Queen Mary
University of London

Lessons from Optimising data transfers at QMUL

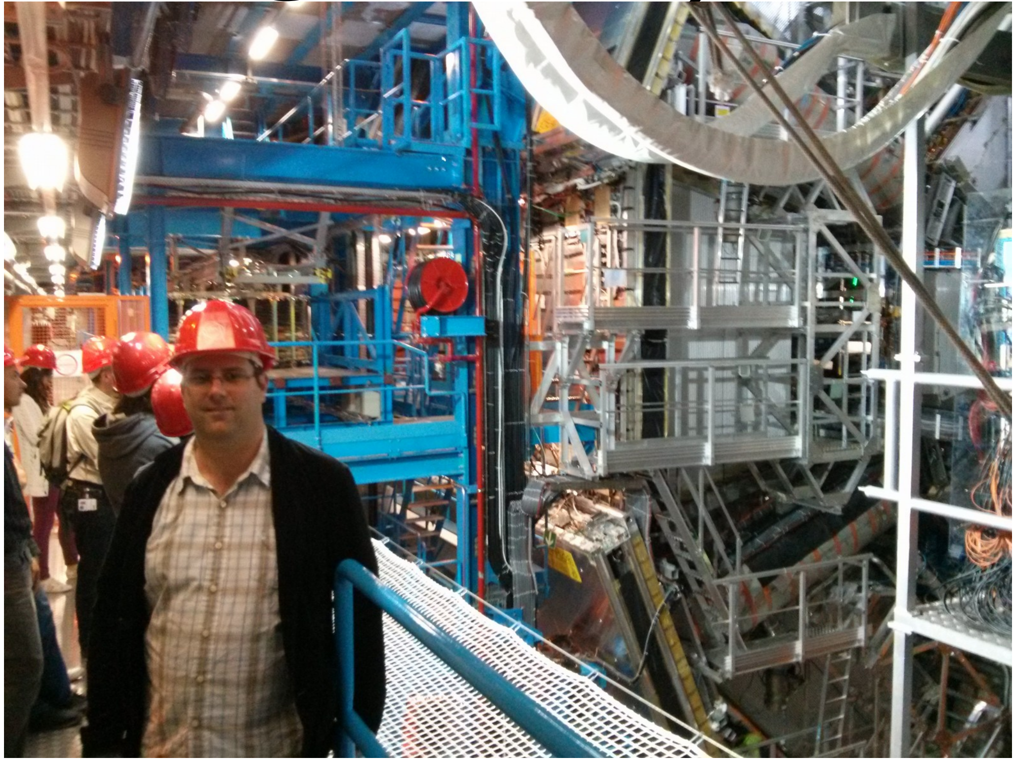
Christopher J. Walker
<C.J.Walker@qmul.ac.uk>

Overview

- What can be achieved
- TCP/IP explained
- Bottlenecks found
- Conclusions

Motivation (LHC@CERN)

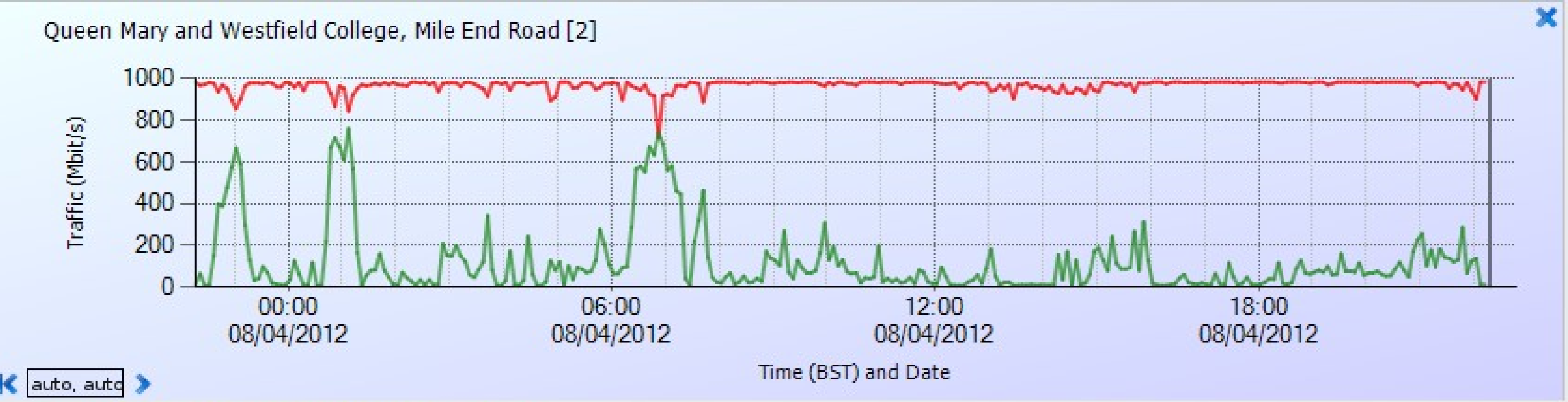
- Collisions 25ns
  - 100 PB/year
- QMUL
  - Small fraction



Network Expectations

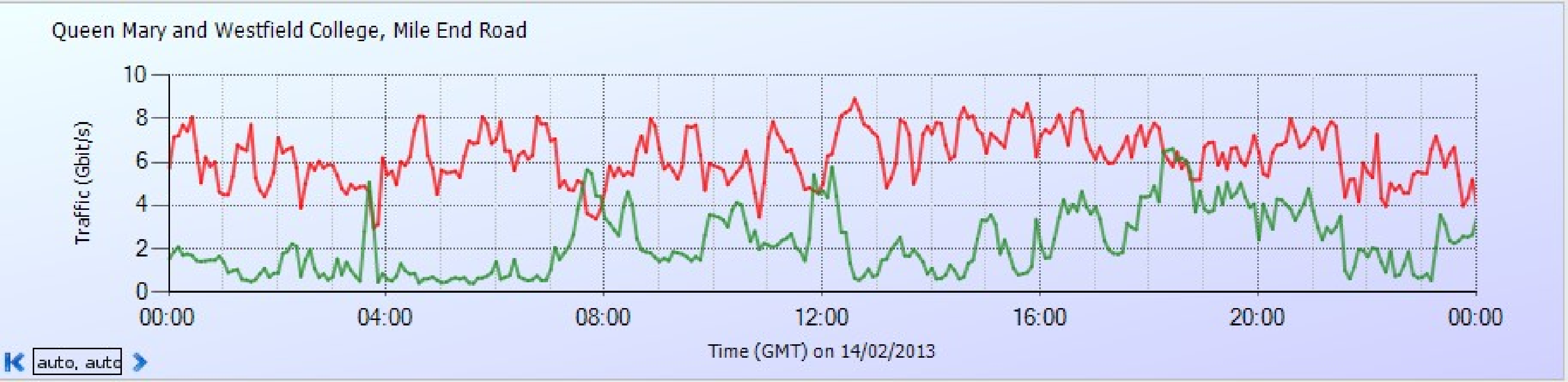
- 1 Terabyte can be transferred in:
  - 100 Mbps network : 30 hrs
  - 1 Gbps network : 3 hrs
  - 10 Gbps network : 20 minutes
- Takes work to achieve this in practice
  - TCP tuning
  - Find and eliminate bottlenecks
  - Reduce packet loss
- Fasterdata.es.net
  - Excellent source of information

WAN Performance



	To Site	From Site
Mean	969Mbit/s	121Mbit/s
Max	985Mbit/s	761Mbit/s
Min	716Mbit/s	9.62Mbit/s
Last	985Mbit/s	12.9Mbit/s

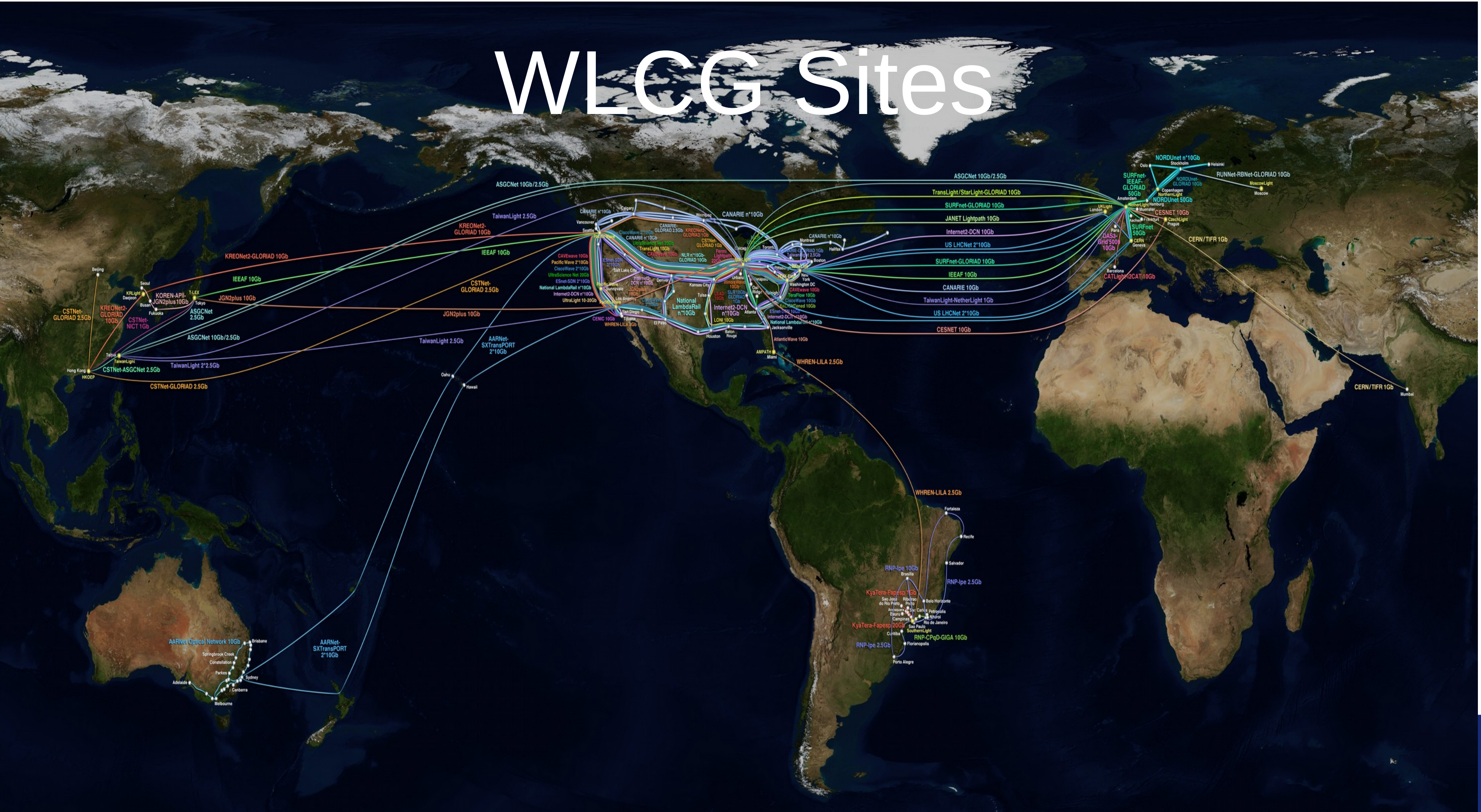
- April 2012:
  - 1 Gbit

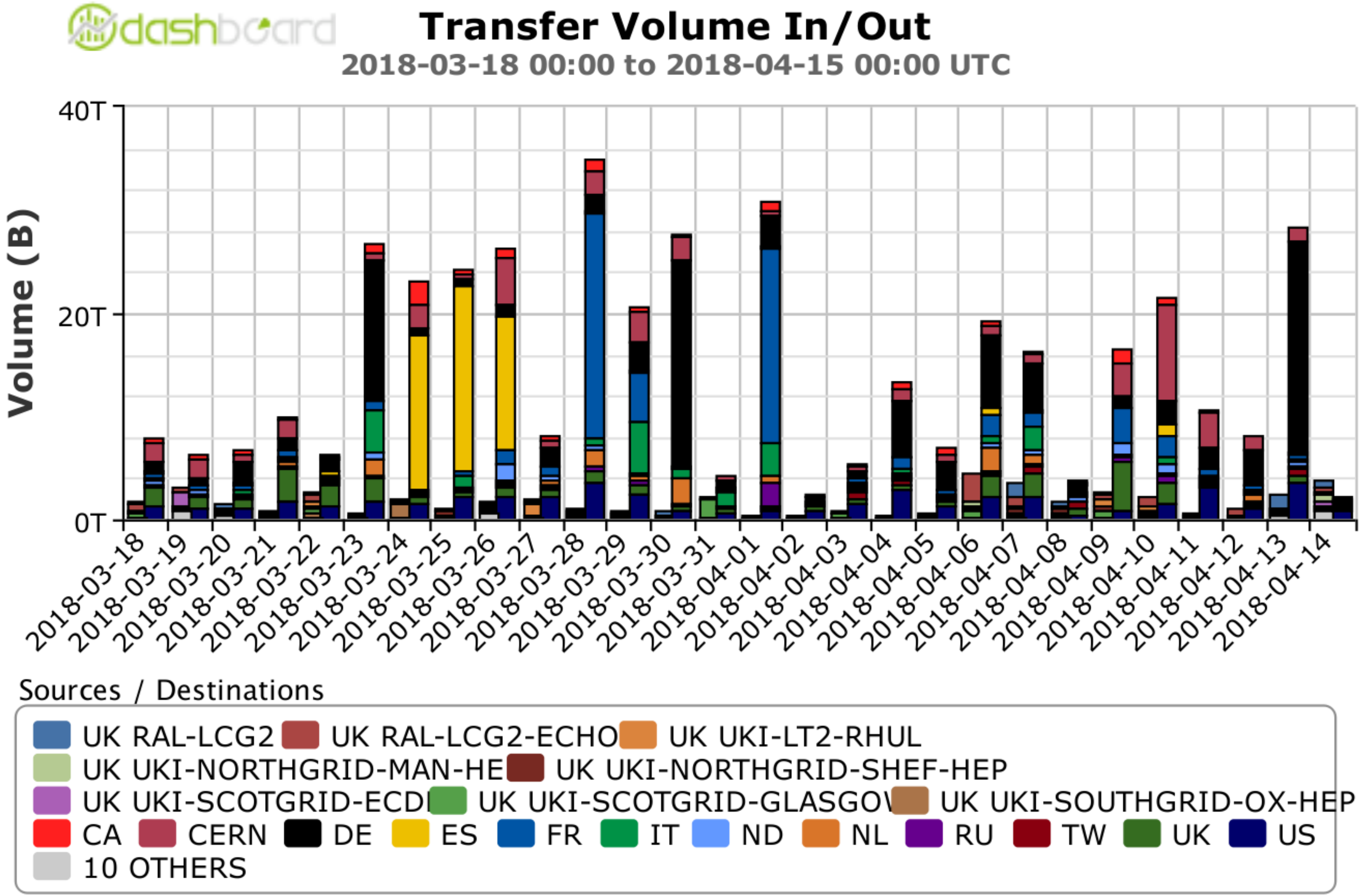


	To Site	From Site
Mean	6.27Gbit/s	2.26Gbit/s
Max	8.93Gbit/s	6.62Gbit/s
Min	2.96Gbit/s	400Mbit/s
Last	5.2Gbit/s	2.65Gbit/s

- Sept 2013
  - 10Gbit

WLCG Sites


ASGCNet 10Gb/2.5Gb
TaiwanLight 2.5Gb
KREONet2-GLORIAD 10Gb
KREONet2-GLORIAD 10Gb
IEEAF 10Gb
IEEAF 10Gb
CSTNet-GLORIAD 2.5Gb
JGN2plus 10Gb
JGN2plus 10Gb
Beijing
Seoul
KRLight
Daejeon
Busan
Fukuoka
KOREN-APII-JGN2plus 10Gb
T-LEX
Tokyo
ASGCNet 2.5Gb
CSTNet-GLORIAD 2.5Gb
KREONet2-GLORIAD 10Gb
CSTNet-NICT 1Gb
ASGCNet 10Gb/2.5Gb
TaiwanLight 2.5Gb
Taipei
TaiwanLight
Hong Kong
HKOEP
CSTNet-ASGCNet 2.5Gb
TaiwanLight 2*2.5Gb
CSTNet-GLORIAD 2.5Gb
AARNet-SXTransPORT 2*10Gb
Oahu
Hawaii
CANARIE n*10Gb
Calgary
Vancouver
Seattle
Pacific Wave
Winnipeg
CANARIE n*10Gb
CANARIE-GLORIAD 2.5Gb
KREONet2-GLORIAD 1Gb
CSTNet-GLORIAD 1Gb
CiscoWave 2*10Gb
CANARIE n*10Gb
UltraScience Net 20Gb
TransLight 10Gb
CAVEwave 10Gb
NLR n*10Gb-GLORIAD 10Gb
Fermi Lightpath 22Gb
CAVEwave 10Gb
Pacific Wave 2*10Gb
CiscoWave 2*10Gb
UltraScience Net 20Gb
ESnet-SDN 2*10Gb
National LambdaRail n*10Gb
Internet2-DCN n*10Gb
UltraLight 10-20Gb
ESnet-SDN 2*10Gb
Salt Lake City
Internet2-DCN n*10Gb
Denver
Pacific Wave
Sunnyvale
JGN2plus 10Gb
Los Angeles
Pacific Wave
CiscoWave 2*10Gb
San Diego
Tijuana
CENIC 10Gb
WHREN-LILA 2Gb
El Paso
Kansas City
Tulsa
National LambdaRail n*10Gb
LEAD 10Gb
SLR10Gb-GLORIAD 1Gb
Internet2-DCN n*10Gb
LONI 10Gb
Houston
Baton Rouge
Chicago
StarLight
Toronto
Urbana
IllinoisWave 10Gb
Indianapolis
Oak Ridge
Raleigh
Atlanta
Jacksonville
Montreal
CANARIE n*10Gb
Halifax
CANARIE-GLORIAD 1Gb
TaiwanLight 2.5Gb
Boston
Pittsburgh
MAN LAN
New York
NGIX-East
Washington DC
CAVEwave 10Gb
TeraFlow 10Gb
CiscoWave 10Gb
EnLIGHTened 10Gb
ESnet-SDN 10Gb
Internet2-DCN n*10Gb
National LambdaRail n*10Gb
AtlanticWave 10Gb
AMPATH
Miami
WHREN-LILA 2.5Gb
ASGCNet 10Gb/2.5Gb
TransLight/StarLight-GLORIAD 10Gb
SURFnet-GLORIAD 10Gb
JANET Lightpath 10Gb
Internet2-DCN 10Gb
US LHCNet 2*10Gb
SURFnet-GLORIAD 10Gb
IEEAF 10Gb
CANARIE 10Gb
TaiwanLight-NetherLight 1Gb
US LHCNet 2*10Gb
CESNET 10Gb
NORDUnet n*10Gb
Oslo
Stockholm
Helsinki
RUNNet-RBNet-GLORIAD 10Gb
MoscowLight
Moscow
SURFnet-IEEAF-GLORIAD 50Gb
NORDUnet-GLORIAD 10Gb
Copenhagen
NorthernLight
NORDUnet 50Gb
Amsterdam
NetherLight
Hamburg
Muenster
UKLight
London
CESNET 10Gb
Aachen
Frankfurt
CzechLight
Prague
Paris
SURFnet 50Gb
DAS3-Grid'5000 10Gb
CERN
Geneva
CERN/TIFR 1Gb
Barcelona
CATLight-i2CAT 10Gb
CERN/TIFR 1Gb
Mumbai
WHREN-LILA 2.5Gb
Fortaleza
Recife
Salvador
RNP-Ipe 10Gb
Brasilia
RNP-Ipe 2.5Gb
KyaTera-Fapesp 1Gb
Sao Jose do Rio Preto
Ribeirao Preto
Belo Horizonte
Araraquara
Sao Carlos
Petropolis
Bauru
Niteroi
Campinas
Rio de Janeiro
KyaTera-Fapesp 20Gb
Sao Paulo
SouthernLight
Curitiba
RNP-CPqD-GIGA 10Gb
RNP-Ipe 2.5Gb
Florianopolis
Porto Alegre
AARNet Optical Network 10Gb
Brisbane
Springbrook Creek
Constellation
Parkes
Sydney
Adelaide
Canberra
Melbourne
AARNet-SXTransPORT 2*10Gb

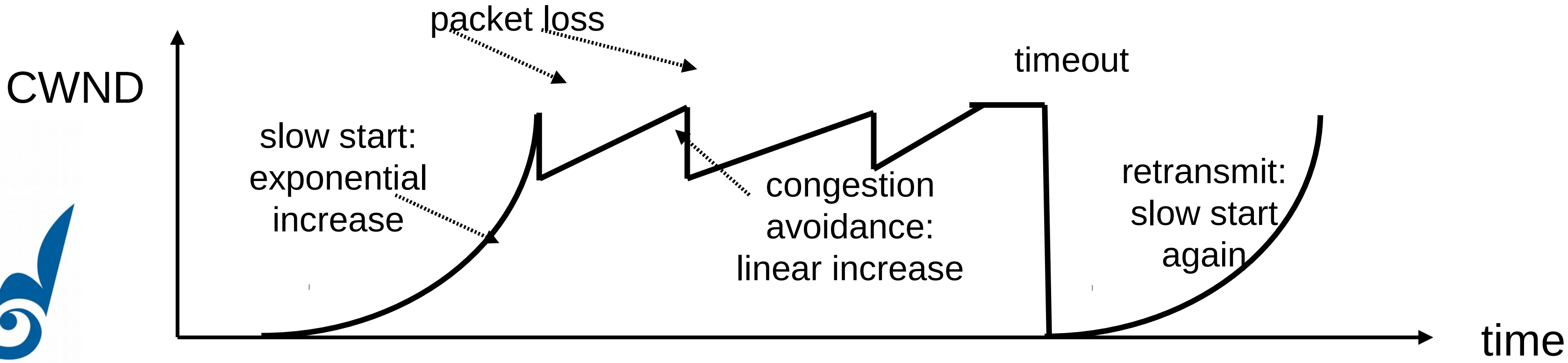

dashboard
Transfer Volume In/Out
2018-03-18 00:00 to 2018-04-15 00:00 UTC
Volume (B)
40T
20T
0T
2018-03-18
2018-03-19
2018-03-20
2018-03-21
2018-03-22
2018-03-23
2018-03-24
2018-03-25
2018-03-26
2018-03-27
2018-03-28
2018-03-29
2018-03-30
2018-03-31
2018-04-01
2018-04-02
2018-04-03
2018-04-04
2018-04-05
2018-04-06
2018-04-07
2018-04-08
2018-04-09
2018-04-10
2018-04-11
2018-04-12
2018-04-13
2018-04-14
Sources / Destinations
UK RAL-LCG2
UK RAL-LCG2-ECHO
UK UKI-LT2-RHUL
UK UKI-NORTHGRID-MAN-HE
UK UKI-NORTHGRID-SHEF-HEP
UK UKI-SCOTGRID-ECDI
UK UKI-SCOTGRID-GLASGOV
UK UKI-SOUTHGRID-OX-HEP
CA
CERN
DE
ES
FR
IT
ND
NL
RU
TW
UK
US
10 OTHERS

scp or ssh+rsync: good enough?

- 100MB testfile copy to QMUL HPC in Slough
  - 4G mobile: 32s
  - Southampton eduroam: 13s
  - 100Mbit desktop: 10s
  - QMUL physics (1Gbit): 2.5s

TCP: A short overview

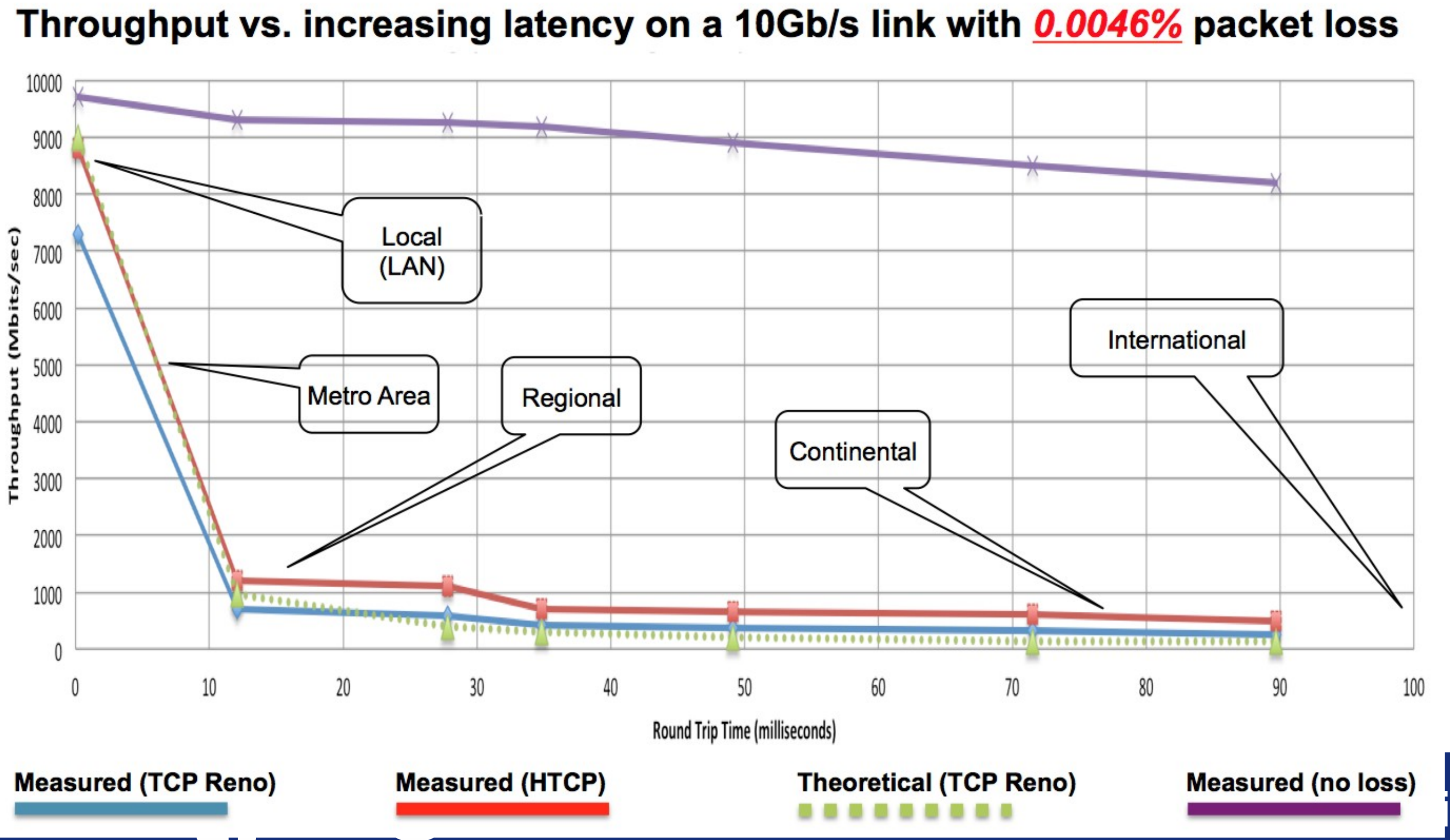
- Congestion window (CWND) = the number of packets the sender is allowed to send
  - The larger the window size, the higher the throughput
  - Throughput = Window size / Round-trip Time
- TCP Slow start
  - exponentially increase the congestion window size until a packet is lost
  - this gets a rough estimate of the optimal congestion window size



TCP Tuning

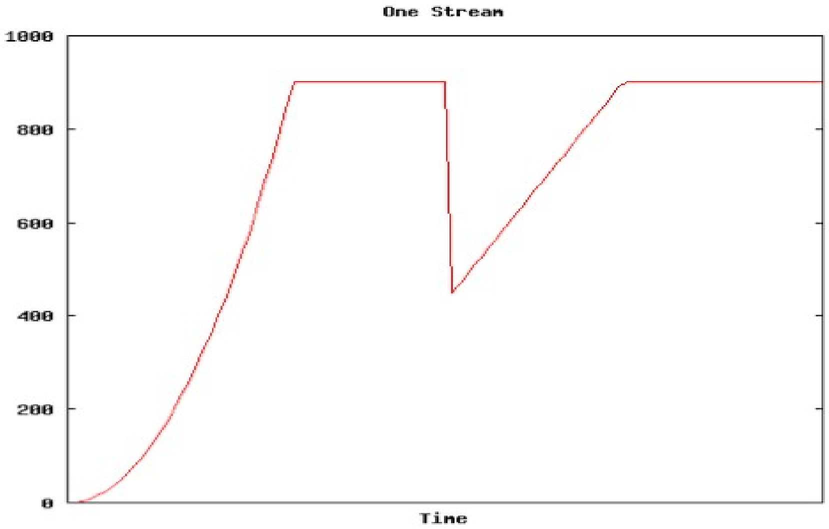
- Latency: time to send 1 packet from the source to the destination
- RTT: Round-trip time
- Bandwidth*Delay = Bandwidth Delay Product
  - The number of bytes in flight to fill the entire path
  - Example: 10 Gbps path; ping shows a 90 ms RTT (QMUL->BNL)
    - BDP = 10 * 0.090 = 0.9 Gbits (112 MBytes)
  - QMUL ->Taiwan 273ms RTT (at 10Gbps path)
    - BDP = 10*0.273 = 2.73 Gbits (340 MBytes)

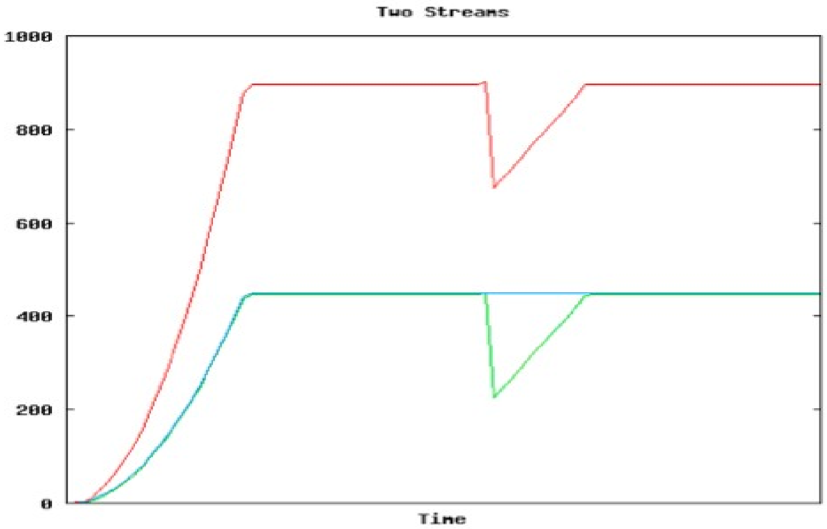
Effect of Packet loss with distance



Multiple streams

- Parallel streams can help
  - Potentially unfair on other users





TCP lessons

- Increase TCP buffers for distant transfers
  - Fasterdata.es.net has good recommendations
- Packet loss needs eliminating
- Application
  - large buffers (not scp)
  - Multiple streams
  - GridFTP has these
- Aspera uses UDP (and GridFTP can)
- Fasterdata.es.net has excellent recommendations

Slough backup – case study

- 10Gbit (*2) link QMUL ↔ Slough
  - 1.5ms RTT
  - 1GB/s expected transfer rate
    - 40MB/s achieved initially

TSM's TCPWindowSize

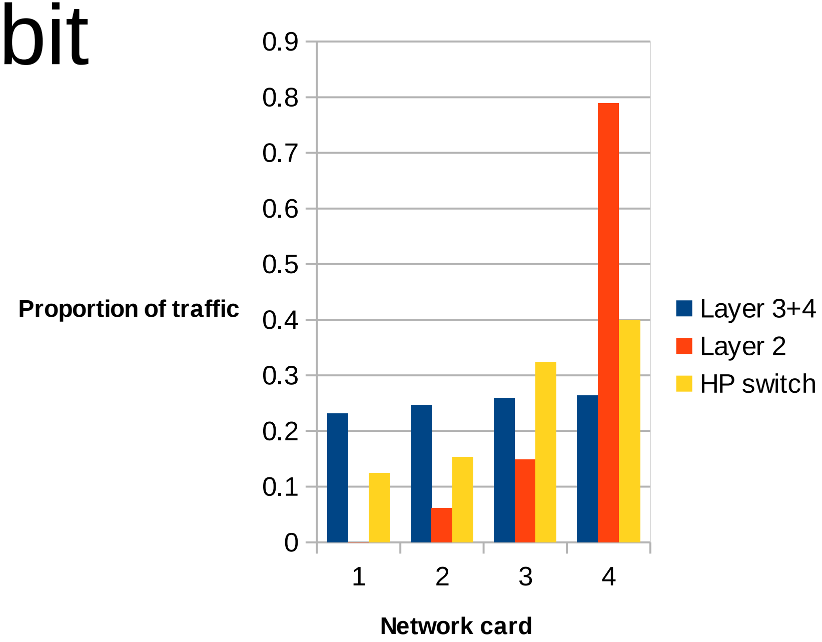
- TCPWindowSize 63 #TSM default
  - 39,389.44 KB/sec
- TCPWindowsize 10 # Reduced for testing
  - 22,591.03 KB/sec
- TCPWindowSize 0 # Use Linux default
  - 129,721.81 KB/sec

Packet loss

- Iperf measurements from TSM to:
  - Backup1 (10.x.y.101): 9.6GBit/s
  - Backup2 (10.x.y.102): 0.6 Gbit/s
    - Adjacent IPs
      - Packets take different legs of bond – packet loss on one of them
- Packet loss: 0.015%
  - http://fasterdata.es.net/network-tuning/tcp-issues-explained/packet-loss/

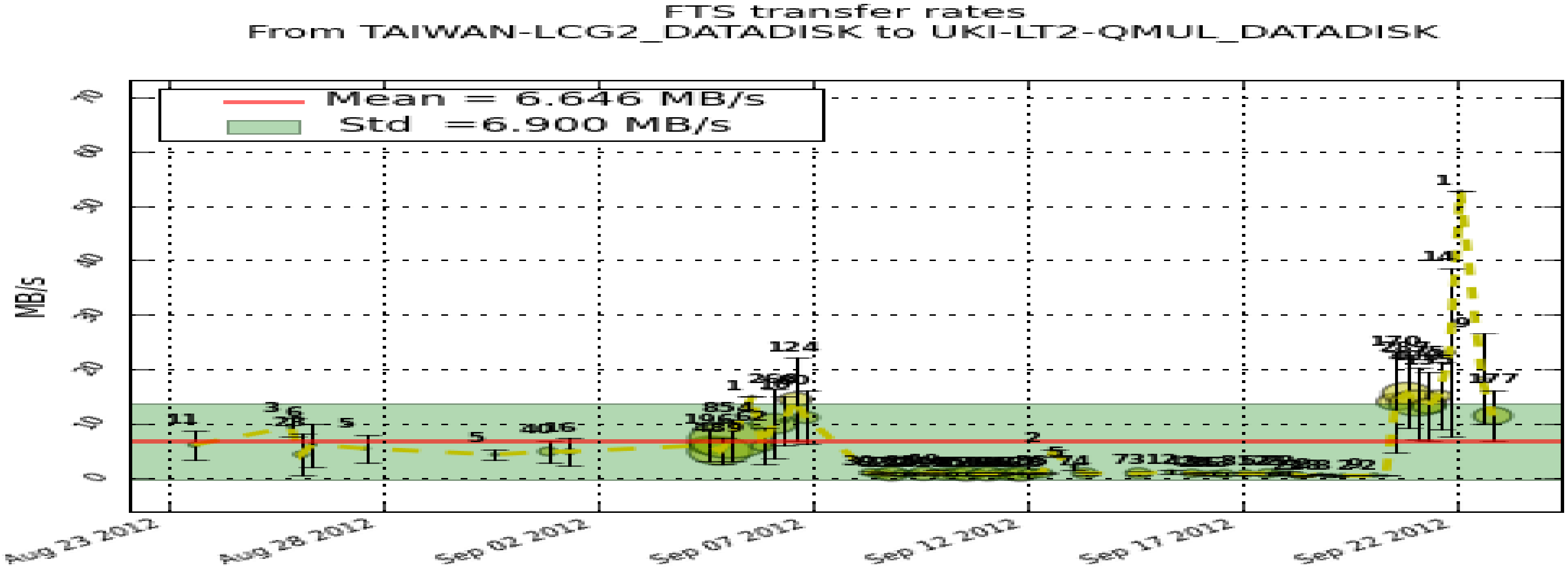
Bottlenecks found

- Gbit connected at 100Mbit
  - GridFTP node
  - Dept
  - College
  - NOC ↔ Archer
- Routing:
  - QMUL ↔ CERN via USA



Routing problems with 10Gbit/s upgrade

- 8th September 10Gbit/s WAN upgrade
- UK sites – increased rates
- ASGC (Taiwan) decrease
  - Route not advertised via GEANT.

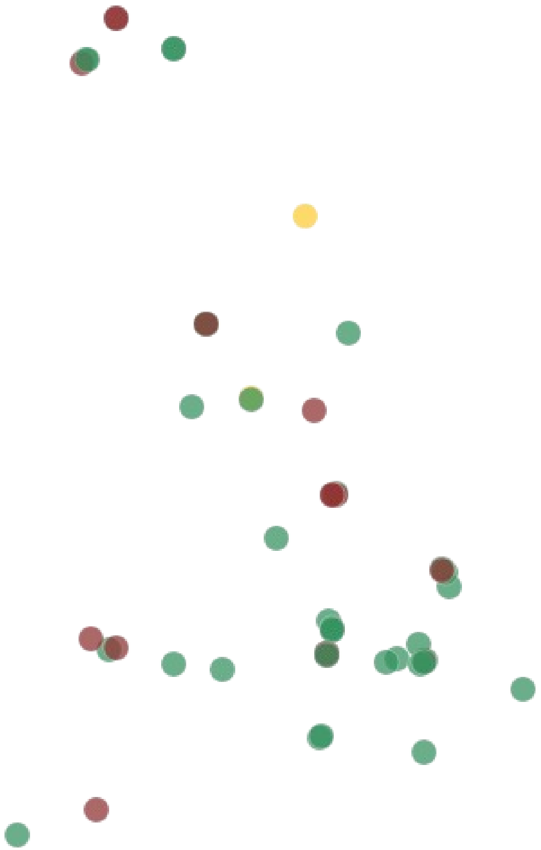


Firewalls

- ICMP
  - Often blocked
    - timeout rather than failure
- IPv6
  - Tracepath6 blocked (ICMP blocking)
- Barclays bank blocked
  - Deep packet inspection and rewriting of http packets (but not https)
- Scp failing half way through transfer
- GridFTP Slow performance
  - 1 MB/s through firewall, 50MB/s avoiding firewall
- GridFTP control connection forgotten

QMUL plans

- Slough Firewall upgrade
  - 10Gig module
- HPC Data transfer node
- Improve monitoring
  - Perfsonar
  - RIPE atlas



Conclusions

- Large transfers routine
  - Need to transfer multiple files at once and deal with failures
  - Takes work (GridPP sites have this experience)
- Many “small” transfers in the UK can work with rsync/ssh
- Monitoring vital
  - Transfers
  - Network
- Network
  - Need Low packet loss (most problems in the last mile)
  - Good relationship with network team useful
- Information
  - Fasterdata.es.net

Acknowledgements

- Fasterdata.es.net (Brian Tierney)
  - Many thanks for the TCP tuning slides

Backup slides

IPv6

- Routes
  - May be different to IPv4
    - Geneva ->QMUL via New York (fixed)
- Software (IPv6) / ASIC (IPv4)
  - Older routers may give poor performance
- Preferred over IPv4
  - If IPv6 address (AAAA record) in DNS, it will be used by machines that think they are IPv6 connected.
- Blocked differently by firewalls