Jisc BBRv3: initial tests

Tim Chown, Raul Lopes, Duncan Rand, Chris Walker (Jisc)

NETWG Packet Pacing call, 3 Oct 2023

DC24 Context

Networking proposals

- The DC24 proposals folder has a description of each activity:
  - https://cernbox.cern.ch/files/link/public/aCITXJenZxpF5qw?tiles-size=1&items-per-page=100&view-mode=resource-table
- Includes BBRv3 testing, packet pacing, use of jumbo frames
- BBRv3 test spec includes:
  - Testing BBRv2/BBRv3/other CCAs (TBD) with WLCG applications
  - Duration of file transfers; explore performance gain
  - Fairness with other traffic
  - Behaviour on congested links
- October: define scope and tests, Nov-Jan: run the tests

BBRv3 and other TCP testing

Evaluating new TCP variants – focus on BBRv3

- BBRv3 was announced at the July IETF meeting
  - Essentially BBRv2 with some bug fixes
  - Given a new version number as impact may be significant
  - See https://datatracker.ietf.org/meeting/117/materials/slides-117-ccwg-bbrv3-algorithm-bug-fixes-and-public-internet-deployment-00 (an update from Google)
  - Available on github:
    - https://github.com/google/bbr/blob/v3/README.md
- Note that BBRv2/v3 can make use of Explicit Congestion Notification (ECN), so support for that may also be of interest
  - Q: Will we consider ECN in scope?

Jisc

Jisc BBRv3 deployment: perfSONAR

Using perfSONAR

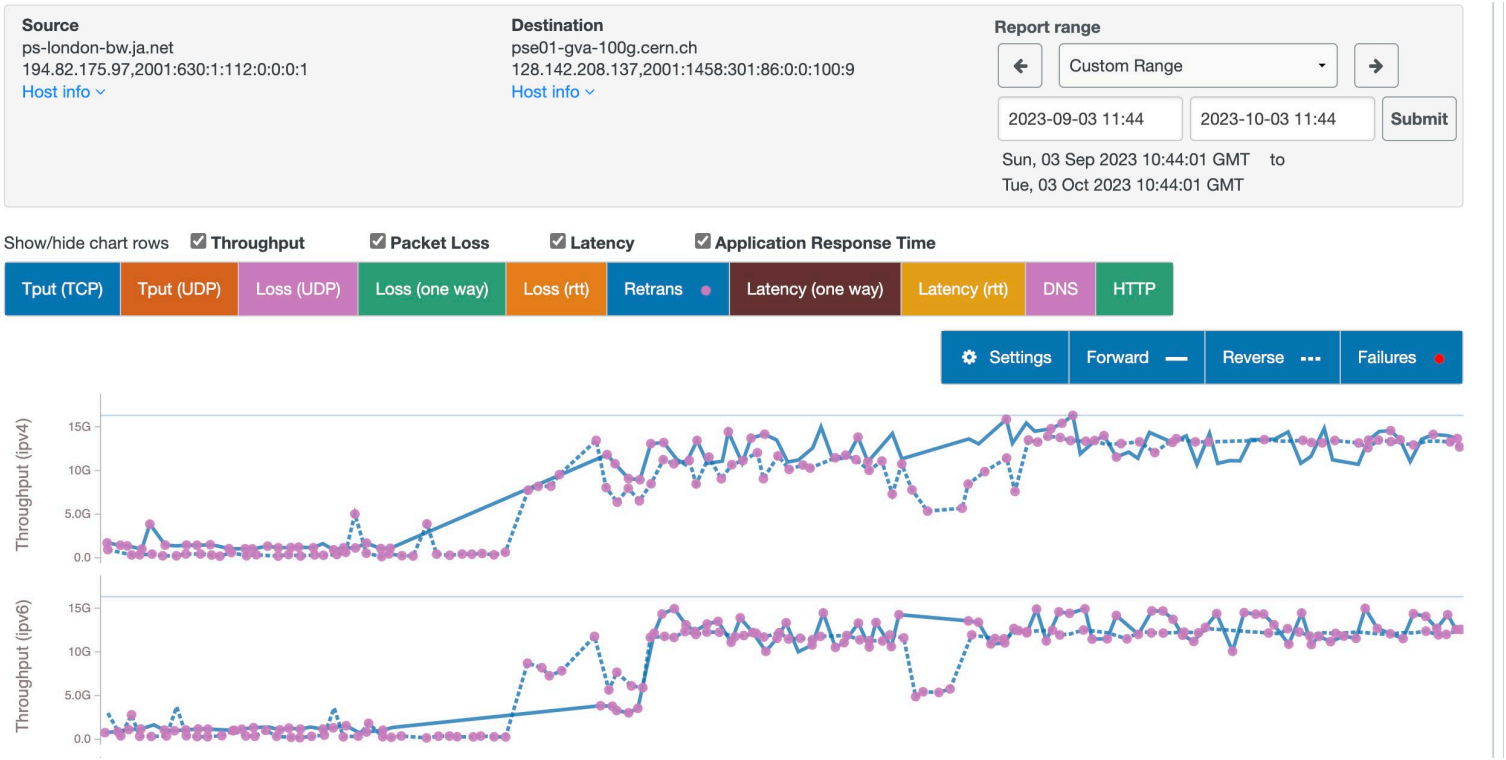
- Jisc has a 100G-connected perfSONAR server in London
  - Interfaces: ps-london-bw.perf.ja.net, ps-london-lat.perf.ja.net
- Raul has upgraded it to run the 6.5 kernel on CentOS 7
  - BBRv3 patch applied, and made the default CCA
  - Other CCAs can be selected via pscheduler –congestion option
- Thus any default perfSONAR tests from ps-london will show the impact of BBRv3 (and the 6.5 kernel)
- Most easily seen on the WLCG 100G mesh:
  - https://psmad.opensciencegrid.org/maddash-webui/index.cgi?dashboard=WLCG%20100G%20Mesh

Jisc

perfSONAR results for discussion

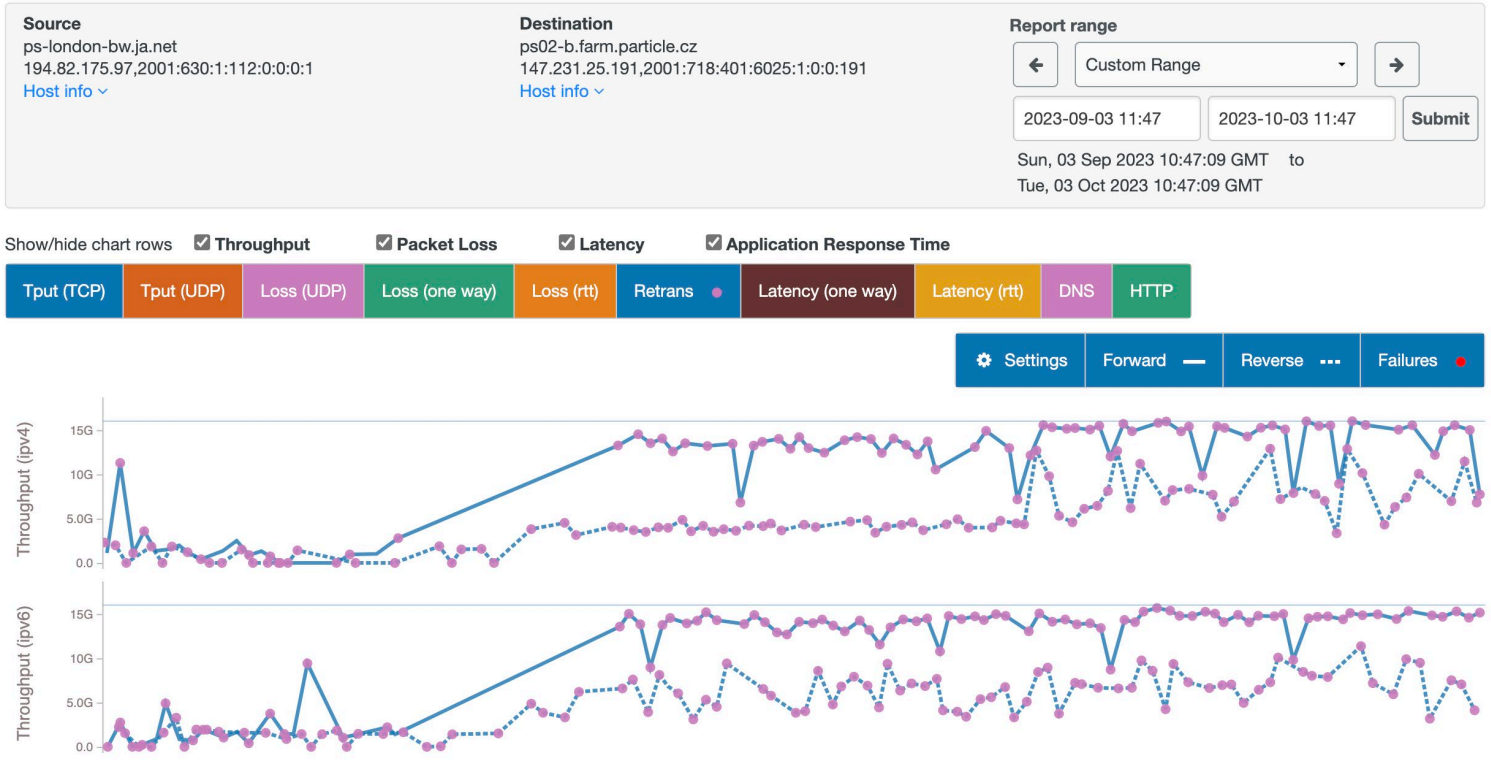
28-day views

ps-london - CERN



https://psmad.opensciencegrid.org/maddash-webui/details.cgi?uri=/maddash/grids/WLCG+100G+Mesh+-+WLCG+100G+IPv4+Bandwidth+-+Throughput/ps-london-bw.ja.net/pse01-gva-100g.cern.ch/Throughput

ps-london - Prague



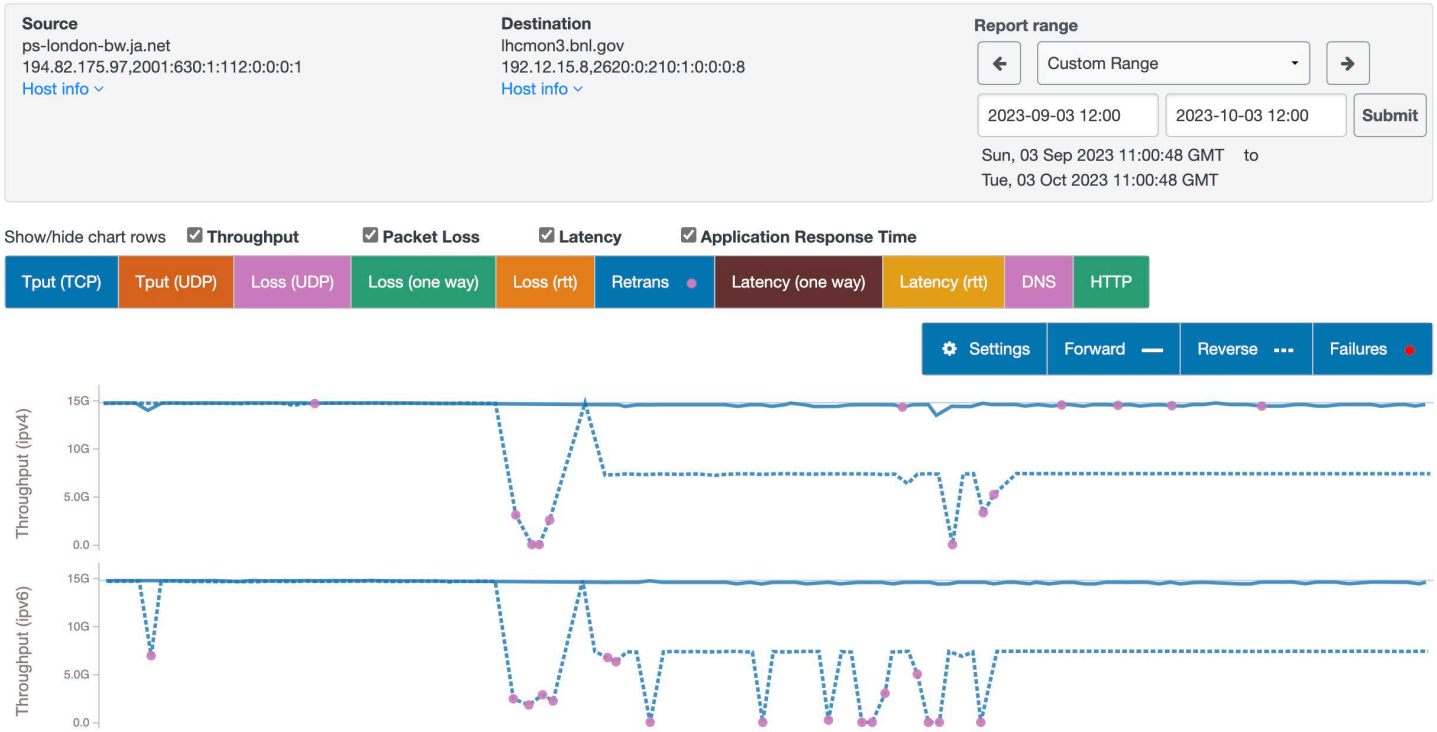
https://psmad.opensciencegrid.org/maddash-webui/details.cgi?uri=/maddash/grids/WLCG+100G+Mesh+-+WLCG+100G+IPv4+Bandwidth+-+Throughput/ps-london-bw.ja.net/ps02-b.farm.particle.cz/Throughput

Jisc

ps-london - QMUL

https://psmad.opensciencegrid.org/maddash-webui/details.cgi?uri=/maddash/grids/WLCG+100G+Mesh+-+WLCG+100G+IPv4+Bandwidth+-+Throughput/ps-london-bw.ja.net/perfsonar-bandwidth.esc.qmul.ac.uk/Throughput

ps-london - BNL



https://psmad.opensciencegrid.org/maddash-webui/details.cgi?uri=/maddash/grids/WLCG+100G+Mesh+-+WLCG+100G+IPv4+Bandwidth+-+Throughput/ps-london-bw.ja.net/lhcmon3.bnl.gov/Throughput

perfSONAR tests

Initial findings

- From the four sample tests:
  - CERN – improved in both directions – was there a change at CERN too?
  - Prague – relatively poor, good improvement but more loss/retransmits
  - QMUL – well-performing site, stayed the same
  - BNL – outbound the same, inbound down with no explanation
- General trends?
  - Improved, more consistent, throughput
  - Performs better where packet loss exists
  - Can cause more retransmissions

Jisc

Iperf3-MT local lab tests

Jisc BBRv3 deployment: iperf3

Using iperf3

- Jisc also has other 100G servers in its London network test facility
  - Raul has been running BBRv3 on these, using iperf3 and iperf3-MT
  - Tests run through the local Edgecore Wedge running RARE/FreeRtr
- Most recent results
  - iperf3-MT gets 55Gbps single stream and 90Gbps on two streams
  - Performs well without requiring pinning - *speculation* is this to do with BBRv3 using *io_uring* to send packets?
  - Raul is tweaking TCP settings like default write/read window, MTU, queueing discipline, …
- Very early stages but Raul would be interested in testing on longer paths

Jisc

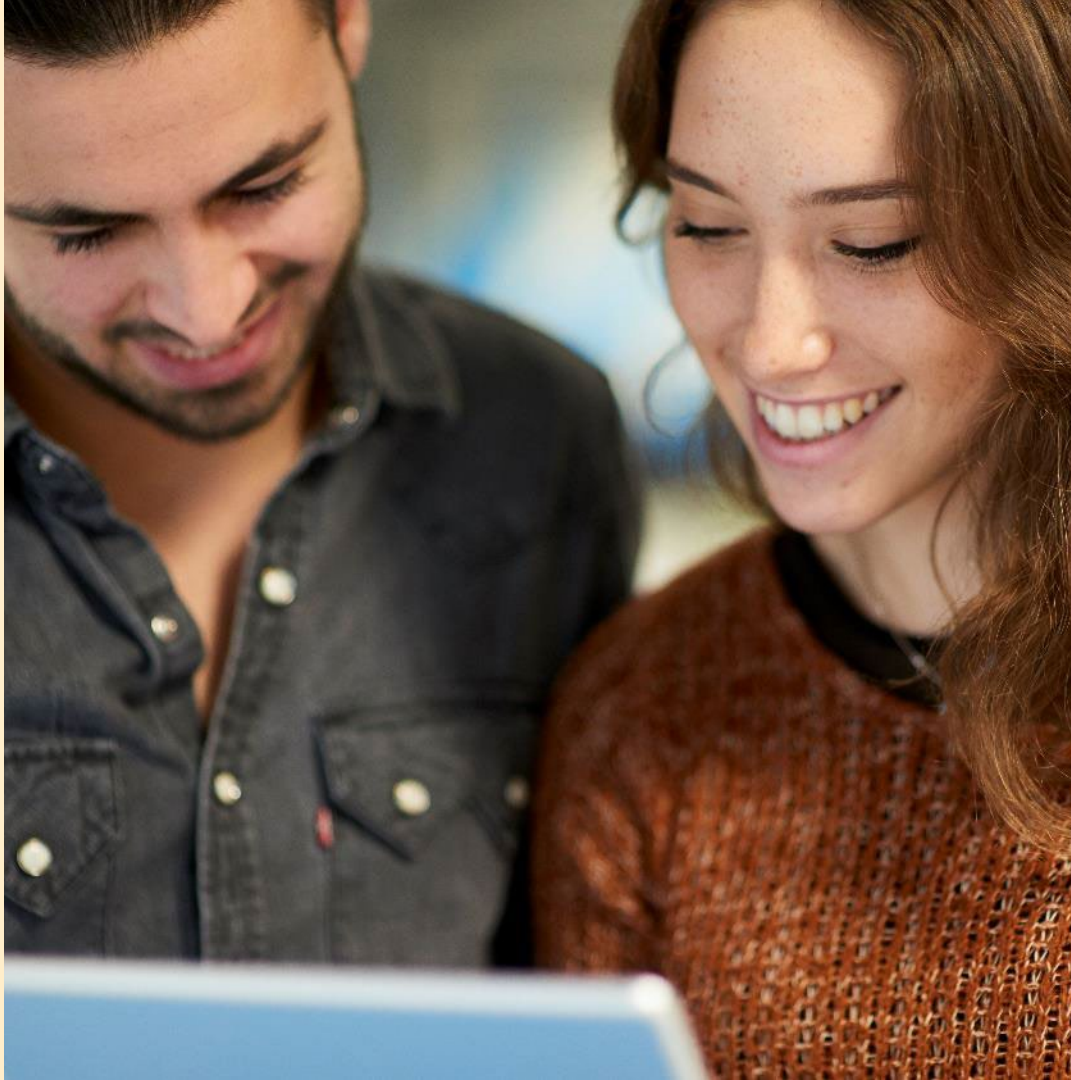
Discuss!

Tim Chown

tim.chown@jisc.ac.uk

netperf@jisc.ac.uk

Tel 01325 822106

customerservices@jisc.ac.uk

jisc.ac.uk