Achieving 100 Gigabit Performance

Richard Hughes-Jones, GÉANT

March 2022

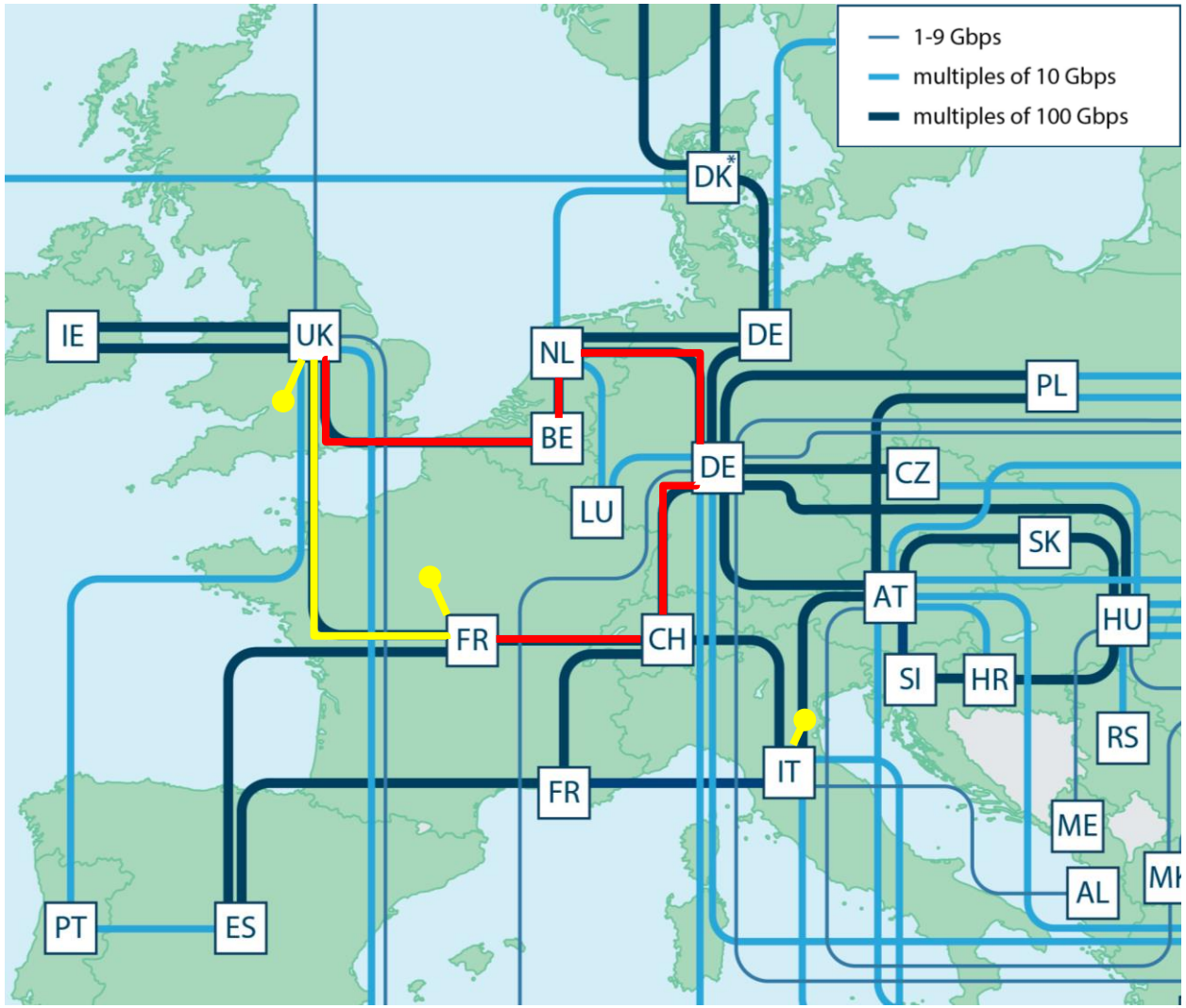
www.geant.org

Agenda

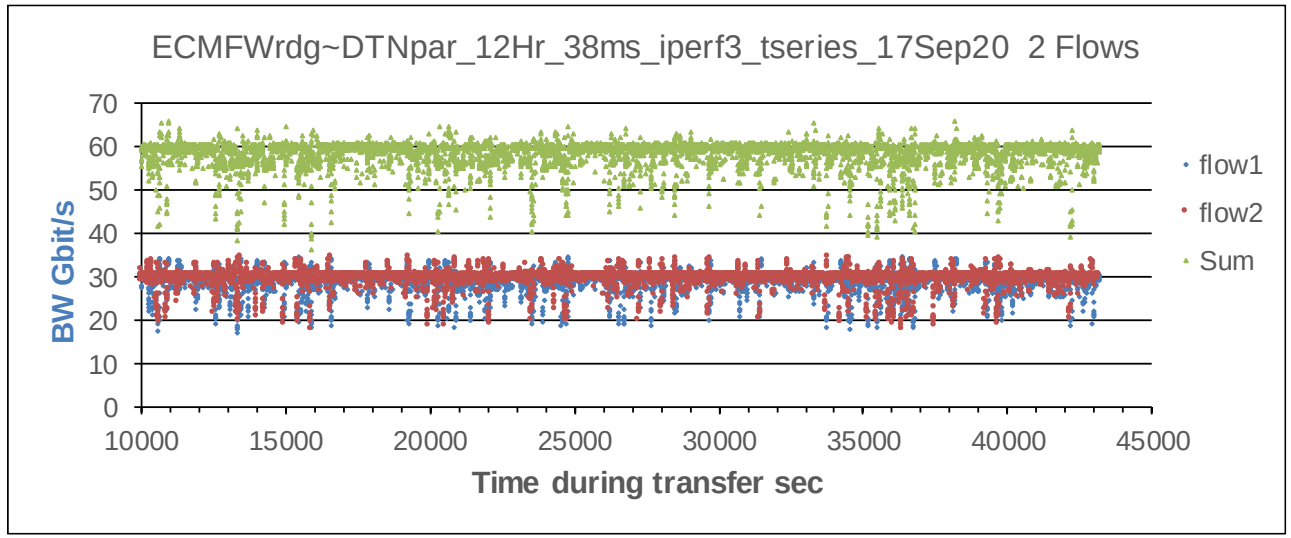
- The talk will present the tests and measurements made on 100 Gigabit network infrastructure using TCP/IP.
- Collaboration with ECMWF. TCP Transfers Between end sites in the UK and Italy using JISC, GÉANT, and GARR.
- TCP Transfers between Latin America and Europe using the EllaLink Submarine Cable.
- Transfers between Australia to Europe to support SKA using:
  - a path via the US and
  - a path on the CAE1 link.

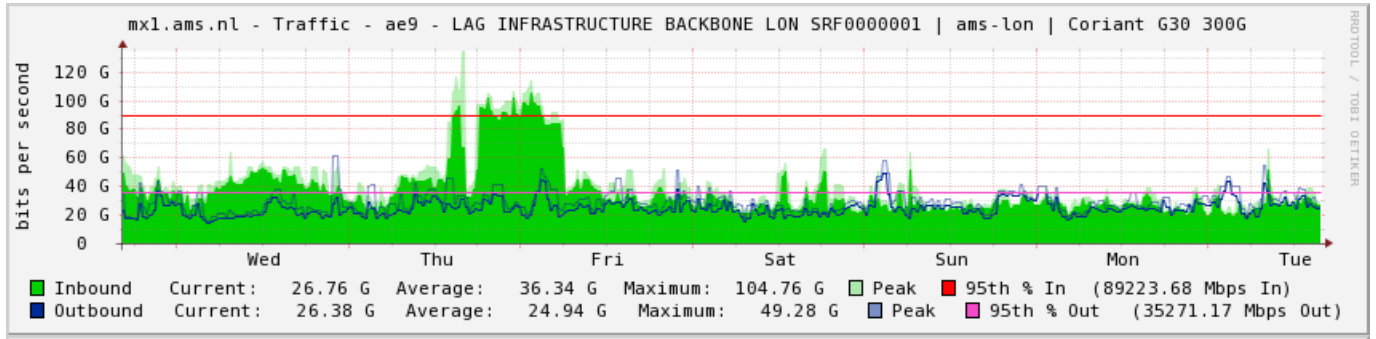
Collaboration with ECMWF

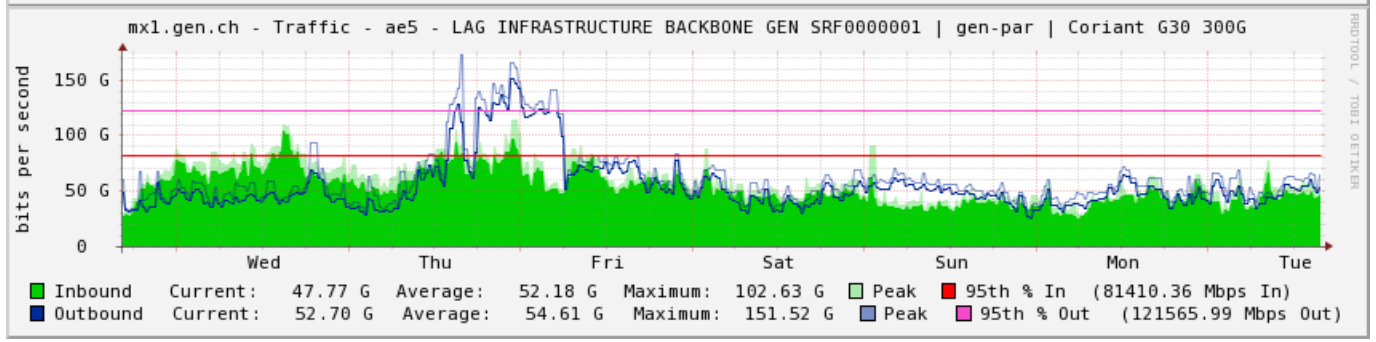
- ECMWF need to move data from Reading to Bologna.
- 100 Gigabit access links from JISC and GARR
- Worked with them to create a DMZ
  - Tests did not interfere with site production traffic
- Installed 2 DTNs
- DTN tuning followed the AENEAS recommendations
- Network & disk-to-disk tests between Reading and GÉANT DTN in Paris
- Two routes set up by the NOC
  - Direct route 16.1 ms
  - Long route 38.4 ms



Performance of Two TCP Flows Reading – Paris RTT 38.4 ms







- Two TCP flows for 12 Hours.
- Used iperf3 on 2 pairs of CPU cores.
- Used TCP auto-tuning.
- Single flows stable ave. 30-32 Gbit/s.
- Sum of both flows 60 Gbit/s

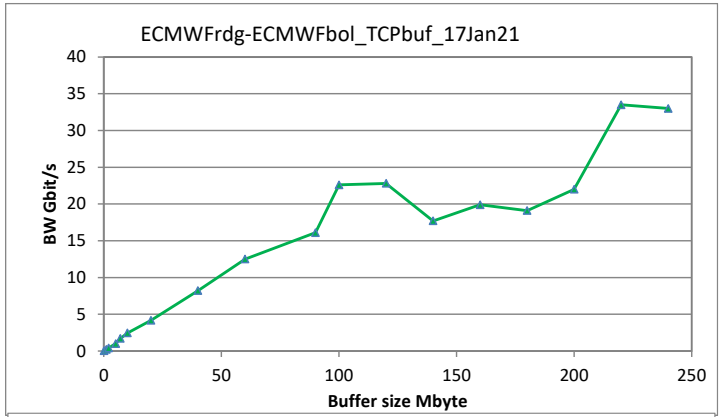
- Network can carry the load.

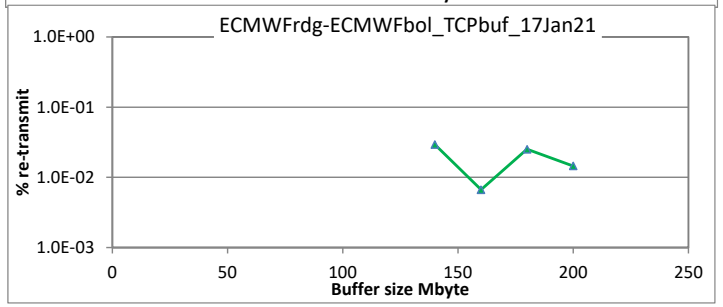
TCP Performance between ECMWF DTNs in Bologna and Reading

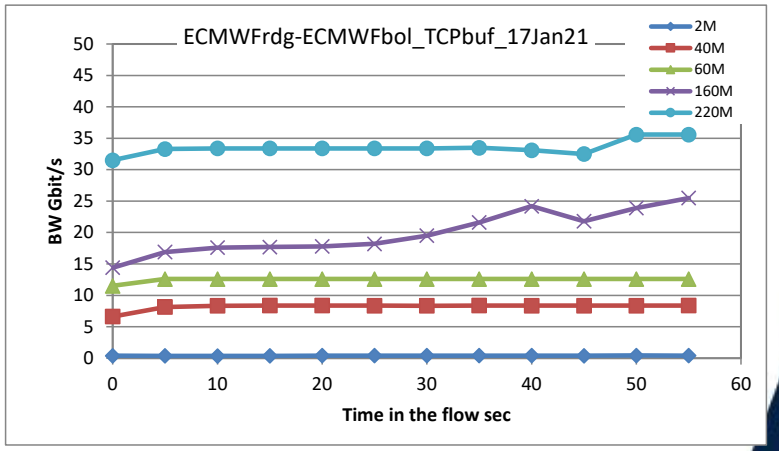
- Route JANET, GÉANT, GARR:
  Reading-London-Paris-Geneva-Milan-Bologna
- RTT 40 ms.
- Delay Bandwidth Product 170 MB for 34 Gbit/s.

- One TCP flow rises to the plateau at ~34 Gbit/s.
- Small amount of packet loss 140 – 200 Mbyte points.

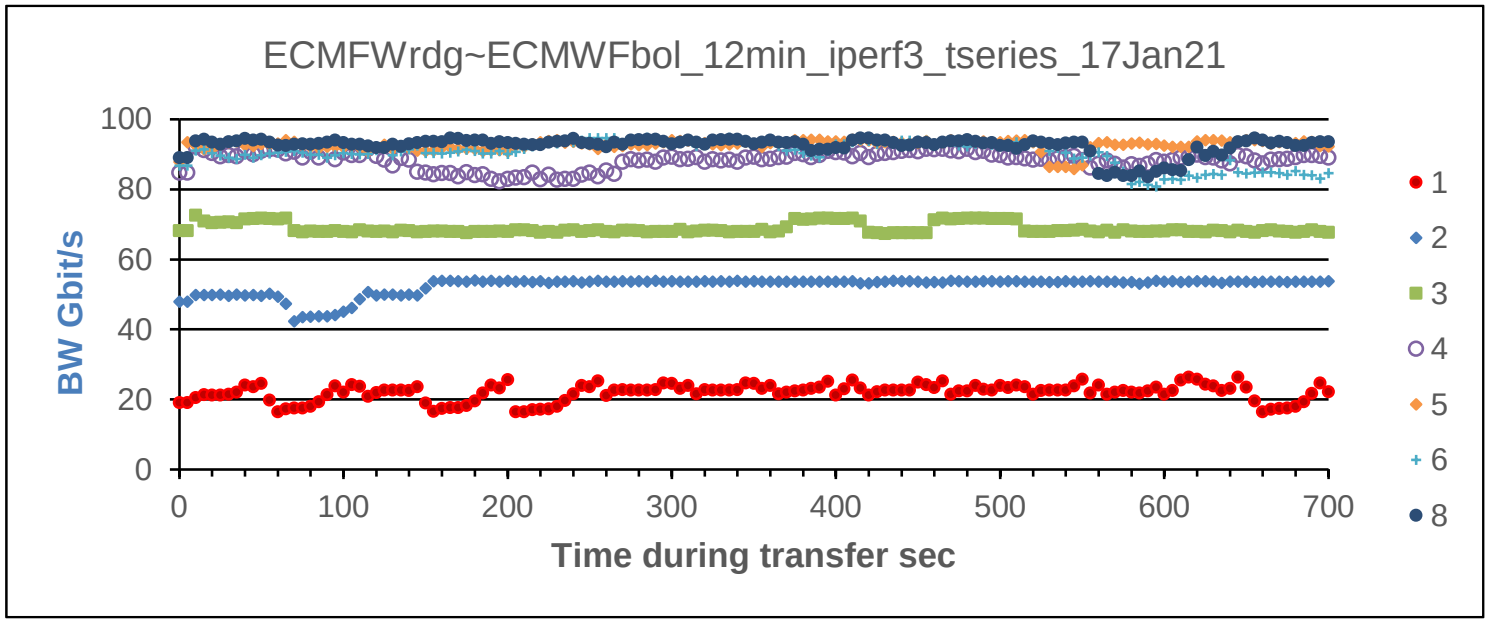
- Throughput throughout the:
  - Smooth as a function of time reaching 33.3 Gbit/s
  - Plateau from 5s onwards.





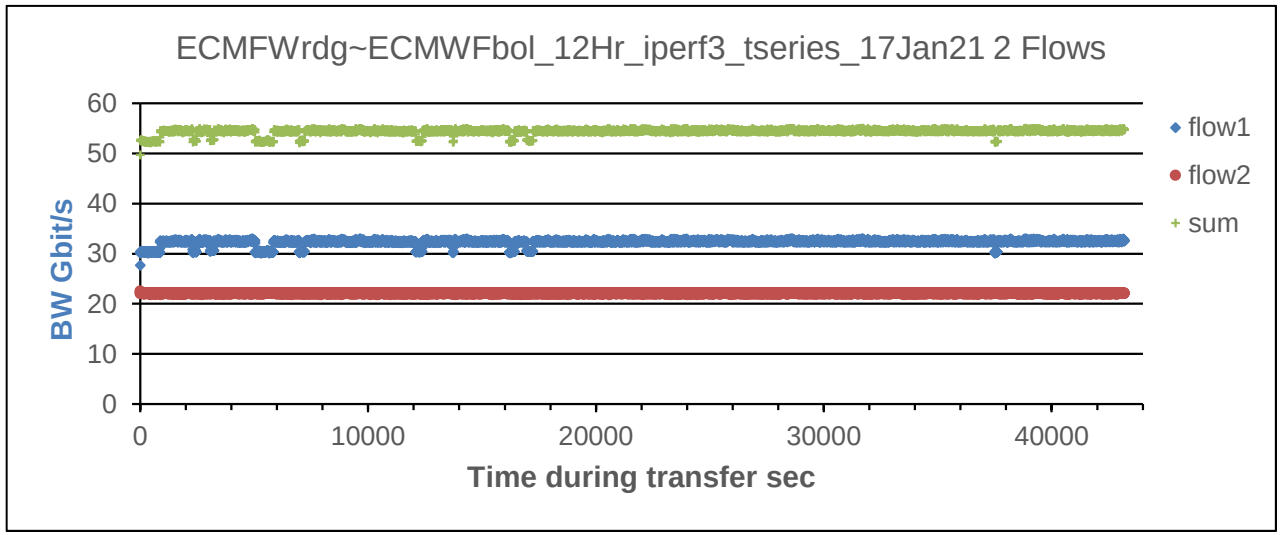


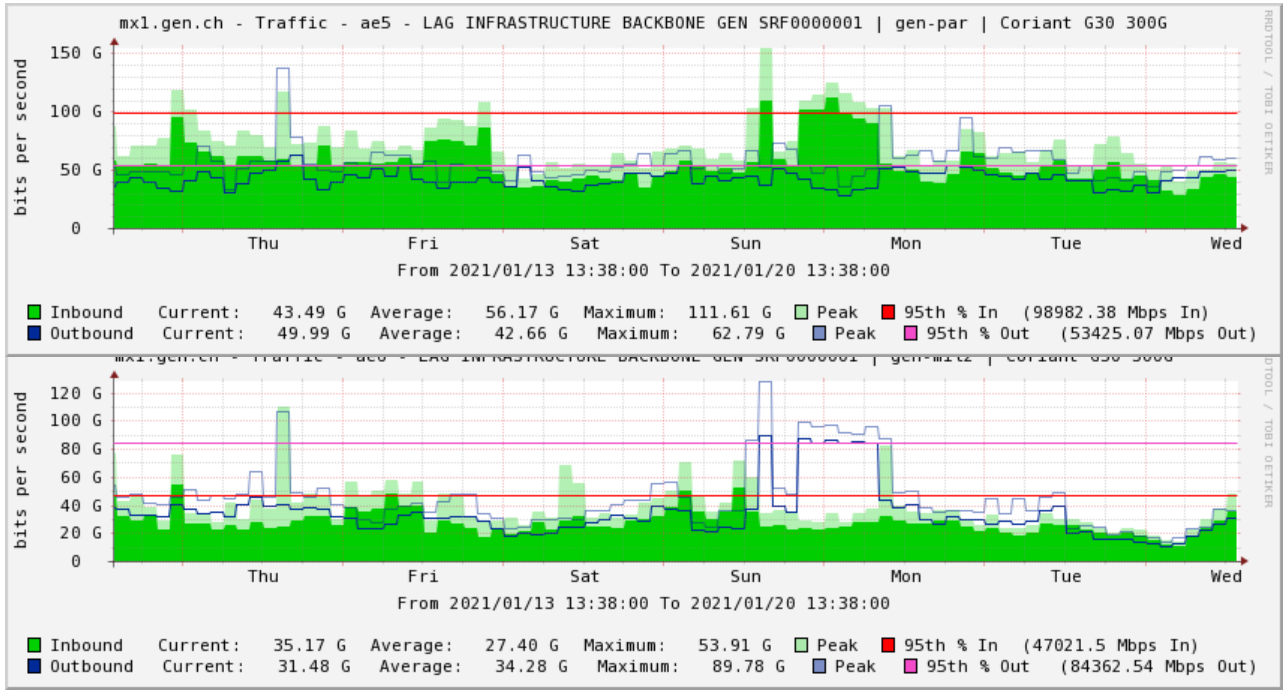
Performance of Multiple TCP Flows Reading – Bologna RTT 40 ms



- Step through 1 then 2 then 3 … TCP flows with 12 min for each configuration.
- Each iperf3 flow is between a separate pair of CPU cores.
- Kernel parameters tuned – using TCP auto-tuning
- Single flow stable ~22.5 Gbit/s
- 3 or more flows reach very stable 67.7 Gbit/s.
- 4 or more flows reach over 82 Gbit/s.
- Demonstrated suitability of DTNs tuning and the network.

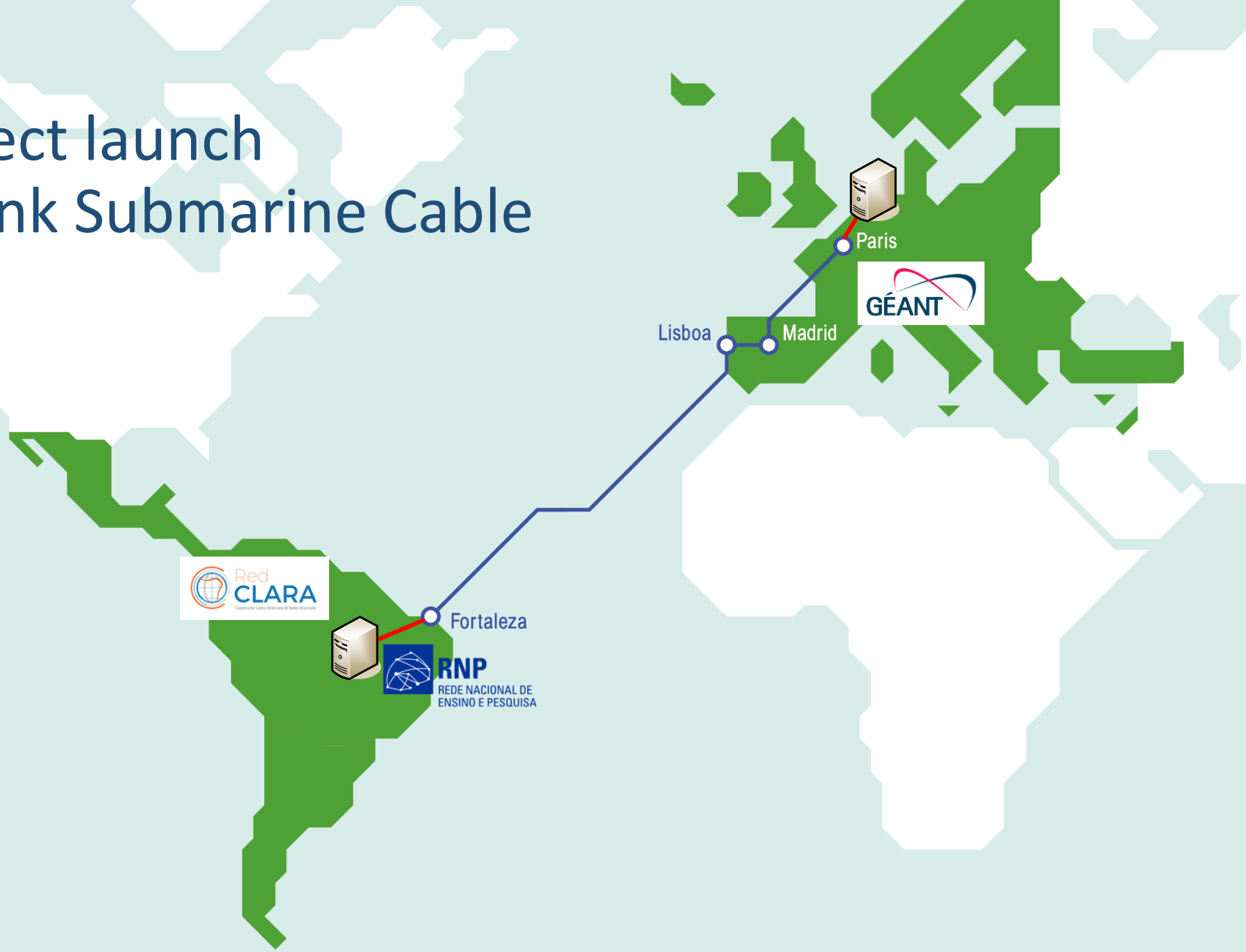
Performance of Two TCP Flows Reading – Bologna RTT 40 ms





- Two TCP flows for 12 Hours.
- Used iperf3 on 2 pairs of CPU cores.
- Using TCP auto-tuning.
- Single flows stable at 22 and 32 Gbit/s
- Sum of flows 55 Gbit/s.
- Network can carry the load.

The Bella Project launch with the EllaLink Submarine Cable

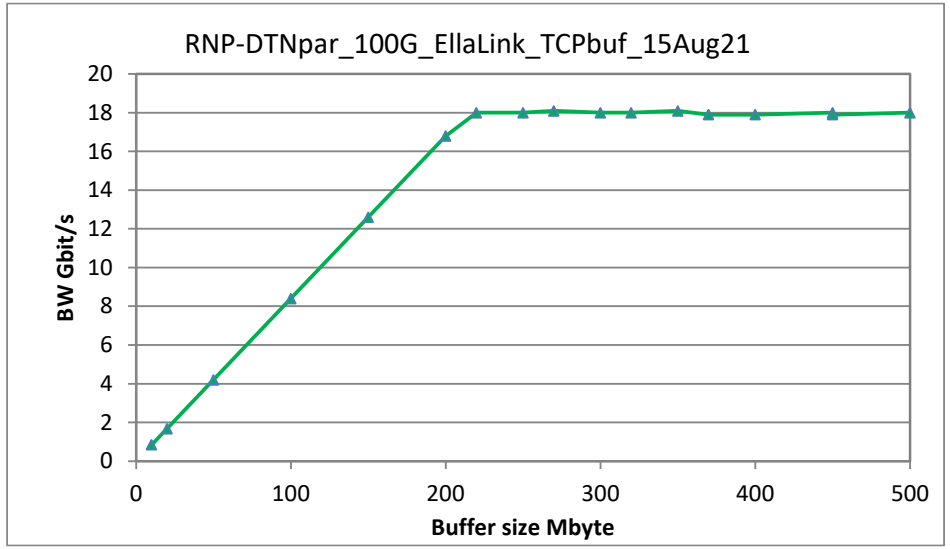


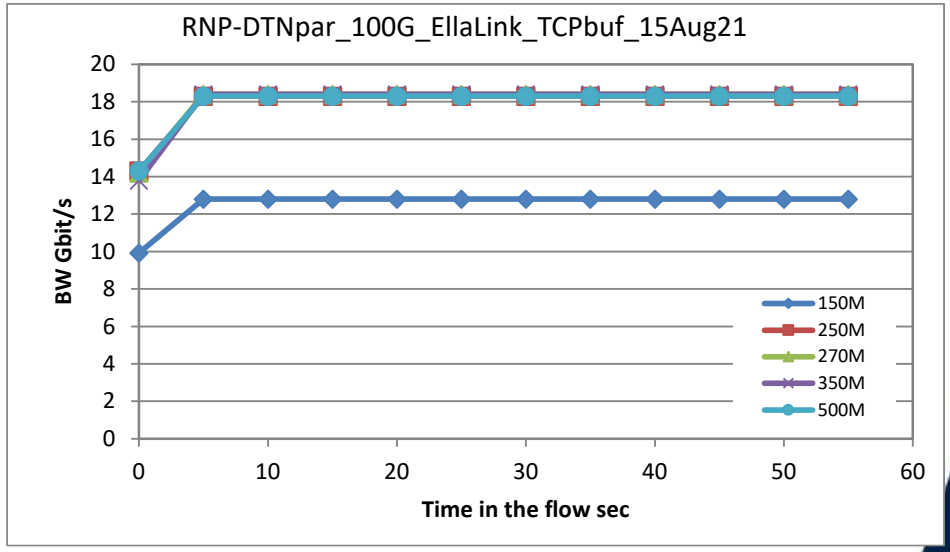
100 Gig TCP performance from RNP Fortaleza to GÉANT Paris via EllaLink

- Route GÉANT, EllaLink, RedCLARA & RNP : **Paris-Madrid-Lisbon-Sines-Fortaleza**
- RTT 86 ms
- iperf3 memory-memory TCP buffer size scan
- Delay Bandwidth Product 194 MB for 18 Gbit/s.
- Set iperf3 affinity to the “good” cores 100 GE NIC

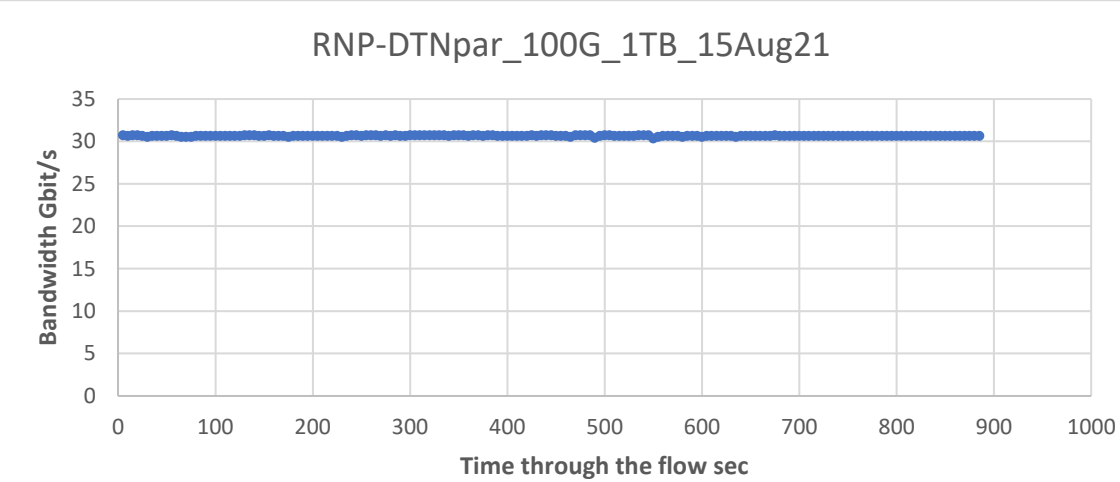
- One TCP flow rises smoothly to plateau of 18 Gbit/s with TCP window of 220 MBytes.
- In good agreement with calculation.
- No packet loss

- Throughput throughout the flow:
  - Smooth as a function of time reaching 18.3 Gbit/s after TCP slow start





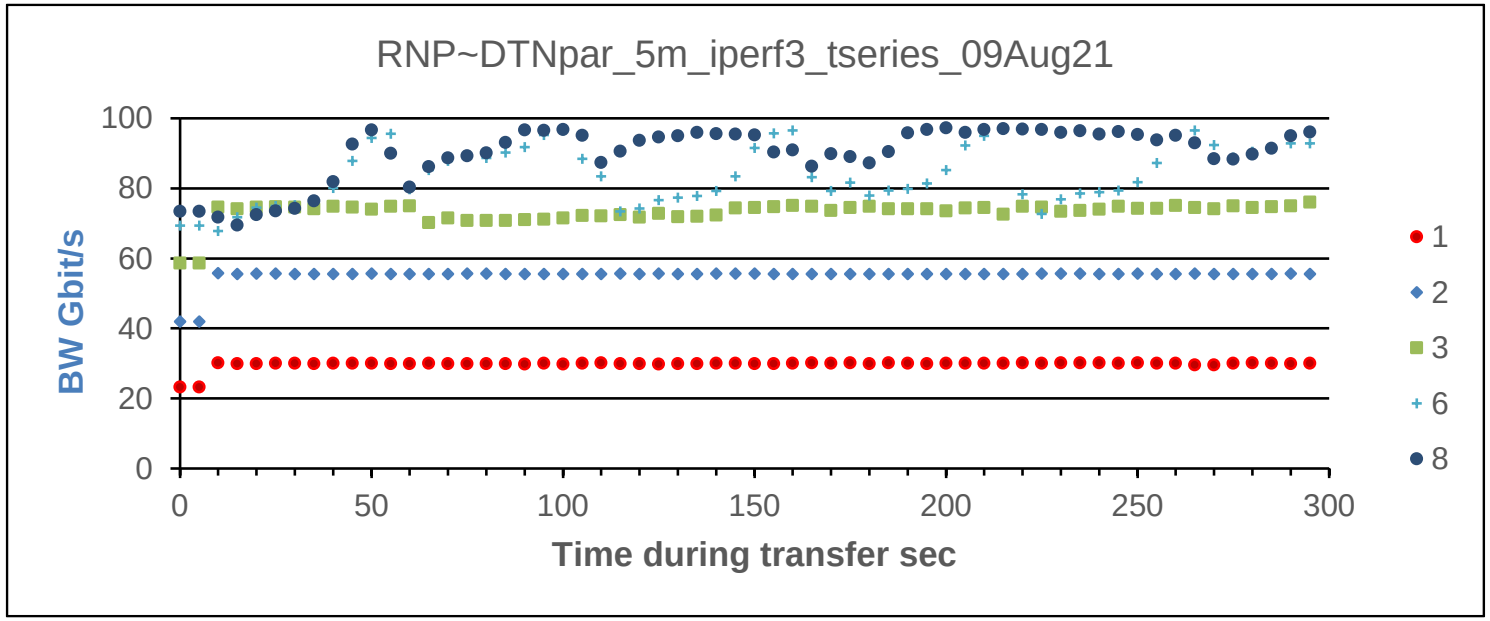
Time to Move 1 T Byte from Fortaleza to Paris via EllaLink RTT 86 ms



- One TCP flow using iperf3 memory to memory for 15 min.
- Measure the transfer every 5 sec.
- Kernel parameters tuned – using TCP auto-tuning.

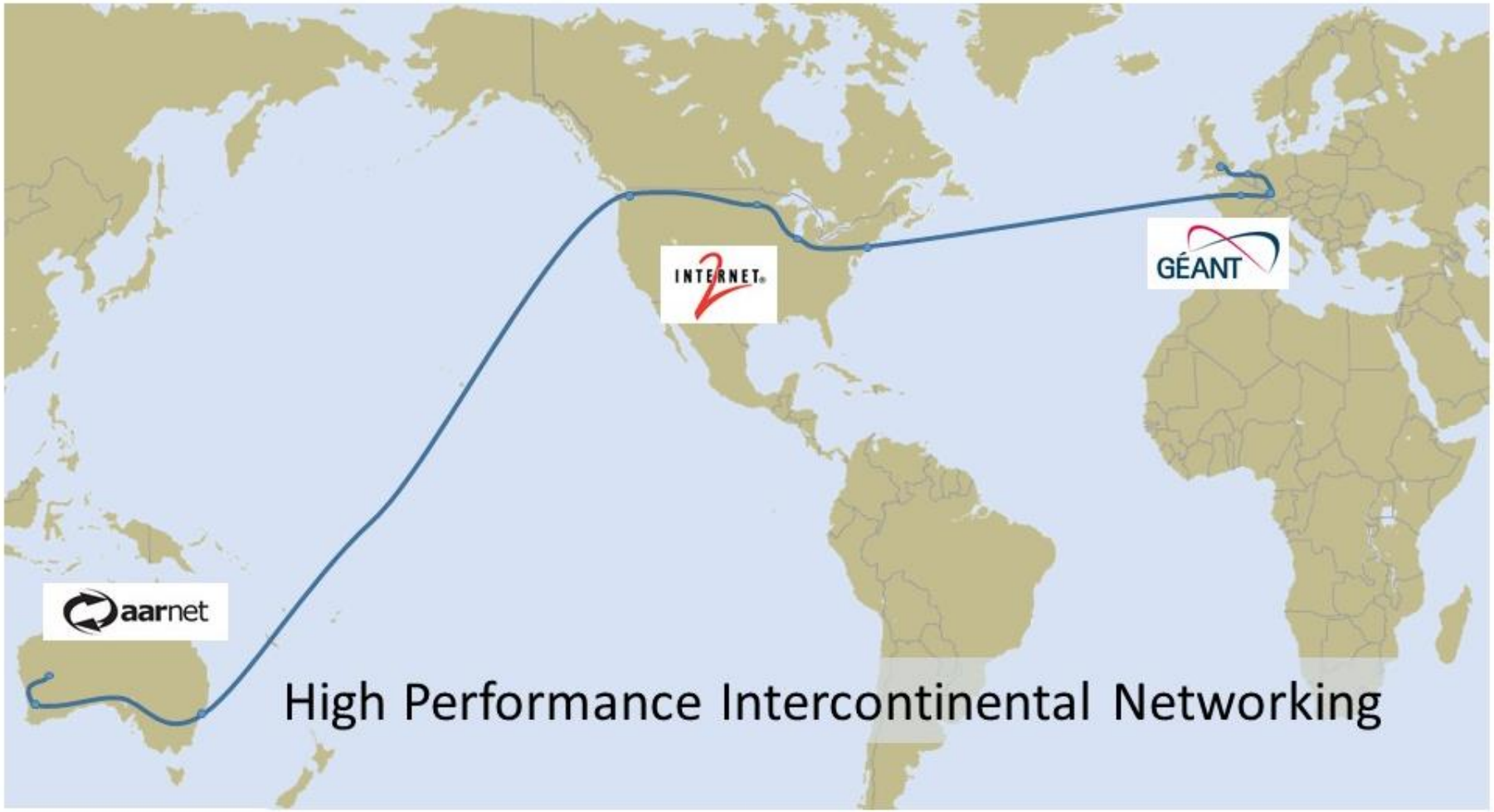
- Single flow stable at 30.5 Gbit/s for 15 min.
- No TCP re-transmits
- **Transfer of 1 TByte took 4 min 41 sec.**
- Theoretical time at 10 Gbit/s 13.3 minutes.

Performance of Multiple TCP Flows Fortaleza to Paris via EllaLink RTT 86 ms



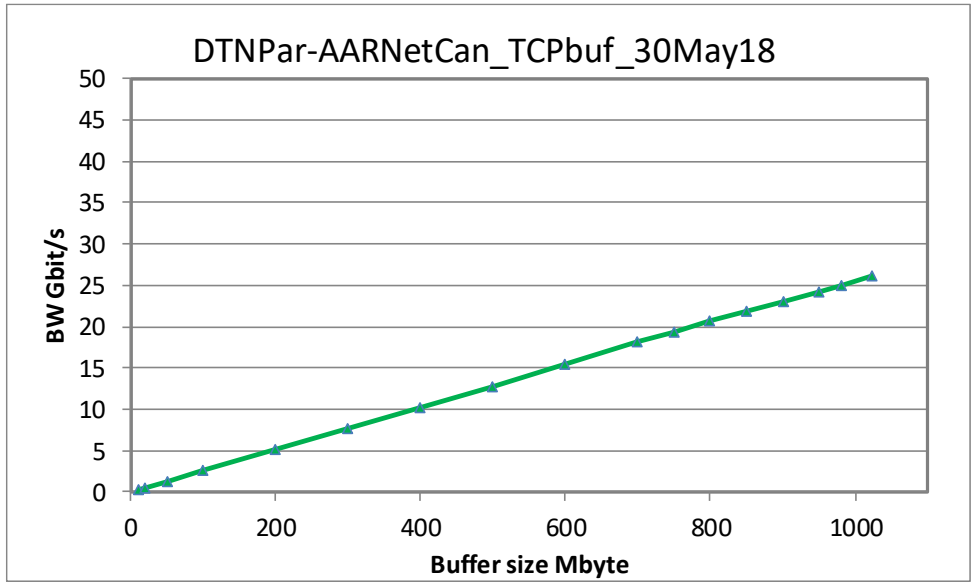
- Step through 1 then 2 then 3 … TCP flows with 5 min for each configuration.
- Each iperf3 flow is between a separate pair of CPU cores.
- Kernel parameters tuned – using TCP auto-tuning.
- Single flow ~30 Gbit/s
- 3 flows reach a very stable 74 Gbit/s.
- 6 or more flows reach over 90 Gbit/s.

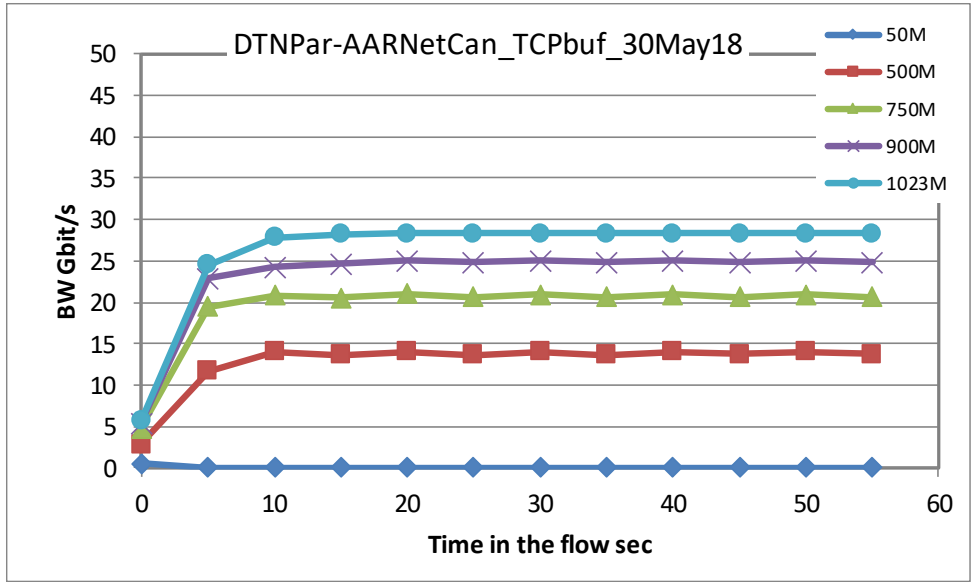
The Network Path between GÉANT (Lon, Par) – AARNet (Canberra, MRO) via the US



Thanks to Karl Meyer

100 Gigabit between GÉANT Paris and AARNet Canberra





- Route GÉANT, ANA300, Internet2, & AARNet:
  Paris-New York-Seattle-LosAngeles-Sydney-Canberra
- TCP offload on, TCP cubic stack
- RTT 303 ms.
- Delay Bandwidth Product 3.78 GB for 100 Gigabit

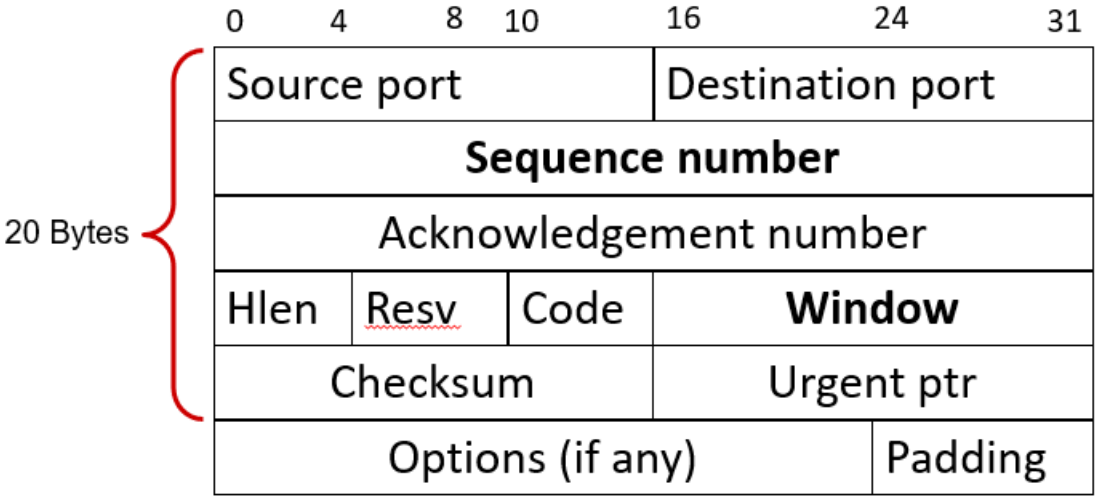
- One TCP flow rises smoothly to 26.1 Gbit/s at 1023 MBytes including slowstart.
- No TCP re-transmitted segments
- Rate after slowstart 28.3 Gbit/s
  - Plateau after ~15s

- Reach the limit of TCP protocol
  Max TCP window is 1 Gbyte
- Rate for RTT 303 ms and TCP window 1023 MB
  28.32 Gbit/s

- CPU core only 75-80 % in kernel mode

The TCP Protocol Limit

TCP Header

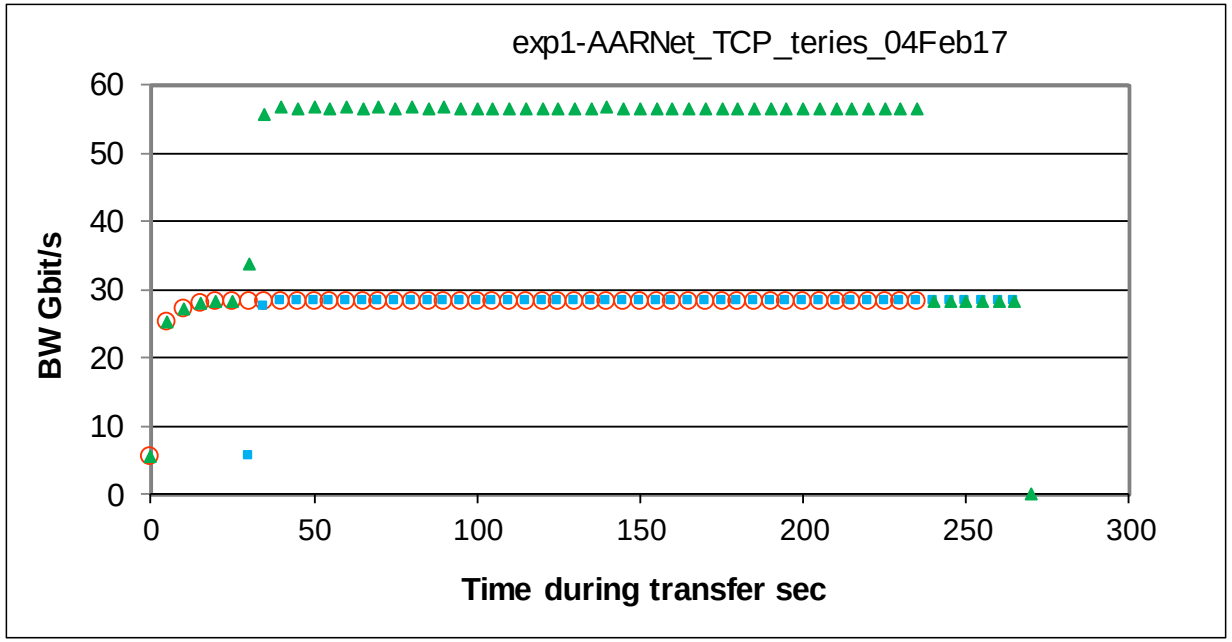


2^{32} → 4096 MB

2^{16} → 64 kB

- To fix the Window size there is the Window Scale factor negotiated at the SYN exchange. RFC 7323 (obsoletes 1323)
- Max value 14 → max Window ($2^{16} + 2^{14}$) → 1024 MB
- Window size < Sequence number
  - Deal with sequence number wrapping
  - Allow to tell is a segment is old or new

100 Gigabit: Multiple flows between GÉANT and AARNet



- Route GÉANT, ANA300, Internet2, & AARNet:
  Paris-New York-Seattle-LosAngeles-Sydney-Canberra
- RTT 303 ms.
- TCP window 1023 MB.

- Two 4 minute TCP flows
- Second flow started 30s after the first
- Each flow stable at 28.3 Gbit/s
- Total transfer rate 56.6 Gbit/s

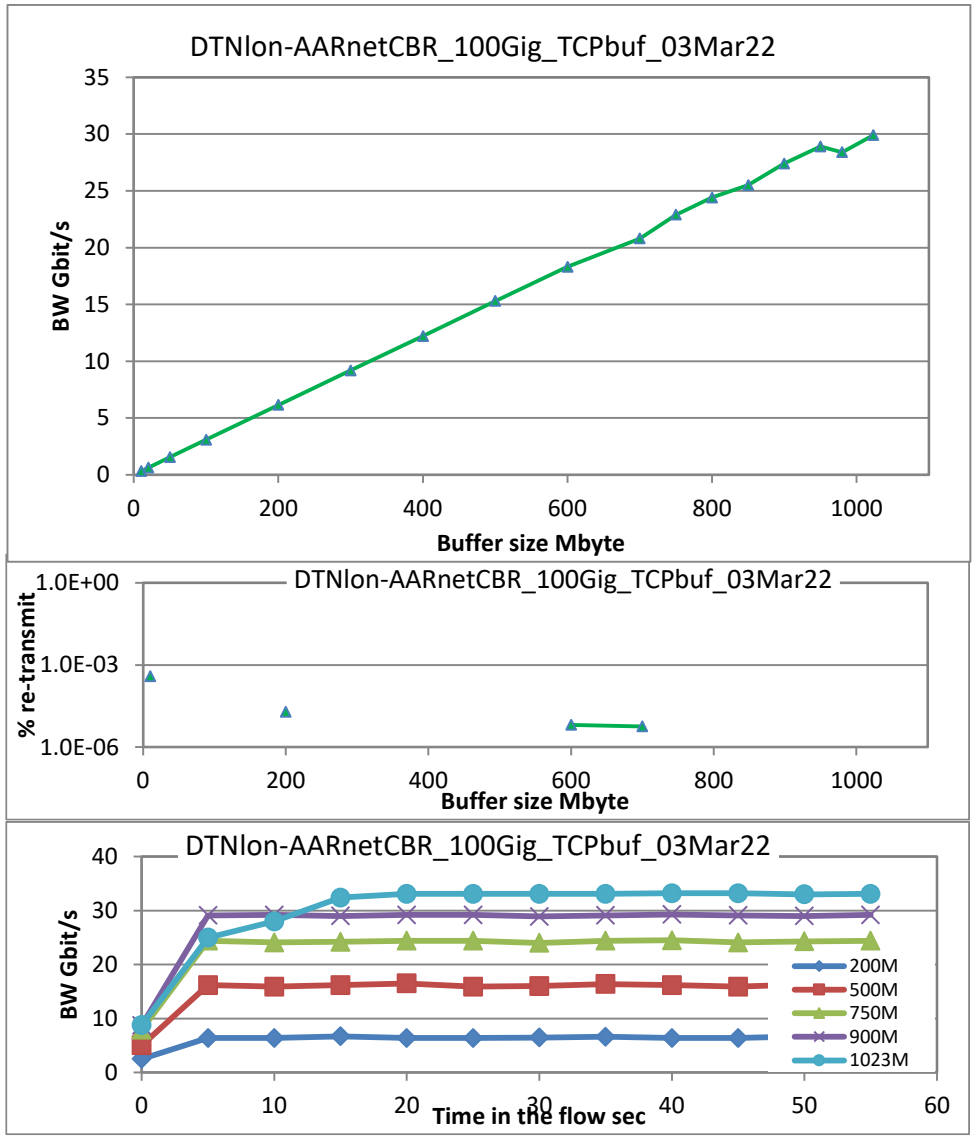
- 1.55 Tbytes data sent in 4.5mins.

- No TCP segments re-transmitted.

The Network Path between GÉANT (Lon, Par) – AARNet (Canberra, MRO) using CAE-1

Thanks to Karl Meyer

100 Gigabit between GÉANT LondonParis and AARNet Canberra

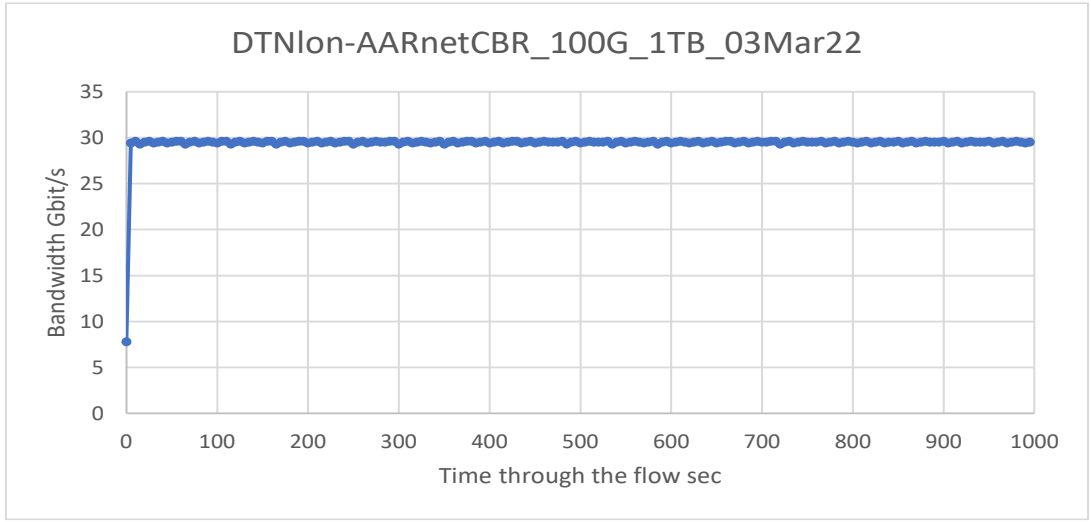


- Route GÉANT, CAE-1 & AARNet:
  London-Singapore-Sydney-Canberra
- TCP offload on, TCP cubic stack
- RTT 259 ms. 44 ms less than via the US
- Delay Bandwidth Product 3.24 GB for 100 Gigabit
  0.971 GB for 30 Gigabit

- One TCP flow rises smoothly to 30 Gbit/s at 1023 MBytes including slowstart. (26.1 Gbit/s via the US)
- Very small number of TCP re-transmitted segments
- Rate after slowstart 33.1 Gbit/s
  - Plateau after ~15s

- Reach the limit of TCP protocol
  Max TCP window is 1 Gbyte
- Rate for RTT 259 ms and TCP window 1023 MB
  33.13 Gbit/s – good agreement

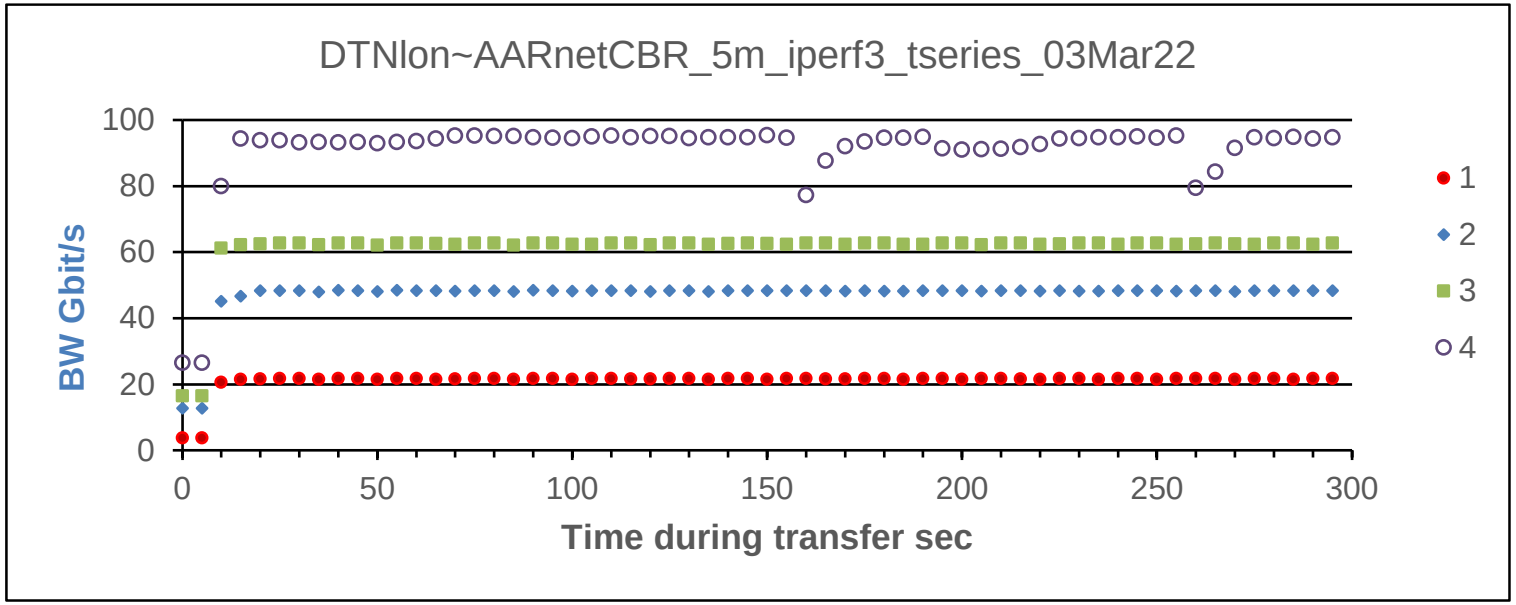
Time to Move 1 T Byte from London to Canberra via CAE-1 RTT 259 ms



- One TCP flow using iperf3 memory to memory for 1000 sec (16.7 min).
- Measure the transfer every 5 sec.
- Kernel parameters tuned – using TCP auto-tuning.

- Single flow stable at 30.5 Gbit/s.
- No TCP re-transmits
- **Transfer of 1 TByte took 4 min 32 sec.**
- Total data transferred 3.4 TB
- Theoretical time at 10 Gbit/s 13.3 minutes.

Performance of Multiple TCP Flows London to Canberra via CAE-1



- Step through 1 then 2 then 3 … TCP flows with 5 min for each configuration.
- Each iperf3 flow is between a separate pair of CPU cores
- Kernel parameters tuned – using TCP auto-tuning
- Single flow ~22 Gbit/s
- 3 flows reach about 62.5 Gbit/s.
- 4 flows reach over 94 Gbit/s. Some TCP re-transmits at the same time for the 4 flows

Summary

- With tuned DTN nodes measured excellent TCP performance between
  - Reading and Paris and Reading and Bologna RTT 40ms
  - Fortaleza and Paris RTT 86 ms
  - London and Canberra RTT 259 ms
- Shown that 1 T Byte can be moved in under 5 minutes with a single TCP flow.
- Demonstrated that multiple TCP flows between the DTN nodes can fill the 100 Gigabit link.

The Keys to good performance are

- Detailed tuning of the DTNs
- Bottleneck free end-to-end network. Core, Access links, Campus, DTN connection
- Minimise packet loss at the receiving NIC

GÉANT
Networks • Services • People

Thank you!

Richard.Hughes-Jones@geant.org

www.geant.org

The research leading to these results has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 731122 (GN4-2).