GÉANT Data Transfer Node Service

Vincenzo Capone

Head of Research Engagement and Support

Joint WLCG & HSF Workshop

Naples, 26 March 2018

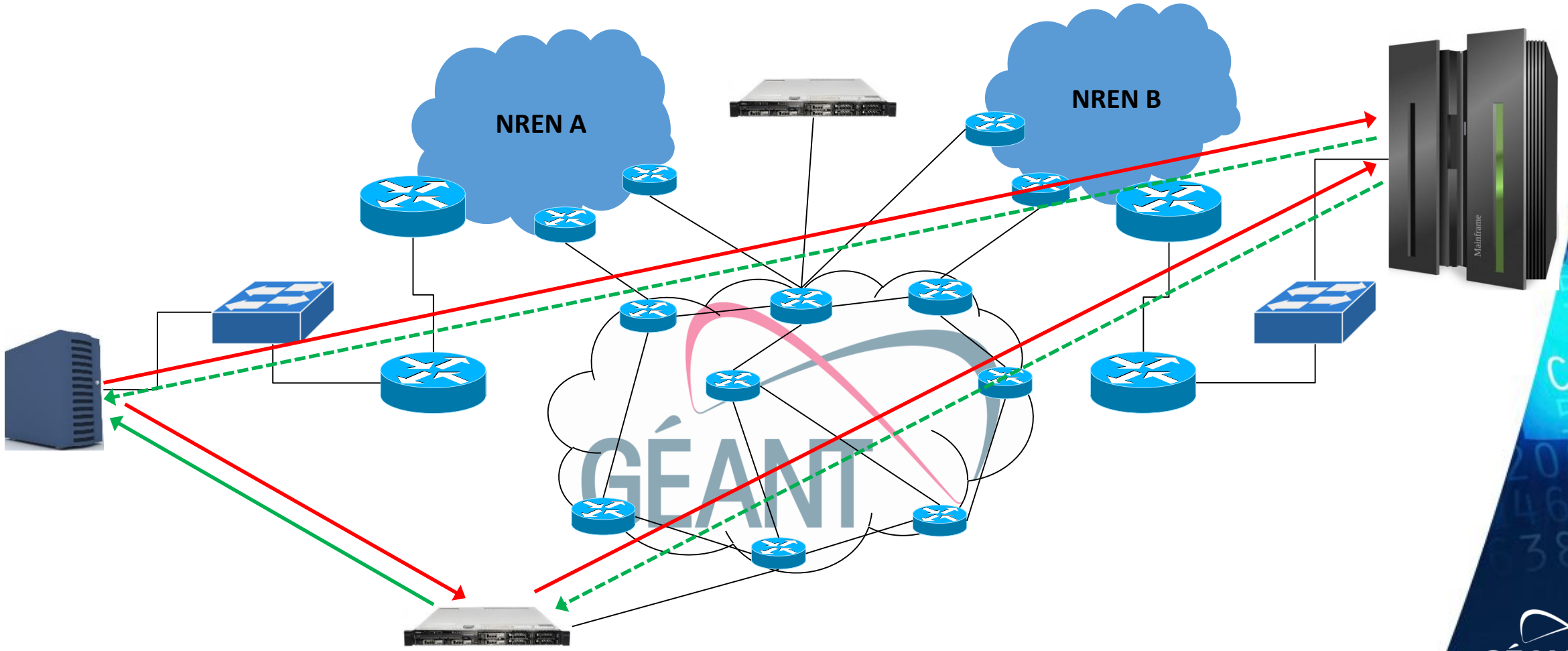
The problem

- Typical TCP's maximum throughput on standard servers is low compared to the available capacity
  - Optimised for general purpose, multi-user, bandwidth sharing mode
- Campus infrastructure is not optimised for large flows
- R&E networks need to engineer for zero packet loss
  - *packet loss is the curse of TCP transfers*
- ~80% of network traffic is using TCP

Some solutions

- Optimised tools and protocols
  - i.e.: GridFTP/FDT + ScienceDMZ
  - vs. basic scp, rysnc, wget
- Bypassing the campus firewalls removes the most typical bottleneck cause
- Fine-tuning the OS provides that extra-push…

DTN service example



Why using a DTN node?

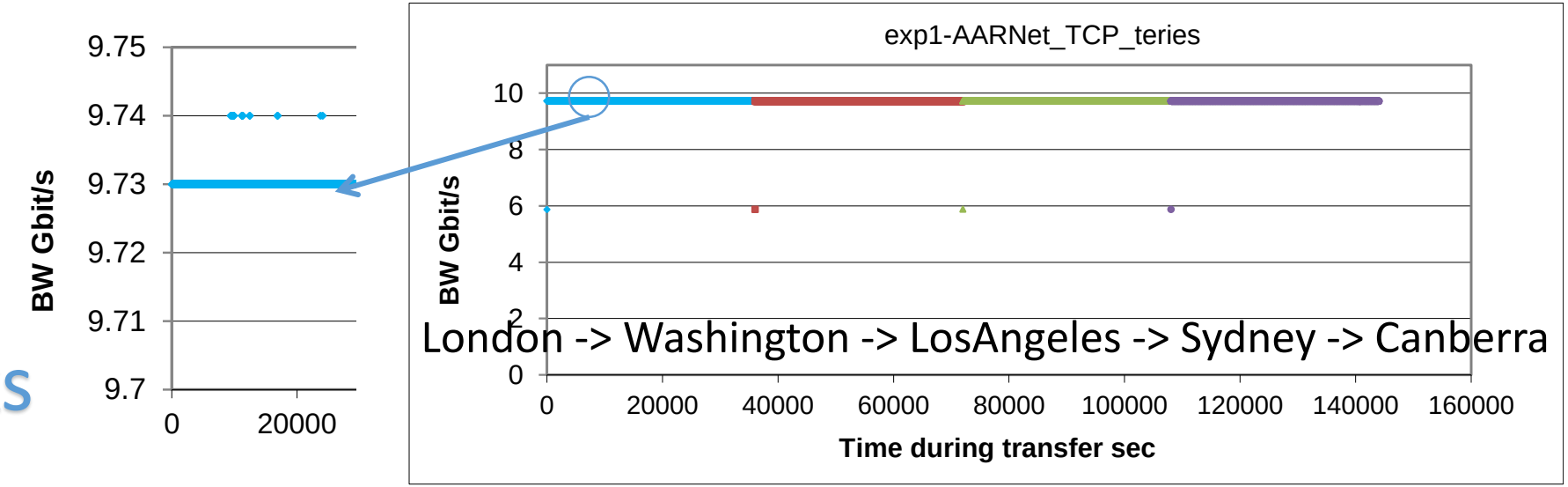
- Dedicated hardware
- Highly-optimized software and configuration
- Optimal network topology/location
  - No firewalling
  - Only simple ACLs
- Bottleneck-free
- What for?
  - Network troubleshooting
  - Network validation
  - Achievable network throughput measurement
  - Test of the whole software/application/storage stack

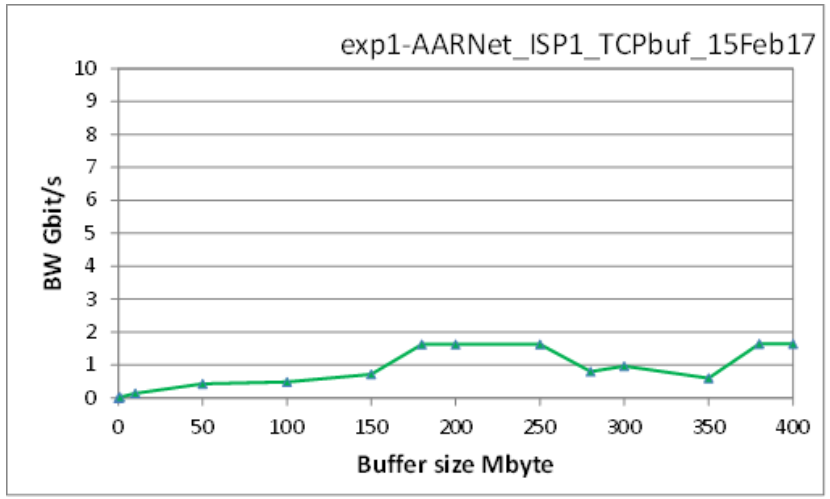
Other low-level diagnostics

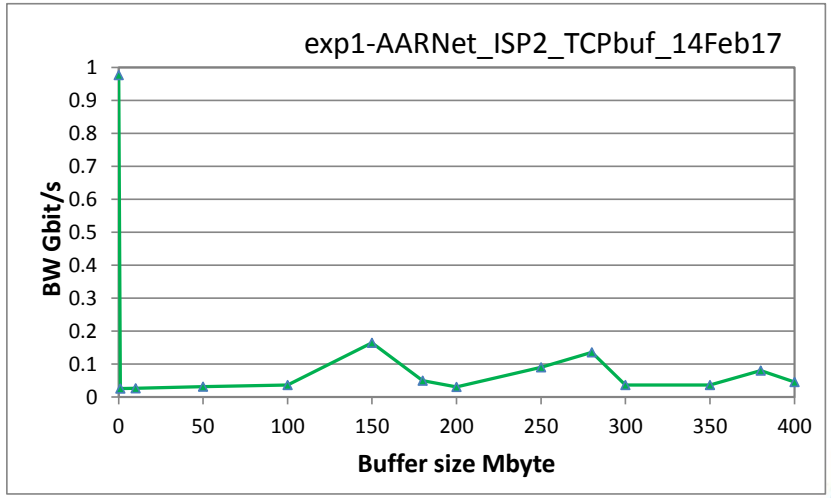
- UDP Achievable Throughput and packet loss as a function of the inter-packet gap (udpmon)
- Histograms of the inter-packet arrival times for UDP packets sent
- Investigation of lost UDP packets (udpmon)
- TCP Achievable Throughput and the number of re-transmitted segments as a Function of TCP Buffer Size (iperf3)
- TCP Achievable Throughput and the number of re-transmissions as a Function of Time (iperf3, logging information every 10seconds)

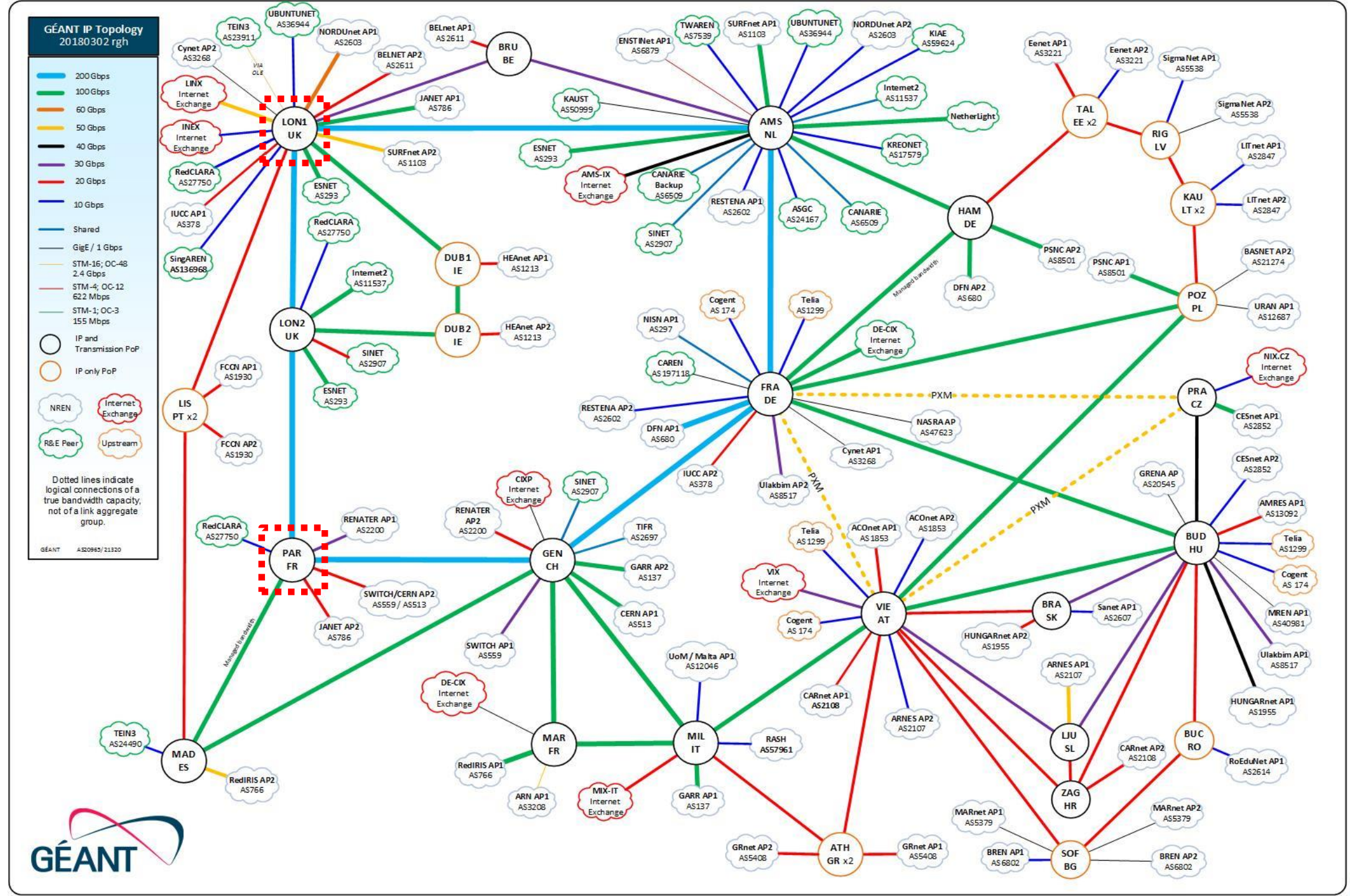
Network validation example

R&E Networks
vs.
Comm. ISPs

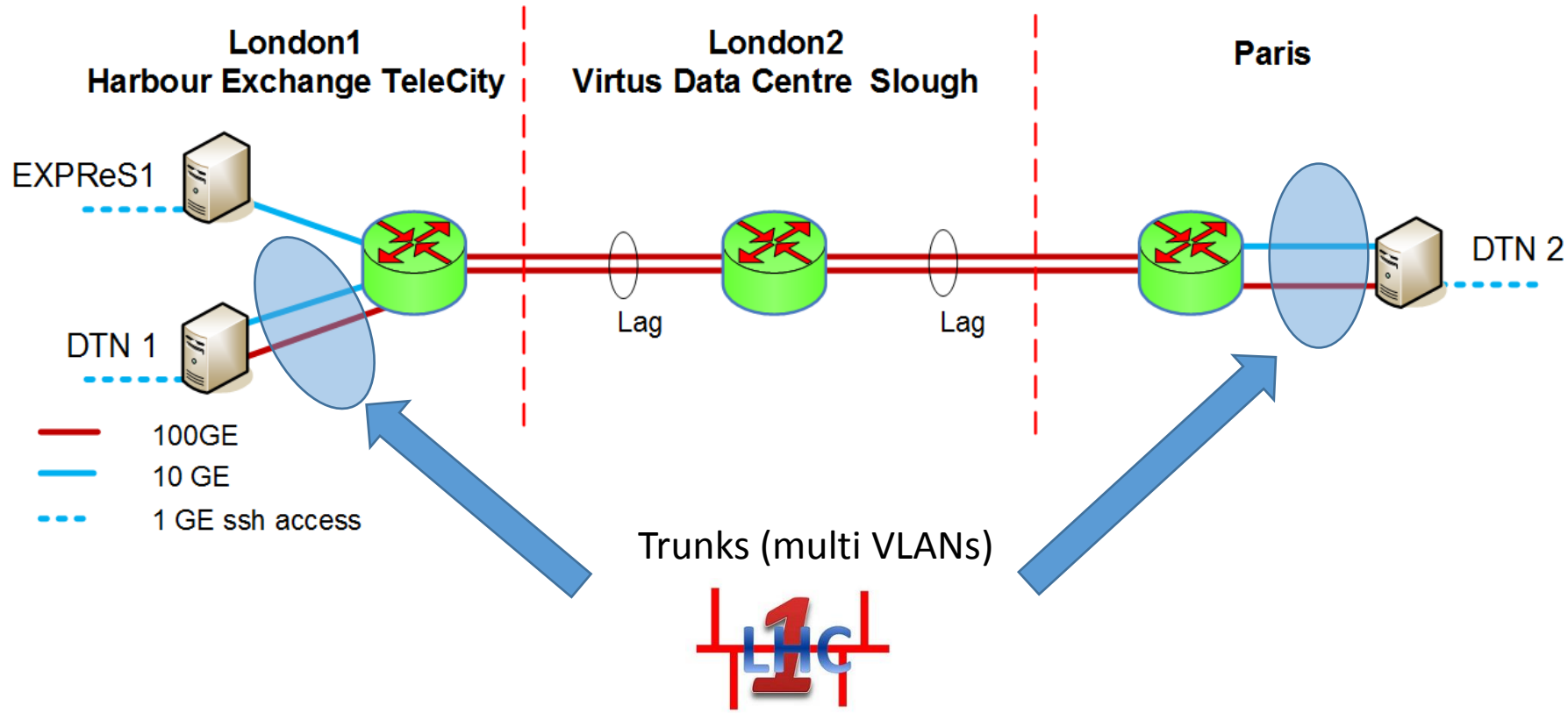







GÉANT IP Topology
20180302 rgh
200 Gbps
100 Gbps
60 Gbps
50 Gbps
40 Gbps
30 Gbps
20 Gbps
10 Gbps
Shared
GigE / 1 Gbps
STM-16; OC-48 2.4 Gbps
STM-4; OC-12 622 Mbps
STM-1; OC-3 155 Mbps
IP and Transmission PoP
IP only PoP
NREN
Internet Exchange
R&E Peer
Upstream
Dotted lines indicate logical connections of a true bandwidth capacity, not of a link aggregate group.
GÉANT AS20965/21320
LON1 UK
LON2 UK
PAR FR
LIS PT x2
MAD ES
GEN CH
AMS NL
BRU BE
FRA DE
HAM DE
DUB1 IE
DUB2 IE
MIL IT
MAR FR
VIE AT
ATH GR x2
ZAG HR
LJU SL
SOF BG
BUC RO
BUD HU
PRA CZ
POZ PL
BRA SK
RIG LV
TAL EE x2
KAU LT x2
SIN SG
TEIN3 AS23911
UBUNTUNET AS36944
NORDUnet AP1 AS2603
NORDUnet AP2 AS2603
Cynet AP2 AS3268
LINX Internet Exchange
INEX Internet Exchange
RedCLARA AS27750
IUCC AP1 AS378
SingAREN AS136968
BELnet AP1 AS2611
BELNET AP2 AS2611
JANET AP1 AS786
SURFnet AP2 AS1103
ESNET AS293
HEAnet AP1 AS1213
HEAnet AP2 AS1213
Internet2 AS11537
SINET AS2907
FCCN AP1 AS1930
FCCN AP2 AS1930
RENATER AP1 AS2200
RENATER AP2 AS2200
SWITCH/CERN AP2 AS559 / AS513
JANET AP2 AS786
RedIRIS AP1 AS766
RedIRIS AP2 AS766
TEIN3 AS24490
CIXP Internet Exchange
TIFR AS2697
GARR AP2 AS137
CERN AP1 AS513
SWITCH AP1 AS559
ARN AP1 AS3208
MIX-IT Internet Exchange
GARR AP1 AS137
DE-CIX Internet Exchange
UoM / Malta AP1 AS12046
RASH AS57961
KAUST AS50999
ENSTINet AP1 AS6879
TWAREN AS7539
SURFnet AP1 AS1103
UBUNTUNET AS36944
NORDUnet AP2 AS2603
KIAE AS59624
Internet2 AS11537
NetherLight
KREONET AS17579
AMS-IX Internet Exchange
CANARIE Backup AS6509
RESTENA AP1 AS2602
ASGC AS24167
CANARIE AS6509
SINET AS2907
ESNET AS293
Cogent AS 174
Telia AS1299
NISN AP1 AS297
CAREN AS 197118
DE-CIX Internet Exchange
RESTENA AP2 AS2602
DFN AP1 AS680
IUCC AP2 AS378
Ulakbim AP2 AS8517
NASRA AP AS47623
Cynet AP1 AS3268
PXM
Managed bandwidth
DFN AP2 AS680
PSNC AP2 AS8501
PSNC AP1 AS8501
Eenet AP1 AS3221
Eenet AP2 AS3221
SigmaNet AP1 AS5538
SigmaNet AP2 AS5538
LITnet AP1 AS2847
LITnet AP2 AS2847
BASNET AP2 AS21274
URAN AP1 AS12687
NIX.CZ Internet Exchange
CESnet AP1 AS2852
CESnet AP2 AS2852
GRENA AP AS20545
AMRES AP1 AS13092
Telia AS1299
Cogent AS 174
MREN AP1 AS40981
Ulakbim AP1 AS8517
HUNGARnet AP1 AS1955
ACOnet AP1 AS1853
ACOnet AP2 AS1853
VIX Internet Exchange
HUNGARnet AP2 AS1955
Sanet AP1 AS2607
ARNES AP1 AS2107
CARnet AP1 AS2108
ARNES AP2 AS2107
CARnet AP2 AS2108
RoEduNet AP1 AS2614
GRnet AP2 AS5408
GRnet AP1 AS5408
MARnet AP1 AS5379
MARnet AP2 AS5379
BREN AP1 AS6802
BREN AP2 AS6802

Network topology



Hardware description

- CPU: 2 x Intel Xeon E5-2643 v3
- M/B: Supermicro X10DRT (Intel C612 chipset)
- RAM: 128 GB DDR4 2133MHz ECC
- Storage: 6 x Intel DC P3700 400GB NVMe (8xPCI-e)
- NICs
  - Mellanox ConnectX-4 100GE 16 x PCI-e 3.0
  - Mellanox ConnectX-5 100GE 16 x PCI-e 3.0
  - Mellanox ConnectX-3 10/40GE 8 x PCI-e 3.0
  - Intel X540-AT2 10GE (integrated, for user access)
  - 1G IPMI (OOB)

Tools currently available

- Fast Data Transfer Service (FDT)
- GridFTP
- iperf (v2.0.8)
- iperf3 (v3.1.2)
- udpmon
- Network monitoring via MRTG/CACTI

FDT

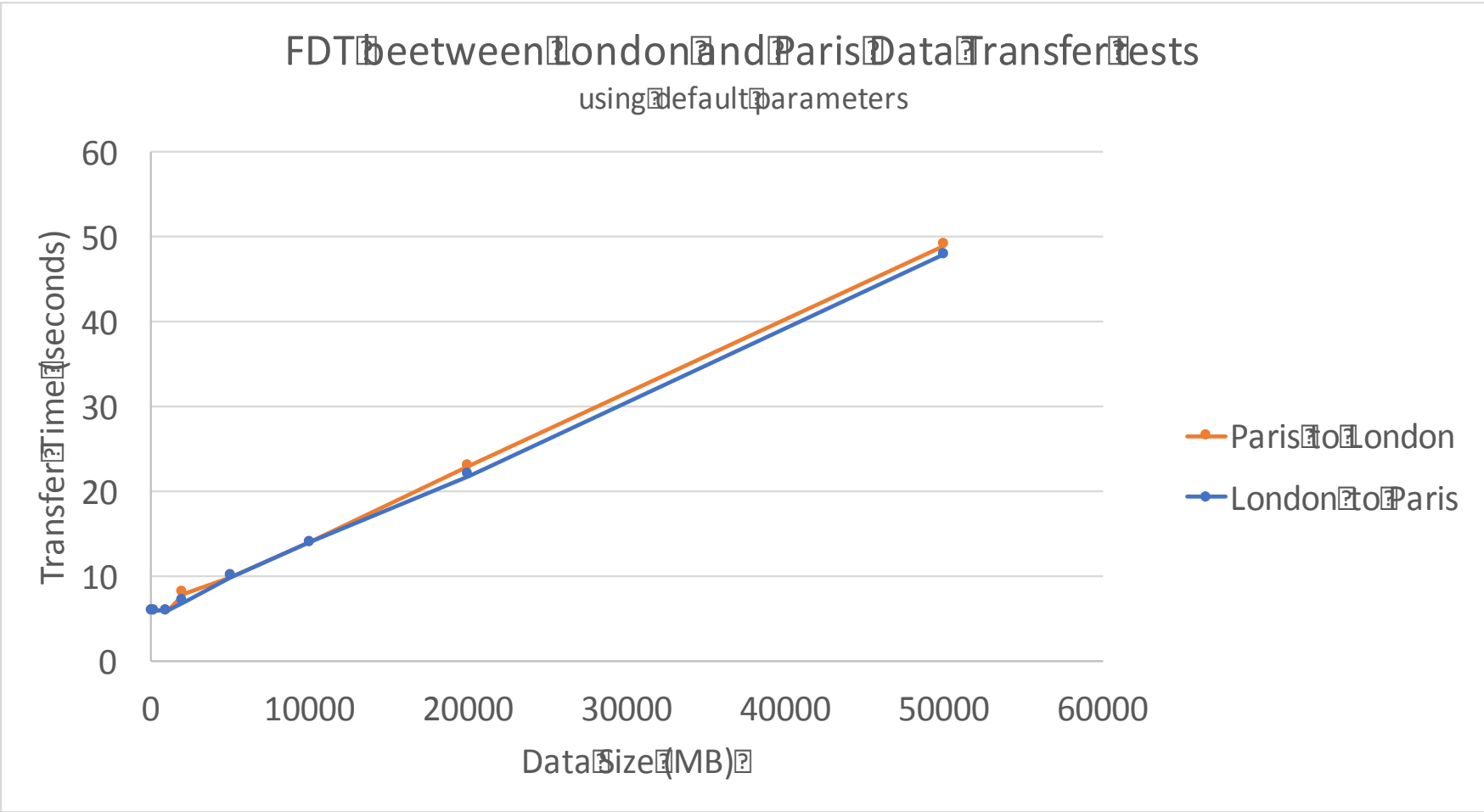
- http://monalisa.cern.ch/FDT
- Based on an asynchronous, flexible multithreaded system using the capabilities of the Java NIO libraries
- Controlled through a command line interface (CLI)
- Binaries are available for all major platforms and it is easy to use
- Does not require root privileges to be installed

FDT Main Features:

- Streams a dataset (list of files) continuously, using a managed pool of buffers through one or more TCP sockets.
- Uses independent threads to read and write on each physical device
- Transfers data in parallel on multiple TCP streams, when necessary
- Uses appropriate-sized buffers for disk I/O and for the network
- Restores the files from buffers asynchronously
- Resumes a file transfer session without loss, when needed

FDT tests between London and Paris

Disk-to-disk over 10Gb/s link



Number of Streams	Achieved Bandwidth (Gb/s)
1	9.311 Gb/s
2	9.889 Gb/s
3	9.89 Gb/s
4	9.894 Gb/s
10	9.897 Gb/s

Current service process

- Schedule tests (Google calendar)
  - **1 test at the time per node – no concurrency**
- Create local user
- Collect IP source address(es)
- Create ACLs
- Perform tests
- Delete ACLs
- Remove user and clean scratch space

Upcoming improvements

- WebDAV
- udpmon
- Storage optimisation for 100G tests
- Process automation
- AAI integration
- Web interface
- Filesender framework
- Additional nodes with different hardware

What would you like?

- Other tools/protocols?
  - mdtmFTP
  - UDP-based
  - webDAV **(planned)**
- Data distribution integration?
  - FTS
  - PHEDEX
  - RUCIO
  - DIRAC **(planned)**
- Hardware changes?

Help us to help you

- Participate in the testing
- Trial new tools
- Provide feedback

- Contact researchengagement@geant.org

GÉANT
Networks • Services • People

Thank you!

researchengagement@geant.org
vincenzo.capone@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).