Jisc

30/09/2016

Challenges in end-to-end network performance

Dr. Tim Chown, Jisc, UK tim.chown@jisc.ac.uk

Overview

- In this talk we look at issues around achieving optimal end-to-end network performance
  - Framed in the context of Janet, as a National Research and Education Network (NREN), operated by Jisc
  - The issues are likely to be ones common with other European and international NRENs and their connected campuses / sites
  - Experiences presented here are drawn from our work to date on the Janet End-to-End Performance Initiative (E2EPI)

- Lots of good work has already done; we're building on that, e.g.,
  - GridPP particle physics community (to handle LHC data)
  - International organisations such as ESnet (see fasterdata.es.net)
  - Data Transfer Zone deployment by CEDA at RAL

Emerging data-intensive science

- Why is this topic becoming more important?
- There are many new use cases emerging in the field of "data-intensive science"
  - Increasing requirement on the network to transfer larger volumes of data to/from compute facility or to/from storage / archive
- Existing research fields, e.g., astrophysics, particle physics, genomics, …
- Plus new types of networked scientific equipment
  - e.g., electron microscopy; where there may be no local compute
  - One such site is seeking to push 50GB data to remote compute with a 30 second turnaround for visualisation, which implies throughout of > 10Gbit/s
- The compute may be an HPC facility on an NREN network, or a commercial cloud facility such as Amazon
- Seeing increased interest in research communities in exploiting remote compute

Janet end-to-end performance initiative

- The Janet e2e performance initiative is:
  - Creating dialogue between Jisc, Janet-connected campus / site computing service groups, and research communities
  - Pro-active in engaging with existing data-intensive research communities and identifying emerging communities
  - Holding workshops, facilitating discussion on e-mail lists, etc.
  - Helping researchers manage expectations
  - Establishing and sharing best practices in identifying and rectifying causes of poor performance
  - Includes consideration of low-latency applications, such as LOLA

- More information:
  - https://www.jisc.ac.uk/rd/projects/janet-end-to-end-performance-initiative

End-to-end performance factors

- Achieving optimal end-to-end performance is a multi-faceted, nuanced problem. It includes:
  - Appropriate provisioning between the end sites by the NRENs and other ISPs
  - Properties of the local campus network (at each end), including capacity of the NREN connectivity, LAN design, and the performance of firewalls and configuration of other devices on the path
  - End system configuration and tuning; network stack buffer sizes, disk I/O, memory management, etc.
  - The choice of tools used to transfer data, and the underlying network protocols

- NB. It's not practical to expect researchers to understand these issues in detail, but a broad understanding is helpful towards managing expectations

NREN and campus provisioning

- Capacity provision and management applies to both the NREN and the campus networks

- 1: How the NREN, such as Janet, provisions capacity on the backbone between sites to ensure latent capacity for existing and emerging applications

- 2: How connected organisations / campuses make optimal use of their connectivity
  - Capacity of the link to the NREN
  - How competing "day-to-day" and data-intensive traffic is handled within the campus network and at its border

The NREN perspective

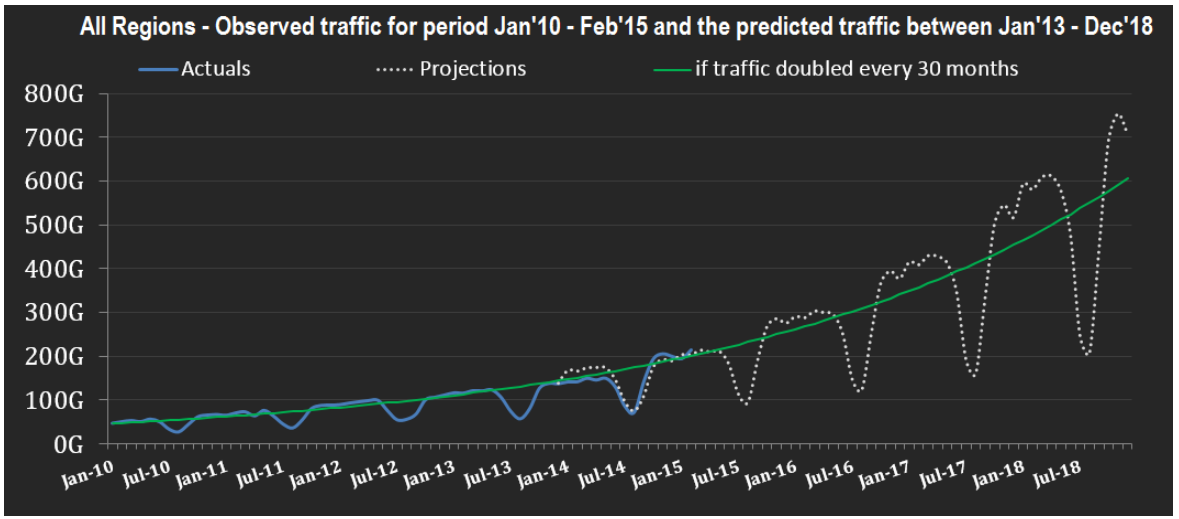
Capacity planning

- From the NREN perspective, it's important to ensure there is sufficient capacity in the network for connected sites
- Ongoing review of utilisation
  - Observe the utilisation, and model growth
  - Provision the core / backbone network
  - Provision external connectivity, to other NRENs and networks

- Janet has no differential queueing for regular IP traffic
  - The Netpath service exists for dedicated / overlay links
  - In general, Jisc plans regular network upgrades with a view to ensuring sufficient latent capacity in the backbone
- Other NRENs may provide differential QoS services; approaches may vary

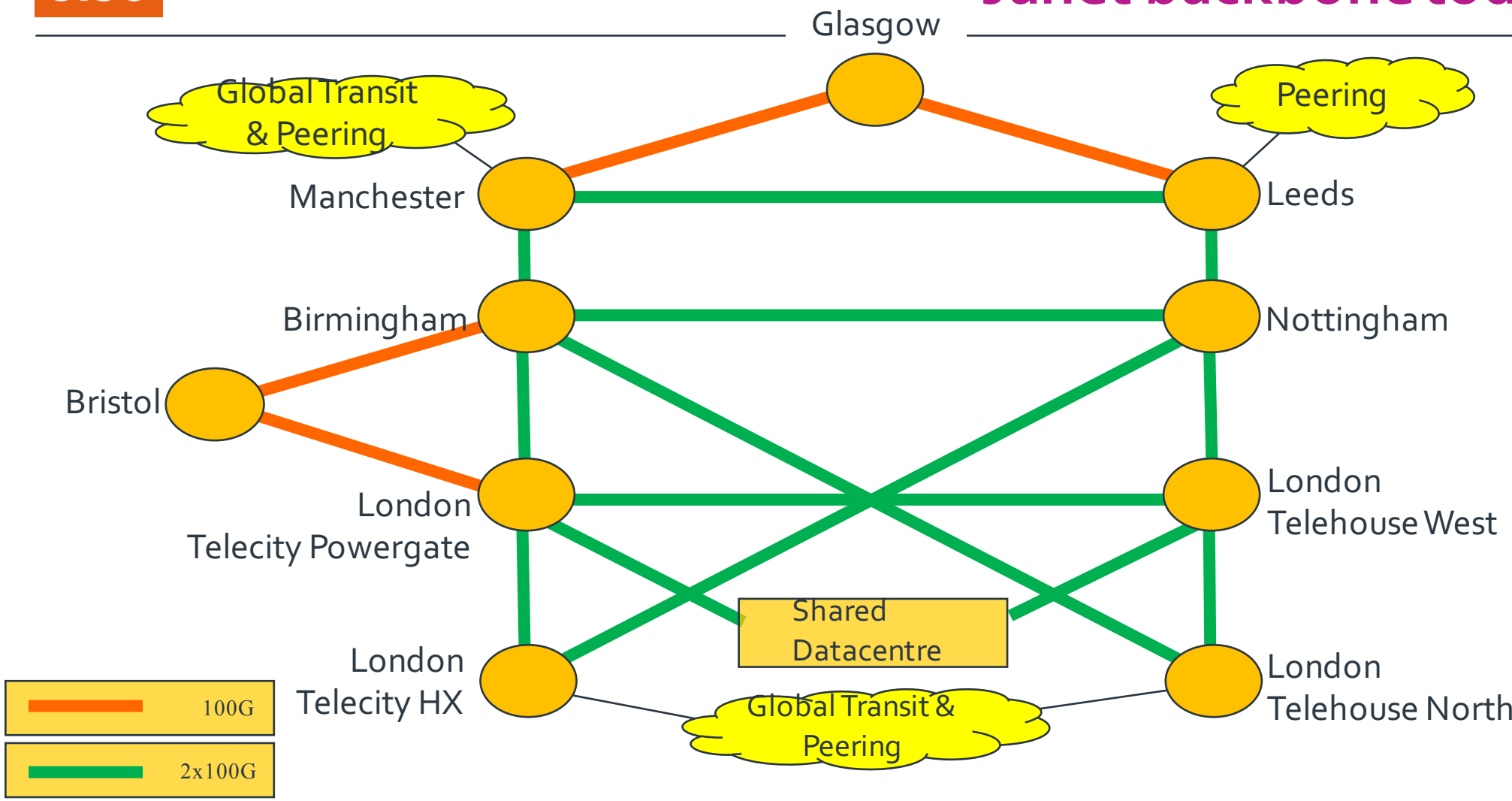
Janet aggregate bandwidth utilisation

- Various modelling done by Jisc
  - Recent trend is that traffic is doubling every 2.5 years
- Desirable to understand the drivers for growth

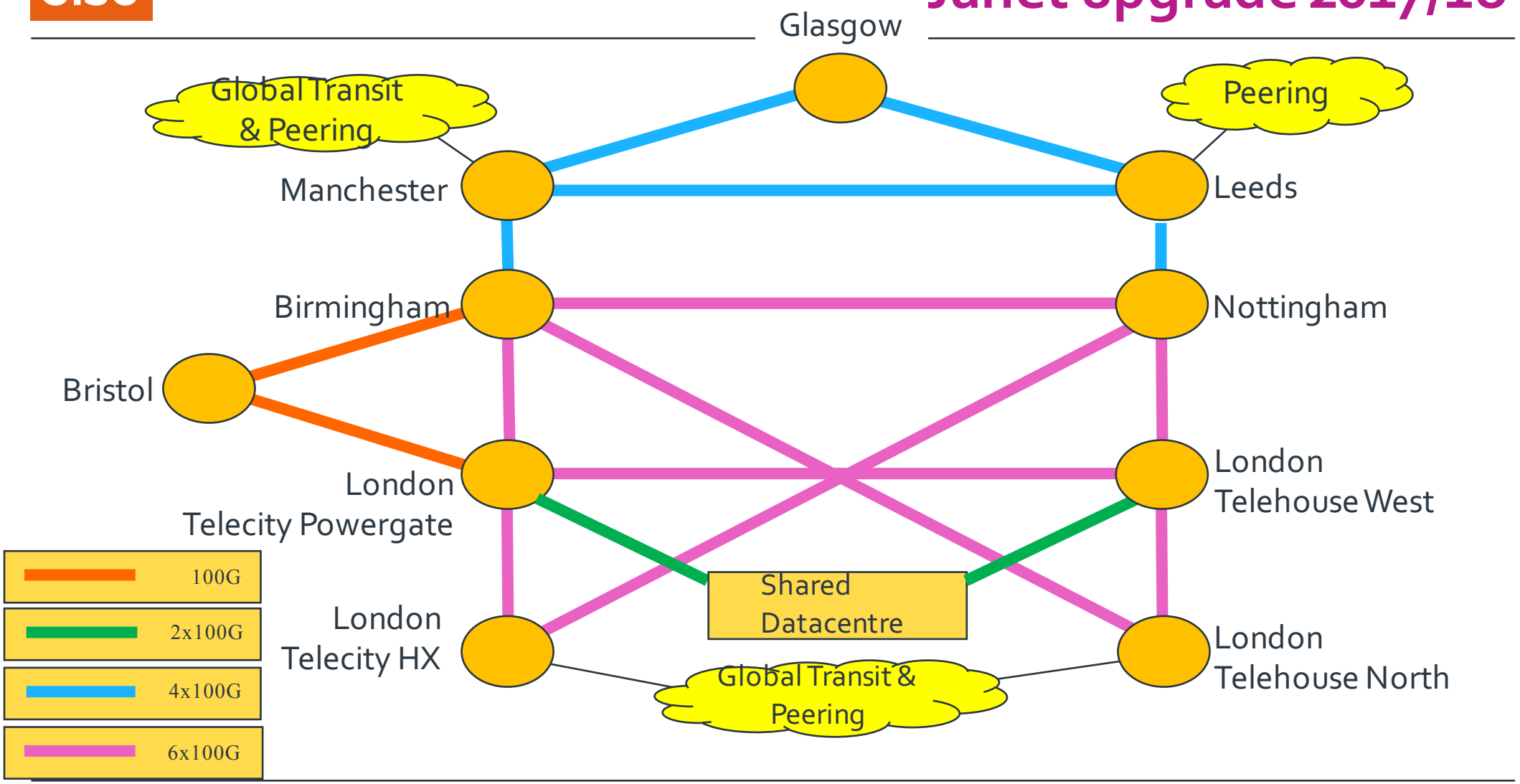


- We expect that other NRENs see a similar pattern?

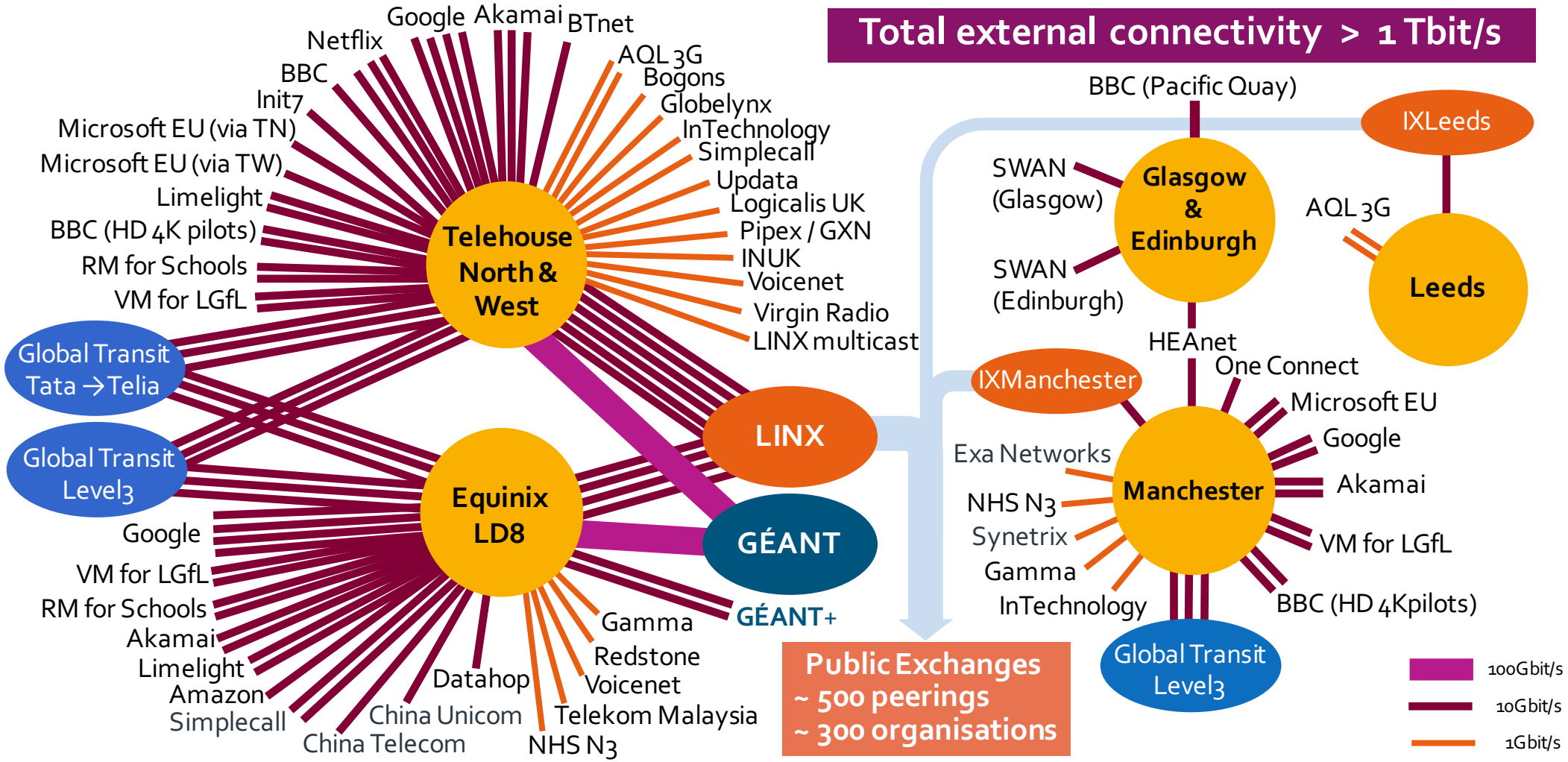
Janet backbone today



Janet upgrade 2017/18



Janet external connectivity



Use of IP network vs overlays?

- Ideally, an NREN will want to minimise the complexity in operating its network and network services
- Janet offers and supports various overlay services
  - Netpath – for L2VPN or dedicated paths
    - See https://www.jisc.ac.uk/netpath
  - Community overlays, e.g., LHCONE, which is used by the GridPP community – see http://lhcone.web.cern.ch/
  - It also includes optical infrastructure for network research communities
- If all emerging data-intensive communities were to each be supported by LHCONE-style overlays, operational complexity would increase
- It is thus desirable to encourage such communities to at least first try to use the Janet IP network

The campus / site perspective

Campus connectivity to Janet

- Janet-connected sites will typically be connected at 10Gbit/s, but the specific capacity will vary site to site
  - Utilisation is monitored; discussions on capacity upgrades will typically be triggered by observed organic growth
  - It is harder to anticipate step changes in site requirements caused by a specific new data-intensive use case; these can surface at relatively short notice
  - A new data-intensive networking requirement may be significant compared to the overall campus Janet connection capacity
- Sites may have resilient links
  - Intended to only be used when the primary site link fails

Handling step changes in requirements

- Ideally a university computing service will perform regular "future looks" of network requirements
  - Implies close collaboration with researchers
  - In practice this is quite hard to do
  - May best be driven at a senior level; e.g., PVC or CIO
- Communities may also do this, e.g. GridPP where the scale of the LHC experiments is planned well in advance; this can help inform NREN capacity planning

- Close co-operation between researchers, their campus computing services, and their NREN is highly desirable
  - Share experiences, best practices, knowledge of what's happening across the community

A quart in a pint pot?

- At some point, a site's NREN connectivity requirement will outgrow its current capacity
  - Ideally, before that happens, the site has a conversation with the NREN about upgrading its capacity
  - The specifics of how this happens may vary
  - There may be different cost models within different NRENs

- But we should in principle want to avoid sites:
  - Rate-limiting their researchers (except as a short-term measure)
  - Using their resilient link for bulk scientific data
  - There are examples of both practices on Janet

End-to-End Performance Issues

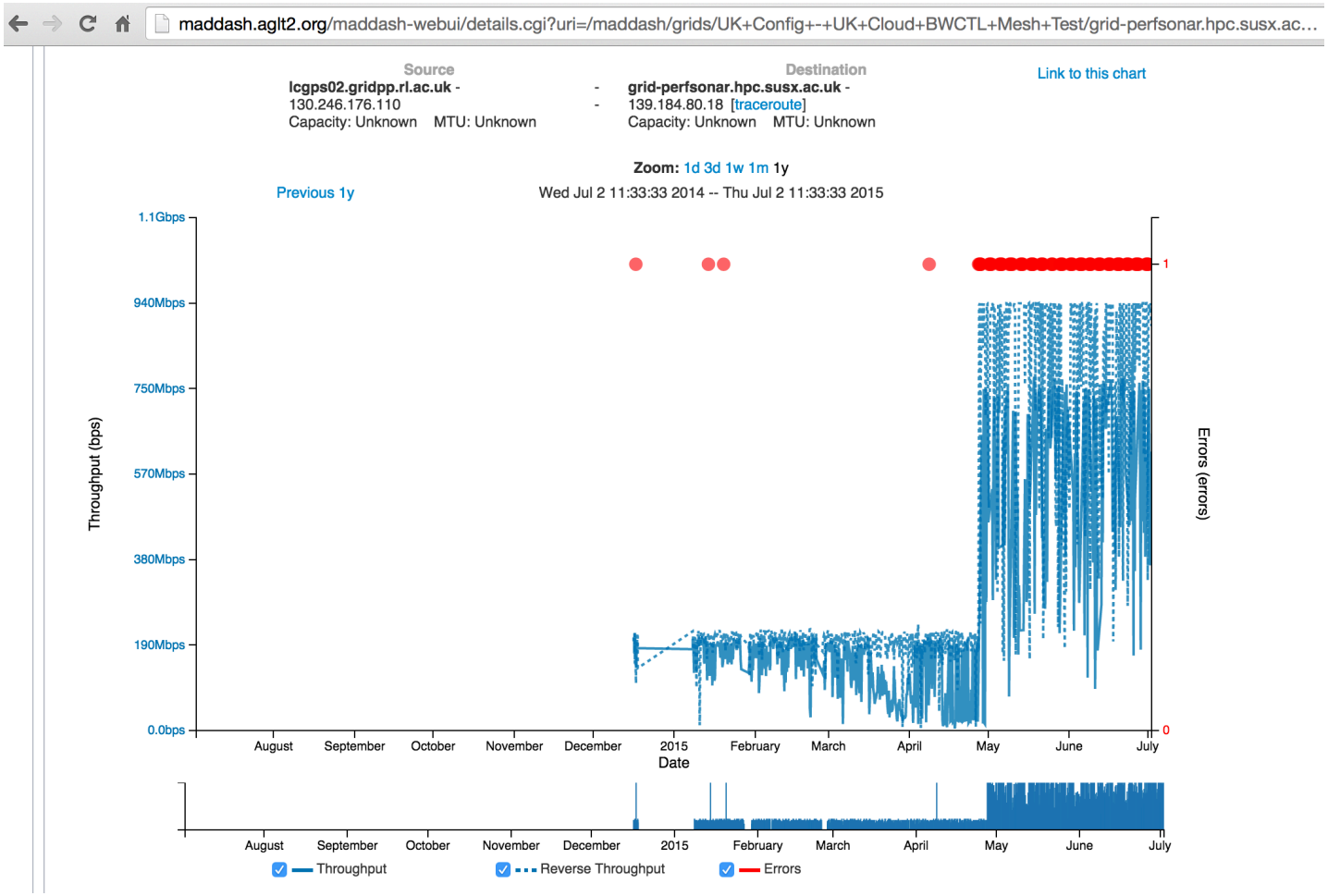
Common e2e issues encountered

- So what type of end-to-end issues commonly arise?
- From our experience on the Janet E2EPI:
  - Limitations in campus firewalls; their internal architecture may not support high throughput flows
  - Choice of transfer tools; researchers expecting / hoping to get good performance from standard Unix tools like *scp* or *ftp*
  - TCP buffer size tuning for higher RTT transfers
  - Issues generally within the campuses, not on the NREN network
- Many issues are recorded in the eduPERT knowledge base
  - More on this in a moment…

Firewall issues

- A stateful campus firewall may be inspecting many tens of thousands of concurrent flows, applying filtering policy, performing intrusion and DoS detection on those flows
  - Such devices are not necessarily designed for high throughput flows
- Recent examples
  - Sussex: 200Mbit/s observed at one HPC site; achieved full 1Gbit/s on path once firewall replaced
  - Durham: limited to 300-400Mbit/s for DiRAC; achieved 3-4Gbit/s once path engineered around the firewall
- Implies it may be prudent to design campus network architectures to avoid high throughut flows passing through generic campus firewalls
  - May still apply security policy, but more efficiently via bespoke ACLs

Sussex HPC firewall example



Measuring network performance

- To diagnose or understand poor network performance, having telemetry on your network is very important
  - Without this, you are blind to potential problems
- The GridPP community have homed in on perfSONAR as their tool of choice
  - Open source package; requires dedicated hardware
  - Built upon a suite of established tools, including *iperf*
  - Allows creation of multi-site dashboard views, with at-a-glance indications of problems via views of throughput, latency and packet loss
  - See http://www.perfsonar.net/
- Requires some technical knowledge to install, but it is worth the effort
  - ESnet guidance, including training videos, can be found at: https://fasterdata.es.net/performance-testing/perfsonar/
- Much more on this in Duncan Rand's talk later in this session

Small node perfSONAR

- The Small Node perfSONAR project offers perfSONAR with 1Gbit/s throughput test capability on devices costing under €200
  - Co-ordinated by Antoine Delvaux and Szymon Trocha at Poznan
  - Current platform being tested is Gigabyte Brix
    - IPv4 and IPv6 test mesh at http://perfsonar-smallnodes.geant.org/maddash-webui/
  - Janet E2EPI plans to build 10-20 such devices to offer to communities for tests; the aim is to make them as "plug and play" as possible; results should be at least indicative of fuller perfSONAR devices
  - Also a stepping stone to a full perfSONAR node

- Further information and TNC2016 slide deck:
  - https://lists.geant.org/sympa/d_read/perfsonar-smallnodes/

perfSONAR small node test mesh

perfsonar-smallnodes.geant.org/maddash-webui/

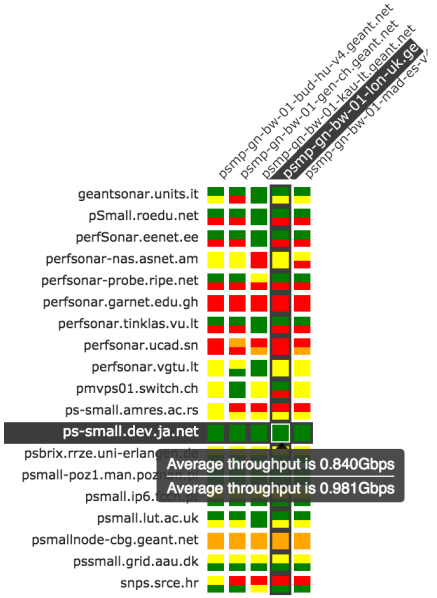
pSmall GÉANT project

Dashboards | Settings | External Resources

pSmall-GEANT Dashboard

pSmall-GEANT - IPv4 throughput

Throughput >= 800Mbps | Throughput < 800Mbps | Throughput <= 500Mbps | Unable to retrieve data | Check has not yet run



perfsonar-smallnodes.geant.org/maddash-webui/details.cgi?uri=/maddash/grids/pSmall-GEANT+--+IPv4+throughput/ps-sma

ps-small.dev.ja.net to psmp-gn-bw-01-lon-uk.geant.net (Throughput)

Status: OK **Last Checked:** September 22, 2016 15:24:17 PM BST **Next Check:** September 22, 2016 23:24:17 PM BST

Summary | History | Check Details

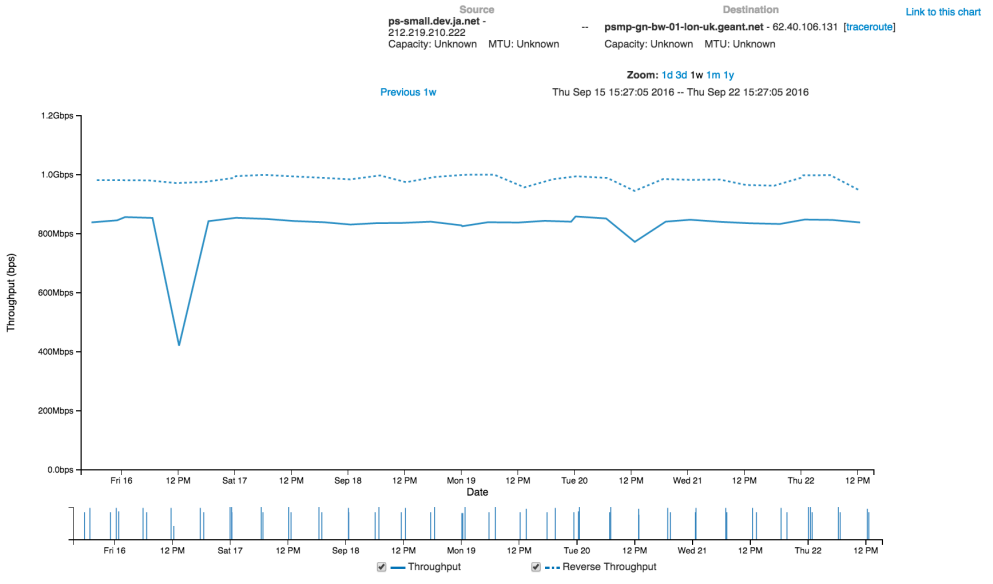
Current Results

Current Status: OK
Result of last check: OK
Message For Current Status: Average throughput is 0.840Gbps
Events:

Name	Description	Start	End
No events currently scheduled.			

Statistics

Graph



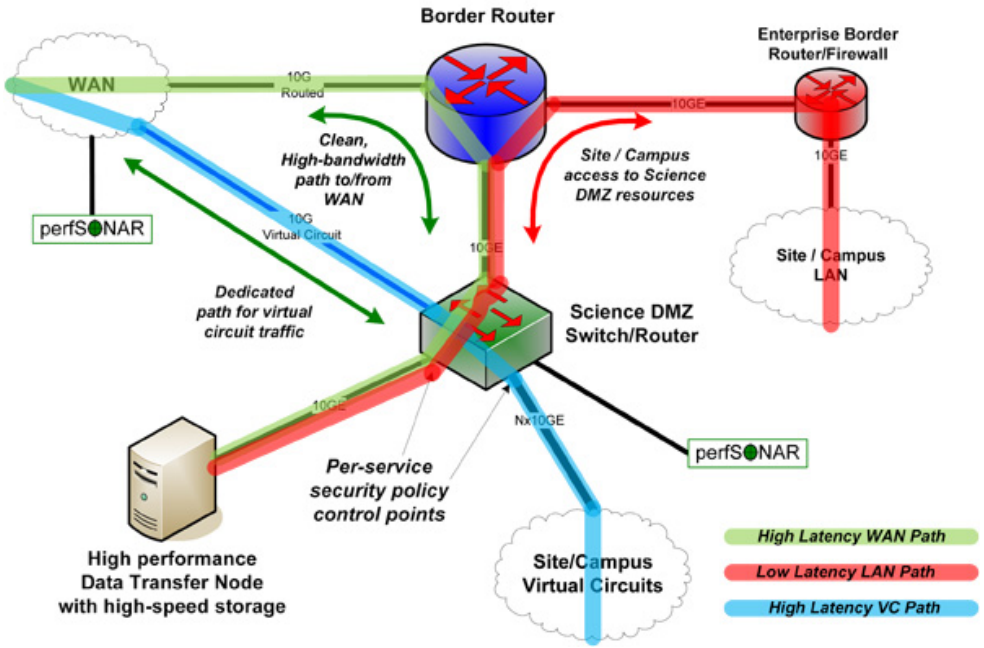
Choice of appropriate transfer tools

- You might get good enough performance from *scp/ftp*
- TCP-based applications can be very sensitive to packet loss
  - See https://www.switch.ch/network/tools/tcp_throughput/ for an online Mathis formula calculator; also includes a TCP buffer size calculator
  - A fraction of 1% packet loss can have a significant effect
  - GridFTP can mitigate this by using multiple parallel TCP streams
  - Google's recent work on TCP-BBR may also help; now in Linux kernel
- UDP-based applications are less sensitive to loss
  - Aspera is an example of a commercial UDP-based solution
  - But UDP is not considerate of TCP applications; TCP flows will back off in the presence of competing UDP
- And there is also likely to be "competing" regular campus traffic

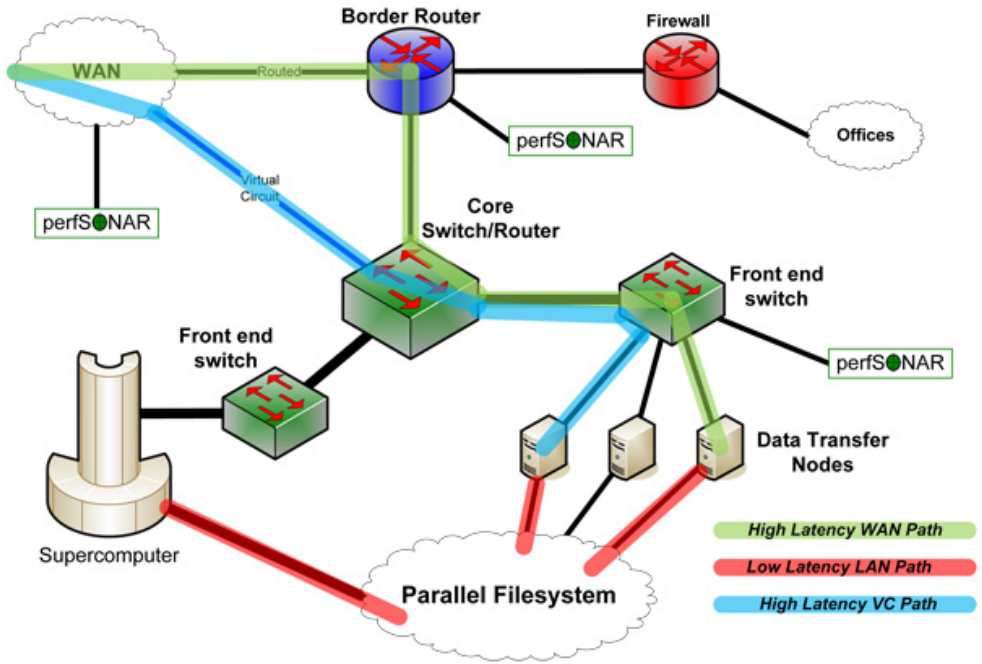
The ESnet "Science DMZ" approach

- ESnet published the Science DMZ design pattern in 2012/13
  - https://www.es.net/assets/pubs_presos/sc13sciDMZ-final.pdf
- Four key elements:
  - Network architecture; avoiding local bottlenecks
  - Network performance measurement
  - Security model
  - Data transfer node design and configuration
- The NSF Cyberinfrastructure (CC*) Program has funded this model in over 100 US universities, and continues to offer awards in similar areas:
  - See http://www.nsf.gov/pubs/2016/nsf16567/nsf16567.htm
  - No current funding equivalent in the UK; down to individual campuses to fund changes to network architectures for data-intensive science

Science DMZ network architecture



Simple Science DMZ



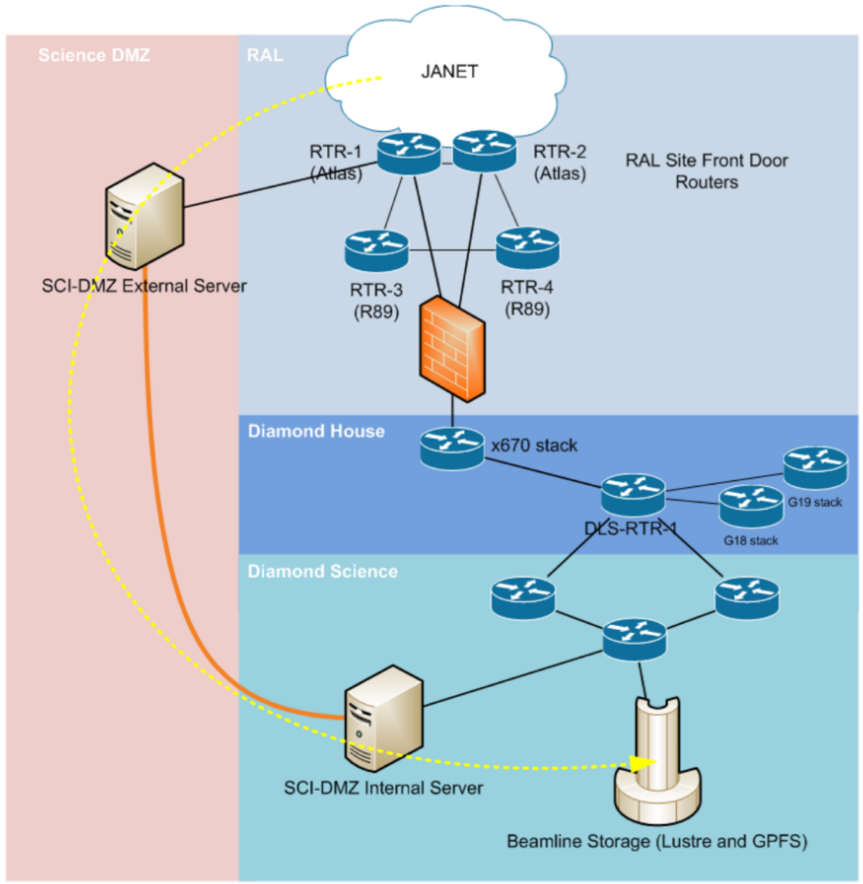
Supercomputer Center

http://fasterdata.es.net/science-dmz-architecture

"Science DMZ" in the UK

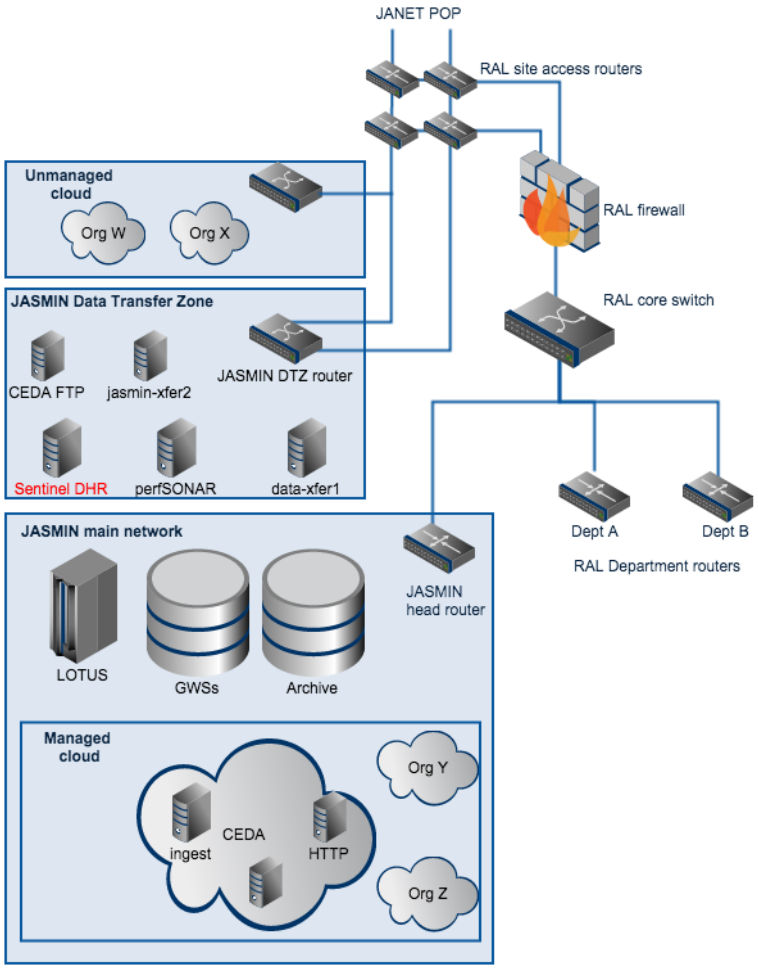
- There are several examples of sites in the UK that have a form of Science DMZ deployment
  - May not be full implementations, e.g. may lack perfSONAR
- In many cases these deployments were made in the absence of knowledge of the Science DMZ model
  - ESnet formalised the approach as a design pattern
  - A set of principles that can be applied to a variety of scenarios
- Examples in the UK:
  - Diamond Light Source
  - JASMIN/CEDA Data Transfer Zone
  - Imperial College GridPP; supports up to 40Gbit/s of IPv4/IPv6
  - To realise the benefit, both end sites need to apply the principles

Diamond Science DMZ



- c/o Alex White
- Investigated network issues between Oxford and Diamond by deploying perfSONAR nodes at three locations
- Identified STFC firewall as the cause of packet loss
- Initial 50Mbit/s goal over 10ms path required < 0.026% loss
- Re-engineered the network to optimise path to the DTN
- Loss significantly reduced
- Achieved 2Gbit/s throughput

JASMIN Data Transfer Zone



c/o Matt Pritchard & Jonathan Churchill

JASMIN/CEDA facility for climate and earth observation science

See http://www.ceda.ac.uk/

(for context, see the talk at http://repository.jisc.ac.uk/6210/1/e2epi-enabling-high-throughput-matt-pritchard.pdf)

Science & Technology Facilities Council

Some related GÉANT activities

GÉANT eduPERT

- eduPERT is a collaborative effort by a variety of campus and NREN participants to document and share experiences in end-to-end performance problems
- Aims to help users get the best from their connectivity
  - See http://services.geant.net/edupert
- Includes the searchable eduPERT knowledgebase, which contains entries added over the last 10 years
  - See http://kb.pert.geant.net/PERTKB/WebHome
- Originally designed to be a coordination point between Performance Enhancement Response Teams (PERTs)
- In practice, it's open to anyone to register and contribute
  - All such contributions are very welcome
  - To join the mail list: https://lists.geant.org/sympa/info/pert-discuss

GÉANT SIG-PMV

- The Special Interest Group for Performance Monitoring and Verification (SIG-PMV) is an open group studying the use of appropriate performance monitoring and measurement tools by researcher, campus and NREN groups
  - Started in Q3 2016
  - Initial activity will be to conduct surveys of communities to identify the existing tools being used, and potential gaps that may exist
  - Will produce recommendations for a variety of scenarios
  - Includes small node perfSONAR and WiFiMon
- See https://wiki.geant.org/display/PMV/SIG-PMV

- Next meeting: November 3rd 2016 at SWITCH offices in Zurich
  - An eduPERT training event follows on the 4th November
  - Details and registration: https://eventr.geant.org/events/2494
- To join the mail list: https://lists.geant.org/sympa/info/pmv-discuss

Proposed GÉANT TF-RED Task Force

- There is a proposal currently being built to create a new Task Force on Research Engagement Development (TF-RED)
- Its aims include:
  - Supporting research collaborations that want to start to collaborate internationally;
  - Coordinating reliable, predictable network behavior in support of individual end-user applications;
  - Establishing a continuous and permanent information flow between RENs and Science and Research communities;
  - Improving performance of network-centric and data-centric workflows.

- The fine details are under discussion; the high-level goal is certainly very important
  - To join mail list: https://lists.geant.org/sympa/sigrequest/tf-red

Summary

Recommendations

- Recommendations:
  - Within an NREN's scope, facilitate and encourage dialogue between the NREN, campus computing services and research communities
  - Undertake periodic networking "future looks"; inform capacity planning
  - Network performance measurement is very important; promote wider deployment of perfSONAR or similar tools
  - Campuses should consider appropriate local network engineering, noting in particular firewall throughput issues; the ESnet Science DMZ model is one such approach to draw upon
  - Share experiences and best practices; look at / contribute to eduPERT, SIG-PMV, and resources such as fasterdata.es.net
  - Draw up guidance to help researchers manage their expectations

Related Jisc links…

- Janet end-to-end performance initiative mail list:
  - Open to anyone to subscribe; focus is on Janet community
  - To join, see:
  - https://www.jiscmail.ac.uk/cgi-bin/webadmin?A0=E2EPI

- Campus Network Engineering for Data-Intensive Science workshop, October 19th 2016, London:
  - Free to attend
  - To register, visit:
  - https://www.jisc.ac.uk/events/campus-network-engineering-for-data-intensive-science-workshop-19-oct-2016

Dr Tim Chown

Senior Network Services Developer

Jisc, UK

tim.chown@jisc.ac.uk

jisc.ac.uk