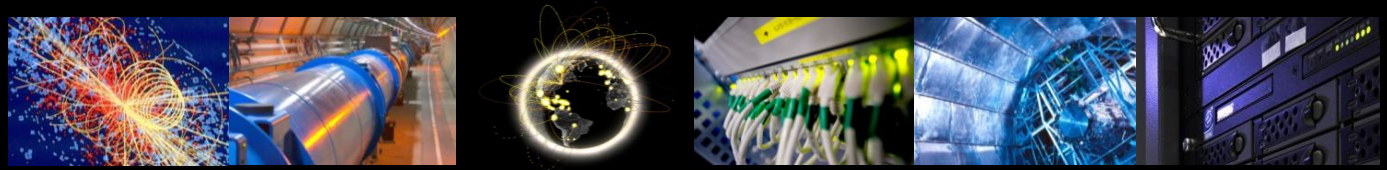
LHCOPN/LHCONE perfSONAR Update

Marian Babik, Shawn McKee
on behalf of WLCG Network Throughput WG



WLCG
Worldwide LHC Computing Grid

Open Science Grid

Outline

- News
  - OSG/WLCG activities and WLCG Network Throughput WG
  - perfSONAR 4.1
- Platform Updates
  - WLCG perfSONAR Infrastructure Status
  - OSG Network Monitoring Platform
- Platform Use
  - Activities and collaborations
  - Network Throughput Support Unit
- New Projects
  - SAND
  - IRIS-HEP
- Plans
- Summary

OSG/WLCG Networking Activities

- OSG is in its 6th year of supporting WLCG/OSG networking focused on:
  - Assisting its users and affiliates in identifying and fixing network bottlenecks
  - **Developing and operating a comprehensive Network Monitoring Platform**
  - Improving our ability to manage and use network topology and network metrics for analytics

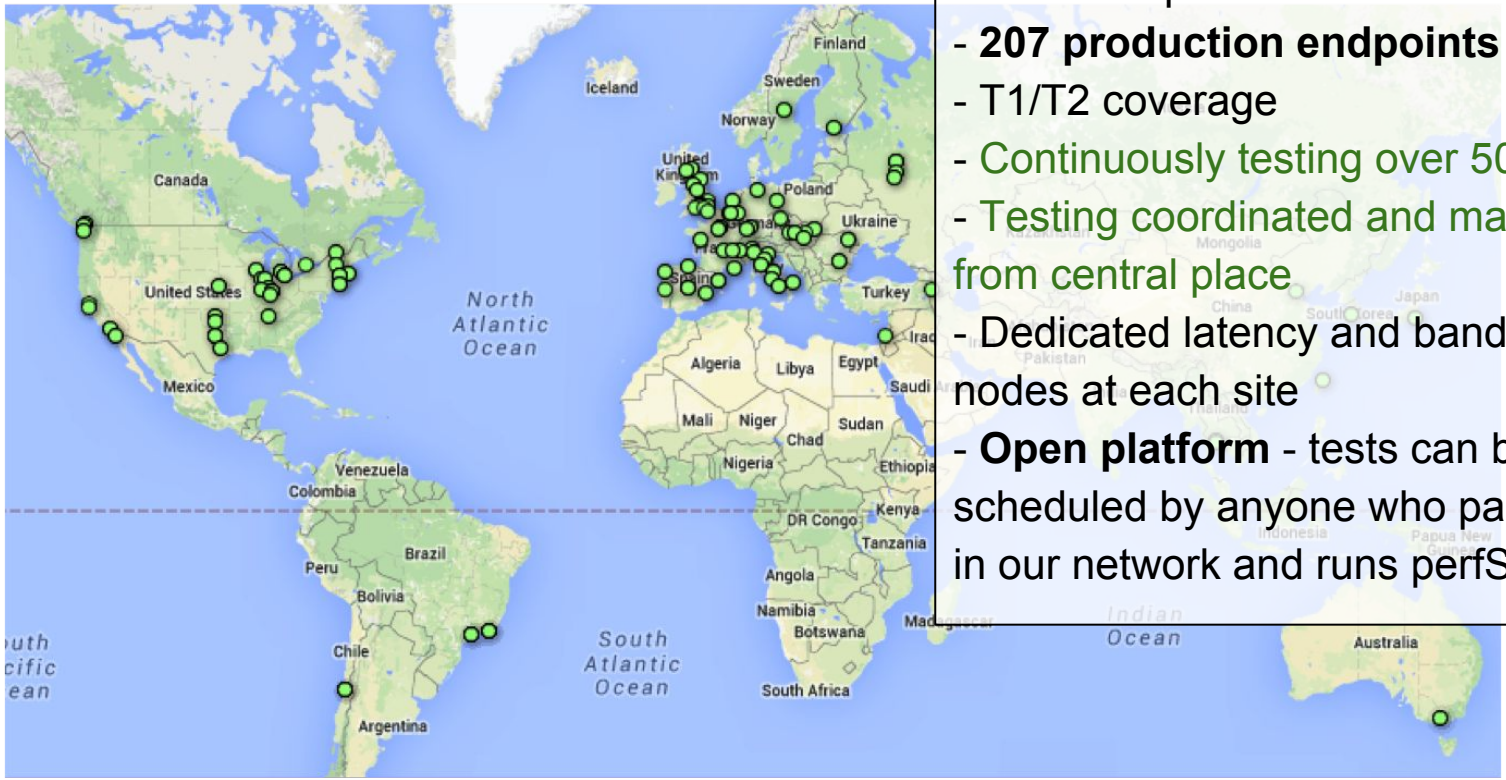
- WLCG Network Throughput Working Group was established to ensure sites and experiments can better understand and fix networking issues:
  - Oversees the **WLCG perfSONAR infrastructure**
    - Core infrastructure for taking network measurements and performing low-level debugging activities
  - **Coordinates WLCG network performance incidents** - runs a dedicated support unit which involves sites, network experts, R&Es and perfSONAR developers
    - Many issues are potentially resolvable within the working group

perfSONAR News

perfSONAR is 4.1 was released at the end of August

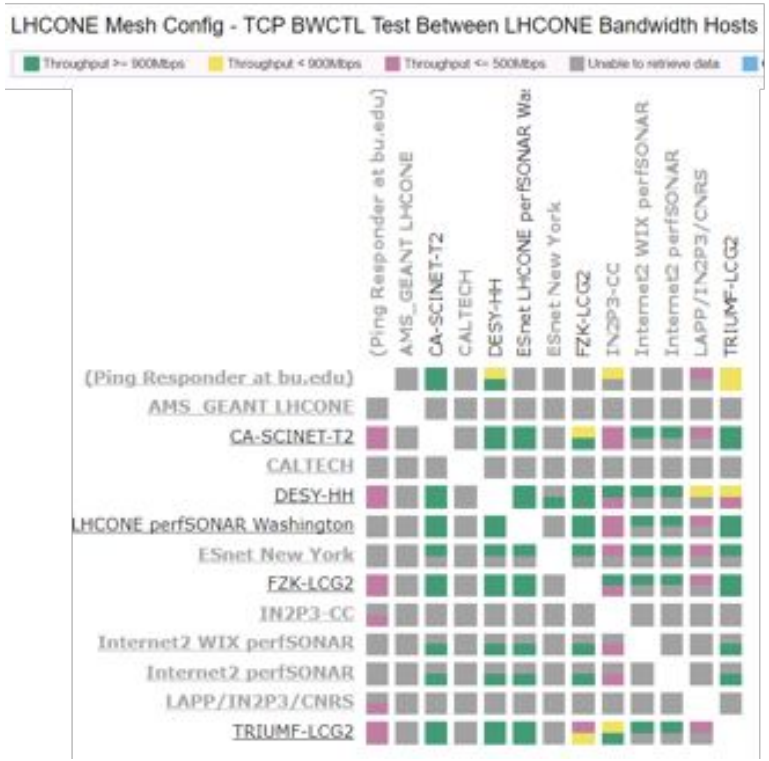
- New plugins
  - Network traffic capture (via ‘snmp’)
  - Application-level (HTTP response times)
  - TWAMP (two-way active measurement protocol) - more accurate round trip measurements than the ones from ping, can test devices not running perfSONAR
- New configuration
  - PWA/PSCONFIG - **new central web interface and toolkit configuration mechanism**
  - Brings a lot more options and better use of pScheduler
- pScheduler adds preemptive scheduling support
  - **Retires BWCTL** - still installed but no longer configured
  - pScheduler **requires** port 443 to be open to all (potential) testing nodes
- **Docker support**
- Drops SL6 support which is the OS for most of our instances
  - Our recommendation: reinstall with CentOS7; **don’t worry about saving data**

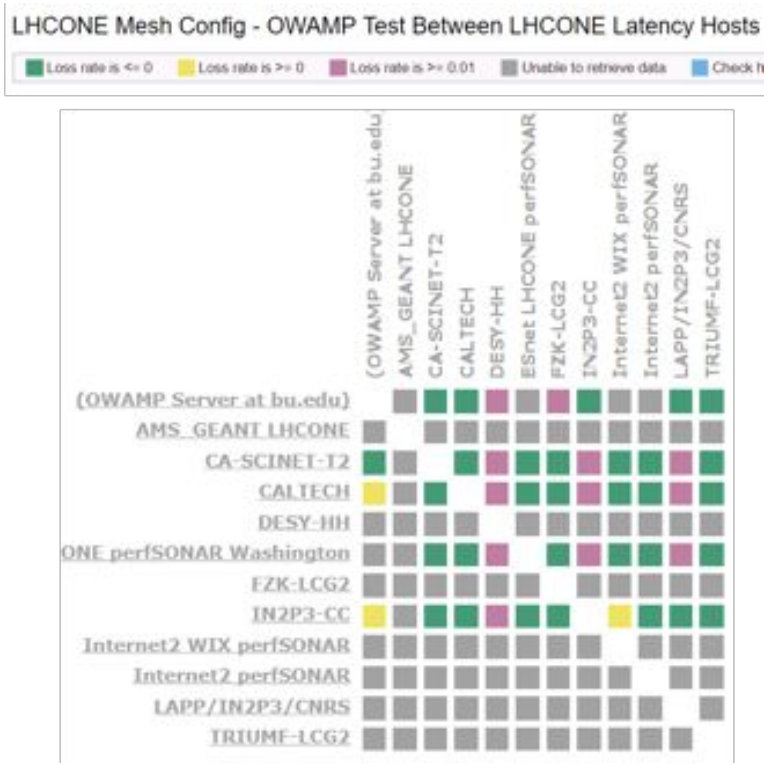
perfSONAR deployment



288 Active perfSONAR instances
- **207 production endpoints**
- T1/T2 coverage
- Continuously testing over 5000 links
- Testing coordinated and managed from central place
- Dedicated latency and bandwidth nodes at each site
- **Open platform** - tests can be scheduled by anyone who participates in our network and runs perfSONAR

LHCONE - 5th March 2018

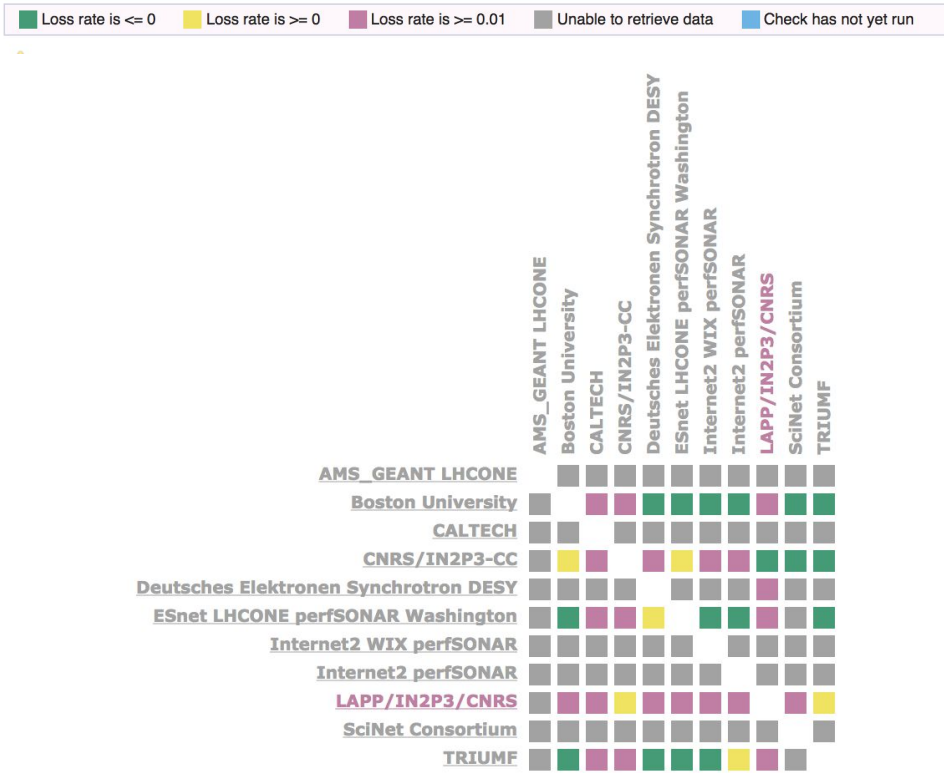




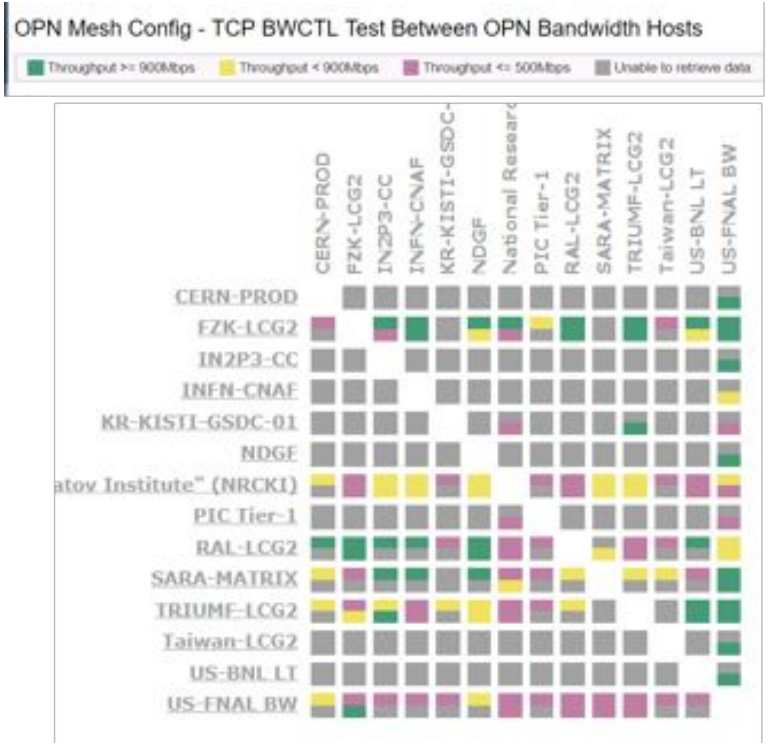
LHCONE 29th October 2018

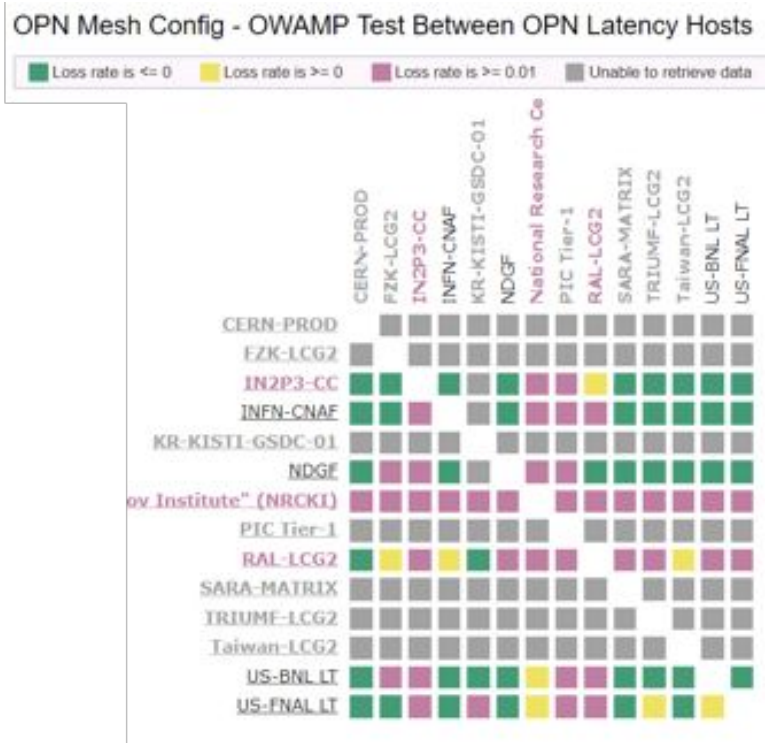
LHCONE Mesh Config - TCP BWCTL Test Between LHCONE Bandwidth

LHCONE Mesh Config - OWAMP Test Between LHCONE Latency Hosts



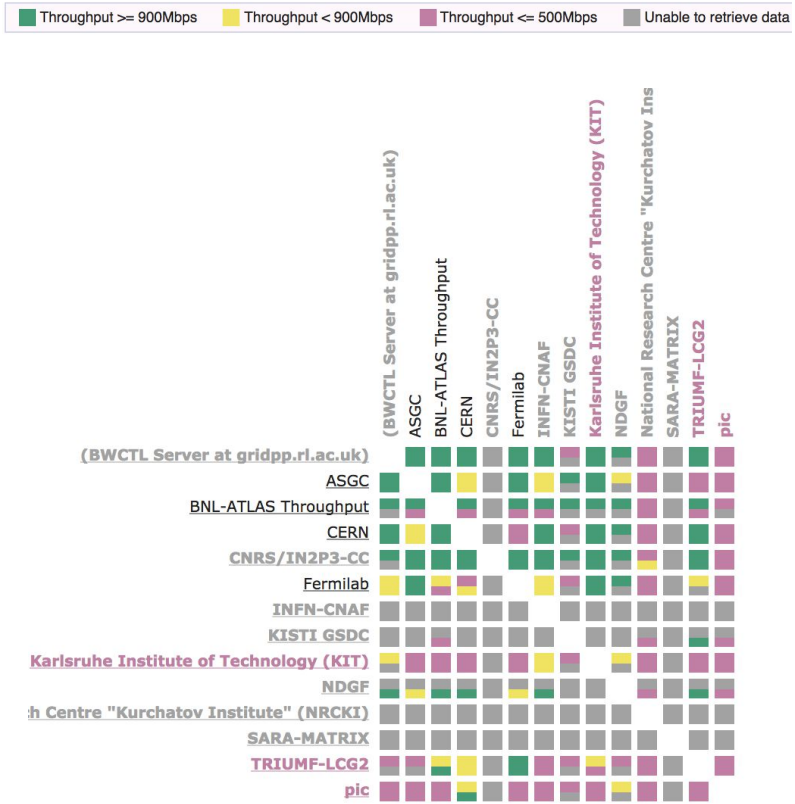
LHCOPN - 5th March 2018



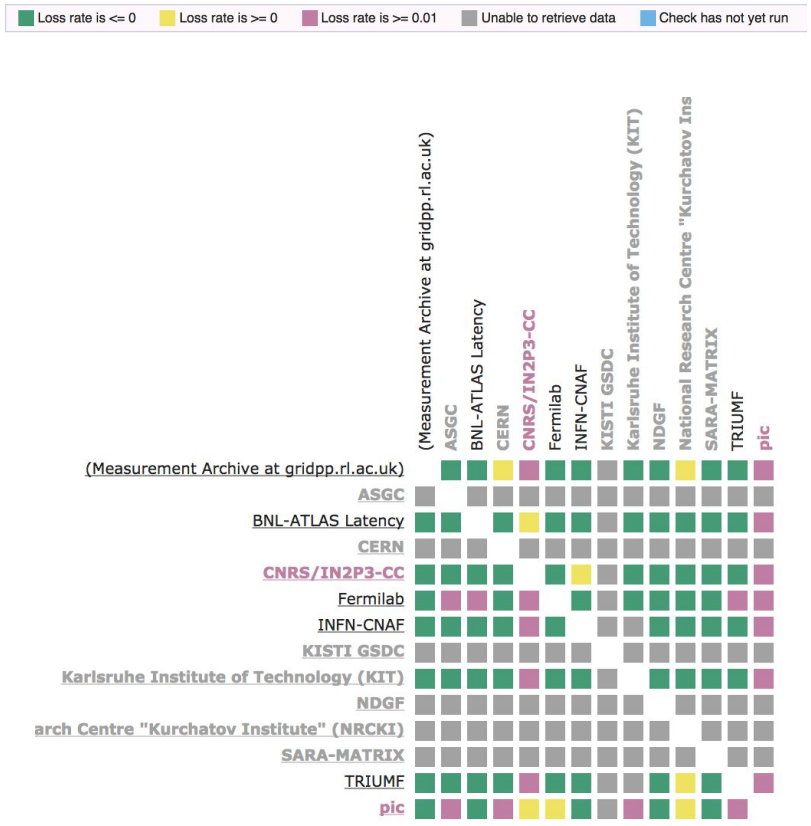


LHCOPN - 29th October 2018

OPN Mesh Config - TCP BWCTL Test Between OPN Bandwidth Hosts

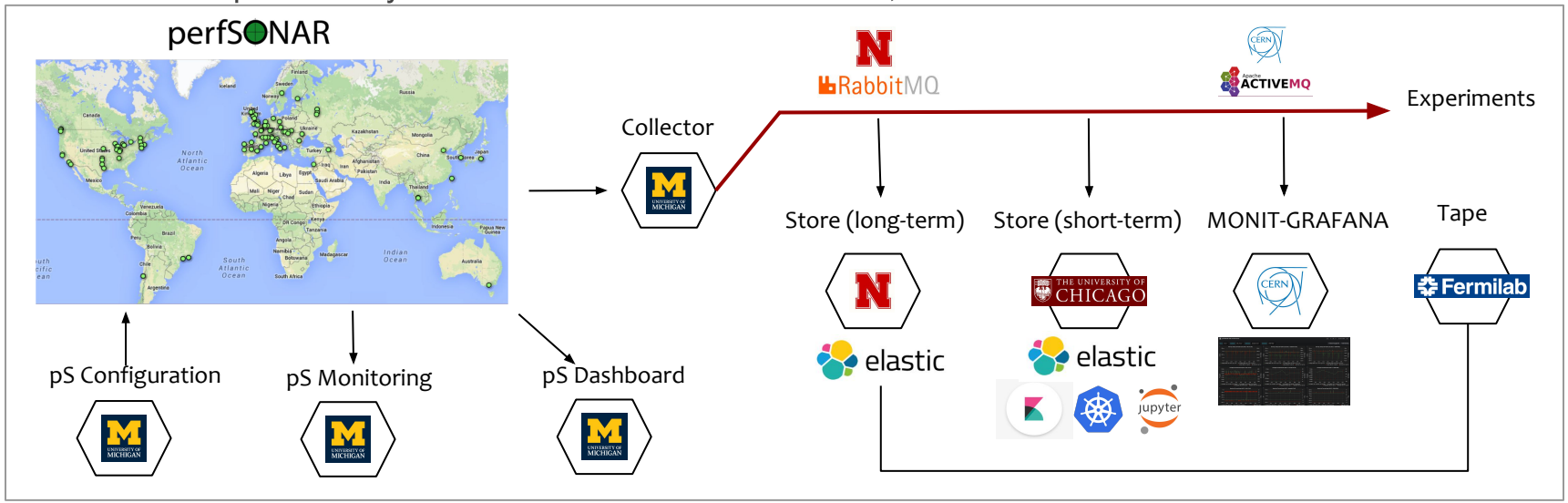


OPN Mesh Config - OWAMP Test Between OPN Latency Hosts

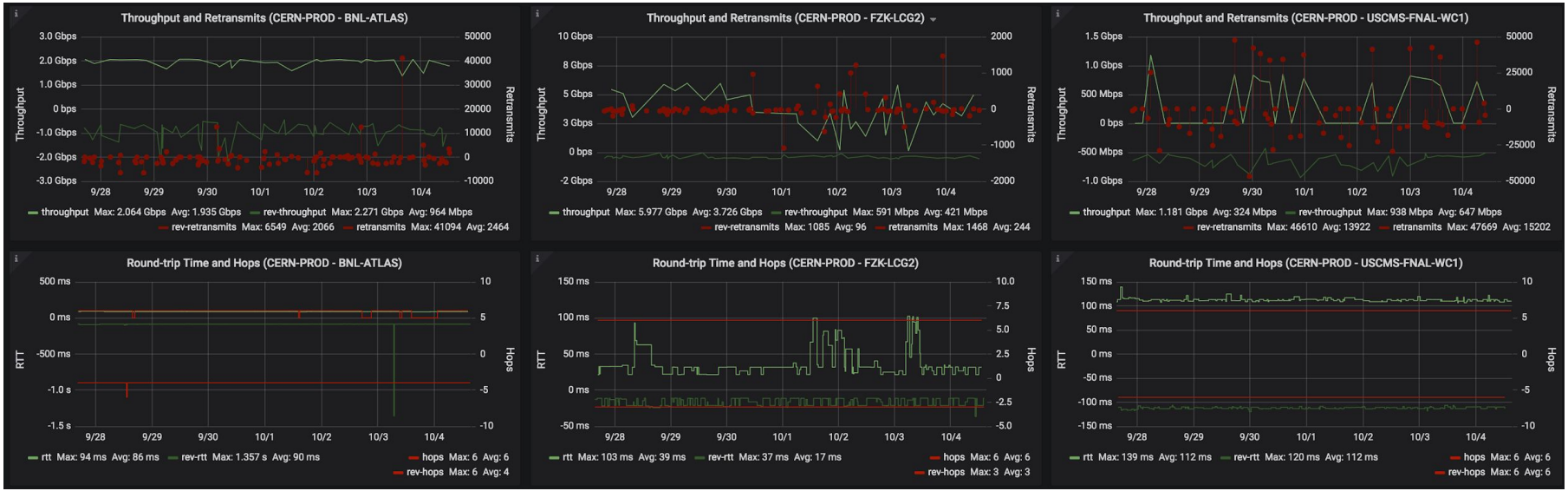


Platform Overview

- Collects, stores, configures and transports all network metrics
  - Distributed deployment - operated in collaboration
- All perfSONAR metrics are available via API, live stream or directly on the analytical platforms
  - Complementary network metrics such as ESNet, LHCOPN traffic also via same channels

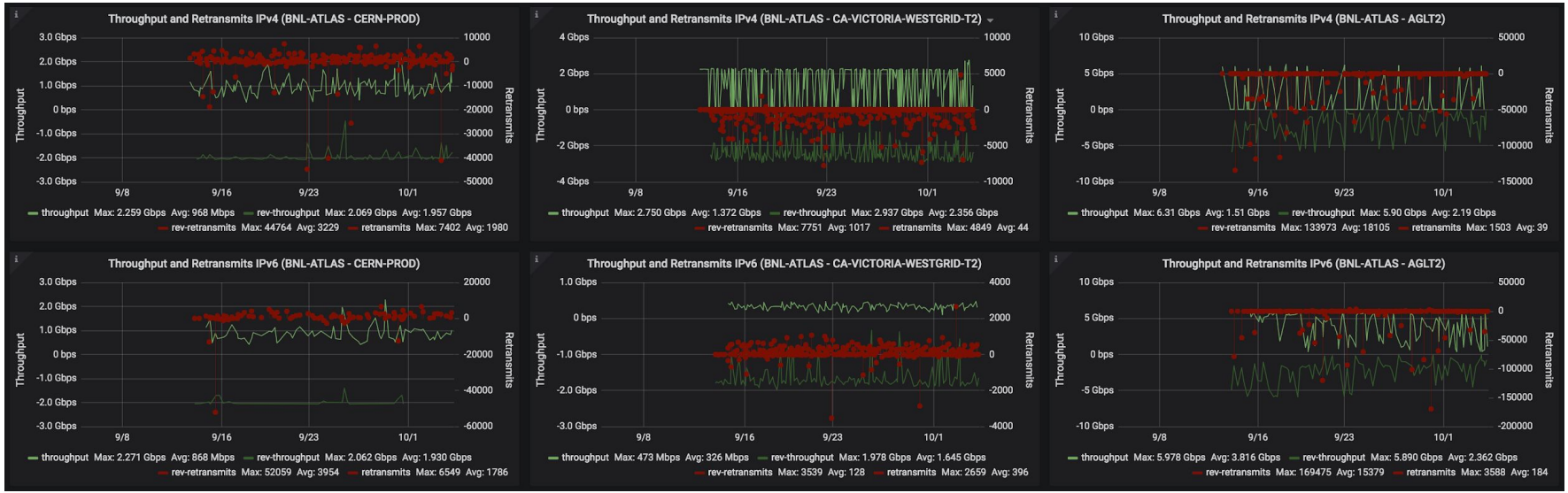


Grafana - perfSONAR dashboard



- Now includes all WLCG sites that run perfSONAR
  - Additional work needed to better filter production nodes
- Added additional row that tracks RTT and number of hops as reported by traceroute/tracepath
- Can you spot the network issue(s) above ?

Grafana - IPv6 dashboard



- Added IPv6 dashboard
  - Side-by-side comparison btw. IPv4 and IPv6 performance
- Currently very sparsely populated - we suspect an potential bug in the pwa/psconfig - needs some additional debugging
- Due to performance limitations it was agreed that won't configure IPv6 latency tests

Platform Use

- **WLCG and OSG operations**
  - Baseline testing and interactive debugging for incidents reported via support unit
  - Regular reports at the WLCG operations coordination and WLCG weekly operations
  - Providing **Grafana dashboards** that help visualise the metrics
- Enabling analytical studies - data stored in the ATLAS Analytics platform
  - Providing an important source for network metrics (bandwidth, latency, path)
- **Cloud testing - HNSciCloud** - testing commercial cloud providers
  - Baselining and evaluating network performance
- HEPiX IPv6 WG
  - Now testing bandwidth and paths over IPv6
- Collaboration with other science domains deploying perfSONAR
  - E.g., US Universities, Pittsburgh Supercomputer Center, European Bioinformatics Institute
  - Also close collaboration with (N)RENs who provide LHCONE perfSONAR coverage

WLCG Network Throughput Support Unit

Support channel where sites and experiments can report potential network performance incidents:

- Relevant sites, (N)RENs are notified and perfSONAR infrastructure is used to narrow down the problem to particular link(s) and segment. Also tracking past incidents.
- Feedback to WLCG operations and LHCOPN/LHCONE community

Most common issues: MTU, MTU+Load Balancing, routing (mainly remote sites), site equipment/design, firewall, workloads causing high network usage

As there is no consensus on the MTU to be recommended on the segments connecting servers and clients, LHCOPN/LHCONE working group was established to investigate and produce a recommendation. (See coming talk :))

T2_PK_NCP case

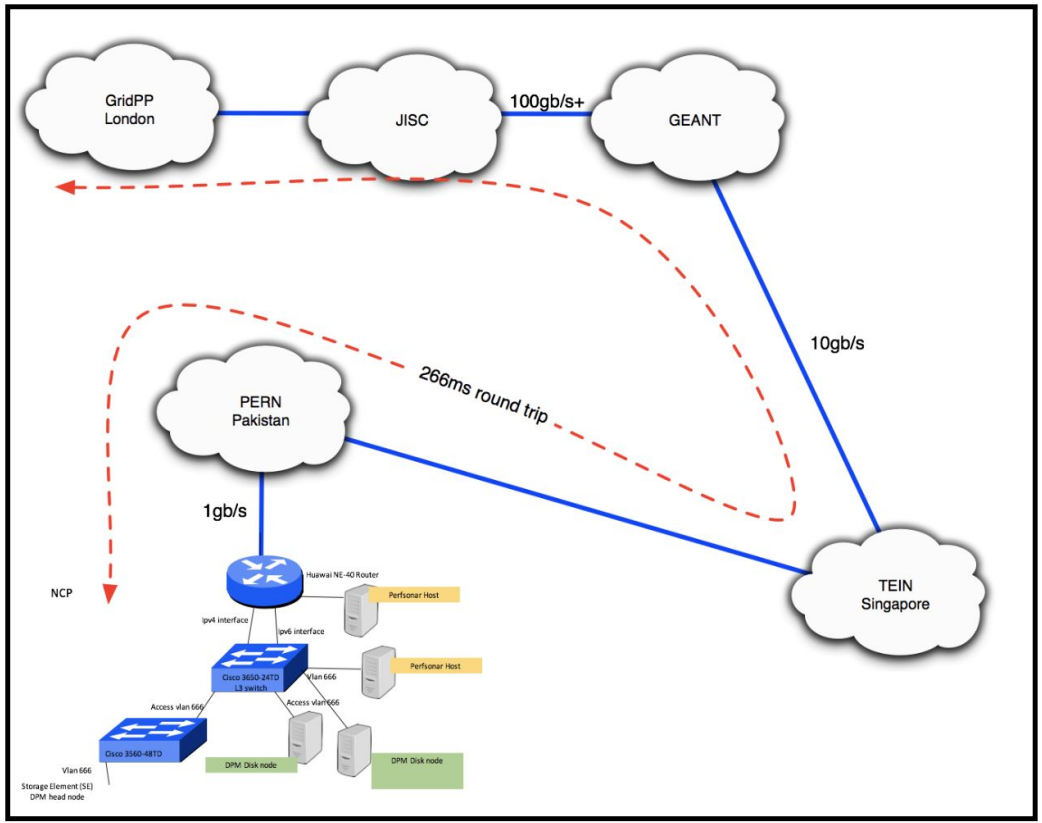


Figure 1: Performance path for data between NCP LHC Tier 2 data collector and data source at Tier1 site at Queen Mary University, London.

Very complex case - asked GlobalNOC Performance Engagement Team for help

https://docs.google.com/document/d/1HHhK9t4PpYPzZOfJUAhupRAodhPT6HNIZi6J8ljpBtw/edit#

NCP is considering commercial route (~100ms latency difference) but this is **not** a good precedent!

IHEP/JINR case

IHEP is currently routed to/from JINR via Internet2 - additional latency negatively impacts performance

```
[root@perfsonar-bw ~]# bwtraceroute -4 -c t1-pfsn1.jinr-t1.ru -s perfsonar.ihep.ac.cn
bwtraceroute: Using tool: traceroute
bwtraceroute: 18 seconds until test results available

SENDER START
traceroute to 159.93.229.150 (159.93.229.150), 30 hops max, 60 byte packets
1 202.122.32.161 (202.122.32.161) 0.363 ms 0.370 ms 0.376 ms
2 192.168.1.25 (192.168.1.25) 1.910 ms 1.877 ms 1.872 ms
3 8.195 (159.226.254.118) 0.462 ms 0.463 ms 0.437 ms
4 8.131 (159.226.254.61) 1.201 ms 8.204 (159.226.254.65) 1.394 ms 8.131 (159.226.254.61) 1.527 ms
5 8.193 (159.226.254.38) 2.255 ms 2.241 ms 2.225 ms
6 210.25.189.65 (210.25.189.65) 1.231 ms 1.386 ms 1.419 ms
7 210.25.187.46 (210.25.187.46) 1.671 ms 2.108 ms 1.878 ms
8 210.25.187.41 (210.25.187.41) 0.871 ms 0.846 ms 0.794 ms
9 210.25.189.50 (210.25.189.50) 158.573 ms 158.458 ms 158.433 ms
10 210.25.189.134 (210.25.189.134) 177.674 ms 177.664 ms 177.651 ms
11 ae-1.4079.rtsw.salt.net.internet2.edu (162.252.70.115) 176.799 ms 176.800 ms 176.842 ms
12 ae-5.4079.rtsw.kans.net.internet2.edu (162.252.70.144) 212.856 ms 212.738 ms 212.703 ms
13 ae-3.4079.rtsw.chic.net.internet2.edu (162.252.70.140) 209.806 ms 209.674 ms 209.689 ms
14 et-0-0-0.4079.rtsw.star.net.internet2.edu (162.252.70.117) 221.139 ms 211.056 ms 211.044 ms
15 198.71.46.34 (198.71.46.34) 212.548 ms 212.557 ms 212.452 ms
16 we.starlight.nlight.ams.as59624.net (144.206.255.182) 370.918 ms 360.758 ms 366.048 ms
17 they.jinr.bgp.as59624.net (144.206.254.34) 376.366 ms 371.588 ms 376.272 ms
18 t1-pfsn1.jinr-t1.ru (159.93.229.150) 355.471 ms 361.170 ms 361.141 ms
```

Looking for alternative routes would be a good topic for Asia Tier Forum

IRIS-HEP

The Institute for Research and Innovation in Software in High Energy Physics (**IRIS-HEP**) project has been funded by National Science Foundation in the US as grant OAC-1836650 as of 1 September, 2018. (Kick-off mtg tomorrow at UC!)

The institute focuses on preparing for **High Luminosity (HL) LHC** and is funded at **$5M** / year for 5 years. There are three primary development areas:

- Innovative algorithms for data reconstruction and triggering;
- Highly performant analysis systems that reduce `time-to-insight' and maximize the HL-LHC physics potential;
- Data organization, management and access systems for the community's upcoming Exabyte era.

The institute also funds the **LHC part of Open Science Grid, including the networking area** and will create a new integration path (the Scalable Systems Laboratory) to deliver its R&D activities into the distributed and scientific production infrastructures. **Website for more info:** http://iris-hep.org/

The NSF funded SAND Project

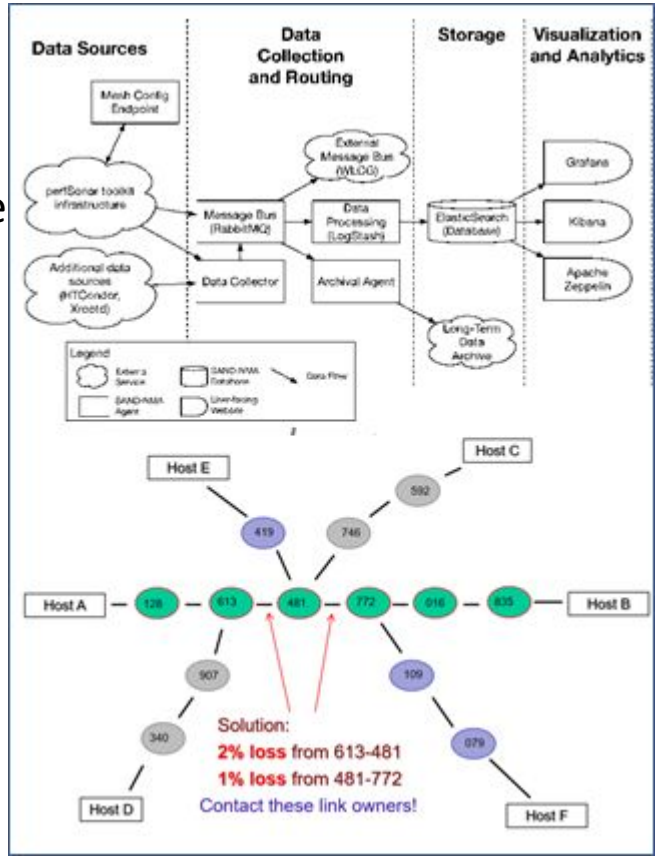
SAND: Service Analysis and Network Diagnosis

This is a newly funded NSF project (award #1827116) focusing on combining, visualizing, and analyzing disparate network monitoring and service logging data

It will **extend** and **augment** the **OSG networking** efforts with a primary goal of extracting useful insights and metrics from the wealth of network data being gathered from perfSONAR, FTS, R&E network flows and related network information from HTCondor and others.

Website https://sand-ci.org/ (Project just started in September 2018 and will last 2 years)

PI: **Brian Bockelman**, **Co-PIs**: **Shawn McKee, Rob Gardner**



perfSONAR near-term releases

- perfSONAR 4.2 (Q1 2019)
  - **GridFTP plug-in** - Significant interest from NRP community and others.
  - Measurement pre-emption - Easier for diagnostic tests to get a slot on busy hosts
  - **Additional pSConfig utilities** - Continuing to make meshes easier to build and manage through command-line and graphical interface
  - Lookup Service improvements - Bulk renewals and record signing

- perfSONAR 4.3 (Q3 2019)
  - User Interface and Visualization Strategy - Seek to improve user experience and operational efficiency within development team by consolidating code
  - **pScheduler Resource Pooling** - Better management of resources like ports, potential gains in environments like Kubernetes where ports may be constrained
  - **Esmond Updates** - Option to run using pure postgresql (no cassandra)

WG near-term Plans

- Finalise campaign to update perfSONARs to CC7
  - Followed up by campaign to fix configuration issues that prevent running regular testing activities
  - What about LHCOPN/LHCONE specific instances (in GEANT, I2, ESnet)?
- pS Lookup was migrated to GCP; work with pS-devs to set cloud policy
- IPv6
  - Migrate all throughput and traceroute meshes to run on both IPv4 and IPv6
  - Create dedicated IPv6 latency mesh that will be used and **retire** dual-stack mesh
- Network Path Analysis projects (UROP and hourly students from SAND)
- CERN plans to deploy 40Gbps/100Gbps perfSONAR
  - Initial testing and commissioning of 40Gbps/100Gbps testing
  - Please let us know if you plan to deploy high bandwidth nodes
- Further improve Grafana dashboards
  - Add (N)REN endpoints/sites

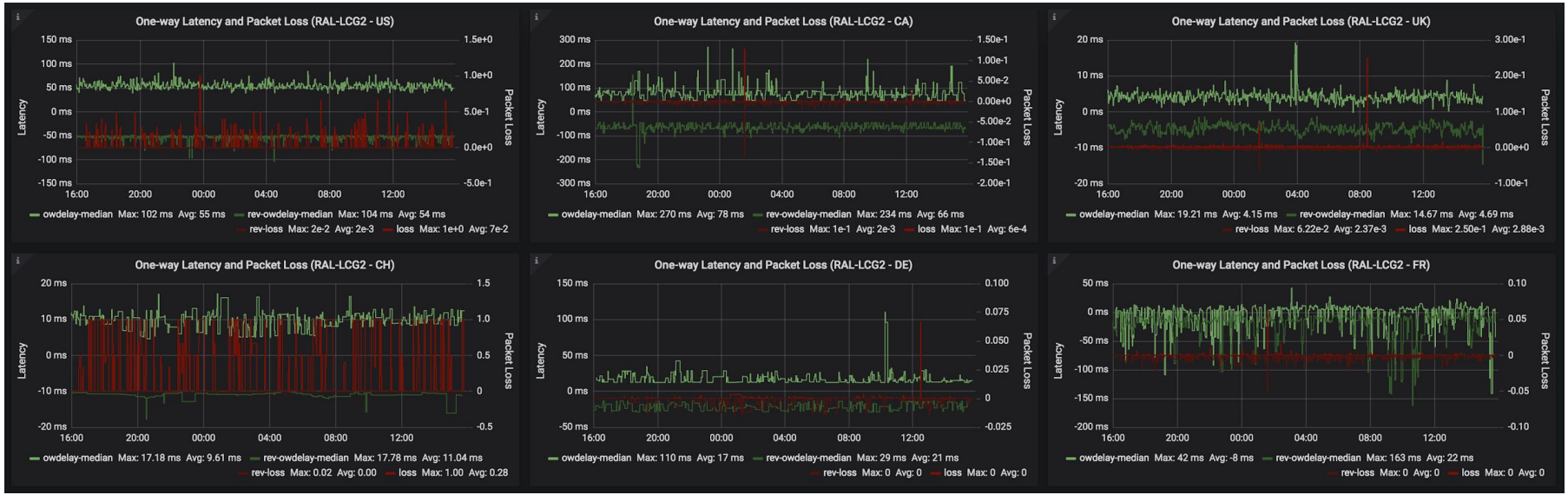
Summary

- OSG in collaboration with WLCG are operating a comprehensive network monitoring platform
- Platform has been used in a wide range of activities from core OSG/WLCG operations to Cloud testing and IPv6 deployment
- Providing feedback to LHCOPN/LHCONE, HEPiX, WLCG and OSG communities
- Next version of perfSONAR will enable additional functionality as well as improve overall stability and performance
- IRIS-HEP and SAND are starting and will contribute to the R&D in the network area
- Further analytical studies are planned to better understand our use of networks and how it could be improved

References

- OSG/WLCG Networking Documentation
  - https://opensciencegrid.github.io/networking/
- perfSONAR Stream Structure
  - http://software.es.net/esmond/perfsonar_client_rest.html
- perfSONAR Dashboard and Monitoring
  - http://maddash.opensciencegrid.org/maddash-webui
  - https://psetf.opensciencegrid.org/etf/check_mk
- perfSONAR Central Configuration
  - https://psconfig.opensciencegrid.org/
- Grafana dashboards
  - http://monit-grafana-open.cern.ch/
- ATLAS Analytics Platform
  - https://indico.cern.ch/event/587955/contributions/2937506/
  - https://indico.cern.ch/event/587955/contributions/2937891/

Grafana - Inter-Regional Latency Dashboard



Networking Challenges

There are number of challenges in the networking, which will require improved collaboration with other sciences as well as HEP-focused R&D:

- Capacity/share for data intensive sciences
  - No issues wrt available technology, however
  - What if N more HEP-scale science domains start competing for the same resources ?
- Remote data access proliferating in the current DDM design
  - Promoted as a way to solve challenges within experiment's DDM
  - Different patterns of network usage emerging
    - Moving from large streams to a mix of large and small frequent event streams
- Integration of Commercial Clouds
  - Impact on funding, usage policies, security, etc.
- Technology evolution
  - Software Defined Networking (SDN)/Network Functions Virtualisation (NFV)

Network Evolution Areas

The following are some of the key areas for HEP Networking R&D:

- Improving efficiency of data transfers
  - TCP BBR - version 2 is in the works with promising improvements
  - Exploring alternative protocols for transfers (UDP)
- Caching
  - Data caches co-located with network hubs in a similar way as on commercial CDNs
- Federations/Clouds
  - Overlay networks spanning multiple domains
  - Multi-clouds - expanding DC networking via L3VPNs
- Technology
  - SDN/NFV approaches - currently looked at by HEPiX NFV WG
    - Compute - Agile service delivery on Cloud Infrastructures (OpenStack, Kubernetes)
    - Data Transfers - Network resource optimisation - dynamically optimising the network based on its load and state (more in Shawn/Ilija)
  - SD-WAN approaches - https://www.mode.net/

Importance of Measuring Our Networks

- End-to-end network issues are difficult to spot and localize
  - Network problems are multi-domain, complicating the process
  - Performance issues involving the network are complicated by the number of components involved end-to-end
  - Standardizing on specific tools and methods focuses resources more effectively and provides better self-support.
- **Network problems can severely impact experiments workflows and have taken weeks, months and even years to get addressed!**
- perfSONAR provides a number of standard metrics we can use
  - Latency, Bandwidth and Traceroute
  - These measurements are critical for network visibility
- **Without measuring our complex, global networks we wouldn't be able to reliably use those network to do science**