OSG Networking: Status, Collaborations and Plans

Shawn McKee

HOW Meeting

Jefferson Labs, March 20, 2019

Newport News, VA, US

Introduction

- OSG has been working on networking for its constituents and collaborators for more than 6 years
  - We have a complete infrastructure to reliably measure, gather and store important network metrics
- In this presentation I want to cover the current status, related collaborations and our near-term plans for the OSG networking area
- **In addition I want to provide sites some things to consider as they think about how they will be updating their infrastructure.**

Current Network Status

Regarding our networking, I want to state a few things about our current status up front

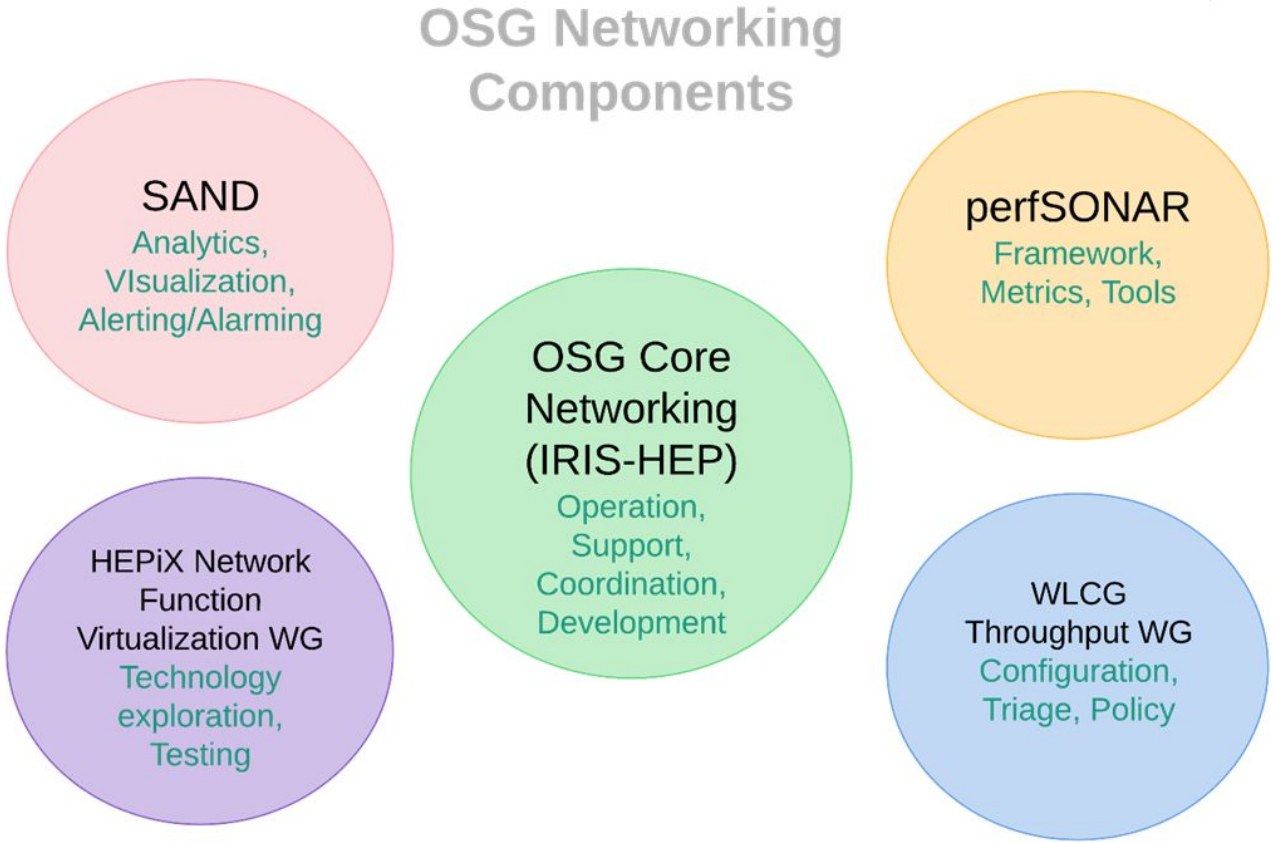
- Our networks have performed very well for our community
- Most users are happy with the networking we have
- Primary concerns exist around our ability to fully utilize **existing** networks
- Visibility is key to understanding, maintaining and fixing our networks

So there continues to be **near-term** work regarding our networking in optimizing, monitoring and fixing network problems, **but** we should also think longer term regarding how the situation may evolve and what that might mean for us.

OSG Networking Collaboration Overview

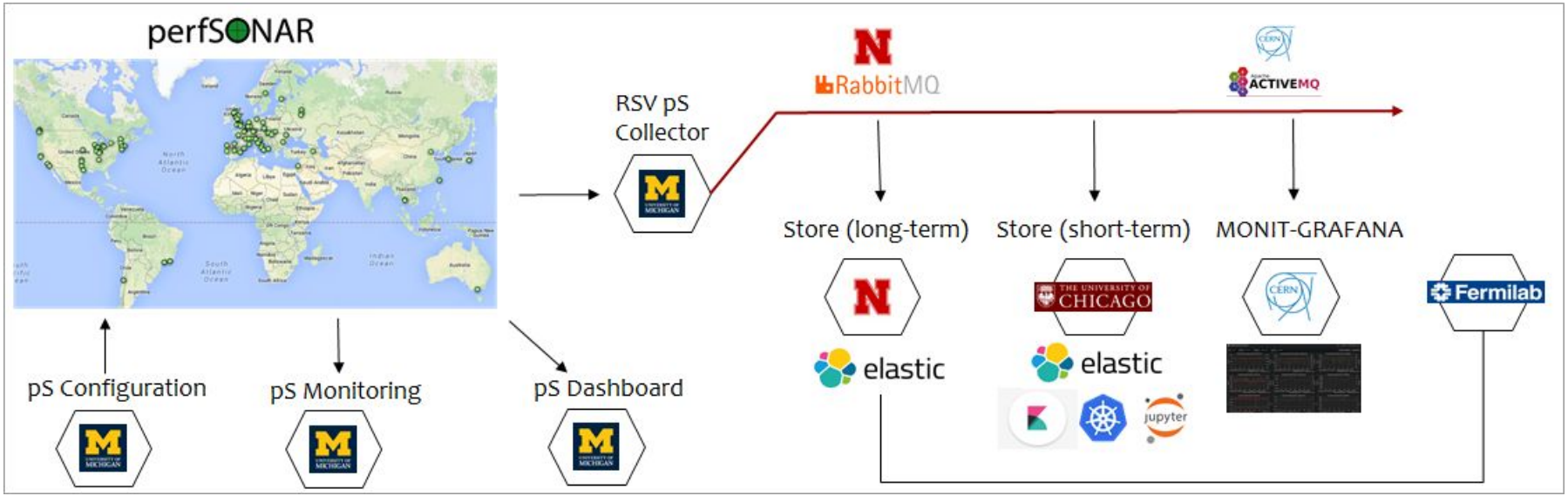
There are 4 coupled projects around the core OSG Net Area

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



The OSG Network Monitoring Data Pipeline

- **Collects, stores, configures and transports all network metrics**
  - Distributed deployment - operated collaboratively
- **All perfSONAR metrics available via API, live stream or on our analytics platforms**
  - Complementary metrics (ESnet, LHCOPN traffic, FTS data) available on same platforms

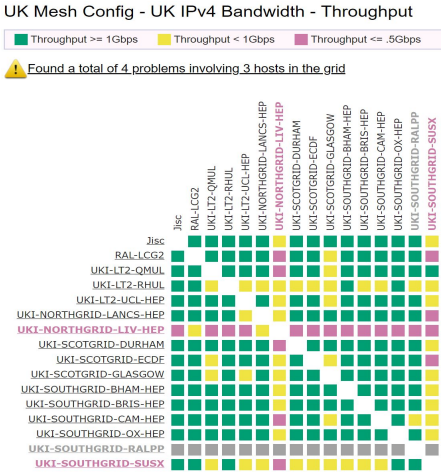


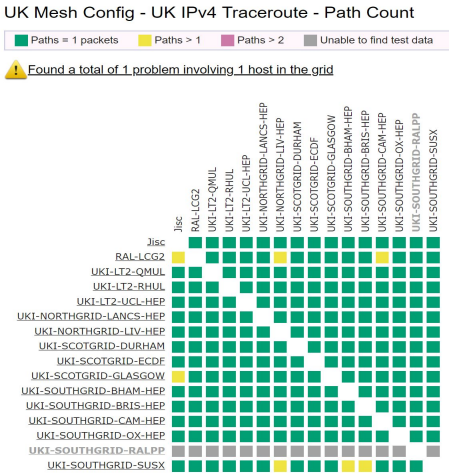
Network Monitoring Considerations

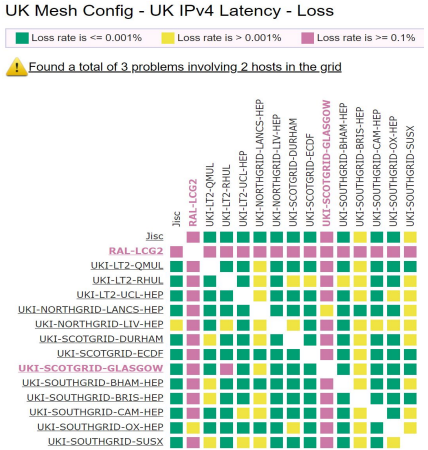
- We need network visibility to understand performance, find problems and enable orchestration
- **All sites should have deployed perfSONAR and have a plan to keep the hardware and software updated**
  - The recommendation is to provide two instances: latency and throughput (which could be on the same server with at least two NICs)
  - The perfSONAR instances should be (co)located with your sites STORAGE, network-wise
  - The throughput instance should use the same NIC capacity as your storage servers
  - Additional perfSONAR instances can be helpful for identifying LAN issues
- https://opensciencegrid.org/networking/perfsonar/installation/#perfsonar-installation-guide

Campaign to Upgrade perfSONAR

- We have recently begun a campaign encouraging sites to upgrade their perfSONAR deployments, both **hardware** and **software**
  - Many sites deployed their perfSONAR systems >5 years ago and the hardware is often just at the minimum (or even below) what is required to run the tests we need
  - With perfSONAR 4.1, all sites running CentOS 6.x need to reinstall using CentOS 7.x since perfSONAR no longer support CentOS 6.x
- It is possible to get robust reliable network metrics using perfSONAR 4.1+ reasonable hardware.
  - Duncan Rand has really helped get the UK sites in shape:







Example perfSONAR Config: Dell R240

Trusted Platform Module (TPM)
Trusted Platform Module 2.0
Chassis Configuration
3.5" Chassis with up to 4 Hot Plug Hard Drives
Processor
Intel® Xeon® E-2146G 3.5GHz, 12M cache, 6C/12T, turbo (80W)
Memory DIMM Type and Speed
2666MT/s UDIMMs
Memory Capacity
(2) 16GB 2666MT/s DDR4 ECC UDIMM
RAID/Internal Storage Controllers
PERC H330 RAID Controller, Adapter, Full Height
Hard Drives
1.2TB 10K RPM SAS 12Gbps 512n 2.5in Hot-plug Hard Drive, 3.5in
Additional Network Cards
On-Board Broadcom 5720 Dual Port 1Gb LOM
Embedded Systems Management
iDrac9, Express

Web price $2451. This system missing 10G+ NIC options. Need 10G SFP+ option here

Internal Optical Drive
DVD +/-RW, SATA, Internal for Hot Plug Chassis
Rack Rails
1U/2U 2/4-Post Static Rails
Bezel
No Bezel
Power Cords
C13 to C14, PDU Style, 12 AMP, 2 Feet (.6m) Power Cord, North America
Power Supply
Single, Cabled Power Supply, 250W
Password
iDRAC,Factory Generated Password
PCIe Riser
PCIe Riser with Fan with up to 1 LP, x8 PCIe + 1 FH/HL, x16 PCIe Slots
Hardware Support Services
3 Years, Basic Hardware Warranty Repair: 5x10 HW-Only, 5x10 Next Business Day Onsite
Deployment Services
No Installation

Network Operations

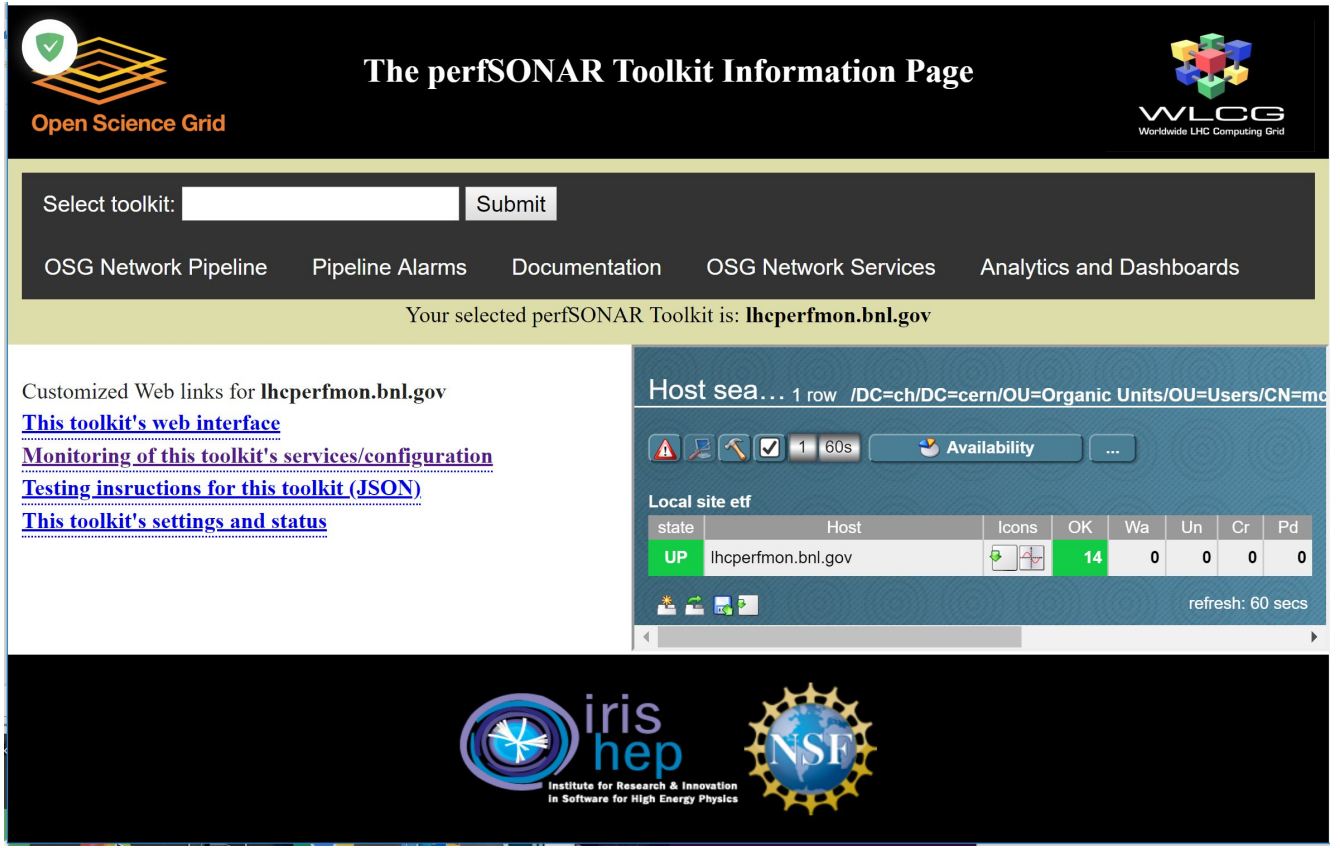
- Deployment of perfSONARs at most OSG/WLCG sites made it possible for us to see and debug end-to-end network problems
  - OSG gathers global perfSONAR data and making it available to collaborators
- We have a group focusing on helping sites and experiments with network performance using perfSONAR - WLCG Network Throughput
  - Reports of non-performing links are actually quite common
- Sites with assumed network problems can open a ticket with OSG to allow us to help diagnose the issue
- Sites experiencing **known** network issues should first contact their local network team, who can pursue the issues with the regional and backbone providers on their behalf

New Toolkit Info Web Page

At a prior OSG All-hands meeting we discussed providing a “front-end” web page the could 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 your feedback:

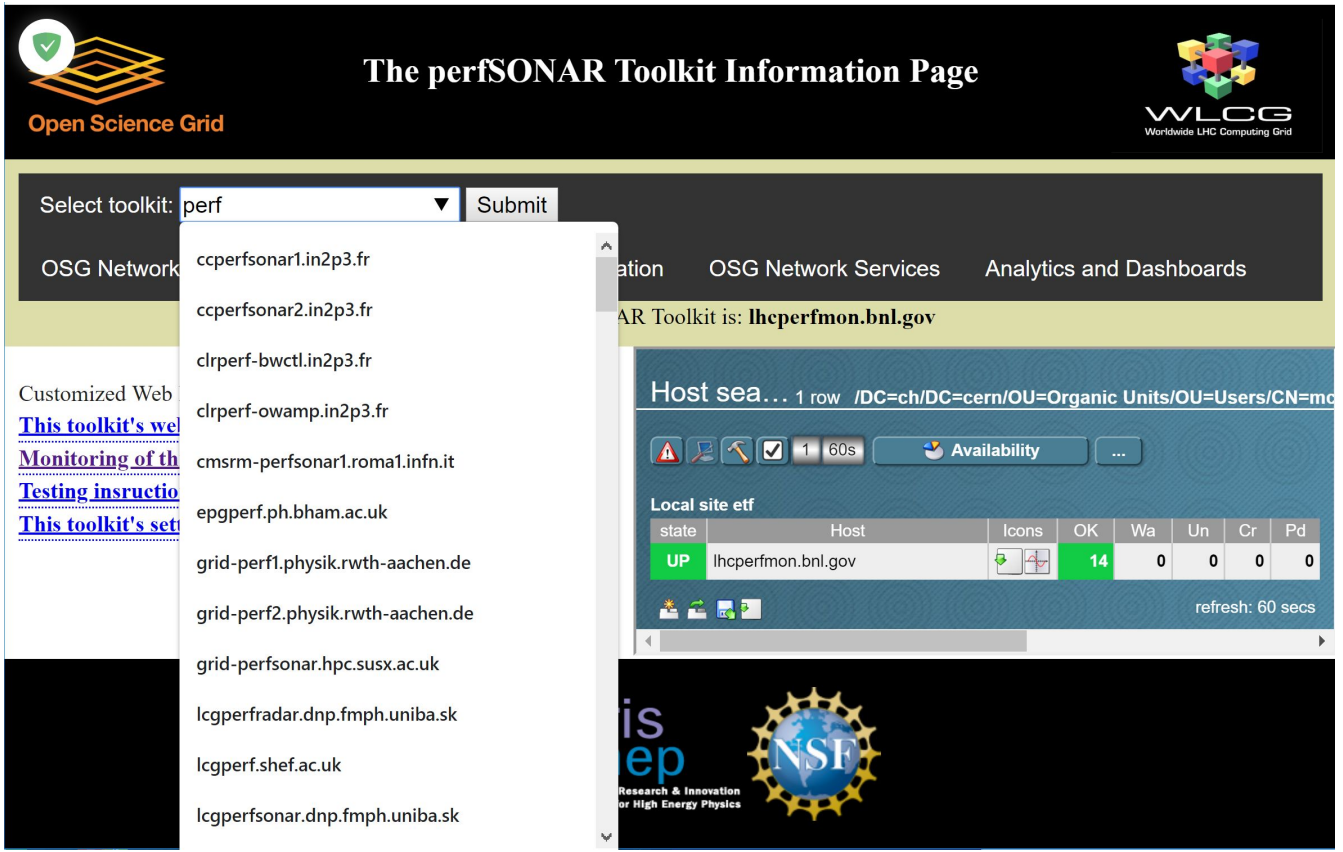
https://toolkitinfo.opensciencegrid.org/



Toolkit Info Web Page (2)

You can select any of the currently registered perfSONAR toolkit instances to get a set of customized links specific to that instance.

If you know part of the DNS name, you can start typing in the box to narrow the selection list.

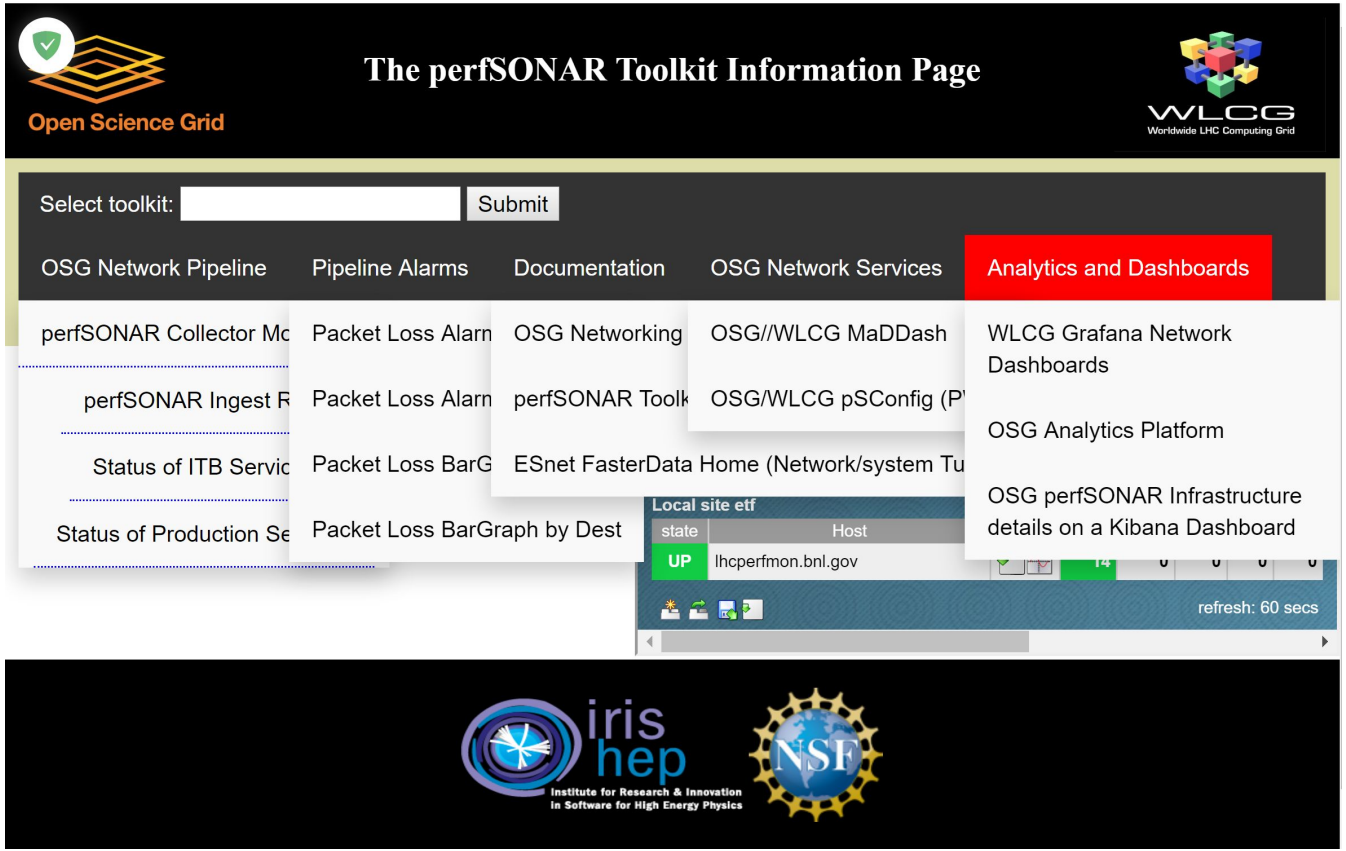


Toolkit Info Web Page (3)

There are additional menus setup to provide one-stop shopping to relevant services, documentation and dashboards.

We are also implementing “hover-over” text boxes to help describe the various links.

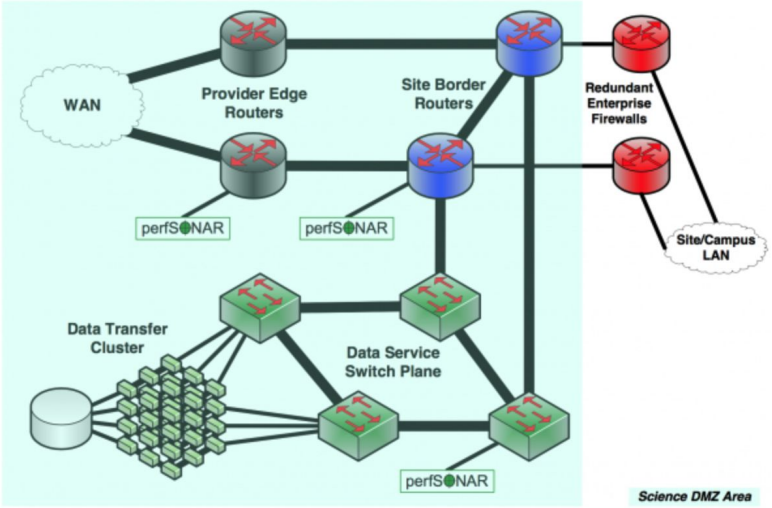
Please try it out and email me with your feedback!



Site Planning Considerations

Network Planning and Design

- Core capabilities to effectively deploy and support high-performance science applications include high bandwidth, advanced features, and capable gear that does not compromise on performance
- Science DMZ architecture is one of the examples that brings together performance, operational and security requirements
- Concerning network components:
  - Make sure routers/switches has sufficient buffer space to handle “fan-in” issues
  - Be wary of routers and switches that are over-subscribed (as this leads to consistent loss)
  - Look for devices that have flexible and performant ACLs support to eliminate need for stateful firewall which impact performance
  - If you’re planning to re-engineer your network, consider SDN/NFV approaches



Cloud Networking Considerations

- If you're planning to re-engineer your network stack and/or planning to deploy **OpenStack** or **Kubernetes** consider Software Defined Networking (SDN) approaches
- Orchestrating network together with Compute is possible and can work in production and at scale. When selecting an approach consider the following aspects:
  - **Multistack** - Connecting multiple orchestration stacks like Kubernetes, Mesos/SMACK, OpenShift, OpenStack and VMware
  - Networking and security across legacy, virtualized and containerized applications
  - Multistack and across-stack policy control, visibility and analytics
  - Multi-cloud support - DCI and Remote Compute
  - **Support for configuration and control** of the network equipment
  - Offloading of virtual networks via physical hardware (or via smart NICs)
- Some of the **key benefits**: self-service networks, auto-provisioning of VPNs, isolation (multi-domain), improved visibility and debugging of the networks, scalability (spanning services across multiple data centers), etc.

- HEPiX SDN/NFV Working Group was formed to bring together sites, experiments, (N)RENs and engage them in testing, deploying and evaluating network virtualization technologies

Software Defined Networks (SDN)

- Software Defined Networking (SDN) a set of new technologies enabling the following use cases:
  - **Automated service delivery** - providing on-demand network services (bandwidth scheduling, dynamic VPN)
  - **Clouds/NFV** - agile service delivery on cloud infrastructures usually delivered via Network Functions Virtualisation (NFV) - underlays are usually Cloud Compute Technologies, i.e. OpenStack/Kubernetes/Docker
  - **Network Resource Optimisation (NRO)** - dynamically optimising the network based on its load and state. Optimising the network using near real-time traffic, topology and equipment. This is the core area for improving end-to-end transfers and provide potential backend technology for DataLakes
  - **Visibility and Control** - improve our insights into existing network and provide ways for smarter monitoring and control
- Many different point-to-point efforts and successes reported within LHCOPN/LHCONE
  - **Primary challenge is getting end-to-end!**
- While it's still unclear which technologies will become mainstream, it's already clear that software will play major role in networks in the mid-term
  - Massive network automation is possible - in production and at large-scale
- HEPiX SDN/NFV Working Group was formed to bring together sites, experiments, (N)RENs and engage them in testing, deploying and evaluating network virtualization technologies

Network Security Considerations

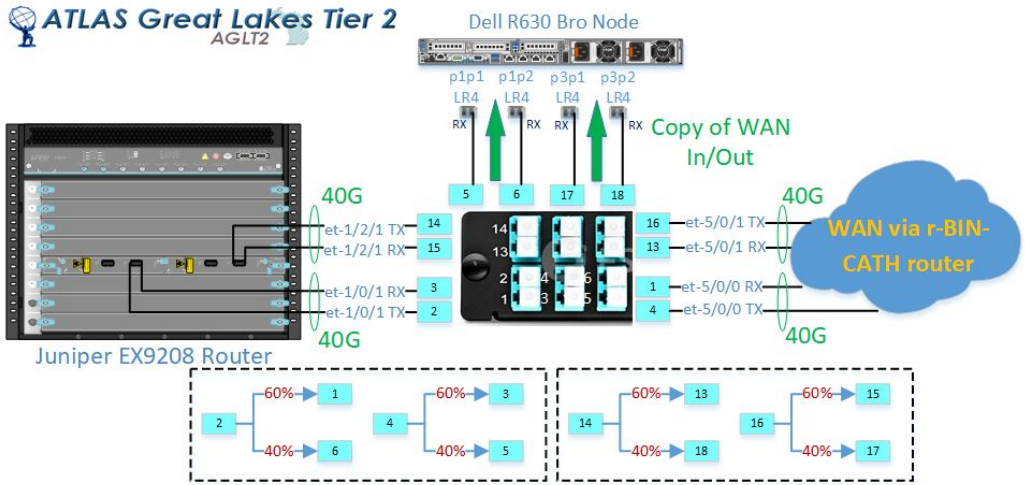
Enabling some additional security via traffic monitoring can help your site identify and defend against attacks

The **WLCG Security Operations Center** working group is advocating a solution involving **Zeek** (formerly "Bro"), **MISP** and **Elastiflow (**See David Crook's talk)

Info at http://wlcg-soc-wg-doc.web.cern.ch/wlcg-soc-wg-doc/

AGLT2_UM has an example deployment of an optical tap and Zeek monitoring for less than **$2K** and the use of a worker node.

As part of site planning it would be useful to consider deploying a similar capability as the network is upgraded.



OSG Networking Near-Term Activities

- We will be working closely with the **SAND (**https://sand-ci.org/) project to:
  - Improve the **robustness** and **efficiency** of the data pipeline
  - Create new analytics capabilities
  - Tune-up the **alerting components** that users can subscribe to
- **Continuing the campaign to get our perfSONAR instances upgraded.**
  - Sites should be running at least version **4.1.6**
- Defining a new, cost-effective hardware recommendation sites can use as an example for renovating their perfSONAR instances
  - It may be possible to get this on some kind of Dell portal for 1-click ordering
- Working with the data in **Elasticsearch** to **correlate** and **visualize** traceroute paths with their related network metrics (packet-loss, latency, bandwidth)

Longer Term Outlook

- Long term outlook (5-10 years) will likely involve:
  - Capacity sharing - other big research domains coming online
  - (Re)organisation - evolution of LHCONE (ASTRONE ?), possibly some form of SD-WAN (dynamic circuits/L3 VPNs on demand)
  - Cloud networking - network federations spanning multiple data centres (inc. commercial clouds), ability to develop and operate services across large number of heterogenous sites easily from one location
- Are we going to be ready ?

Conclusion

OSG has a working network monitoring infrastructure measuring our sites and research and education networks.

The near-real-time network data we are gathering is a unique resource we need to exploit to proactively identify network problems and provide network capacity information to users and applications.

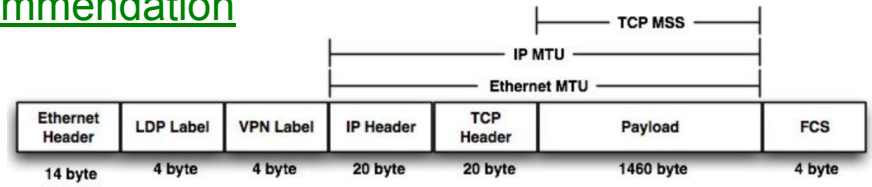
Sites should be aware of network technology evolution and be planning how best to take advantage of them as they evolve their infrastructure

Questions?

Additional Slides

Impact of MTU, Protocols, Routing

- MTU is one of the top operational issues impacting performance/connectivity
  - LHCOPN/LHCONE is working on a recommendation
  - MTU issues combined with load-balancing are very challenging to debug



- IPv4 and IPv6 performance could be very different
  - IPv6 is likely to be processed by different branches of code (QoS, firewalls, IPv6 TCP stack, etc.) or even different equipment
  - Check network path first before looking any further
  - Establishing expectations and test them for both IPv4 and IPv6 is important

- Network paths are dynamic and while sites usually have limited control over this, change of route can have major impact on capacity. This applies also to Commercial Clouds (which will likely take commercial routes unless you have direct peering)

Cloud Networking

- Commercial cloud providers already operate big networks at global scale with significantly higher capacities that are available in R&E
- Cloud computing is also becoming an important topic and eventually we'll need to find ways how to effectively bridge commercial and R&E networks
  - ATLAS/Rucio project with Google is one example going in this direction
- Will commercial WAN become available in a similar manner we are now buying cloud compute and storage services ?
  - The underlying cost will be decisive
  - Transit within major cloud providers such as Amazon/Google currently not possible and likely challenging in the future, limited by regional business model - but great opportunity for NRENs

Tech Trends: Containers

- Recently there has been a strong interest in the container-based systems such as Docker
  - They offer a way to deploy and run distributed applications
  - Containers are lightweight - many of them can run on a single VM or physical host with shared OS
  - Greater portability since application is written to container interface not OS
- Obviously networking is a major limitation to containerization
  - Network virtualization, network programmability and separation between data and control plane are essential
  - Tools such as Flocker or Rancher can be used to create virtual overlay networks to connect containers across hosts and over larger networks (data centers, WAN)
- Containers have great potential to become disruptive in accelerating **SDN** and **merging LAN and WAN**
  - But clearly campus SDNs and WAN SDNs will evolve at different pace