Jumbo Frames

Chris Walker, Jisc, christopher.walker@jisc.ac.uk
Raul Lopes, Jisc, raul.lopes@jisc.ac.uk
Duncan Rand, Jisc duncan.rand@jisc.ac.uk
Tim Chown, Jisc, tim.chown@jisc.ac.uk

HEPiX IPv6 WG, CERN, 23 February 2023

Overview

Jumbo

- Larger packets (MTU=9000, rather than 1500)
- Potential performance advantage of larger MTU
  - Higher link capacities
  - CPU clock speed not increasing
  - Larger frames intuitively make sense
- WLCG recommendation in 2018
  - MTU (“jumbo frames”) recommendation for LHCONE and LHCOPN (cern.ch)
  - Goal was to get NRENs to support jumbo frames
- Is the time right to try it out?

Jisc

What is the benefit?

In principle, higher throughput

- Fewer packets to process means less load on CPU
- Larger frame means faster ramp up / recovery for most TCP algorithms after a congestion event
- The TCP calculator provided by SWITCH gives a theoretical (Mathis) perspective:
  - https://www.switch.ch/network/tools/tcp_throughput
  - Plug in MSS, RTT and estimated loss rate
  - Rate <= MSS/RTT * 1/(sqrt(loss))
- But real data trumps theory…

Jisc

Network test data

- Iperf (Raul from Jisc)

Source	Destination	RTT	9000	1500
SURF (NL)	RNP (Brazil)	100ms	31 Gbit/s	20 Gbit/s
Jisc (London)	BNL (USA)	100ms	14 Gbit/s	6 Gbit/s
SURF (NL)	Jisc (London)	7.2 ms	23 Gbit/s	6 Gbit/s

- Tcpmon (Richard Hughes-Jones – Geant)

Source	Destination	RTT	9000	1500
London	Cambridge	3ms	37 Gbit/s	15.8 Gbit/s
London	AARnet	262ms	21 Gbit/s	3.4 Gbit/s

Jisc

Jumbo Frames

What is a jumbo Frame?

- MTU=9000 (IP layer)
  - VLAN, VxLAN, MPLS etc all add extra overhead that switches routers need to provide
  - https://indico.cern.ch/event/725706/contributions/3120030/attachments/1743507/2821722/LHCONE-MTU-recommendation.pdf
- Can all NRENs carry this?
  - Jisc can for IP, but may not be able to for Netpath links

Jisc

Concerns?

What genuine concerns are there?

- All hosts on a LAN must run 9000 MTU
  - But is that a problem? Put non-jumbo hosts in another LAN?
- Path MTU discovery (PMTUD) needs to work (to non jumbo sites)
  - PMTUD (ICMP) packets may be blocked by
    - Firewalls
    - non routable addresses
    - net.ipv4.tcp_mtu_probing=1 for IPv4?
  - *For IPv6, RFC4890 should be followed – don't drop ICMPv6 PTB!*

Jisc

Thoughts?

- Do we want to have another push on this?
  - QMUL, RALPP already doing this
  - Data transfer tests desirable
- Is MTU=9000 agreed (at least at NREN level)?
  - Do we need to test this?
- What to advocate?
  - (e.g., tips like net.ipv4.tcp_mtu_probing=1)
- Next steps?

Jisc

BACKUP SLIDES

Previous WLCG recommendation - 1

Context given in a LHCONE/LHCOPN meeting in October 2018

- What is meant by MTU – "*largest layer 3 (IP) data unit that can be communicated in a single network transaction*"
- "Jumbo frame" is ethernet frame with (IP) payload > 1500 bytes
- Goal is "*end-sites to be able to set their NIC MTU=9000 and have those packets be able to traverse the LHCONE/LHCOPN networks without fragmentation*"
  - Implicit choice of 9000MTU for hosts
- There would usually be 14 to 22 bytes of framing
  - MPLS adds additional 8 bytes, VXLAN adds 50 bytes

Previous WLCG recommendation - 2

The proposal (as written in 2018)

- LHCONE/LHCOPN network paths should allow MTU size up to 9000 bytes and not block PMTUD packets (RFCs 1911, 1981 and 4821)
- In practice this means that the frame size should be at least 9080 bytes for all devices on the path
- ICMP “Fragmentation Needed” (Type 3, Code 4) should not be blocked by any devices on the path
- https://indico.cern.ch/event/725706/contributions/3120030/attachments/1743507/2821722/LHCONE-MTU-recommendation.pdf

Jisc

How jumbo is a jumbo frame?

What specific MTU do we mean?

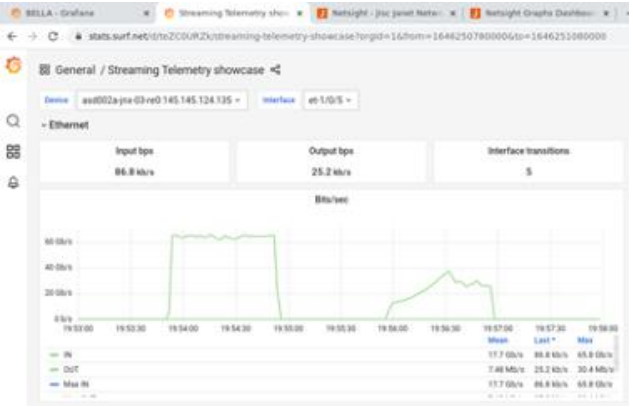
- The WLCG proposal used 9000MTU
- Chris found a 2003 vintage statement by the Internet 2 Joint Engineering Team (JET) and US Federal government on adopting 9000MTU as the “jumbo” size:
  - https://noc.net.internet2.edu/i2network/jumbo-frames/rrsum-almes-mtu.html
  - Various reasons given, mostly technical but also ”*9000 is an easy number to remember*” (!)
- Not clear how widely supported this is within NREN networks

Jisc

Experimental results - 1

SURF – RNP (Brazil)

- Raul tested using perfSONAR's *pscheduler* harness as part of a report for the GÉANT GN4-3 project
- H-TCP, 4 streams, 256 MB windows
- Results shown in SURF's prototype streaming telemetry platform
  - *Top: iperf2: 65Gbit/s vs 40Gbit/s*
  - *Bottom: iperf3: 31Gbit/s vs 20Gbit/s*
- **9000MTU has 50% more throughput**

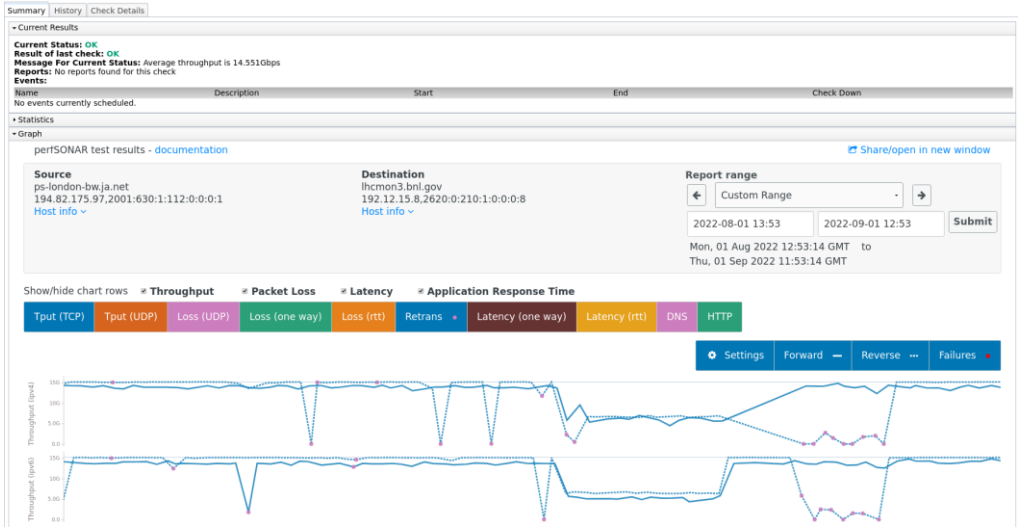


See https://resources.geant.org/wp-content/uploads/2023/02/GN4-3_White-Paper_Network-Performance-Tests-Over-100G-BELLA-Link.pdf

Experimental results - 2

Jisc London - BNL

- Here Raul changed the tuning between two perfSONAR servers and noted the plotted results over time
- The MTU is dropped from 9000 to 1500 for Jisc London to BNL (US), then raised again
- Throughput falls 14Gbit/s to 6Gbit/s
- (The second dip on the reverse path is where Raul sets the London pS node to default OS tuning)



Experimental results - 3

SURF to Jisc London – 9000 MTU

Interval	Throughput	Retransmits	Current Window
0.0 - 1.0	24.38 Gbps	0	90.12 MBytes
1.0 - 2.0	27.57 Gbps	0	90.73 MBytes
2.0 - 3.0	22.58 Gbps	0	90.73 MBytes
3.0 - 4.0	25.98 Gbps	0	90.73 MBytes
4.0 - 5.0	23.03 Gbps	0	90.73 MBytes
5.0 - 6.0	22.75 Gbps	0	90.73 MBytes
6.0 - 7.0	22.41 Gbps	0	90.73 MBytes
7.0 - 8.0	21.82 Gbps	0	90.73 MBytes
8.0 - 9.0	21.93 Gbps	0	90.73 MBytes
9.0 - 10.0	20.06 Gbps	0	90.73 MBytes

- Summary

Interval	Throughput	Retransmits	Receiver Throughput
0.0 - 10.0	**23.25 Gbps**	0	23.23 Gbps

SURF to Jisc London – 1500 MTU

Interval	Throughput	Retransmits	Current Window
0.0 - 1.0	8.91 Gbps	14145	10.54 MBytes
1.0 - 2.0	8.57 Gbps	0	10.68 MBytes
2.0 - 3.0	8.54 Gbps	263	5.41 MBytes
3.0 - 4.0	4.41 Gbps	0	5.55 MBytes
4.0 - 5.0	4.52 Gbps	0	5.69 MBytes
5.0 - 6.0	4.62 Gbps	0	5.86 MBytes
6.0 - 7.0	4.82 Gbps	0	6.18 MBytes
7.0 - 8.0	5.14 Gbps	0	6.65 MBytes
8.0 - 9.0	5.56 Gbps	0	7.25 MBytes
9.0 - 10.0	6.13 Gbps	0	8.02 MBytes

- Summary

Interval	Throughput	Retransmits	Receiver Throughput
0.0 - 10.0	**6.12 Gbps**	14408	6.09 Gbps

The above are tests Raul ran in March from a perfSONAR server in SURF to a Jisc server in London. The results were similar for 12 x 10 second tests and 60 x 1 second tests. In both cases the 1500MTU tests had several instances of retransmissions throughout the test, which may have affected the window size and thus performance. Do larger frames reduce retransmissions? Further tests to follow!

Jisc