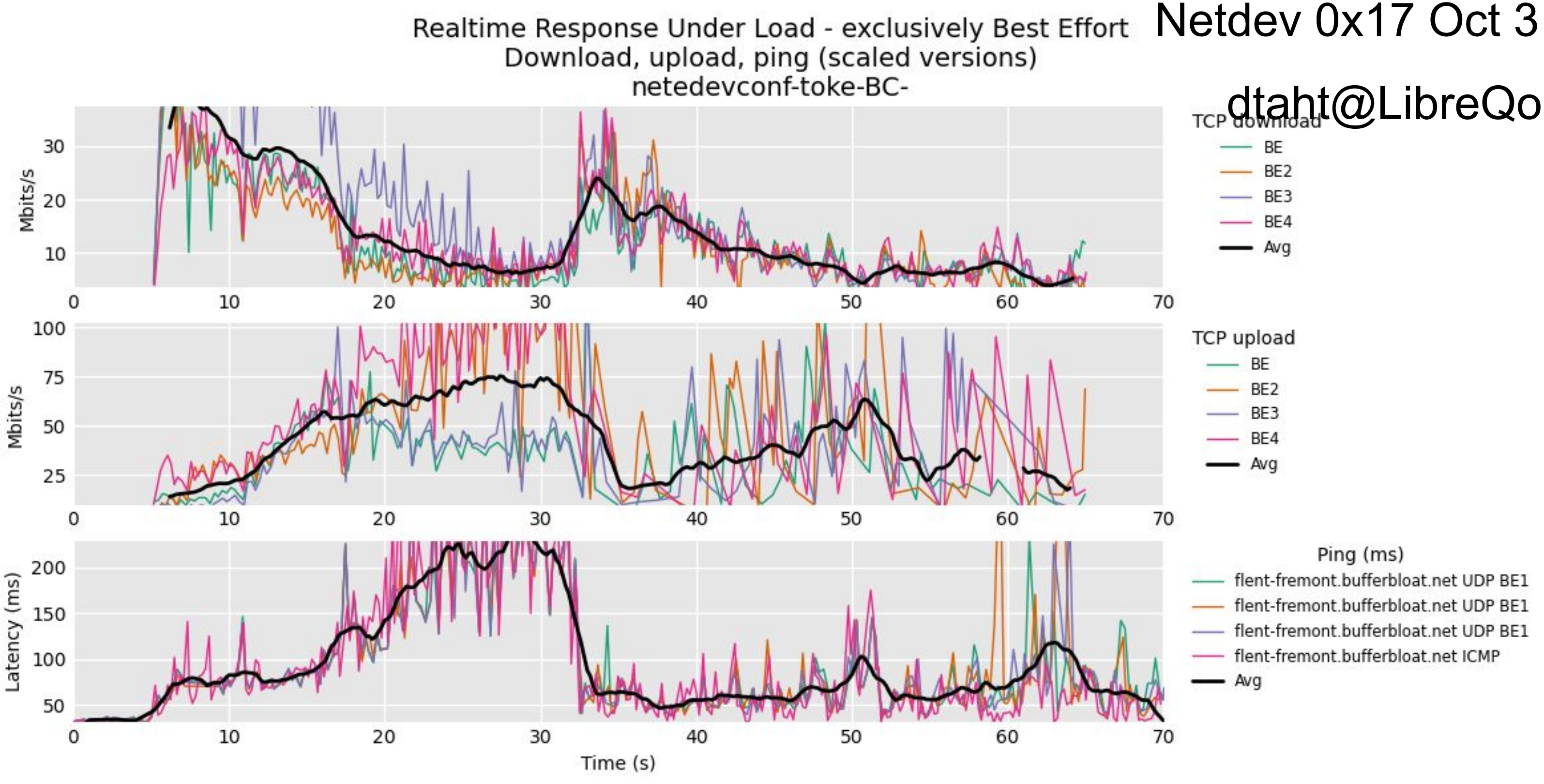
# Low Latency Life Lessons Learned w/CAKE & fq_codel

Dave Taht
Chief Science Officer
Netdev 0x17 Oct 31, 2023

dtaht@LibreQoE.io



Local/remote: penguin/flent-fremont.bufferbloat.net - Time: 2023-10-30T08:59:27.456110 - Length/step: 60s/0.20s

# Who here runs fq_codel?

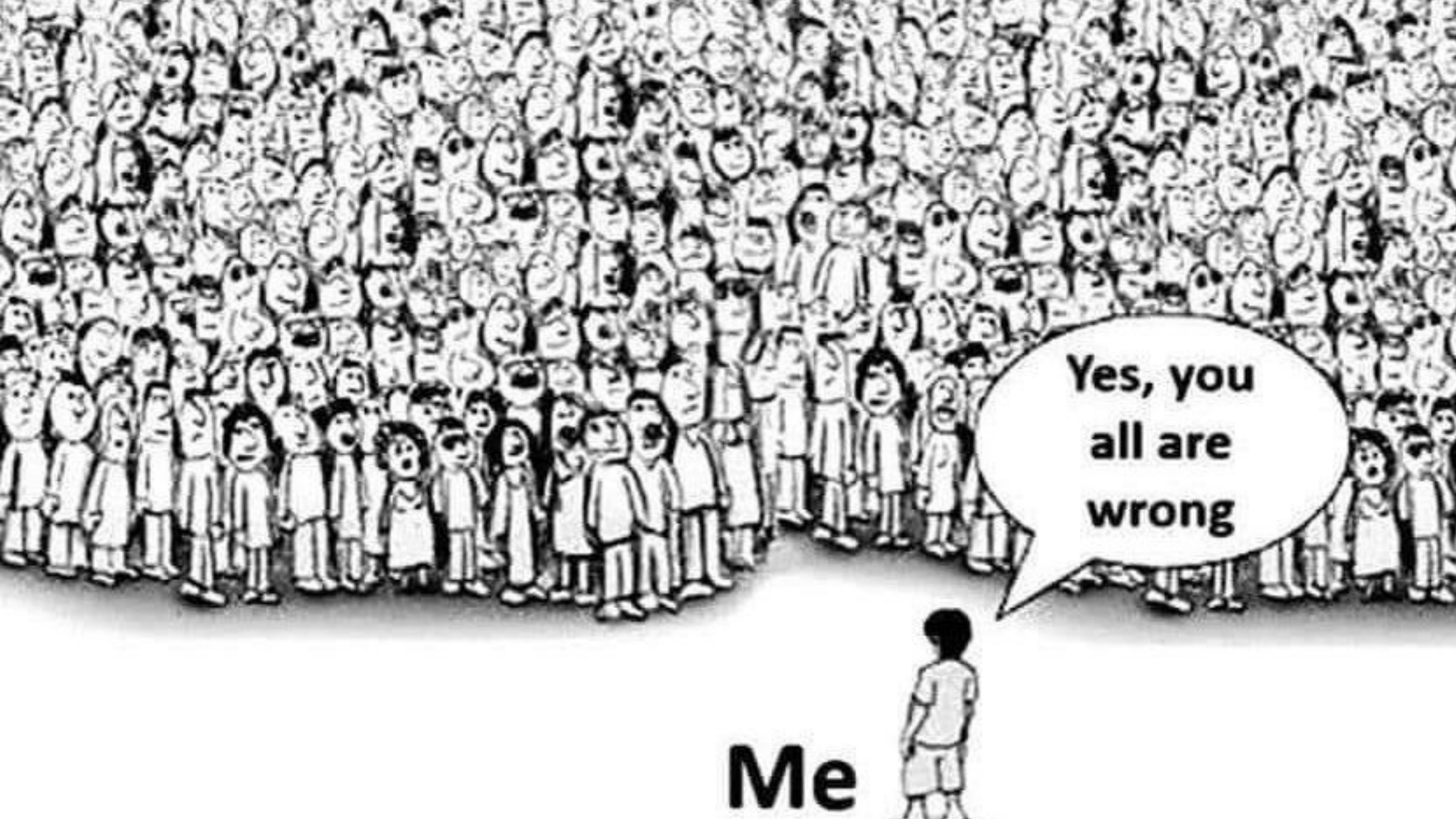
- LINUX and Windows WSL
  - tc -s qdisc show | grep fq
- Apple
  - Netstat -I en0 -qq
- Who knew this? Who has heard of CAKE? How about Flow Queuing? Active queue management? SQM?
- tc qdisc add dev your_device cake bandwidth 40mbit
- Bufferbloat is fixed - but fixes unevenly distributed
- Everybody run: https://www.waveform.com/tools/bufferbloat

# Bufferbloat is Beaten: uploads

- DOCSIS 3.1 modem config
  - Add “AQM 1”. That’s it. RFC8033 (PIE) deployed
  - Only comcast did this!
  - Everyone else sold PLANS
- SQM’s (more than) equivalent: tc qdisc add dev eth0 root cake bandwidth 40mbit ack-filter docsis
  - Outperforms DOCSIS PIE by a lot (IMHO)
  - Comcast, RDK-B is OpenWrt based
  - Starlink, also

# The CAKE MQ CakeMQ project

- Cake upstreamed in 2018- Re started in July 2023. It seemed like making cake and BQL multicore was a path forward!!
- Nope! it is inbound shaping, mirred, and the linux read path is more the problem. EBPF just bypasses a lot of useful stuff like firewalling that you actually need! And then you need to manage congestion!
- NAPI(64) does not make sense to me on a multicore
- BQL was not in half the arm64s we looked at
- The RISCV was especially crappy!
- Still think CakeMQ would help… but...



Yes, you all are wrong
Me

# BQL MIA

- BQL – invented by Tom Herbert – is the basic building block for “backpressure” from the device
  - 6 lines of code needed – “rg netdev_sent_queue”
  - Many Arm64s missed it
  - Many devices have 1 or more queues per core
  - BQL does not scale sideways like this – adding latency per core
    - MIAD
    - No cross CPU locks
    - Core counts are going up
    - BQL CAN BE IMPROVED

# Flow Queueing, defined

- “If the arrival rate of a flow is less than the departure rate of all the other flows (and it is well paced), it goes out first. Otherwise, it is mixed more or less evenly with all the other flows.”
    - Invented by Eric Dumazet
- You don’t need “QoS” for non-behaved applications
- Packet Drops are good!
- RFC3168 ECN can help
- FQ-Codel, FQ-PIE, and CAKE all do Flow Queuing
- So does sch_fq for TCP servers
- It might help userspace VPNs too

# CAKE re-done right

- Make it multi-core -
- Switch to diffserv4 nat by default
  - Yay! Let’s make it slower but do better work
  - Diffserv4 matches wifi and zoom markings better
  - Nat was on by default in OpenWrt
- Submit to -stable maybe?
- Bunch of other potential improvements!

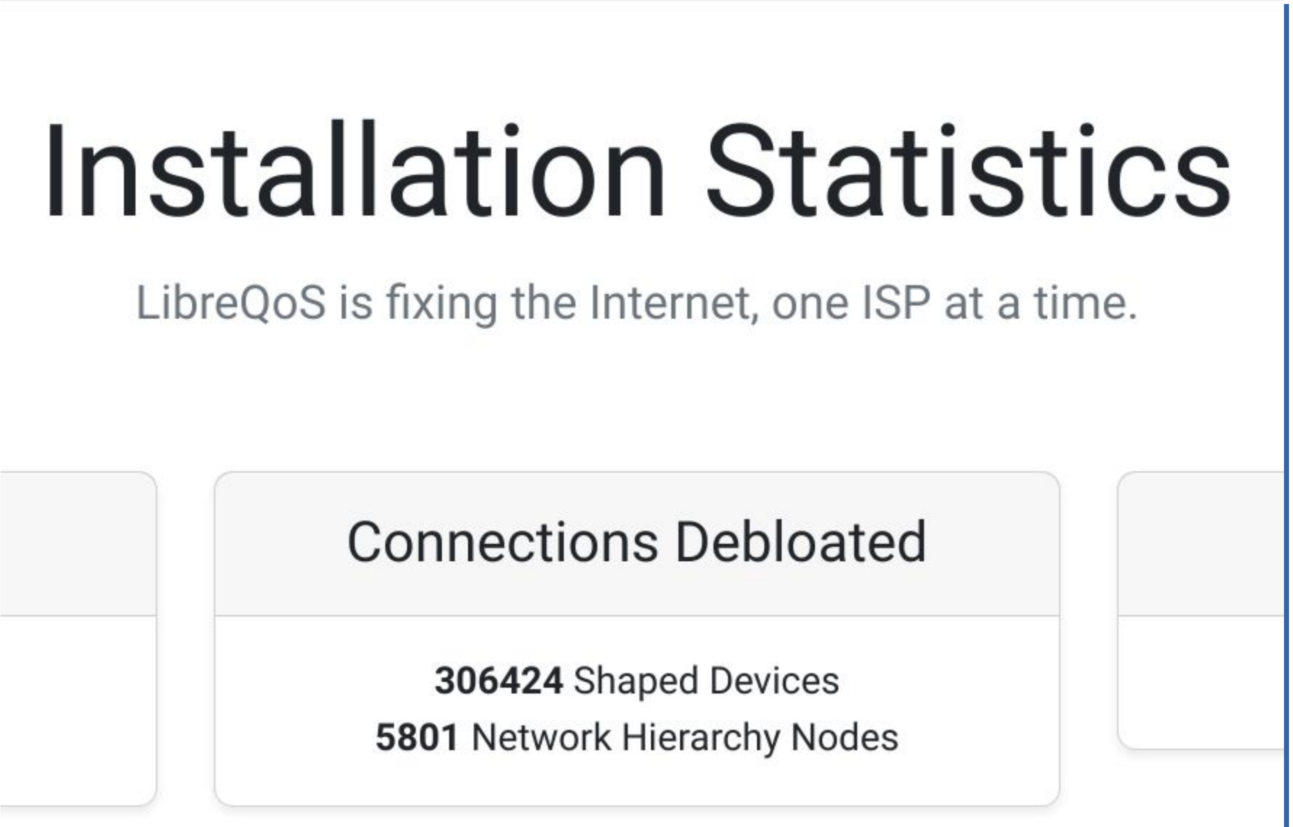
# Real life: Physics vs ISP Plans

- US: Solving wildly variable rate wireless technologies like wifi and 5G
- ISPs: “Selling bandwidth plans”
  - Preseem first to market (2017) with stats and shaping
  - Cambium/Bequant with tcp proxy
  - Paraqum (CAKE)
- Secret Sauce for *all* of them: FQ_Codel – “Now: a commodity”
  - But they failed to leverage ECN
  - And they failed to track the research into CAKE and COBALT
  - Or eBPF. Or dozens of other things we had spent 7 more years on!

# Sad Conclusions

- Middleboxes can deploy rapidly
  - And be rapidly updated
  - EBPF is another example of a workaround
- Embedded gear, can’t
  - It presently takes 10+ years
  - This broken dev/deployment cycle has got to get fixed!
  - It is terrifying to see new products with 2.6.31
- OpenWrt is the sole ray of hope.
- So is federal regulation mandating OTA security on home routers
- The truth about 5G and wifi… is in the packet captures

# LibreQoE Middlebox w/Cake



## Installation Statistics

LibreQoS is fixing the Internet, one ISP at a time.

### Connections Debloated

**306424** Shaped Devices

**5801** Network Hierarchy Nodes

Anyone have an ethernet card with a LPM->CPU_TO_INTERRUPT feature?

Anyone have an ISP with bufferbloat?
1 cheap 20 core Xeon middlebox can handle 10k subscribers at 25gbit without breaking a sweat.

# Real World Traffic

- Apple TV traffic: https://www.youtube.com/watch?v=AXFzJd5BfGQ&t=27s
- Netflix: https://www.youtube.com/watch?v=C-2oSBr2200
- 4k Traffic: https://www.youtube.com/watch?v=KfzHScTwYEw

# Yay! SQM everywhere!

- But: Everyone is doing inbound shaping, which is both CPU intensive, in the wrong place
- Few (I am talking to you eero and firewalla) set the docsis parameter, or nat (very important), or turn on the ack-filter
- We could use a qdisc-defaults

# Buggy fq_codel implementations

- BSD stops at count 400 for no good reason
- Apple’s FQ_Codel has no Codel in it
  - It’s OK!
    - They missed BQL
    - Their stack is slow as hell
    - They have a networkQuality test that measures anything but
- Everybody missed the fixes for codel we put in CAKE
- The latest ubuntu has a race between modprobe and fq_codel

# FQ_Codel for wifi bug

- The wifi code
  - Could use some cobalt (CAKE) features
  - Nobody has published a gang scheduler for OFDMA
- Serious bug – left over for 6 years!
  - The wifi code STILL tunes down Codel too much with more than 4 stations present
  - Testing the fix requires a lab
  - There are 2/3s of the original make-wifi-fast project left to run, like tuning TXOPs to the load
  - Nobody understands wifi could be 5x better than it already is

# GPL vs Cambium and Ubiquiti

- Cambium put out a product – called “elevate” – with a modern OpenWrt, for ubnt’s product line.
- Lawyers descended from all directions!!!
- Net Result
  - Both companies stopped publishing their GPL sources
  - “Elevate” got taken off the market
  - They lost, we lost, their users lost
  - Cambium has been privately demoing cake for 2 years but has not shippe it
    - Because it would compete with their QoE product, when it needs to **have backpressure** and on the CPE most of all.
- +10 – So Og here just did the same thing to Cambium –
- updating one of their products to OpenWrt 23.05… Bwhahahahahahahaha….

# BEAD Boondoggle

- $70B program for better “broadband” shipped to the USA states
- Kids in a candy store
- Lousy metrics – like “Passings”
- Governors thinking they need 6 week programs to train people to cut fiber, not configure internet
- IXPs not even thought of
- Broadband.io founded to educate these folks
- City-bred folk thinking “rural fiber” does not have latency...

# MikroTik and CAKE

- A ray of hope! Mikrotik FINALLY updates their kernel to 5.7 in 2022
- Missed adding in BQL to multiple devices
- Still shipping a FIFO by default. GPL violations a-galore
- Factory wifi drivers
- Market thinks “Shaping” is the answer, not backpressure
  - Me: FQ + RFC7567 on everything
  - They: What’s an RFC?
- We **own* the whole WISP market. We have people abandoning fiber for wisp stuff.**
- **Not a peep from the fiber folk**

# CAKE v ISP Background

- Cake: “A system so simple that even an ISP could configure it”... :Crickets: - LibreQos started
- 2.5 years, 25K lines of C, Rust, and Python later
  - An ISP actually *can* configure it
  - Some really weird topologies in the field
  - I gained empathy for legacy networks
- Please walk in your ISP’s shoes for a while.

# MDT – Micro Delay Transport

- Speeds up slow start by 1-3% for sharded web traffic
- Works well at short RTTs especially
- Does not overshoot (as much) as Hystart++ (RFC)
- 4 years of work - 56 lines of code (currently)
- Pico-Quic based (currently)
- Intensely patentable. I do not want to do that!!
  - Unpatening (dead tree publication) takes time and money too.
  - Testing takes time and money
  - So does IETF standardization
- Plan: Test it with libreqos in my copious spare time

# BufferBloat Resources

- Mailing lists: https://lists.bufferbloat.net
- My grumpy blog: https://blog.cerowrt.org
- Thanks NLNET & Equinix for this round!
- Memorize this song: https://genius.com/Steve-savitzky-the-mushroom-song-lyrics