Integration and evaluation of QUIC and TCP-BBR in long-haul data transfers

Raul H. C. Lopes[1]
Tim Chown[2]
Duncan Rand[3]

[1] Jisc and Brunel University London, UK
[2] Jisc, UK
[3] Jisc and Imperial College, UK

July 9, 2018

Table of Contents

1. Motivation
2. TCP: BBR and Cubic
3. QUIC
4. Then... 100G
5. Acknowledgments

Initial disclaimer

- **First:** I am not a network man.
- PhD in Computational logic and parallel algorithms.
- Example: my previous submiissions to CHEP all related to Parallel Computing.
- Sysadmin at LT2 Brunel London.
- Half-time rented to Jisc (UK academic network) to help research institutions to improve performance in data transfers.

  Meaning... Transfer WLCG sysadmin experience?
- First task: Singaren...

First case

- A testbed
  - Singaren DTN: *filesender.singaren.net.sg:2811*
  - An old disk server at Brunel: *dc2-grid-e6-000.brunel.ac.uk*
    - Dual stack
    - RAM: 8 GB
    - Network: dual 10 gbps
    - CentOS 7 on kernel 4.16
  - 20 gbps link from Brunel to Jisc.
  - 10 gbps link from Singapore to Eurasia link.
  - A reference:
    DTN at CERN: *ftp://cern-dtn.es.net:2811/data1*
- Tools
  - The usual suspects: iperf3, traceroute, ping
  - globus-url-copy to serve disk-to-disk transfers
- Target: transfer files at 10 gbps

Framing the target

- How can we improve on

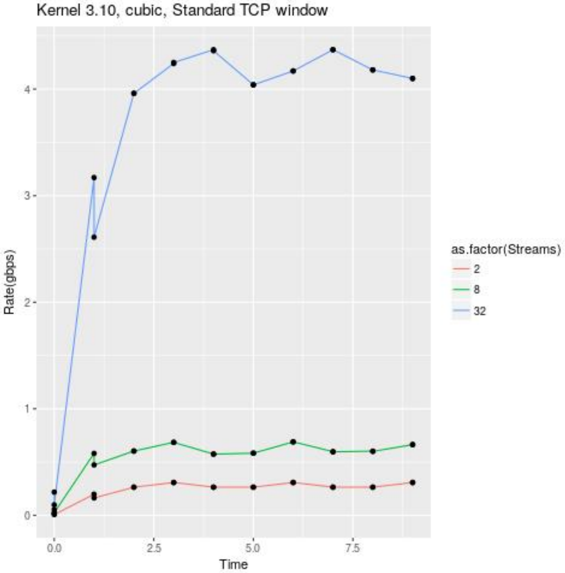
```
        [root@dc2-grid-e6-000 ~]#  globus-url-copy    -p 4 -vb \
        ftp://gridftp-user@filesender.singaren.net.sg:2811/data1/1G.file /dev/null
Source: ftp://gridftp-user@filesender.singaren.net.sg:2811/data1/
Dest:   file:///dev/
  1G.file  ->  null

   1071382528 bytes        42.57 MB/sec avg        78.00 MB/sec inst
```

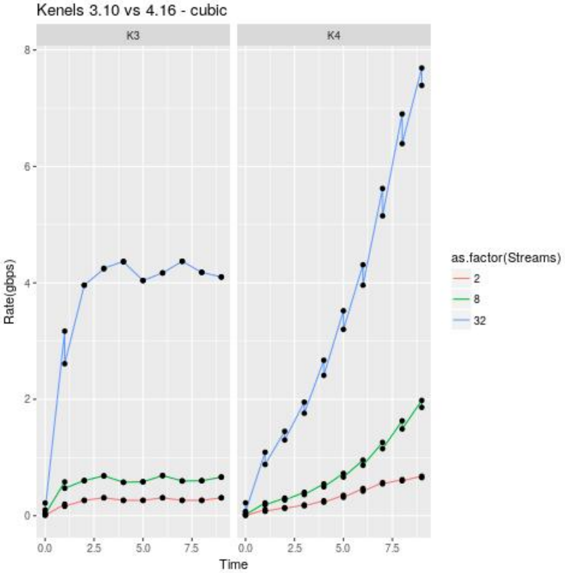
- What is that saying?
  - 0.6 gbps on a 10 gbps link;
  - We are testing 100-200 gbps links at Jisc, 100 gbps to Australia.
  - Optical transmission speeds are approaching Terabit capacity;
  - yet, peak TCP session speeds are not keeping up.

Table of Contents

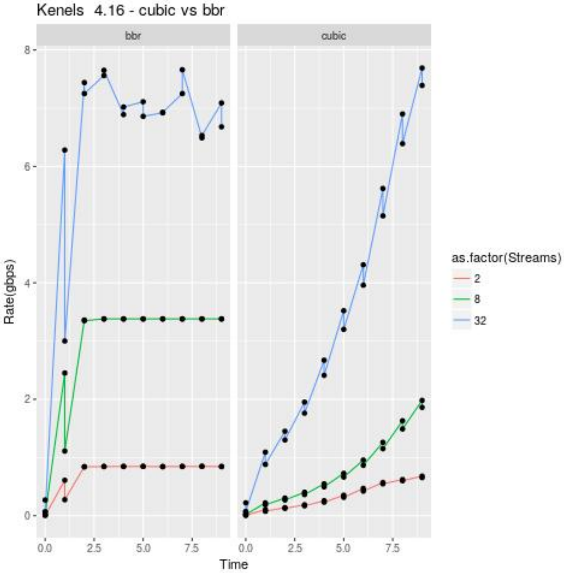
Status after link repair



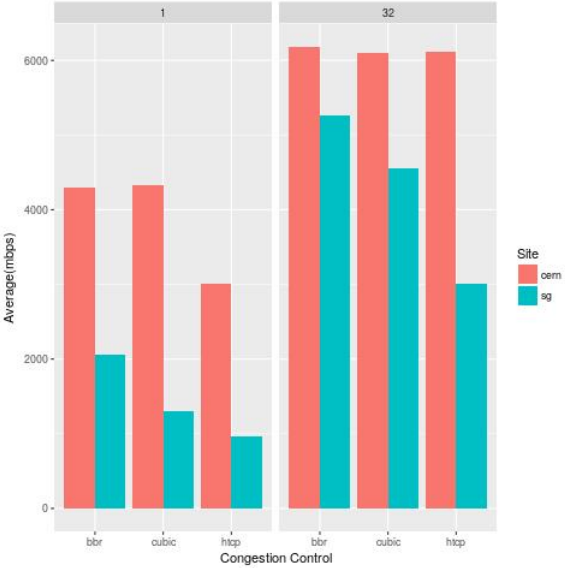
Move to kernel 4.16



BBR and Cubic in Kernel 4.16



BBR, Cubic and htcp



TCP protocol

- It's an ACK-pacing protocol: it reacts to past network states
  - An ACK signal shows the rate of data that left the network at the receiver that occurred at $\frac{1}{2}$ RTT back in time
  - if there is data loss, the ACK signal of that loss is already $\frac{1}{2}$ RTT old
  - If there is no data loss, that is also old news TCP should react conservatively to good news
- TCP control
  - Use ACK for $\frac{1}{2}$ RTT old event to increase/decrease the sending rate;
  - Minimise packet re-ordering;
  - Minimise packet loss.

Table of Contents

1. Motivation
2. TCP: BBR and Cubic
3. QUIC
4. Then... 100G
5. Acknowledgments

QUIC goodies

- QUIC is defined on top of UDP with user space implementation.
- Data sent directly on connection establishment.
  - 0-RTT connection establishment;
  - great in long haul transfer and in presence of packet loss.
- Flexibile to mix with a variety of congestion control approaches: Cubic available, BBR in use at Google(?).
- Encrypted connections by default.
  - It will make some UK biomed community happy.
  - It makes it harder for network middle boxes tampering with traffic. We end network people could be happy.
- Multi-streaming protocol

QUIC in test

- QUIC is used in Google services (claims of 7%) of traffic.
- IETF workgroup points to a dozen implementations.
- Most won't compile.
- Prototype built on top of: quicr, and ngtcp2.
- Debug stage
- Local 10 Gbps testbed
- Rates below 4 Gbps

Table of Contents

1. Motivation
2. TCP: BBR and Cubic
3. QUIC
4. Then... 100G
5. Acknowledgments

100G network

- Brunel HPC clusters
  - Omnipath and Infiniband
  - Kernels 3.10 and 4.16
  - TCP limits below 90 gbps
- Geant
  - Richard-Hughes Jones (Geant and SKA) testing 100 gbps link Europe to Australia.
  - Linux kernel 3.10
  - Lmited TCP rates below 90 gbps.
  - Using Cubic with varied TCP windows, but up to 40 MB
- Jisc testbed
  - 100 gbps, Melanox based
  - Hierarchy of NVMes, SSDs, HDDs targeting 100 gbps between persistent storage.
  - Based on Netflix architwecture: only 100 Gbps success?
  - First step: local tests only.
  - To be available for testing by research institutions.

On going work

- Brunel storage slowly moving Kernel 4.16
  - BBR to evaluated
  - Network performance counters under investigation.
- Half of Brunel compute nodes already on kernel 4.16 and using BBR.
- Brunel/Jisc QUIC tool (proof of concept) should be available soon (4 weeks?)
- Jisc new 100 Gbps testbed to be available in August

Table of Contents

1. Motivation
2. TCP: BBR and Cubic
3. QUIC
4. Then... 100G
5. Acknowledgments

On going work

Lopes, and Rand are members of the GridPP collaboration and wish to acknowledge funding from the Science and Technology Facilities Council, UK.