Security Challenges for High Throughput Data Transfers

Dr Tim Chown, Network Development Manager, Jisc

WISE Meeting – The Cosener’s House – 28 Feb 2018

Overview

- There is growing interest in moving research data around the network
  - Data capture or generation to computing facility, and perhaps back
  - Remote visualisation
  - Data replication / distributed storage / backups
  - To / from cloud

- Data set volumes are increasing
  - 100 TB is no longer 'large'
  - Moving 100 TB takes 10Gbps of throughput for 24 hours

- How do we do this securely, AND with the necessary performance?
  - This deck has some thoughts on this topic…

What are we not talking about?

- » Security embraces many perspectives and methods
- » WISE has no doubt covered many topics at this meeting

- » We're not talking here about the more classic security challenges
  - › Authentication, authorisation, certificates, etc.
  - › Or how to do a GDPR audit!

- » The question to consider here is how do we achieve high throughput on the end-to-end data transfer path, while applying appropriate security measures to the traffic in flight
  - › The path needs to be performant, else the research suffers
  - › Can argue that necessary performance is a **requirement**

Theoretical throughput

The following table, taken from a publication by ESnet[3], shows the *theoretical* throughput required to transfer a given size of data set in a range of example time periods.

	1 Min	5 Mins	20 Mins	1 Hour	8 Hours	1 Day	7 Day	30 Days
10 PB	1,333Tbps	266.7Tbps	66.7Tbps	22.2Tbps	2.78Tbps	926Gbps	132Gbps	30.9Gbps
1 PB	133.3Tbps	26.7Tbps	6.67Tbps	2.2Tbps	278Gbps	92.6Gbps	13.2Gbps	3.09Gbps
100 TB	13.3Tbps	2.67Tbps	667Gbps	222Gbps	27.8Gbps	9.26Gbps	1.32Gbps	309Mbps
10 TB	1.33Tbps	266.7Gbps	66.7Gbps	22.2Gbps	2.78Gbps	926Mbps	132Mbps	30.9Mbps
1 TB	133.3Gbps	26.67Gbps	6.67Gbps	2.22Gbps	278Mbps	92.6Mbps	13.2Mbps	3.09Mbps
100 GB	13.3Gbps	2.67Gbps	667Mbps	222Mbps	27.8Mbps	9.26Mbps	1.32Mbps	309Kbps
10 GB	1.33Gbps	266.7Mbps	66.7Mbps	22.2Mbps	2.78Mbps	926Kbps	132Kbps	30.9Kbps
1 GB	133.3Mbps	26.7Mbps	6.67Mbps	2.22Mbps	278Kbps	92.6Kbps	13.2Kbps	3.09Kbps
100 MB	13.3Mbps	2.67Mbps	667Kbps	222Kbps	27.8Kbps	9.26Kbps	1.32Kbps	0.31Kbps

Thus, in principle, if you need to move 100GB in 20 minutes, you will need at least a 1Gbit/s capacity, end to end. Or, if you have a 10Gbit/s link, you can in principle move 100TB in a day (at a rate of 9.26Gbit/s).

Understanding the factors affecting E2E

- Achieving optimal end-to-end performance is a multi-faceted problem.
- It includes:
  - Appropriate network capacity provisioning between the end sites
  - Properties of the local campus network (at each end), including capacity of the external connectivity, internal LAN design, the performance of firewall / IDS devices, and the configuration of other devices on the path
  - End system configuration and tuning; network stack buffer sizes, disk I/O, …
  - The choice of tools used to transfer data, e.g. scp, Globus, rsync, Aspera, …
- To optimise end-to-end performance, you need to address each aspect
- There will inevitably be a bottleneck somewhere

Campus network engineering

» Question: how to design the local campus network for optimal end-to-end inter-site data transfer performance?

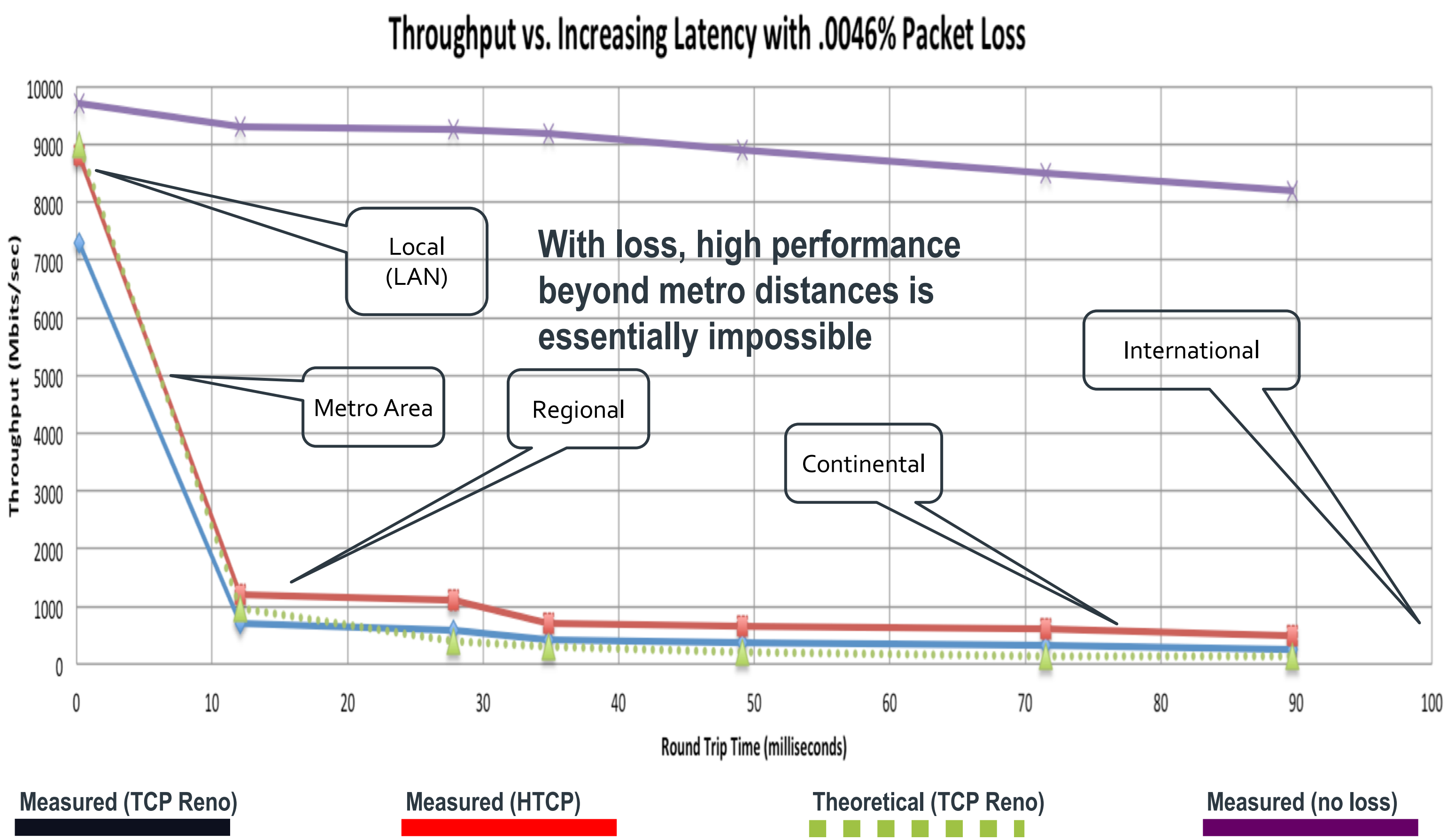
» Problem: An application using TCP will see its performance degrade if packets are lost, with more degradation the higher the path's RTT
  › Very small loss can have a surprisingly significant impact
  › Therefore we need to engineer towards zero packet loss

» Zero loss implies both sufficient capacity and performant network elements

» The challenge is that many campus security appliances, esp. corporate firewall/IDS, are designed for 1000's of small flows, not tens of very large flows, and they can thus drop packets

» Answer? The Science DMZ

TCP with a small amount of packet loss…

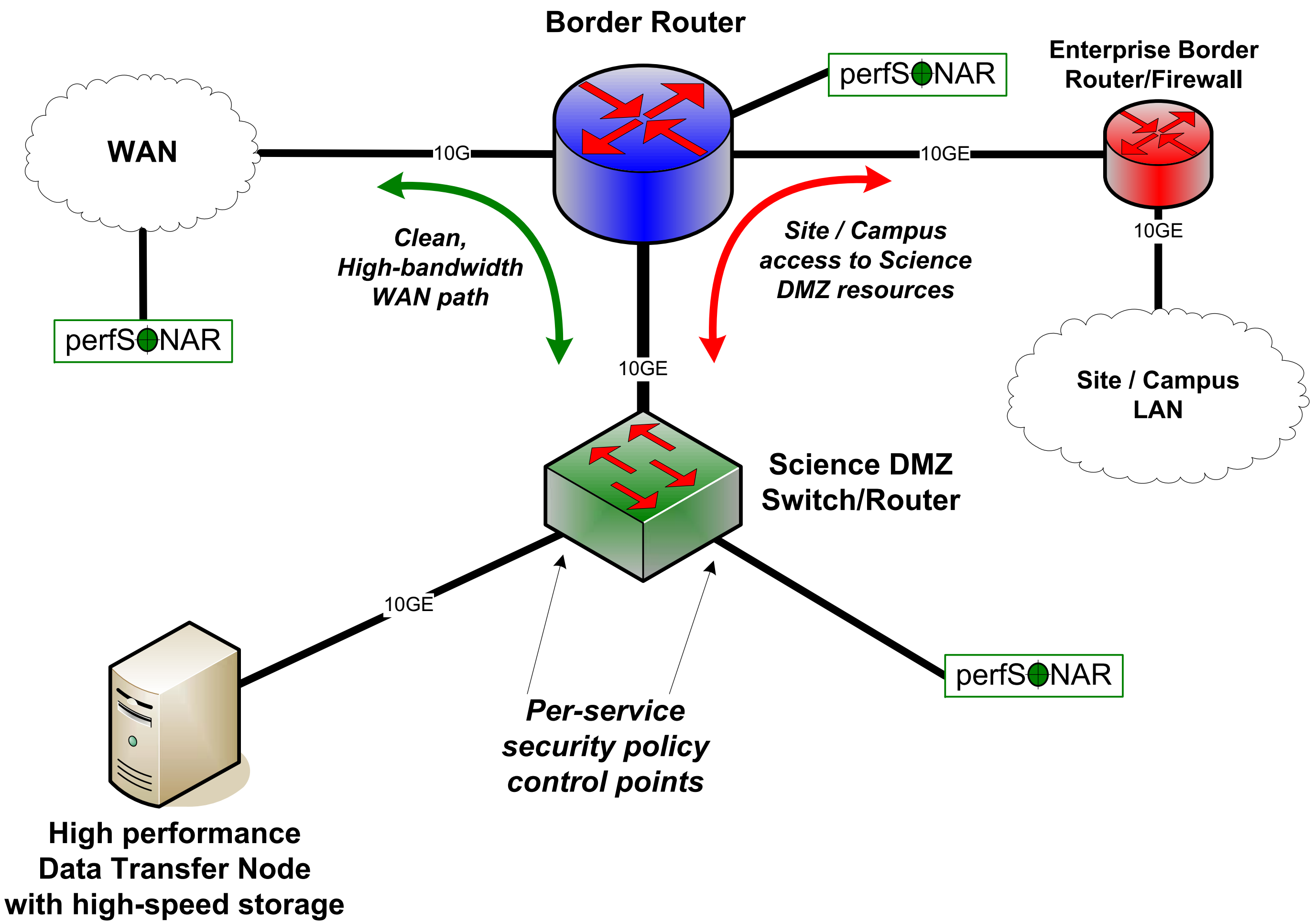


Site network engineering – the Science DMZ

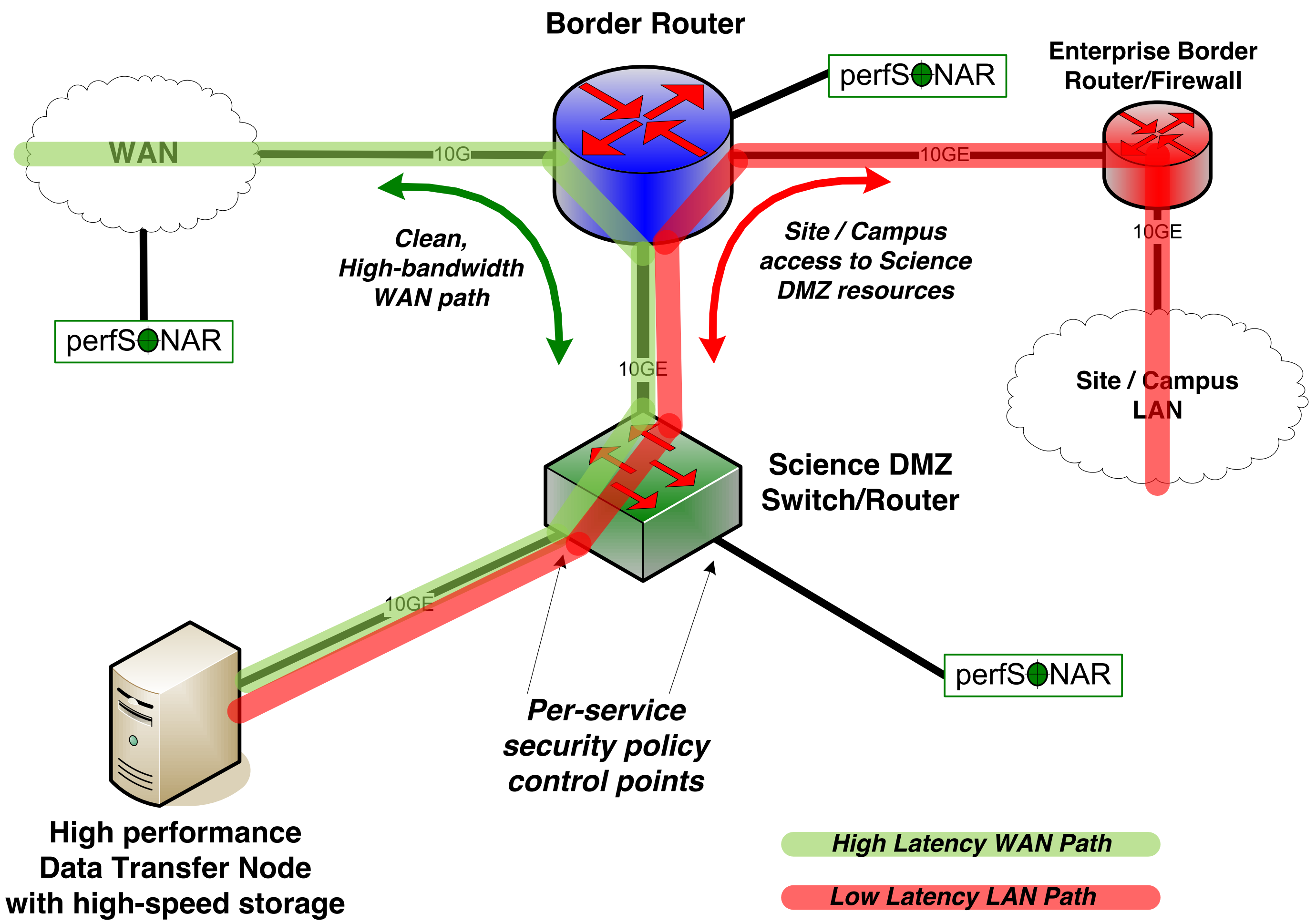
- ESnet published the Science DMZ 'design pattern' in 2012/13
  - https://www.es.net/assets/pubs_presos/sc13sciDMZ-final.pdf
- Three key elements:
  - Network architecture improvements; avoiding local bottlenecks
  - Network performance measurement
  - Data transfer node (DTN) design and configuration
- Also termed a "high speed on-ramp" to the campus storage
  - Splits the internal and external latency domains

- The NSF Cyberinfrastructure (CC*) Program funded this model in over 100 US universities:
  - See https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504748

Science DMZ Design Pattern (Abstract)



Local and wide area data flows

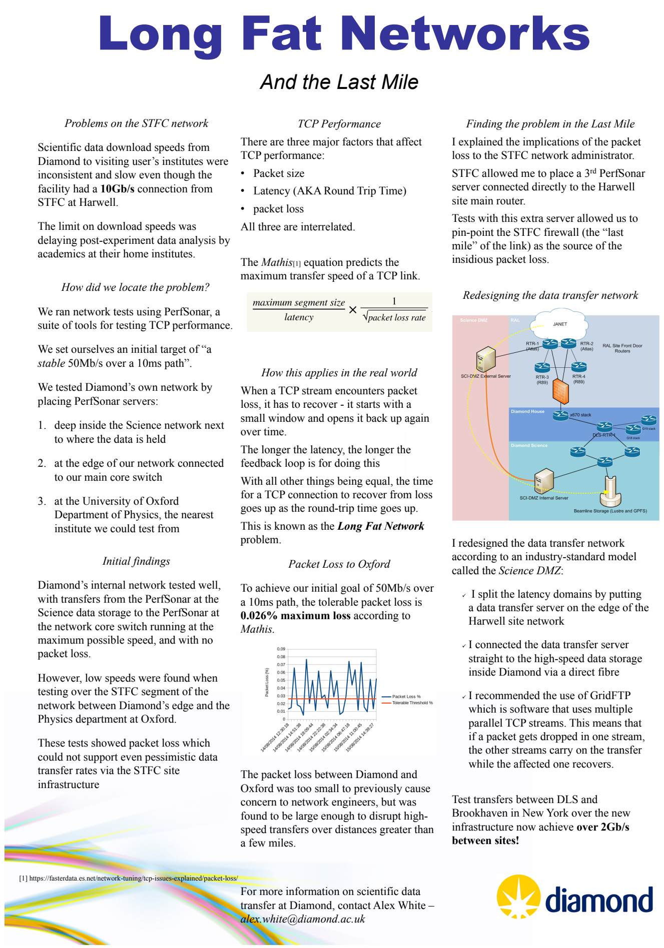


Jisc

Examples of Science DMZ in use at UK sites

- » There are several examples of sites in the UK that have some form of Science DMZ deployment
- » In many cases the deployments were made without knowledge of the Science DMZ model!
- » Science DMZ is just a set of good principles to follow, so it's not surprising that some Janet sites were already doing it, especially the GridPP sites

- » Examples in the UK:
  - › Diamond Light Source
  - › JASMIN/CEDA Data Transfer Zone
  - › Imperial College GridPP; supports up to 35Gbit/s of IPv4/IPv6
  - › BUT, to realise the end-to-end benefit, both ends need to apply the principles



Long Fat Networks

And the Last Mile

Problems on the STFC network

Scientific data download speeds from Diamond to visiting user's institutes were inconsistent and slow even though the facility had a **10Gb/s** connection from STFC at Harwell.

The limit on download speeds was delaying post-experiment data analysis by academics at their home institutes.

How did we locate the problem?

We ran network tests using PerfSonar, a suite of tools for testing TCP performance.

We set ourselves an initial target of "a *stable* 50Mb/s over a 10ms path".

We tested Diamond's own network by placing PerfSonar servers:

1. deep inside the Science network next to where the data is held
2. at the edge of our network connected to our main core switch
3. at the University of Oxford Department of Physics, the nearest institute we could test from

Initial findings

Diamond's internal network tested well, with transfers from the PerfSonar at the Science data storage to the PerfSonar at the network core switch running at the maximum possible speed, and with no packet loss.

However, low speeds were found when testing over the STFC segment of the network between Diamond's edge and the Physics department at Oxford.

These tests showed packet loss which could not support even pessimistic data transfer rates via the STFC site infrastructure

[1] https://fasterdata.es.net/network-tuning/tcp-issues-explained/packet-loss/

TCP Performance

There are three major factors that affect TCP performance:

- Packet size
- Latency (AKA Round Trip Time)
- packet loss

All three are interrelated.

The *Mathis*[1] equation predicts the maximum transfer speed of a TCP link.

$$\frac{\textit{maximum segment size}}{\textit{latency}} \times \frac{1}{\sqrt{\textit{packet loss rate}}}$$

How this applies in the real world

When a TCP stream encounters packet loss, it has to recover - it starts with a small window and opens it back up again over time.

The longer the latency, the longer the feedback loop is for doing this

With all other things being equal, the time for a TCP connection to recover from loss goes up as the round-trip time goes up.

This is known as the ***Long Fat Network*** problem.

Packet Loss to Oxford

To achieve our initial goal of 50Mb/s over a 10ms path, the tolerable packet loss is **0.026% maximum loss** according to *Mathis*.

The packet loss between Diamond and Oxford was too small to previously cause concern to network engineers, but was found to be large enough to disrupt high-speed transfers over distances greater than a few miles.

For more information on scientific data transfer at Diamond, contact Alex White – *alex.white@diamond.ac.uk*

Finding the problem in the Last Mile

I explained the implications of the packet loss to the STFC network administrator.

STFC allowed me to place a 3rd PerfSonar server connected directly to the Harwell site main router.

Tests with this extra server allowed us to pin-point the STFC firewall (the "last mile" of the link) as the source of the insidious packet loss.

Redesigning the data transfer network

I redesigned the data transfer network according to an industry-standard model called the *Science DMZ*:

- ✓ I split the latency domains by putting a data transfer server on the edge of the Harwell site network
- ✓ I connected the data transfer server straight to the high-speed data storage inside Diamond via a direct fibre
- ✓ I recommended the use of GridFTP which is software that uses multiple parallel TCP streams. This means that if a packet gets dropped in one stream, the other streams carry on the transfer while the affected one recovers.

Test transfers between DLS and Brookhaven in New York over the new infrastructure now achieve **over 2Gb/s between sites!**

diamond

Science DMZ as a Security Architecture

- » Rationale?
- » It allows for better segmentation of risks, and more granular application of controls to those segmented risks.
  - › Limit risk profile for high-performance data transfer applications
  - › Apply specific controls to data transfer nodes (DTNs)
  - › Avoid including unnecessary risks, unnecessary controls
- » Remove degrees of freedom – focus only on what is necessary
  - › Easier to secure
  - › Easier to achieve performance
  - › Easier to troubleshoot
- » Performance is a key requirement; e.g., use efficient ACLs
  - › See https://www.slideshare.net/JISC/science-dmz-security (Kate Mace)

Other examples of campus network engineering

- Many Janet sites split their external connectivity (their choice…)
  - e.g., 40G total; 1x10G campus, 1x10G research science data, 2x10G resilience
  - And then apply Science DMZ principles to the dedicated research data path
  - Or employ Science DMZ in their data centre

- The Worldwide LHC Computing Grid (WLCG) has used physical / virtual overlays
  - LHCOPN (private optical network) / LHCONE (virtual network)
  - LHCONE implicitly becomes a 'trusted' network
- But how should campuses cater for multiple data-intensive science disciplines?
  - Would one new overlay network per research community scale?

- Some sites are exploring SDN, to direct traffic dynamically and efficiently on campus
  - The classic on-ramp 'Science DMZ' is a rather static architecture

Aside 1: One person's (D)DoS…

- … is another person's research data transfer!

- What might you see in/out of a campus?
  - High volume UDP data transfer flows, e.g., Aspera
  - New protocols, e.g., QUIC (UDP/HTTP2)
  - 'Smarter' TCP algorithms, e.g., TCP-BBR
  - Highly parallelised flows, e.g., Globus / Grid FTP

- Applications behaving this way might seem to be out of profile, and thus potential (D)DoS
  - Need to keep abreast of application protocol developments
  - May white-list certain applications / address space

Aside 2: Encryption

- If you need to encrypt sent data, this might be implemented by
  - Pre-encrypting the data
  - Encryption on the fly as you transmit
  - Encryption between gateways on the path
- All these have potential performance issues or limitations

- Some NRENs offer solutions in this space
  - e.g., Jisc has Safe Share
  - https://www.jisc.ac.uk/safe-share
  - But by default only up to 1Gbit/s; higher throughput costs £££
  - Genomics project data sets can easily be 100-200TB.

Measuring network characteristics

- Important to have telemetry on your network

- The Science DMZ model recommends perfSONAR for this
- Collects telemetry over time
  - Throughput, loss, latency, path
  - Allows retrospective viewing of data
  - Uses proven tools under the hood such as iperf
- Can run tests between two perfSONAR systems, or build a mesh

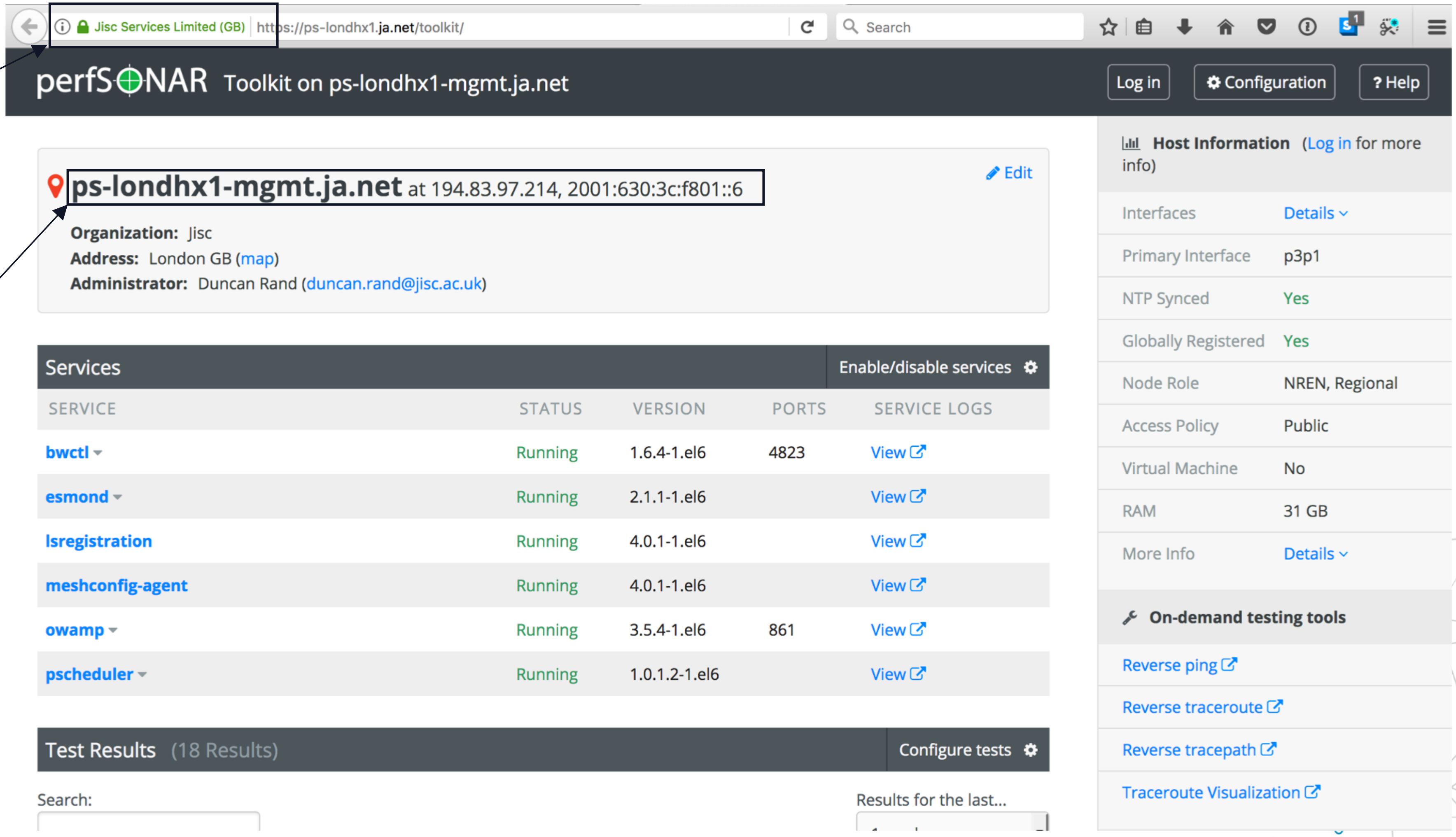
- Helps you assess the impact of changes to your network or systems
- It can highlight poor performance, but doesn't troubleshoot per se
- May indicate impact of security appliances on performance

Janet - London perfSONAR node

Added a Jisc certificate

Dual-stack

Possible to set up tests manually, but better to set up a mesh…



Example perfSONAR mesh – UK GridPP

- GridPP = UK Particle Physics computing collaboration
  - UK component of the WLCG (Worldwide LHC Computing Grid)
- Nineteen sites forming one Tier-1 and four distributed Tier-2 sites

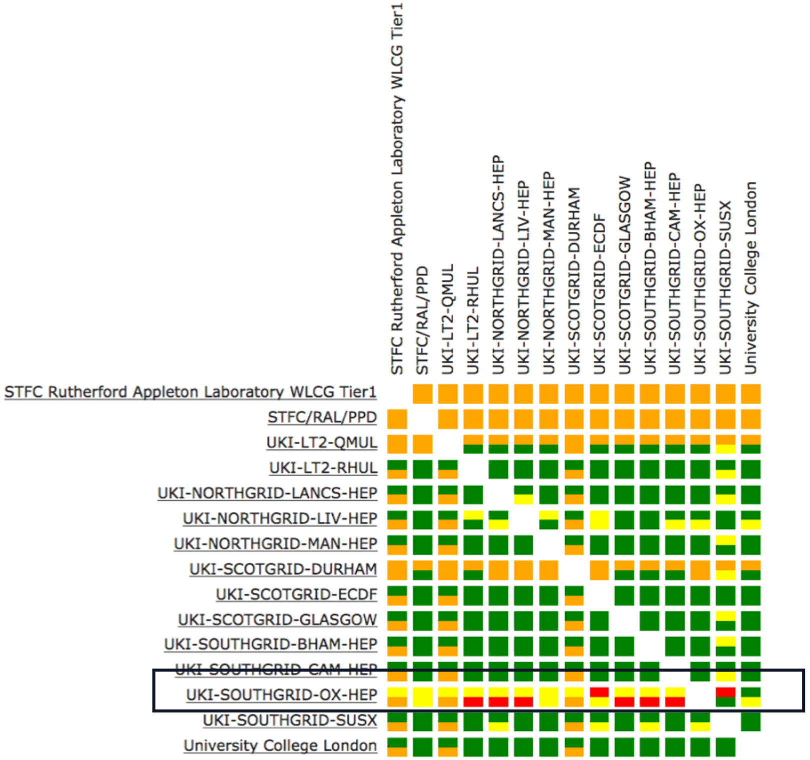
- Most sites have perfSONAR nodes next to their storage servers
- They are running a dual-stack mesh
  - Measure IPv4 and IPv6 performance independently
- Provides an insight into network performance across the sites
- Live version:
  - http://ps-dash.dev.ja.net/maddash-webui/index.cgi?dashboard=UK%20Mesh%20Config

GridPP mesh: Traceroute

UK Mesh Config - IPv4 Traceroute Tests

Number of Paths is <= 1 | Number of Paths is >= 1 | Number of Paths is >= 2 | Unable to retrieve da

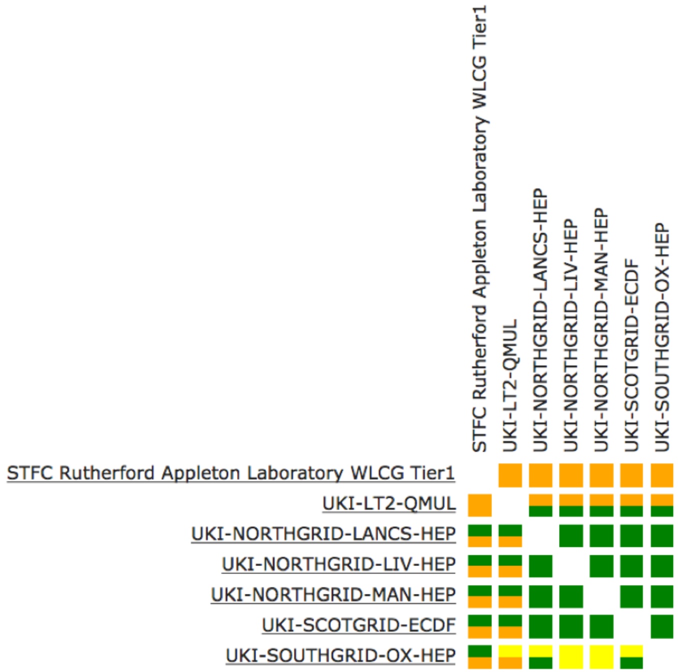
No problems found in grid



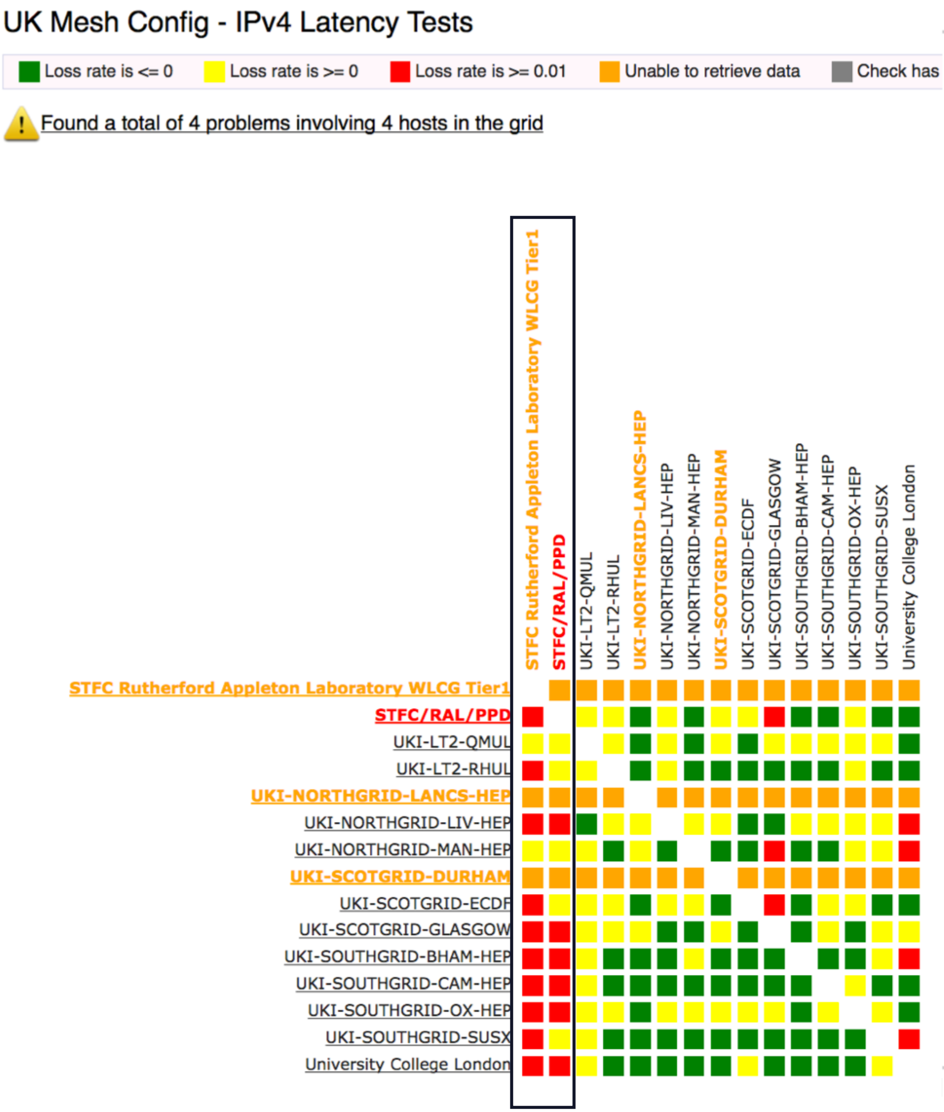
UK Mesh Config - IPv6 Traceroute Tests

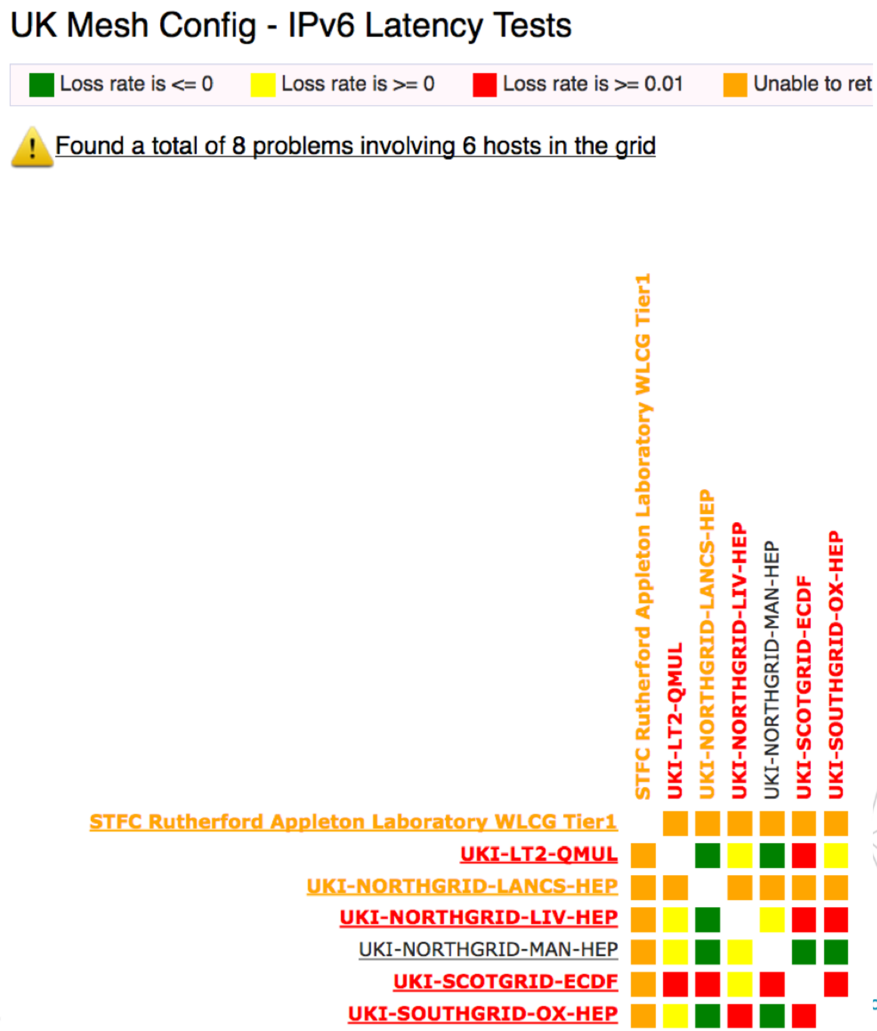
Number of Paths is <= 1 | Number of Paths is >= 1 | Number of Paths is >= 2

No problems found in grid



GridPP mesh: Latency and Loss



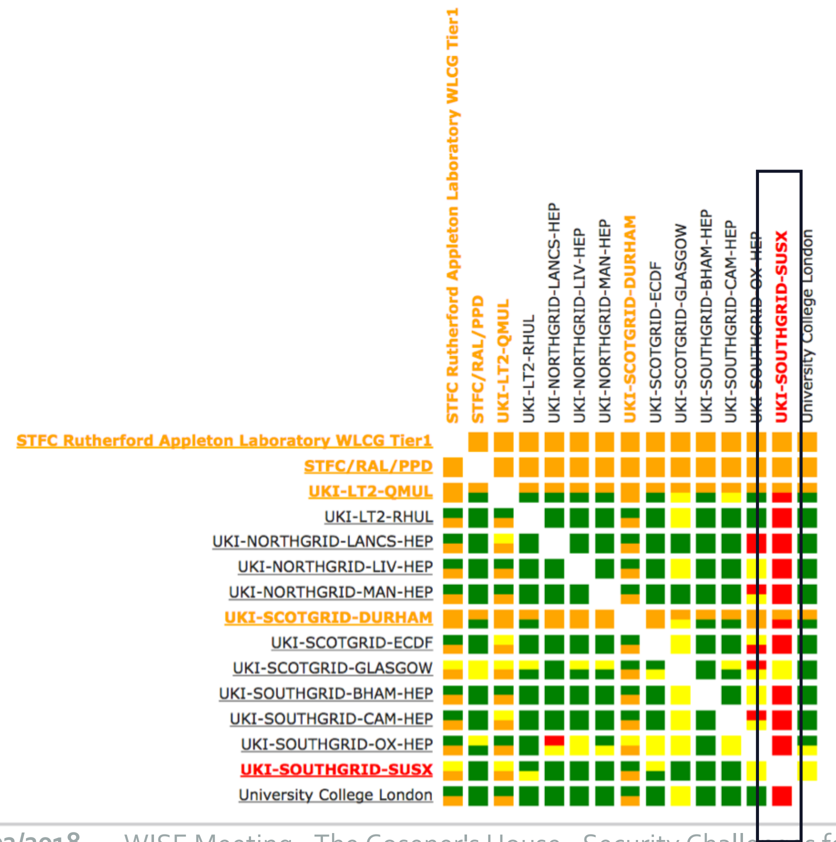


GridPP mesh: Throughput

UK Mesh Config - IPv4 Bandwidth Tests

Throughput >= 900Mbps | Throughput < 900Mbps | Throughput <= 500Mbps | Unable to retrieve da

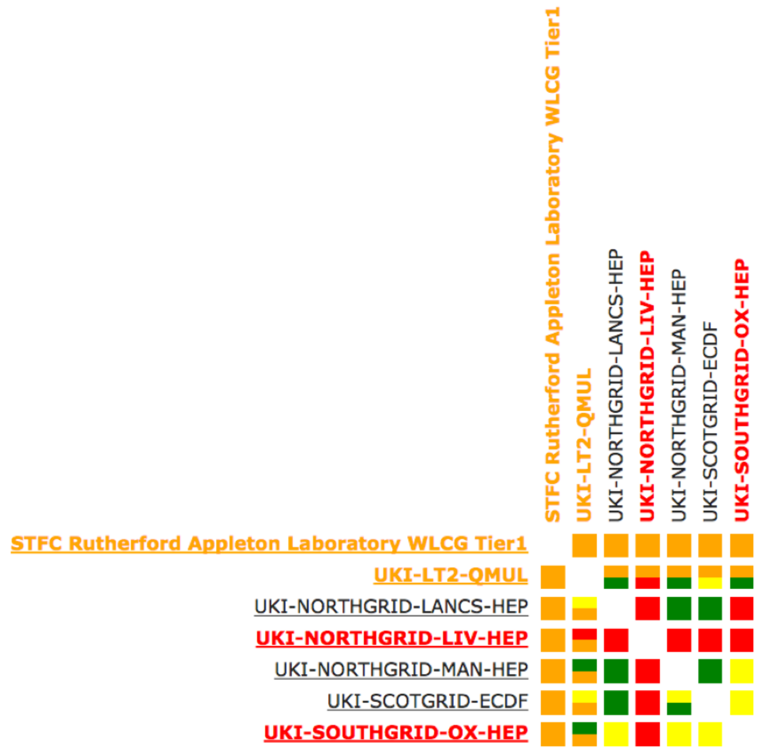
Found a total of 5 problems involving 5 hosts in the grid



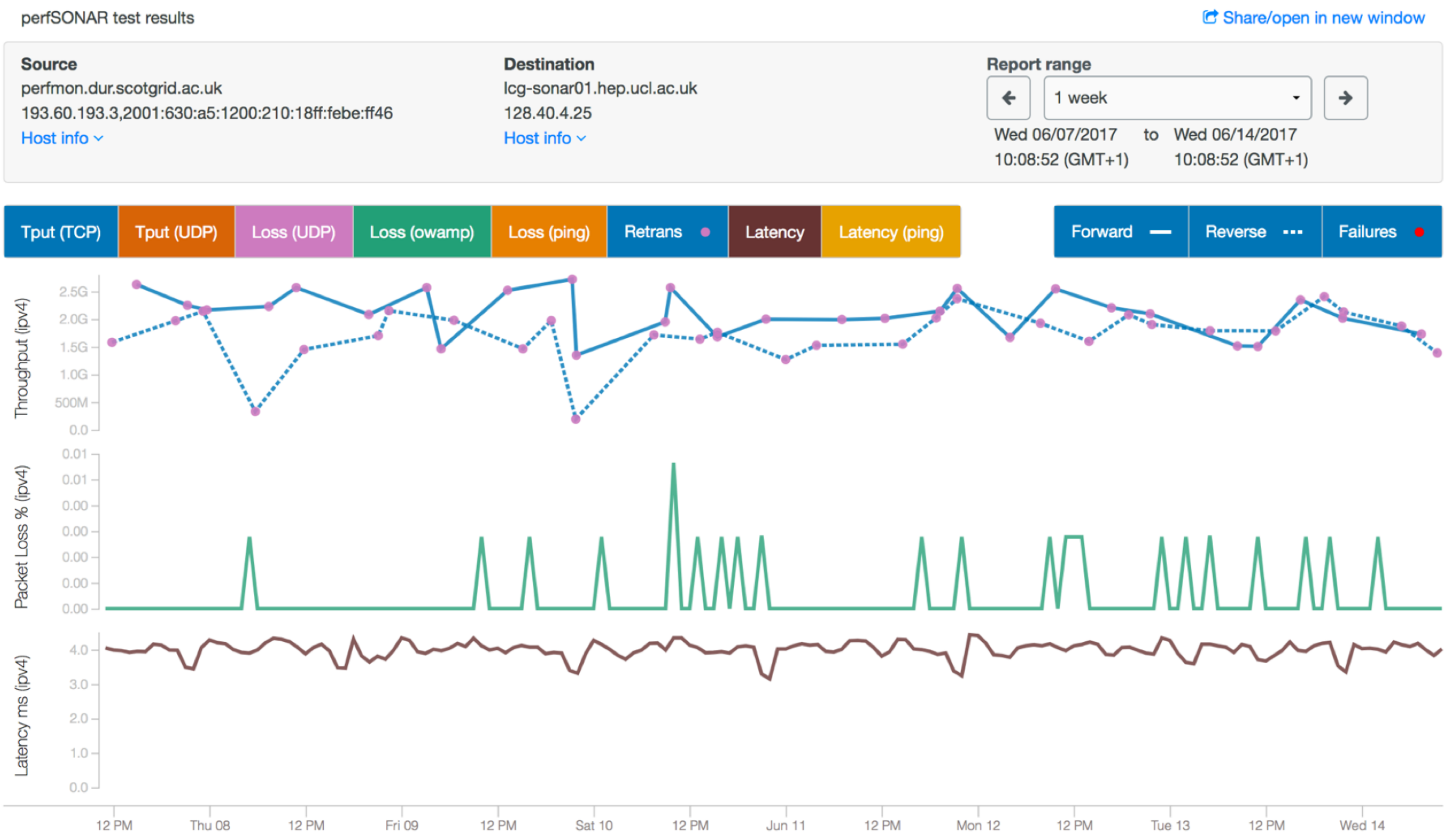
UK Mesh Config - IPv6 Bandwidth Tests

Throughput >= 900Mbps | Throughput < 900Mbps | Throughput <= 500Mbps

Found a total of 5 problems involving 4 hosts in the grid



perfSONAR – performance visualization over time



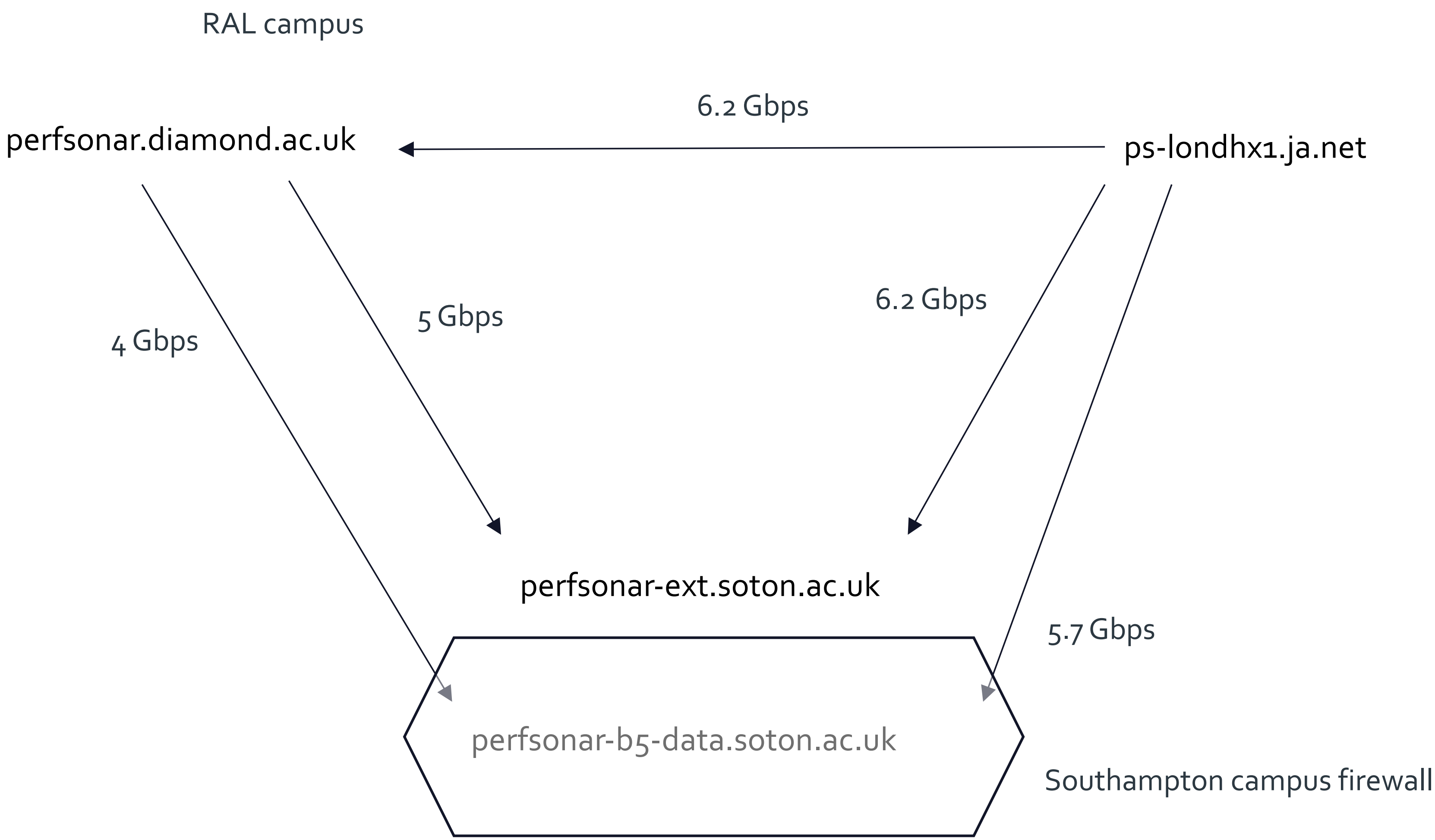
Example: Diamond LS <-> Southampton

- Southampton researchers go to DLS to run experiments and collect data which they wish to transport back home
  - Rather than carrying disks, they ought to be able to use the Janet network

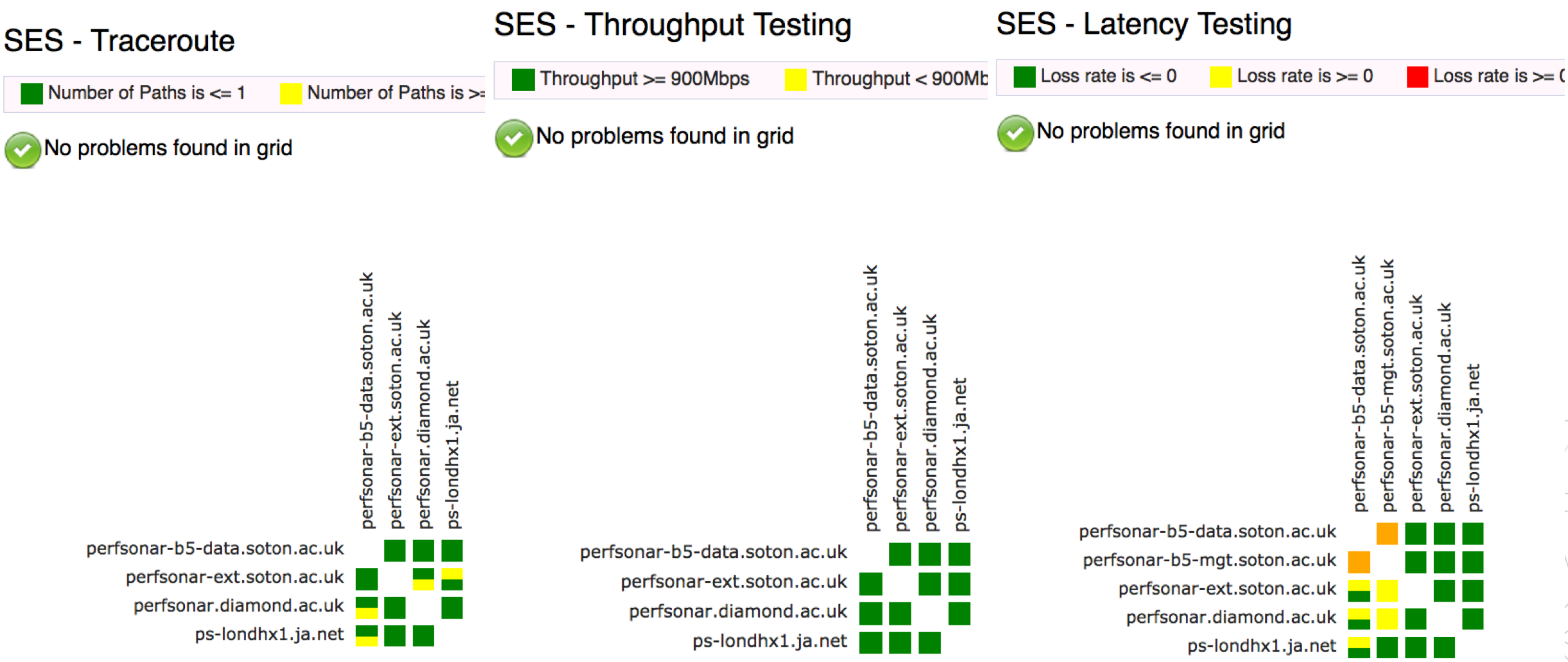
- Initial work at Southampton involved adopting Globus Connect to transfer files from DLS
- Achieved a significant improvement up to the then available 1 Gbps
- Also installed a perfSONAR host (*perfsonar-b5-data.soton.ac.uk*) on campus next to the data storage
- Network to storage then upgraded to 10 Gbps
- Later a perfSONAR host (*perfsonar-ext .soton.ac.uk*) was installed at the Southampton border, outside the firewall

- See http://ps-dash.dev.ja.net/maddash-webui/index.cgi?dashboard=SES

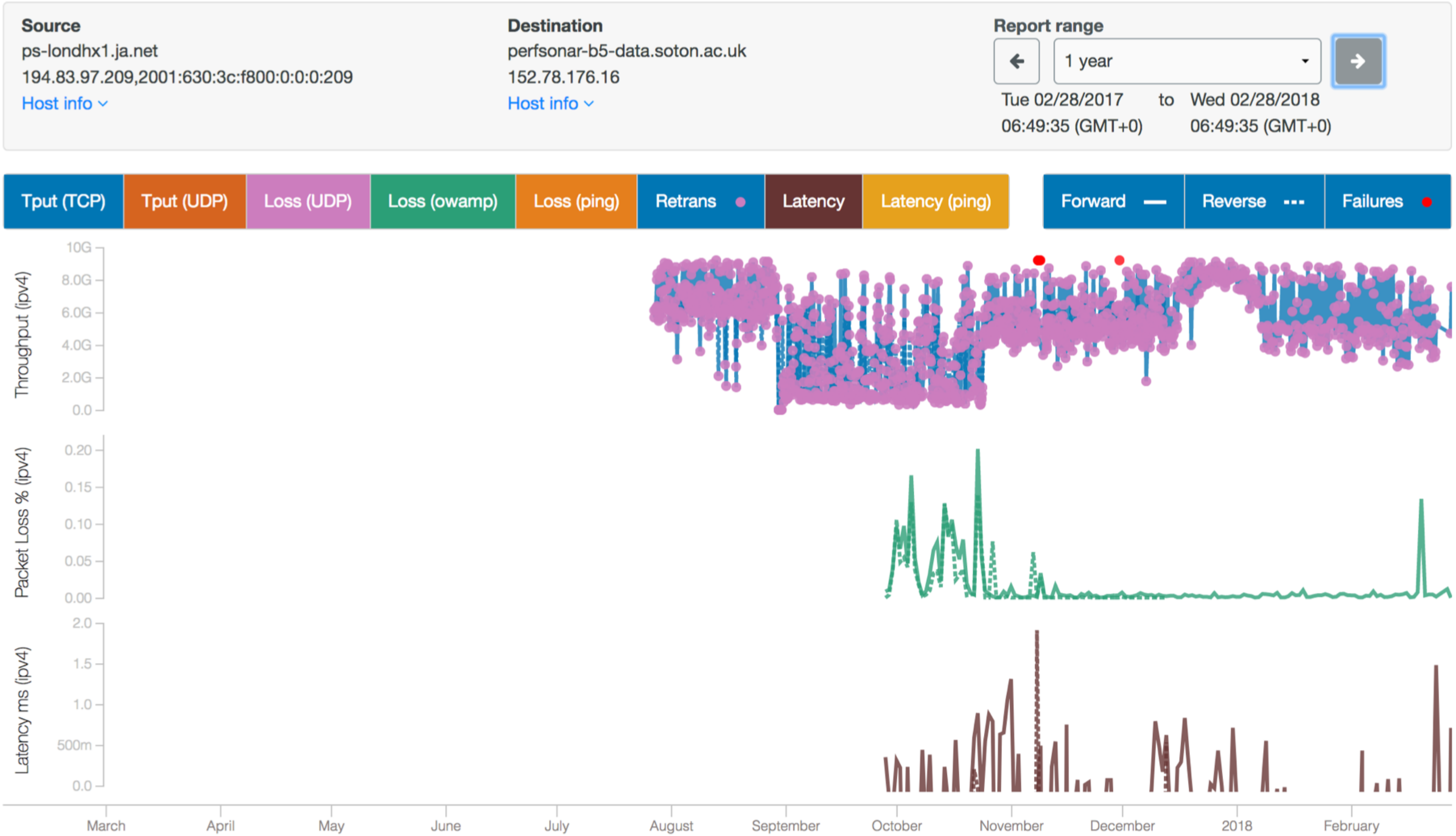
Current observed average throughput



Throughput mesh



Last 12 month view – London -> B5 network



Other examples

- University wanting to backup data to RAL
  - Obtaining 300-400Mbit/s
  - Introduced Science DMZ principles
  - Now sustaining 3-4 Gbit/s

- University on WLCG; IDS imposed
  - Was obtaining up to 18Gbit/s
  - IDS throughput maximum 8 Gbit/s
  - An example of appropriate application of policy

- University with ‘bug’ on firewall
  - Capacity reduced to 1Gbit/s or less on any flow
  - Normal campus users did not report the issue; perfSONAR detected it

Conclusions

- Consider how to apply the necessary policy efficiently
- Is this an area that interests the WISE community?

- Design in appropriate network engineering
- The classic 'Science DMZ' model has value; many were doing it anyway
  - Well-tuned DTNs with host-based security
  - SDN may provide a more agile 'DMZ' architecture
- Consider emerging data-intensive application network protocols; QUIC, etc.
- Measure performance over time (perfSONAR)

- Don't disrupt research; performance is a requirement
  - Just like the confidentiality, integrity and availability (CIA) principles

Some useful links

- Janet E2EPI project page
  - https://www.jisc.ac.uk/rd/projects/janet-end-to-end-performance-initiative
- JiscMail E2EPI list (approx 100 subscribers)
  - https://www.jiscmail.ac.uk/cgi-bin/webadmin?A0=E2EPI
- Camus Network Engineering for Data-Intensive Science workshop slides
  - https://www.jisc.ac.uk/events/campus-network-engineering-for-data-intensive-science-workshop-19-oct-2016
  - https://www.slideshare.net/JISC/science-dmz-security (Kate Mace, ESnet)
- Fasterdata knowledge base
  - http://fasterdata.es.net/
- eduPERT knowledge base
  - http://kb.pert.geant.net/PERTKB/WebHome

Please feel free to get in touch!

Dr Tim Chown

Network Development Manager

tim.chown@jisc.ac.uk

jisc.ac.uk