L4S: Ultra-Low Queuing Delay for All

Bob Briscoe, Olga Bondarenko
[ simula . research laboratory ]
Koen De Schepper, Inton Tsang

NOKIA Bell Labs

Jul 2017

The authors' contributions are part-funded by the European Community under its Seventh Framework Programme through the Reducing Internet Transport Latency (RITE) project (ICT-317700). The views expressed here are solely those of the authors.
application profile is evolving

- increasingly nearly *all* apps require low delay (and often high bit rate too)

  - interactive web, web services
  - voice,
  - conversational video, interactive video, interactive remote presence
  - instant messaging
  - online gaming
  - virtual reality, augmented reality
  - remote desktop, cloud-based apps
  - video assisted remote control of machinery & industrial processes
Main contributions to delay

- Delay: multifaceted problem [Briscoe14]
  1) Caches have cut base (speed-of-light) delay, where they can
  2) Remaining major component of delay: queuing
     - intermittent – solely under load
     - at best, doubles the base delay
Problem characterization

- bottlenecks typically at access edge – per 'customer'
- low statistical multiplexing (1 or a few apps at once)
- during peak, typically 100% utilization 1-3% of the time

- The new norm: all of a customer's apps at any one time need low delay
Solution: L4S  
Low Latency, Low Loss, Scalable throughput

- Demo schematic
  - multiple demanding applications over the same broadband line (40Mb/s downstream)
  - plus ~4 large file downloads
  - 7 ms base (speed-of-light) round trip time from Data Centre to the Home and back
- mean per-packet L4S queuing delay ~500μs (½ ms)

See video of Panoramic Interactive Video + TCP downloads:  
https://riteproject.eu/dctth/#1511dispatchwg
Simplified dashboard

L4S

Flow rates (window)

Classic

Link Utilization

Distribution of queuing delay
Other solutions - in context

• Priority classes (Differentiated Services)?
  • favouring certain traffic requires policing and management
  • non-solution when all of a customer's apps at one time need low latency

• Active Queue Management
  • a solution 'for all' – promising direction
  • but TCP is (literally) the elephant in the room – min queue 5-20ms

• Per-flow queuing?
  • isolates each flow from the delay of others, but overkill...
    1. removes control from variable bit rate apps (network schedules the queues)
    2. individual app flows not always visible to network (e.g. IPsec VPN)
    3. computationally expensive
    4. anyway, doesn't protect a flow from the delay it inflicts on itself

• BBR (Google research)
  • Attempt to reduce queuing delay without changing network
  • Queuing delay intermittent - similar to AQM
  • Problems interacting with AQM: toggles between starving others or itself

• 'Classic' (standard) ECN
  • A congestion 'mark' is equivalent to a packet drop
  • Removes round trip delay delay to repair congestion loss, but not queuing delay
Business Implications of “Low Delay for All”

- can sell L4S per customer rather than per app
  - e.g. small businesses, premium users, enterprises (where migrating from VPN to Internet)
- eventually (or from the start) to all users
The Solution

Remove the Root-Cause of Queuing Delay
The Cause of Queuing Delay

• TCP's capacity-seeking behaviour: 'saw-toothing'

Note on terminology:
- Capacity seeking is called 'congestion control'
- the outcome is called 'congestion' even when just one flow
Active Queue Management (AQM) dilemma: delay vs. utilization

Today (typical)
TCP on end-systems
Drop-tail buffers

Today (at best)
TCP on end-systems
AQM at bottlenecks

delay-utilization dilemma

TCP saw-teeth seeking capacity

TCP saw-teeth seeking the operating point

shallower operating point

cuts delay, but poorer line utilization

buffer occupancy

time

line utilisation

buffer size

AQM operating point

delay amplitude:
~1 'typical' base RTT (round trip time)
Resolving the dilemma:
Finer saw-teeth of a 'Scalable' TCP

Today (at best)
TCP on end-systems
AQM at bottlenecks
if change bottlenecks alone
A Scalable TCP change bottlenecks and TCP

buffer kept for bursts

buffer occupancy

TCP saw-teeth seeking the operating point

AQM operating point

shallower operating point

lower queuing delay and more predictably low

smooth TCP: more smaller saw-teeth

cuts delay but poorer line utilisation

good line utilisation

utilization insensitive to config'n of operating point
Finer Sawteeth → more frequent sawteeth

• Packet drop: no longer feasible as a congestion signal – too much
  • use Explicit Congestion Notification (ECN)
  • standard part of the Internet Protocol (v4 & v6) since 2001

ECN: an opportunity to remove much more delay...

Drop
• state-of-the-art AQMs defer dropping for ~100ms (worst-case RTT)
  • in case the burst clears itself
  • no response for 5 CDN RTTs, or 25 media server RTTs

ECN
• AQM can signal ECN immediately – no risk of impairment
• the sender can smooth out ECN signals (over its own RTT)
  - can react without smoothing if appropriate

RTT: Round Trip Time
Finer Sawteeth → more frequent sawteeth

• Only feasible:

1) with modified TCP/ECN feedback
   • standard TCP source only responds to one signal /RTT
   • so when ECN was added to TCP, only one feedback /RTT
   • IETF now standardising accurate TCP ECN feedback AccECN
     [RFC7560, Briscoe17]

2) if coexistence with existing traffic is solved
   • a Scalable flow with fine saw-teeth
     looks like high congestion to a 'Classic' TCP flow
   • so Classic TCP starves itself relative to a Scalable TCP
Coexistence: Solution

- per customer site (home, office, campus or mobile device)

- DualQ Coupled AQM: a 'semi-permeable membrane' that:
  - isolates latency (separate queues for L4S & Classic)
  - but pools bandwidth (shared by apps/transport, not by network)

Details: [l4s-arch], [dualq-aqm]
Premium Service vs. Default?

<table>
<thead>
<tr>
<th>Codepoint</th>
<th>ECN bits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not-ECT</td>
<td>00</td>
</tr>
<tr>
<td>ECT(0)</td>
<td>10</td>
</tr>
<tr>
<td>ECT(1)</td>
<td>01</td>
</tr>
<tr>
<td>CE</td>
<td>11</td>
</tr>
</tbody>
</table>

• Classifier on 2-bit ECN field in IP header (v4 or v6)
  - if ECT(1) or CE, forward to L4S
  - adopted for standardisation by IETF

• Classifier on any other field
  - source IP address
  - dest. IP address
  - VLAN ID, ...

• ECN field works end-to-end
  - network could solely enable L4S for certain addresses
  - later, could enable for all addresses

• in all cases, no packet inspection deeper than IP
  - compatibility with all privacy technology

Details: [l4s-id]
L4S Deployment

• Coexistence of Classic and L4S traffic
  • DualQ AQM

• Scalable TCP deployment?
  • DCTCP (Data Centre TCP) is a Scalable TCP
  • already in Linux & Windows
  • works over wide area round trip times
  • good enough for controlled trials of L4S

• for production
  • DCTCP needs safety / performance enhancements
  • 'the TCP Prague requirements' [l4s-id]
# L4S Deployment Sequences

Significant benefit realized at each deployment stage

<table>
<thead>
<tr>
<th>servers or proxies</th>
<th>access link</th>
<th>client</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. DCTCP (existing)</td>
<td>DualQ AQM downstream</td>
<td>DCTCP (existing)</td>
</tr>
<tr>
<td></td>
<td>works downstream for controlled trials</td>
<td></td>
</tr>
<tr>
<td>2. TCP Prague</td>
<td>AccECN (already in progress – DCTCP/BBR)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>works downstream</td>
<td></td>
</tr>
<tr>
<td>3.</td>
<td>DualQ AQM upstream</td>
<td>TCP Prague</td>
</tr>
<tr>
<td></td>
<td>works upstream &amp; downstream</td>
<td></td>
</tr>
</tbody>
</table>

Where a stage involves 2 moves:
- The benefit after the 2\(^{nd}\) move has to be worth the 1\(^{st}\) mover's investment risk
- new services or products, not just incremental performance improvement
Maturity status

- IETF: L4S adopted for standardization (experimental status)
- Numerous companies (often research) involved
  - equipment vendors
  - operators
  - OS developers
  - hardware developers
- Some working on related scenarios
  - e.g. coexistence of TCP and DCTCP in data centres
further research / open issues

• TCP Prague
  • safety & performance enhancements to DCTCP

• L4S over radio (cellular, WiFi)
  • initial positive results
  • potential to solve TCP's glacial response to radio dynamics

• redesign of rate limiters/policers (e.g. in PoN)
  • currently just use loss – need an ECN-based warning stage

• radio transmission losses
  • watch this space :)
Further engineering

• Standardization: to be completed
• Network traversal of ECN
  • recent measurements over mobile shows bugs
  • being cleared as (Classic) ECN becomes used
• Implementation of DualQ AQM
  • production software in progress: Linux, NFV (Intel DPDK)
  • compatibility with each vendor's hardware TBA
more info

https://riteproject.eu/dctth/


Conclusions

- Enables previously infeasible interactive apps
- Need low delay for all of customer's applications at once
  - Differentiated service: an incomplete solution
- Technical problem: 'Classic' TCP
- Technical solution:
  - "Scalable" TCP with L4S variant of ECN
  - Incremental deployment path
- Business solution:
  - Premium user
  - L4S can become the default Internet service for all users and apps
- Open issues seem to be only around peripheral problems
  - basic solution is ready for take-up
large saw teeth can ruin the quality of your experience