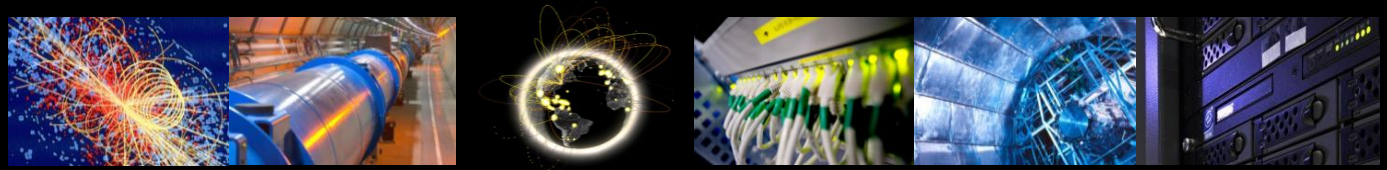
OSG/WLCG Networking Update

Marian Babik[1], Shawn McKee[2]
[1] CERN, [2] University of Michigan



WLCG
Worldwide LHC Computing Grid

Open Science Grid

Outline

- Activity and News
- Collaborating Projects
- Platform and Services Updates/Additions
- 100 Gbps Testing
- Platform Use
- Plans
- Summary

OSG/WLCG Networking Activities

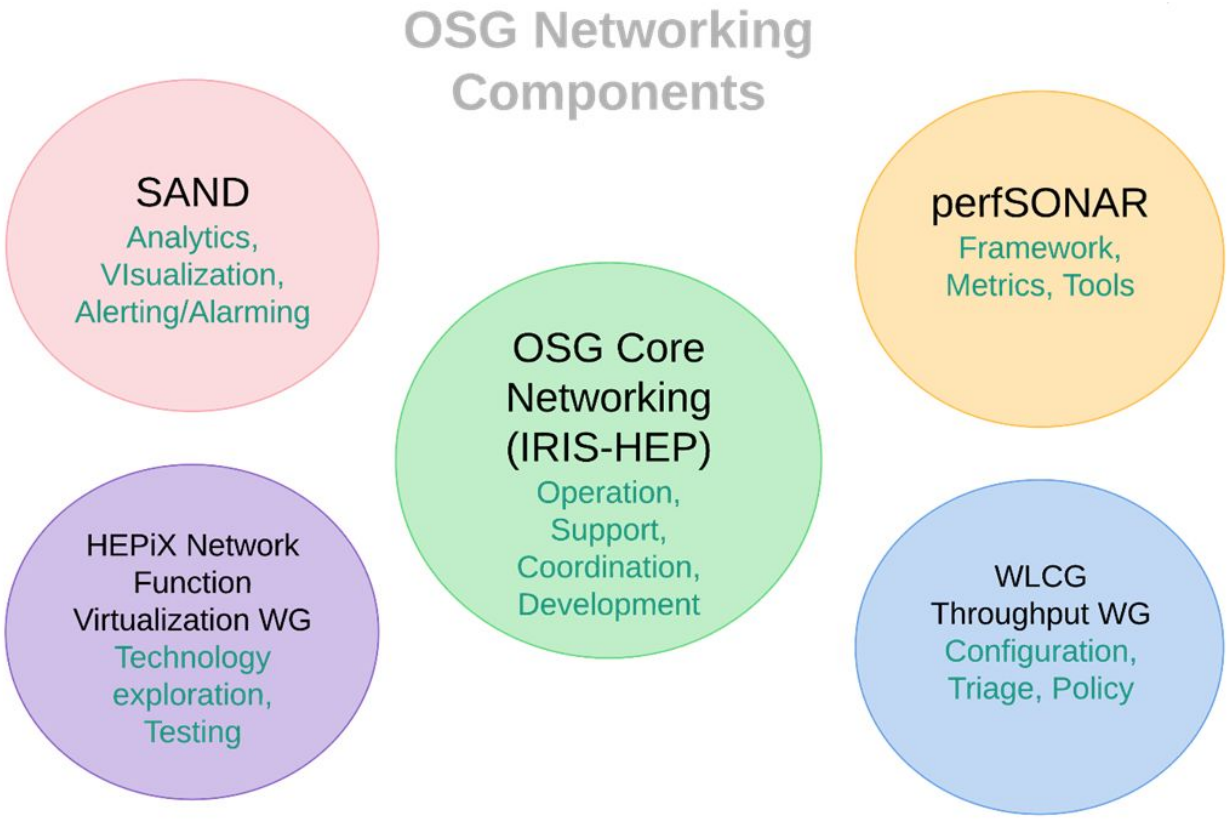
- OSG has entered its 7th year of supporting WLCG/OSG networking:
  - Assisting its users and affiliates in identifying and fixing network bottlenecks
  - **Developing and operating a comprehensive Network Monitoring Platform**
  - Improving our ability to manage and use network topology and network 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

Networking Projects

There are now 4 coupled projects around the core OSG Network Area

1. **SAND** (NSF) project for analytics
2. **HEPiX** NFV WG
3. **perfSONAR** project
4. **WLCG** Throughput WG

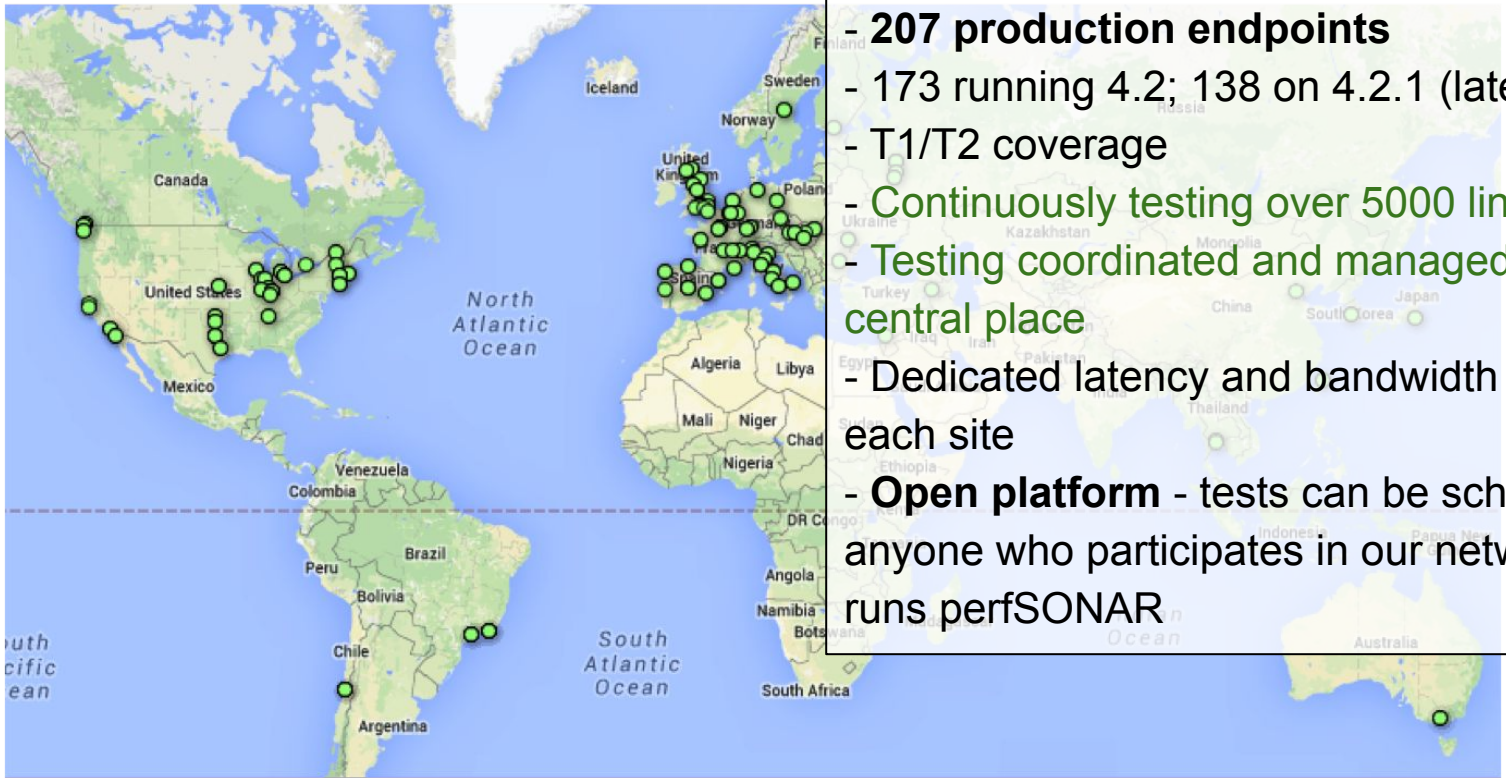


perfSONAR News

perfSONAR is 4.2 was released (4.2.2 is the latest release)

- New plugins
  - GridFTP plug-in - Significant interest from NRP community and others.
  - **Test schedule pre-emption** - Easier for manual 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
- 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 (for “testpoint” deployment)
- SL6 no longer supported
  - Our recommendation: **reinstall** with CentOS7 ASAP; **don’t worry about saving data**

perfSONAR deployment



261 Active perfSONAR instances

- **207 production endpoints**
- 173 running 4.2; 138 on 4.2.1 (latest)
- 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

LHCOPN - 14th Oct 2019

OPN Mesh Config - OPN Latency - Loss

Loss rate is <= 0.001% | Loss rate is > 0.001% | Loss rate is >= 0.1%

⚠ Found a total of 2 problems involving 2 hosts in the grid

OPN Mesh Config - OPN IPv4 Bandwidth - Throughput

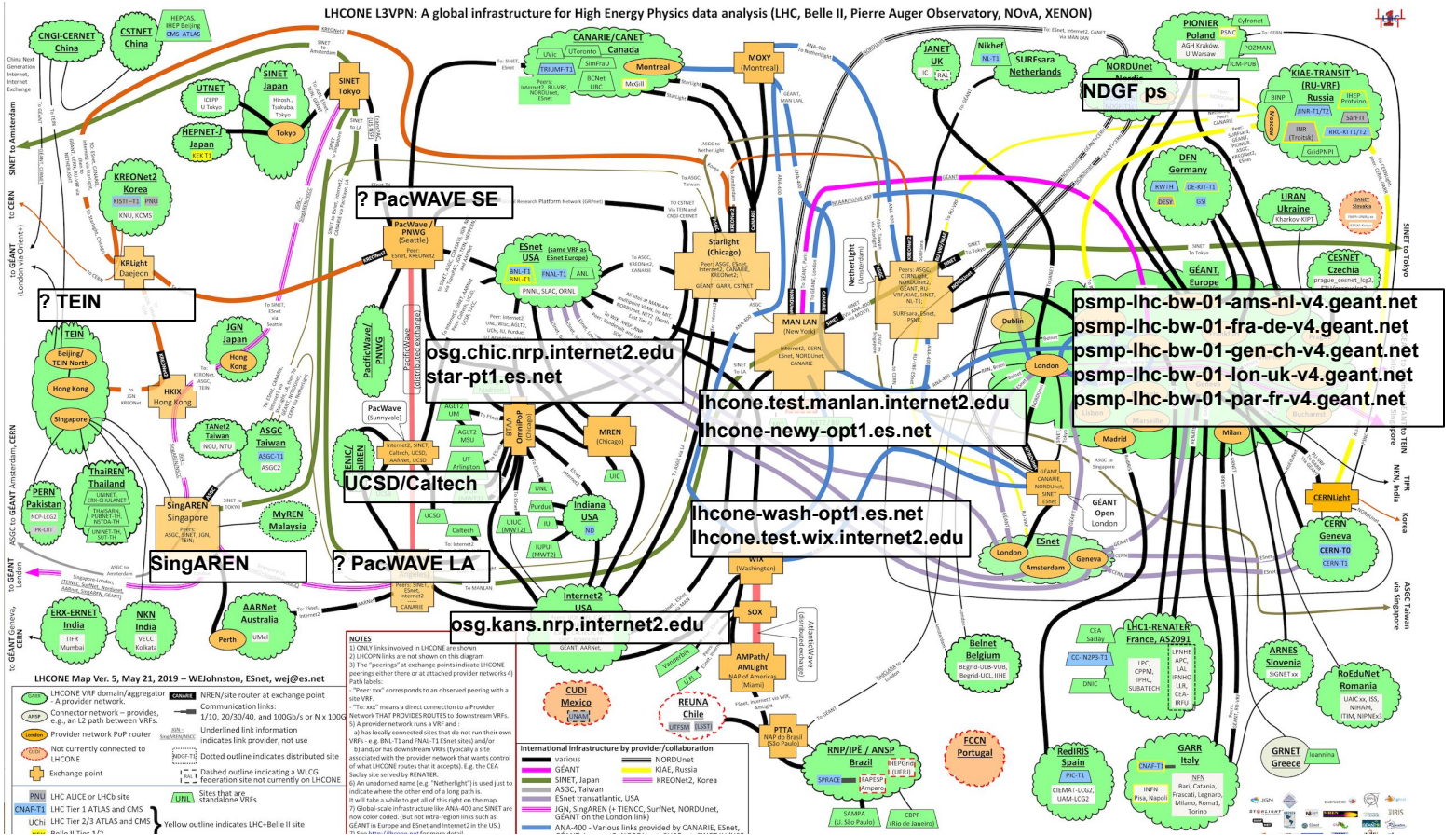
Throughput >= 1Gbps | Throughput < 1Gbps | Throughput <= .5Gbps

⚠ Found a total of 6 problems involving 4 hosts in the grid

New LHCONE mesh proposal

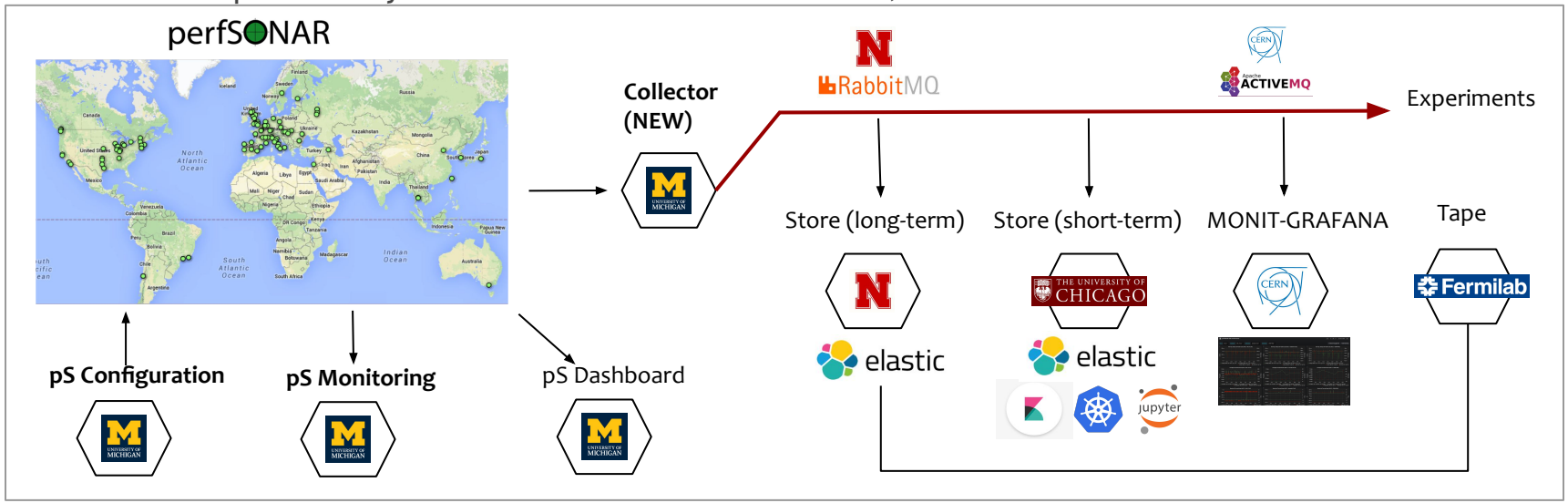
- The main purpose for LHCONE mesh is to check connectivity/performance (possibly reachability) within LHCONE
  - Important to have endpoints to test within R&Es (on LHCONE)
    - We have good coverage in some regions, but not others. TBD (next slide)
  - Testing both directions (from R&Es to sites and vice versa) would be optimal, but it's technically challenging as we don't control R&E endpoints
    - We have relied on perfSONAR feature called bi-directional testing (test B->A from A), but it hasn't been working as expected and future support is unclear
  - Moving to uni-directional testing means we test from sites to R&E endpoints
- Proposal agreed at the last LHCOPN/LHCONE meeting is to replace the current mesh with one that does uni-directional tests from set of selected sites to set of R&E endpoints
  - In some cases where R&E endpoints don't exist we can use particular site close to a network hub
  - Initially over IPv4, IPv6 needs IPv6 R&E endpoints

New LHCONE mesh proposal



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

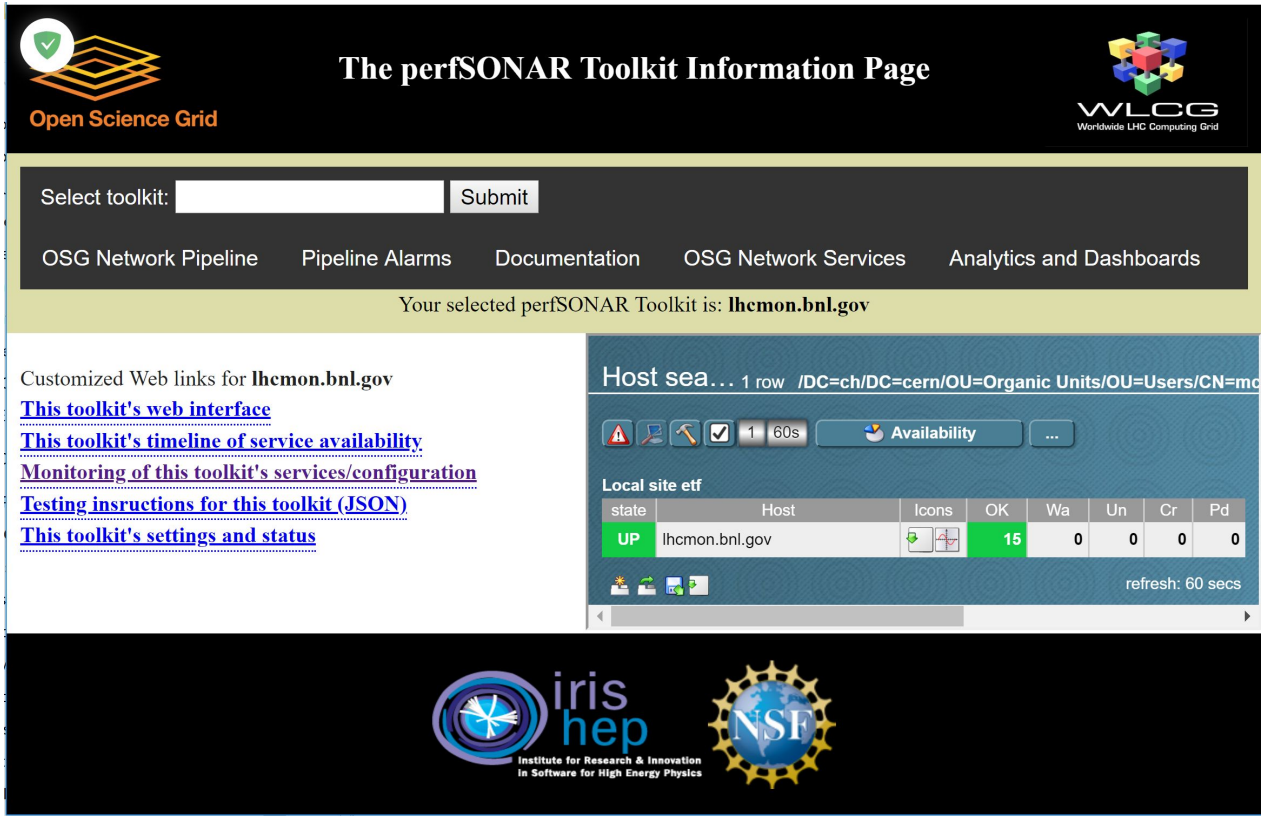


Toolkit Info Web Page

Web page to help toolkit owners in managing and fully utilizing the various resources and services OSG provides.

We now have a **prototype** running that we plan to evolve based upon feedback:

https://toolkitinfo.opensciencegrid.org/



100Gbps Testing

- 100Gbps NICs are now affordable as are 40/100Gbps ToR switches
  - Significant interest in the perfSONAR community to deploy/test 100Gbps
- LHCOPN/LHCONE 100Gbps perfSONAR infrastructure:
  - SARA, CSCS already have 100 Gbps; CERN has now multiple 40Gbps and 100Gbps, BNL has 80Gbps (plans 100Gbps), KIT has 100Gbps in testing
- Some initial tests performed with the following test setup
  - Hosts:
    - Supermicro X10DRi, Intel Xeon E5 2620, 2 sockets, 8 cores each
    - CentOS 7.6 running kernel 3.10.0-957.12.2.el7.x86_64
    - Mellanox ConnectX-5 in 16xPCIExpress (using stock drivers)
  - Topology:
    - LAN had 120usec RTT; irqbalance, tuned, numad were off
    - Core affinity was set to cores 8-15 (on NUMA node 1 - closest to the NICs)
    - All testing via IPv4/9000 MTU unless otherwise stated
    - WAN testing used 100Gbps LHCOPN link to SARA

100Gbps Results

Best TCP single flow results:

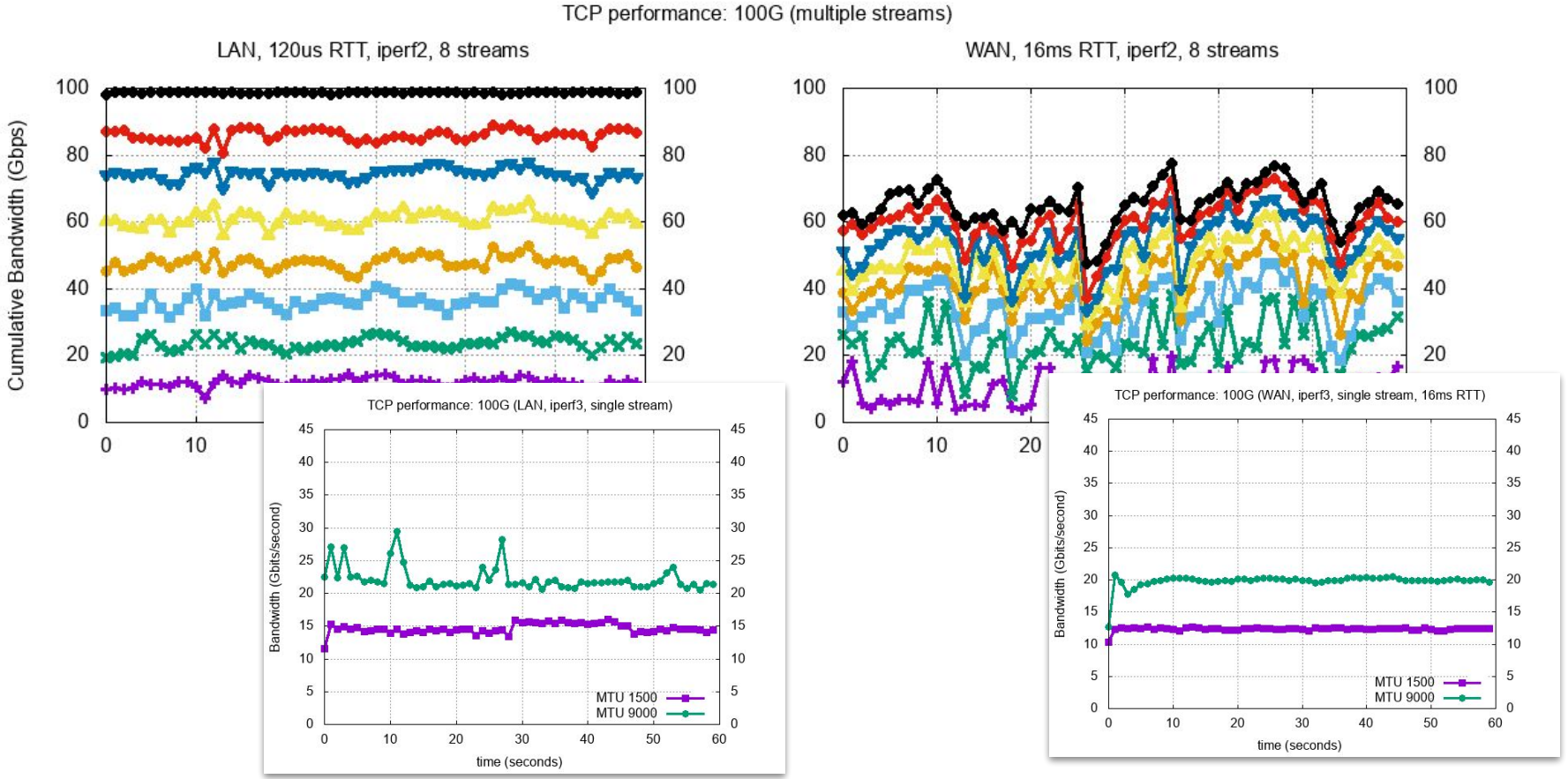
- **LAN**: 22.19 Gbps (25.6* Gbps with tuning)
- **WAN** (16ms RTT): 19.8 Gbps (31.95 Gbps with tuning)
- All tests done using regular perfSONAR pscheduler commands
  - No special setup, no remote assistance required

Best TCP multiple flows results:

- **LAN**: 98.7 Gbps (8 streams)
- **WAN** (16ms RTT): 65.1 Gbps (8 streams, peak 77.4 Gbps)

Others have reported up to **85 Gbps** single stream - most likely bottleneck is single core CPU performance in our system

100Gbps LAN vs WAN, single vs multiple flows



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
  - **perfSONAR** part of the standard benchmarking tools, developed as part of OCRE project
- UK and FR have dedicated regional meshes running
- **Collaboration with GridPP and CCSTAR communities**
- OCRE Cloud testing (https://github.com/cern-it-efp/OCRE-Testsuite/)
- HEPiX IPv6 WG - now testing bandwidth and paths over IPv6
- Collaboration with other science domains deploying perfSONAR
  - E.g., US Universities (via CC*), US HPCs , EBI, etc.

Network Operations

- Sites experiencing network issues should **first** contact their local network team or directly their regional and backbone (R&E) providers
- A group focusing on helping sites and experiments with network performance using perfSONAR - **WLCG Network Throughput**
  - Please include as many details as possible (include any existing tickets with R&Es)
  - For list of existing and recently resolved issues see this link

Operations

- **LHCONE operations** - support for establishing and operating LHCONE infrastructure - regular meetings and support mailing list

- **LHCOPN/LHCONE community** - organised bi-annually - good place to meet R&Es and discuss architecture and plans
  - Next one is at CERN (https://indico.cern.ch/event/828520/)

Architecture, Infrastructure

perfSONAR near-term releases

- Plan is to re-scope the release process - do smaller and more frequent releases
- perfSONAR 4.3 (Q1 2020)
  - Transition to Py3 (https://pythonclock.org/)
- perfSONAR 4.4 (Q2 2020)
  - Add support for **other archiving backends** (non Esmond) - like Grafana and ElasticSearch
  - Also discussing:
    - 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

WG near-term Plans

- Complete campaign to update perfSONARs to CC7 and latest release
  - Now moving on to T2s
- Finalise re-organisation of the LHCONE mesh
  - Add uni-directional testing to perfSONARs hosted by R&Es on LHCONE
- **Add traceroutes between all latency instances**
  - This is to extend the existing coverage; and make it easier to correlate path and latencies
- EU projects ARCHIVER and ESCAPE are planning to setup tests
- **ALICE, StashCache and CC* communities already established**
- **100G testing** - planning regular test mesh as well as network experiments/evaluations (TCP BBR2, IPv4/IPv6, etc.) - subscribe at:
  - http://cern.ch/simba3/SelfSubscription.aspx?groupName=wlcg-perfsonar-100g
- Planning new features in monitoring
  - Self-service notifications - enable alerts from check_mk by joining specific community
  - Provide generic container that starts our check_mk monitoring for any community
- Working closely with the **SAND (**https://sand-ci.org/) project on analytics

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 will contribute to the operations and R&D in the network area
- Further analytical studies are planned to better understand our use of networks and how it could be improved
- More on SAND by Shawn later today

Acknowledgements

We would like to thank the **WLCG**, **HEPiX**, **perfSONAR** and **OSG** organizations for their work on the topics presented.

In addition we want to explicitly acknowledge the support of the **National Science Foundation** which supported this work via:

- OSG: NSF MPS-1148698
- IRIS-HEP: NSF OAC-1836650
- SAND: NSF CC* OAC-1827116

References

- OSG/WLCG Networking Documentation
  - https://opensciencegrid.github.io/networking/
  - https://toolkitinfo.opensciencegrid.org/
- perfSONAR Stream Structure (API)
  - 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/