, to make Jim and I happier.

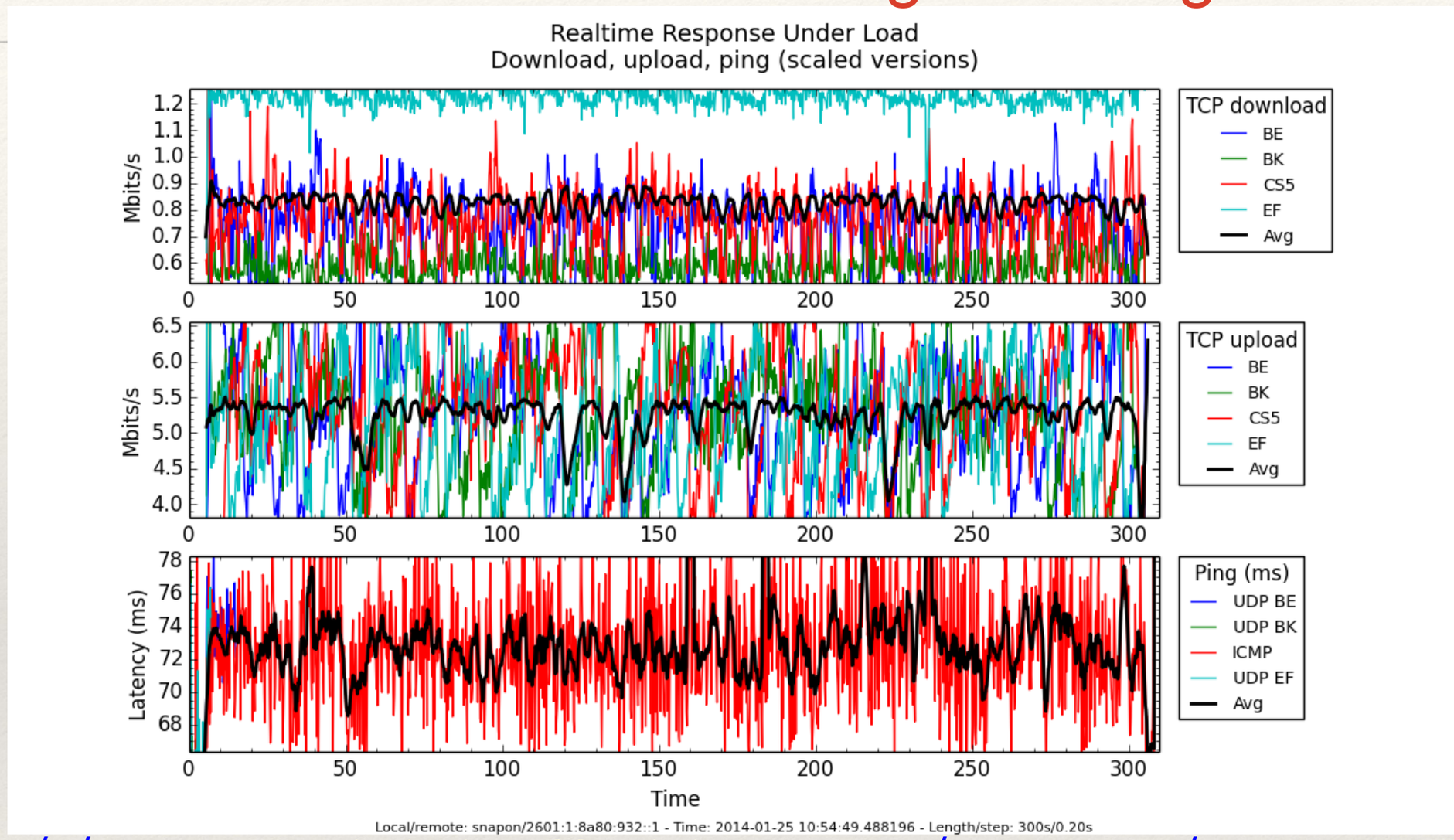
Applying fq_codel instead of “Policing” to Verizon & Comcast etc.

- Opinion: It is unlikely that the CMTSs, DSLAMs and other head ends of the world will evolve towards having aqm or packet scheduling algorithms faster than the CPE can.
- Typical headend buffer sizes are very high
- Can be fixed on the CPE. Should it be?
- Examples at: https://www.bufferbloat.net/projects/codel/wiki/RRUL_Rogues_Gallery

CeroWrt “Smart Queue Management” Designed for extensive experimentation

- Variety of asymmetric rates available from 384kbit to whatever your hardware can support - using packet fifo, byte fifo (DSLAM and CMTS emulations), sfq, sfb, red, ared, sfqred, codel, fq_codel, with inbound and outbound shaping supported also.
- Multiple diffserv based three tier classification systems
- Open Source: works on openwrt, cerowrt, homewrt, and debian derived systems.
- Principal tool I have to explore new technologies

Policing, Classification, Rate Shaping, and wAQM/Packet scheduling “done right”



- http://snapon.lab.bufferbloat.net/~cero2/jimreisert/results.html

Four Questions

- Are packet scheduling with rate limiting techniques (HFSC, HTB, CBQ, DOCSIS-PIE, SQM) within the scope of this Working Group?
- Are we designing something that will only work on ethernet or are we trying to address all layer 2 technologies?
- Are applying various forms of classification to any form of fq and/or aqm within scope?
- Can we come up with something less cumbersome than aqm and packet scheduling as a name for this wg?