Making the Most out of Bandwidth: Networking in OSG

March 16, 2016

Shawn McKee

Networks Supporting Science

- While some of us are interested in (or worried about!) networks it is fair to say most scientists would rather not have to think about them.
  - Ideally networks are "transparent" and always do the right thing, allowing data to move as fast as possible, from anywhere to anywhere at anytime ☺
  - As Irene noted this morning: data must be able to move!
- The challenge is twofold:
  - Networks underlie all our distributed infrastructures and must work well for us to use our grids, clouds and HPC resources.
  - Problems in the network can be very hard to identify, isolate and fix
- OSG is working to better monitor, manage and diagnose our networks for all our benefit.

OSG Networking Area Mission

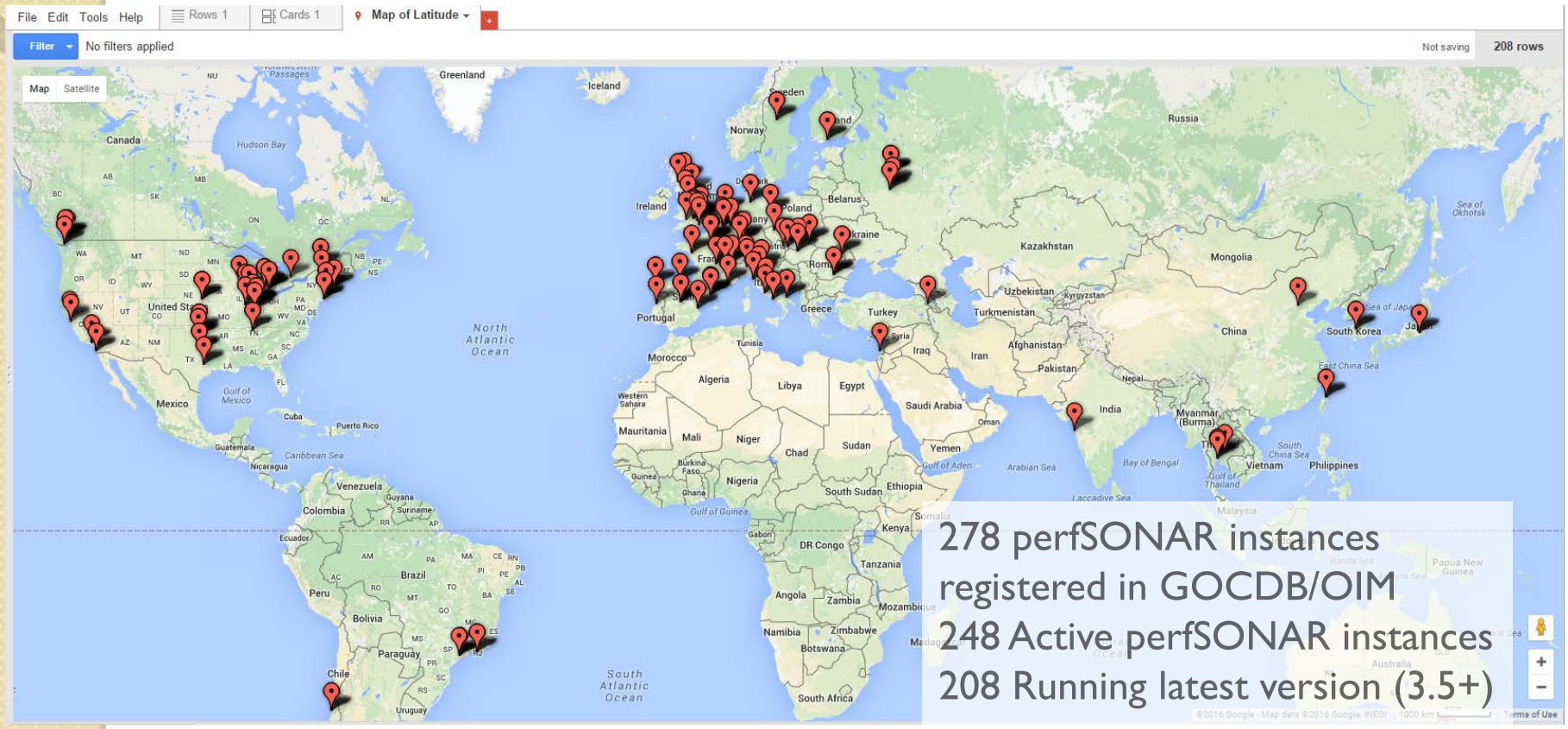
- OSG Networking was added at the beginning of OSG's second 5-year period in 2012
- The "**Mission**" is to have OSG become the network service data **source** for its constituents
  - Information about network performance, bottlenecks and problems should be easily available.
  - Should support our **VOs**, **users** and **site-admins** to find network problems and bottlenecks.
  - **Provide network metrics** to higher level services so they can make informed decisions about their use of the network *(Which sources, destinations for jobs or data are most effective?)*
- ***The GOAL: to make the most out of the bandwidth (network) we have! How?***

Components of OSG Networking

- **Network Monitoring via perfSONAR**
  - Having perfSONAR fully deployed with a global dashboard is giving us powerful options for better management and use of our network
- **A network datastore host all network metrics**
- **Tools to manage and maintain our infrastructure**
  - A Modular dashboard (MaDDash); critical for "visibility" into networks. We can't manage/fix/respond-to problems if we can't "see" them.
  - OMD/Check_mk (used to monitor and verify the state of many globally distributed perfSONAR services); required to maintain the overall proper functioning of the monitoring infrastructure.
  - The development of the "mesh-configuration" and corresponding GUI interface; critical to creating a scalable, manageable deployment for WLCG/OSG
- **Documentation --- Installation, debugging, How-tos**
- **Outreach and Support**
  - With the network R&E community, VOs, software developers
  - OSG Support provides network ticket triage and routing

A Global Monitoring Infrastructure

- We enabled a global deployment of perfSONAR to instrument our networks (both IPv4 & IPv6)

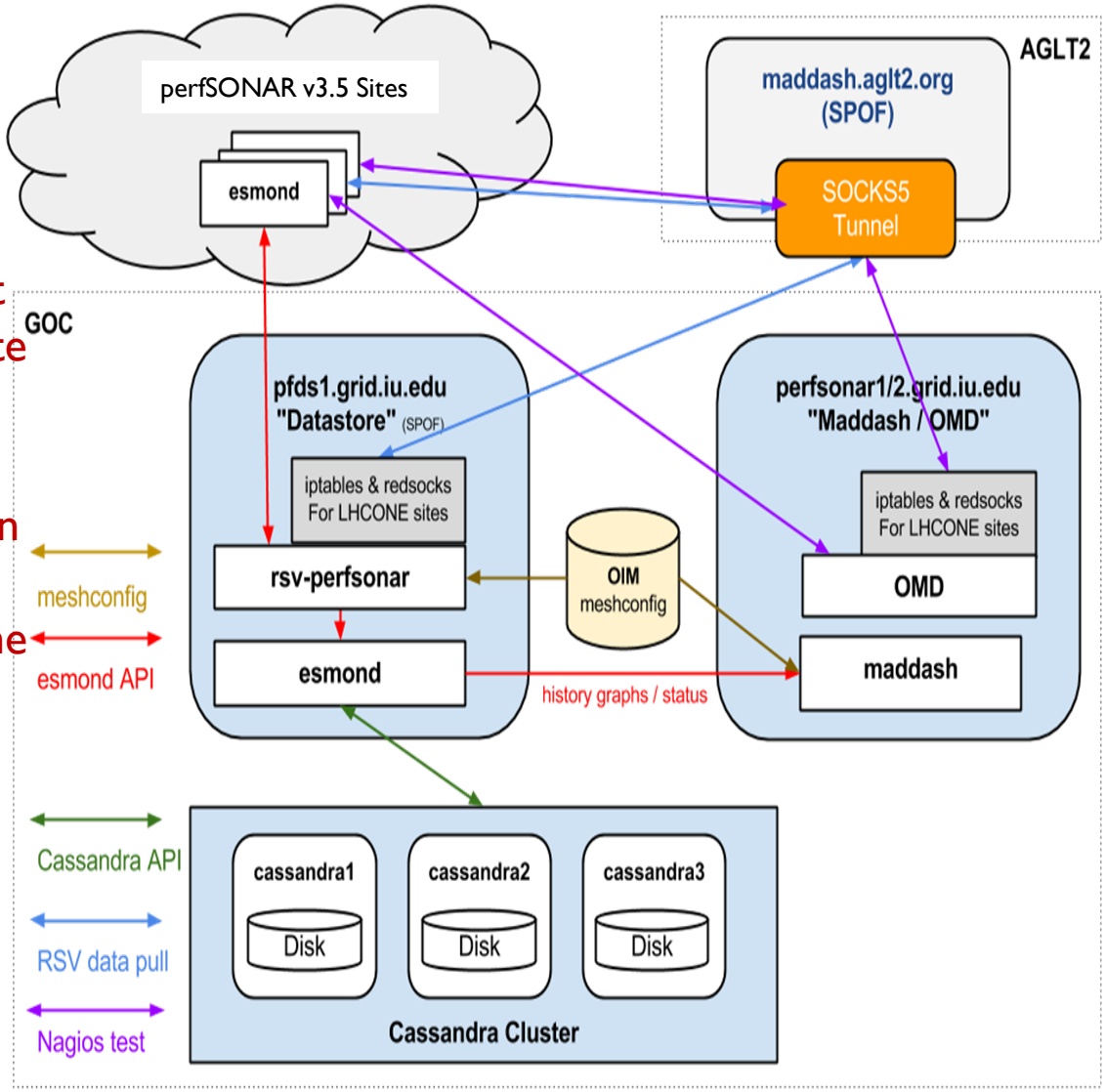


http://grid-monitoring.cern.ch/perfsonar_report.txt for stats (updated daily)

OSG Network Datastore

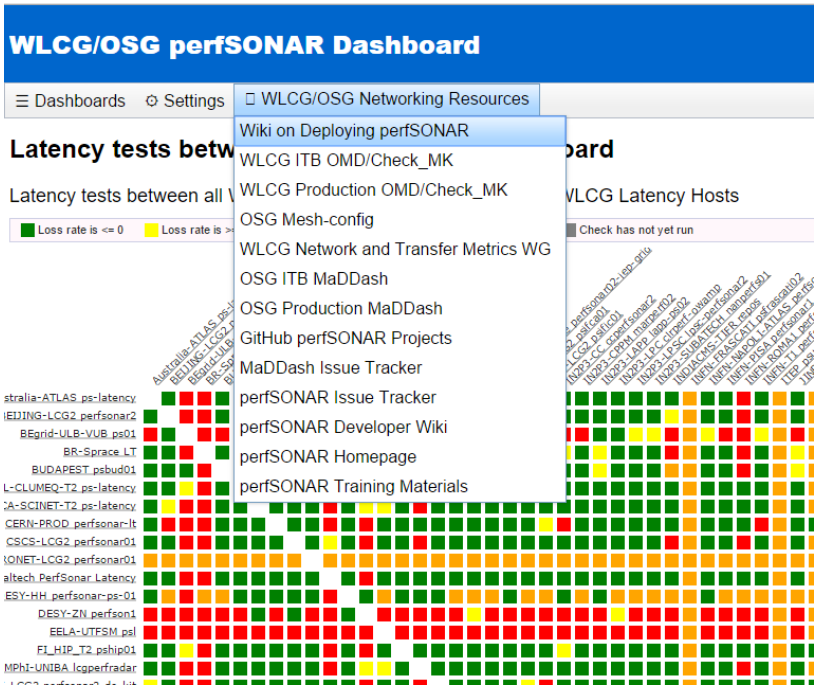
A critical component is the datastore to organize and store the network metrics and associated metadata

- OSG is gathering relevant metrics from the complete set of OSG and WLCG perfSONAR instances
- This data is available via an API, can be visualized and is organized to provide the **"OSG Networking Service"**
- **In production since September 2015**



Monitoring Metrics

- Use MaDDash to view metric summaries
  - Provide quick view about how networks are working
- OSG hosts a production instance at: http://psmad.grid.iu.edu/maddash-webui/
- Metrics are displayed via source-destination matrix
- Multiple dashboards (meshes) can be selected
- Custom menus link to relevant resources
- New release (2.0) will incorporate MadAlert



http://maddash.aglt2.org/madalert.html

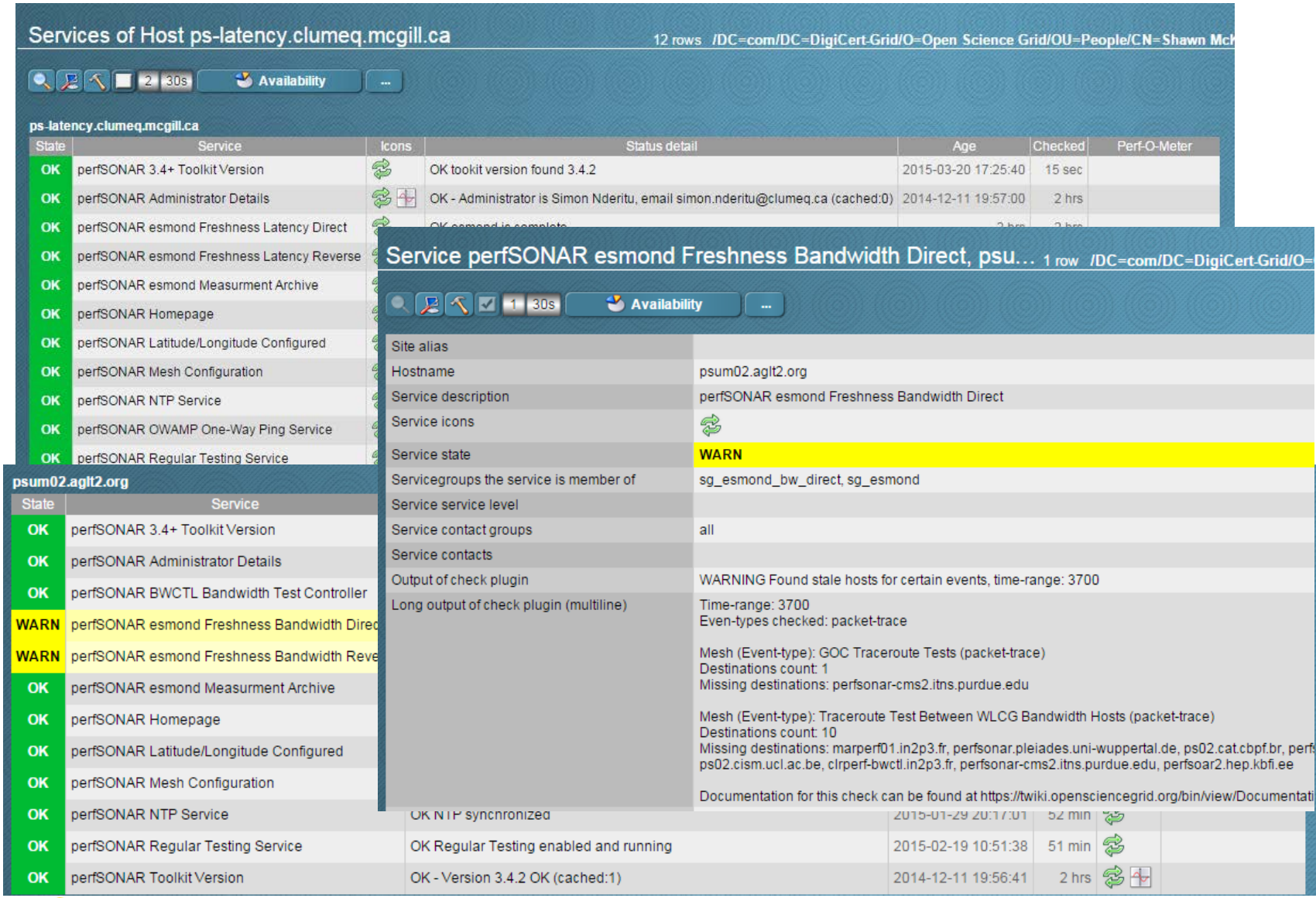
OMD/Check_MK Service Monitoring

We are using OMD & Check_MK to monitor our perfSONAR hosts and services. Provides useful overview of status/problems

https://psomd.grid.iu.edu/WLCGperfSONAR/check_mk/

[Requires x509 in your browser; update to InCommon authentication soon]

Detailed Service Checks



Existing Tools

- We have a number of tools deployed and available to help debug and understand network problems.
- There are very good presentations on these tools in the training materials provided by perfSONAR: http://www.perfsonar.net/about/training-materials/
- While I don't have time to cover all the details (see http://www.perfsonar.net/about/training-materials/201601-ps-training/ and especially the Measurement Tools, Use Cases and Debugging presentations from Jason Zurawski) I do want to note that **command line tools exist to allow you to create on-demand 3rd party tests** (between two remote instances) for bandwidth, latency and traceroute.
  - Follow the debugging strategy as a guide to finding and fixing OSG network issues using perfSONAR capabilities
- **As for new tools….**

Managing perfSONAR Deployments

- OSG originally developed a "mesh-config" GUI built within the OIM/MyOSG framework
  - We provided a GUI to define and organize the regularly scheduled tests between specific sets of perfSONAR instances.
  - The mesh-config was a **huge** benefit; no longer need to use email to hundreds of system admins to make changes to network tests and their organization. The GUI made changes easy and consistent.
- **Problem**: not able to be made easily available to others within or outside OSG.
  - Campuses deploying many perfSONARS
  - Science VOs wanting to organize/customize their perfSONARs
- Soichi Hayashi/OSG hasproduced a new standalone package which provides an even more feature-rich mesh-configuration GUI
  - Will be integrated with perfSONAR v3.6
  - See beta release info at http://soichi7.ppa.iu.edu/pdoc/mca.html#
  - Eventually will replace current GUI

MadAlert: A new project to analyze meshes

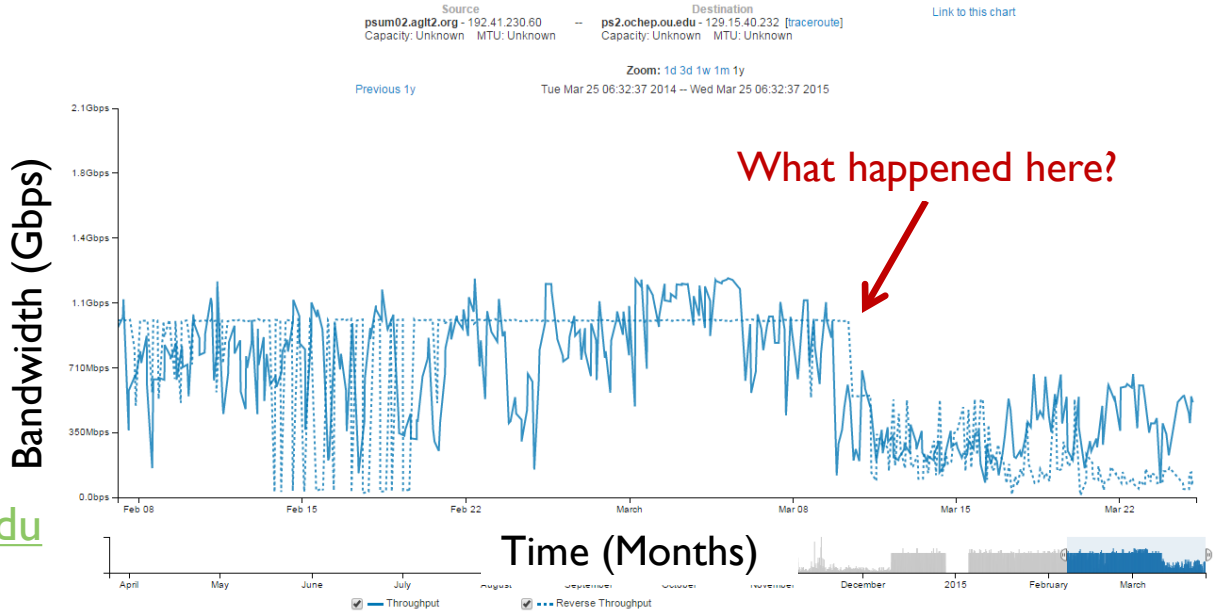
- Gabriele Carcassi/UM has been working with me on creating a new utility to analyze meshes: MadAlert
- See details at http://madalert.aglt2.org/madalert/index.html
  - You can see meshes and reports from the page
  - Reports find both **infrastructure** and **network** problems
- We are now working with Andy Lake/ESnet to incorporate this into the next major release of MaDDash (v2.0) due out later this year.
- Now testing a "diff" to allow us to compare meshes; e.g., IPv4 vs IPv6, testing vs production, mesh(t1) vs mesh(t2)
  - http://madalert.aglt2.org/madalert/testDiff.html
  - Could be really helpful for understanding new software versions or changes in time. ***Time based comparison will require some modifications to MaDDash to allow specifying point-in-time meshes.***

OSG Network Alerting

- What kinds of capabilities can we enable given a rich datastore of historical and current network metrics?
  - Users want "someone" to tell them when there is a network problem involving their site or their workflow.
  - Can we create a framework to identify when network problems occur and locate them? (**Must** minimize the false-positives).

Targeting this in the PuNDIT OSG satellite project

http://pundit.gatech.edu



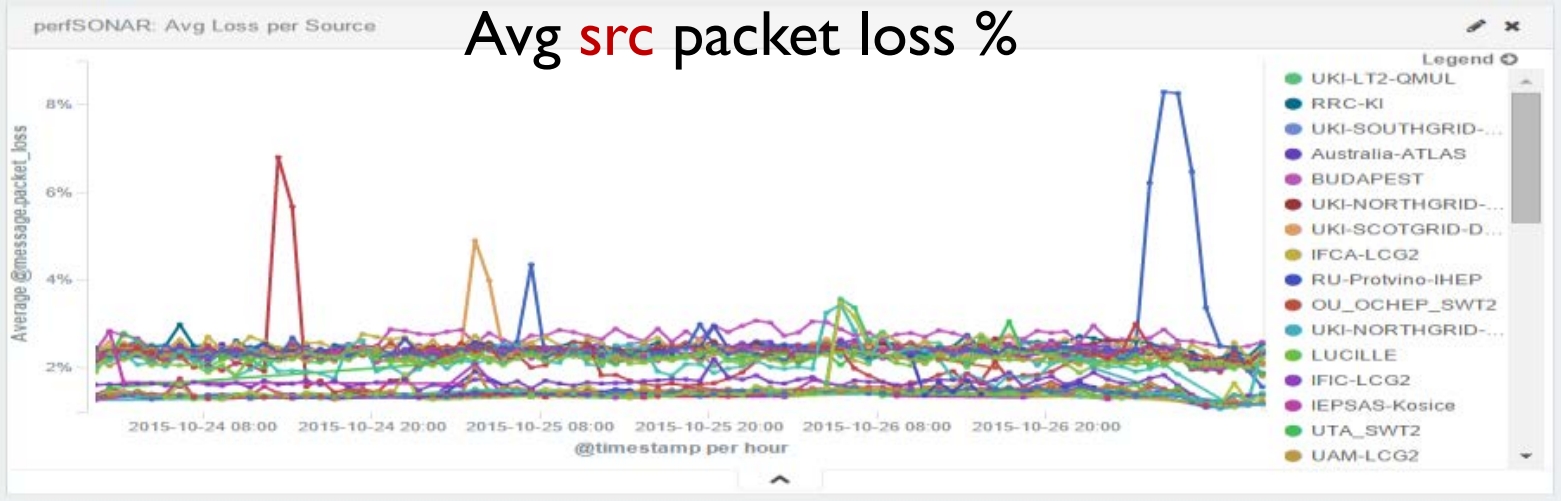
OSG Networking and End-to-end

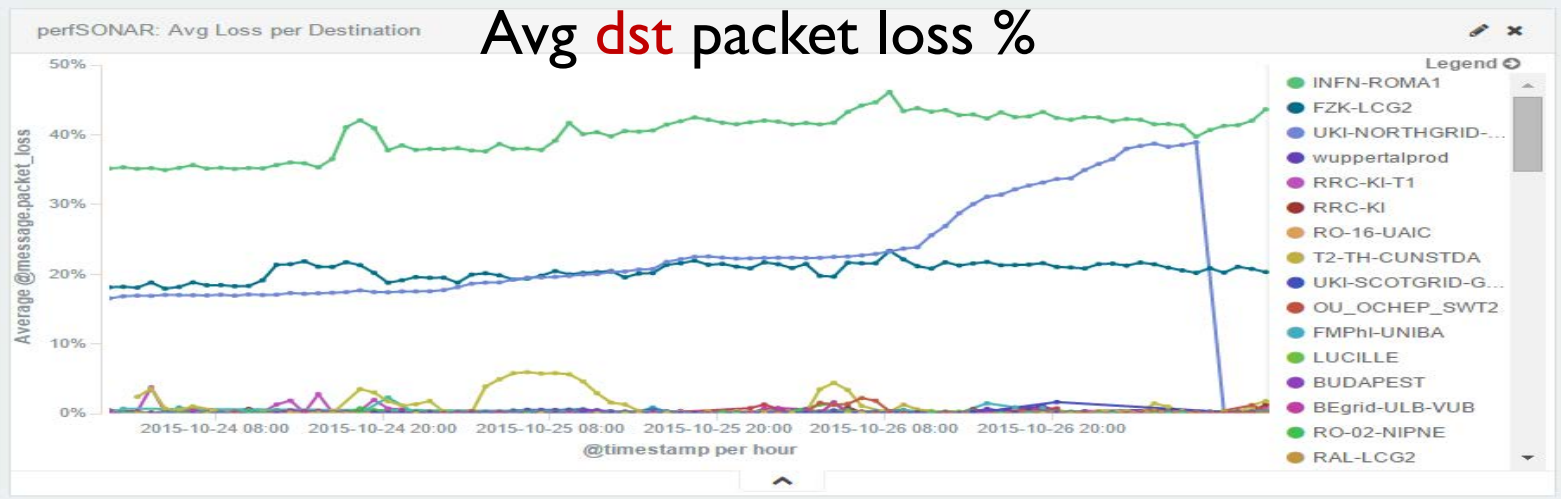
- Most scientists just care about the **end-to-end** results:
  - How well does their infrastructure support them in doing their science?
- Network metrics allow OSG to **differentiate** end-site issues from network issues.
- There is an opportunity to do this better by having access to **end-to-end** metrics to compare & contrast with network-specific metrics.
  - What end-to-end data can OSG regularly collect for such a purpose?
  - Is there some kind of common instrumentation that can be added to some data-transfer tools? (NetLogger in GridFTP, having transfers "report" results to the nearest perfSONAR-PS instance?, etc)
  - Let's try to put all the information we have together…

ATLAS Network Metrics Pipeline

- Ilija Vukotic, Kaushik De, Rob Gardner and Jorge Batista are working with the Network and Transfer Metrics WG to make OSG perfSONAR metrics available to PANDA
  - See Ilija's presentation at http://tinyurl.com/gt92zwb
- **Pipeline**: OSG Network Datastore to CERN Active MQ to Flume to ElasticSearch to PANDA
- Pipeline is operating and analysis has been performed in ElasticSearch to validate our data.
- Working on a network source-destination cost-matrix PANDA can use to evaluate options
  - Actual interface details being discussed/developed with the PANDA team
- **Potentially a very powerful, customizable analysis framework to use all available metrics from the network**

perfSONAR Data in ElasticSearch





http://tinyurl.com/OSGNetES for other examples using OSG data

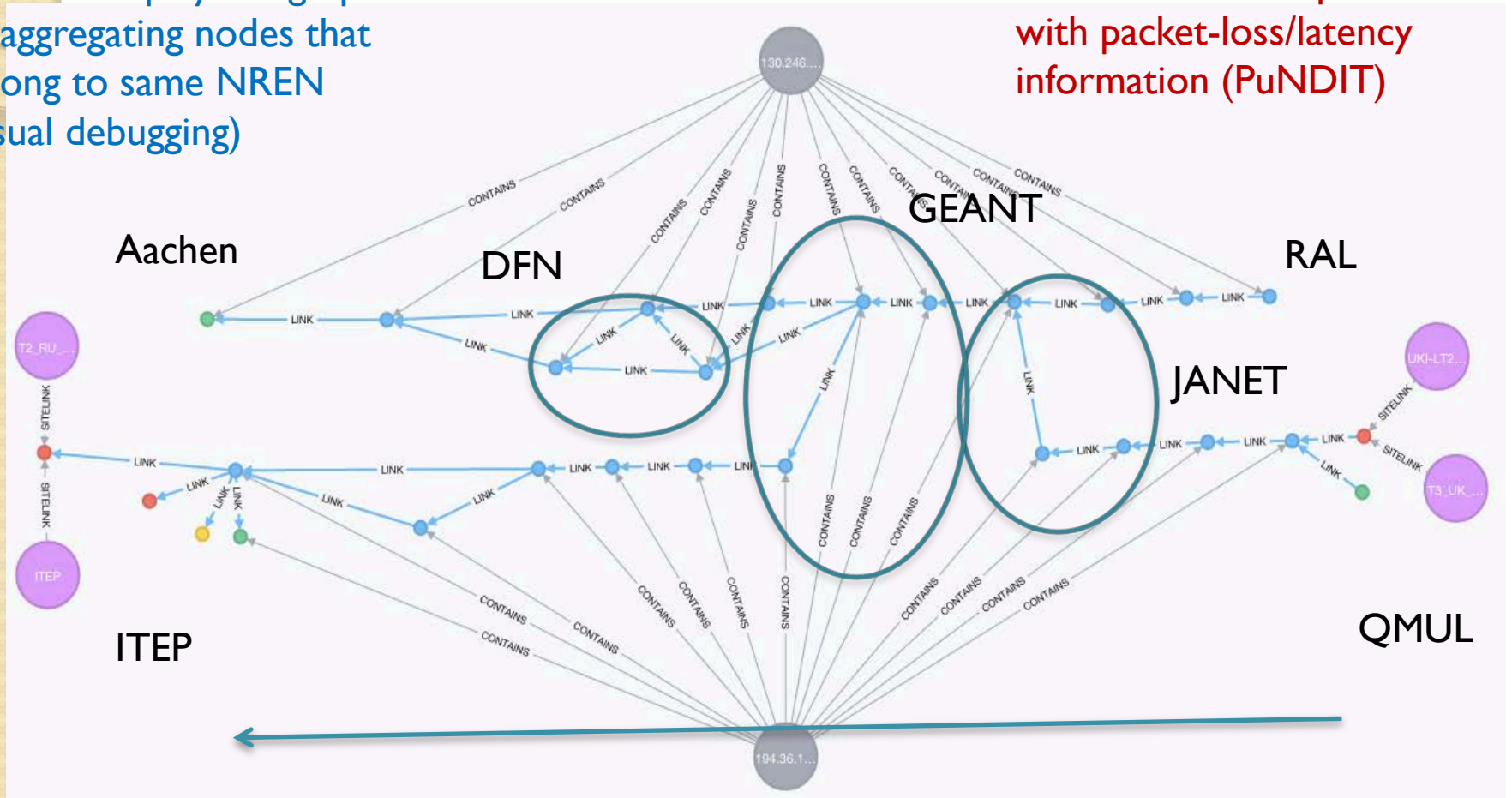
Understanding Network Topology

- One thing that may not be obvious is that all the network measurements are only really useful if we know the path being measured…topology is very important to find/fix problems.
  - OSG measures topology via traceroute and tracepath
- Can we create tools to **manipulate**, **visualize**, **compare** and **analyze** network topologies from the OSG network datastore contents?
- Can we build upon these tools to create a set of next-generation network diagnostic tools to make debugging network problems easier, quicker and more accurate?
- Even without requiring the ability to perform complicated data analysis and correlation, basic tools developed in the area of network topology-based metric visualization would be very helpful in letting users and network engineers better understand what is happening in our networks.
- This area is under active investigation in various projects. Lots of work to do here.

Exploring Path Analysis

We can simplify the graph by aggregating nodes that belong to same NREN (visual debugging)

We can correlate paths with packet-loss/latency information (PuNDIT)



latency, packet-loss, throughput all available

Software Defined Networks and OSG

- Within the next few years evolving technology in the area of **Software Defined Networking(SDN)** may be able to provide OSG researchers with the ability to construct their own Wide-Area networks with specified characteristics.
- What will OSG be able to do to integrate this type of capability with the rest of the OSG infrastructure?
- We are planning for how best to enable evolving capabilities in the network for OSG users and admins. We need to address:
  - What is the impact on the OSG software stack?
  - What strategic modifications/additions/tools are useful?

Summary

- We have made significant strides in making the network "visible"
  - ~250 OSG/WLCG perfSONAR deployments globally
    - All monitored, managed and orchestrated by OSG
  - Tools to manage and maintain our infrastructure are in place
  - We have production datastore providing long term access to all metrics
- New opportunities to improve our networks and their use are possible because of the unique set of data we have
  - Exploiting the rich dataset we have is underway.
  - To identify network problems and do "targeted alerting" on them is a high priority.
  - To inform and enable higher level services, researchers and users

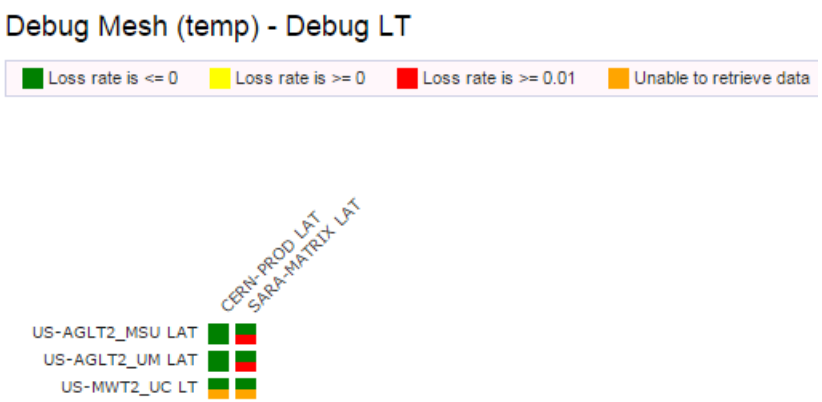
Questions or Comments?

Thanks!

Extra Slides

Network Problem Debugging(1/4)

- Problem with transfers from SARA to AGLT2 noted February 20th.
  - FTS transfers failing. ATLAS asked about network. Dataset had large files; transfers failed because of timeout (**1 MB/sec=3.6GB/hour**; 1 hour timeout)
  - Setup 'Debug' mesh using OSG tools to track SARA, CERN to AGLT2, MWT2
  - Jason Zurawski ran tests using perfSONAR. Bad throughput ~few hundred kbits/sec SARA->AGLT2 (real network issue!)



Network Debugging (2/4)

- Ticket opened by AGLT2 with Internet2 NOC on February 20th
  - Poor iperf results also between SARA and MWT2
  - Routes provided both ways by perfSONAR

Topology beginning at Tue Mar 10 00:01:17 2015 (UTC -4)

Hop	Router	IP	Delay	MTU
1	ge-0-2-0-1020.grid-r1.grid.sara.nl	145.100.17.1	0.16ms	
2	geant-lhcone-gw.mx1.ams.nl.geant.net	62.40.126.161	0.279ms	
3	et-10-0-0.3019.rtr.newy32aoa.net.internet2.edu	64.57.30.225	74.946ms	
4	192.17.10.74	192.17.10.74	105.154ms	
5	64.57.30.154	64.57.30.154	120.308ms	
6	psum02.aglt2.org	192.41.230.60	107.806ms	

Topology beginning at Tue Mar 10 13:46:17 2015 (UTC -4)

Hop	Router	IP	Delay	MTU
1	ge-0-2-0-1020.grid-r1.grid.sara.nl	145.100.17.1	0.158ms	
2	geant-lhcone-gw.mx1.ams.nl.geant.net	62.40.126.161	0.268ms	
3	et-10-0-0.3019.rtr.newy32aoa.net.internet2.edu	64.57.30.225	87.094ms	
4	192.17.10.74	192.17.10.74	102.803ms	
5	requestTimedOut	requestTimedOut		
6	psum02.aglt2.org	192.41.230.60	108.808ms	

Topology beginning at Tue Mar 10 00:40:32 2015 (UTC -4)

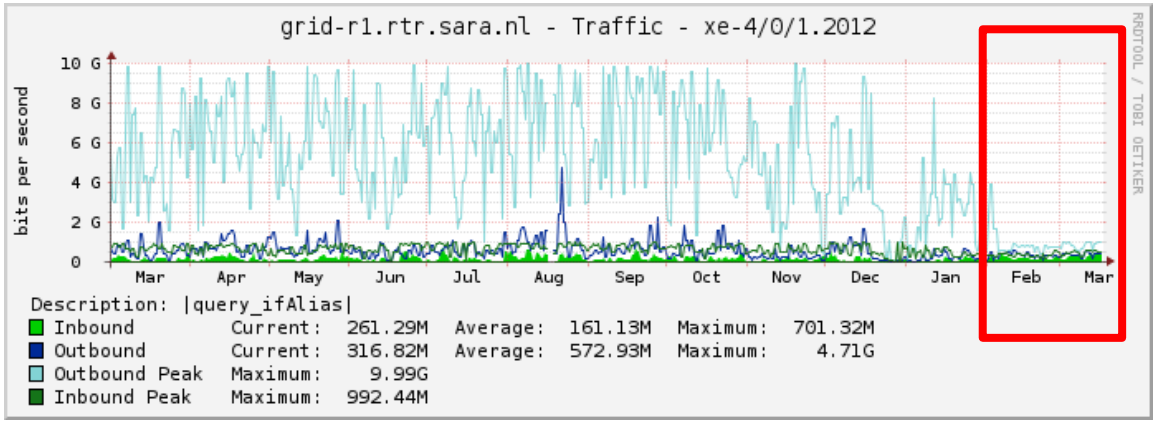
Hop	Router	IP	Delay	MTU
1	192.41.230.1	192.41.230.1	0.266ms	
2	et-8-0-0.3151.rtr.chic.net.internet2.edu	64.57.30.155	5.796ms	
3	et-10-0-0.402.rtr.newy32aoa.net.internet2.edu	64.57.30.142	33.651ms	
4	64.57.30.228	64.57.30.228	119.977ms	
5	surfnet-lhcone-gw.ams.nl.geant.net	62.40.126.160	119.973ms	
6	ps.lhcopn-ps.sara.nl	145.100.17.9	119.803ms	

Topology beginning at Tue Mar 10 13:43:23 2015 (UTC -4)

Hop	Router	IP	Delay	MTU
1	192.41.230.1	192.41.230.1	0.235ms	
2	esnet-lhc1-a-aglt2.es.net	198.124.80.53	6.266ms	
3	62.40.126.149	62.40.126.149	109.678ms	
4	62.40.126.148	62.40.126.148	108.904ms	
5	surfnet-lhcone-gw.ams.nl.geant.net	62.40.126.160	119.086ms	
6	ps.lhcopn-ps.sara.nl	145.100.17.9	108.313ms	

Network Debugging (3/4)

- Internet2 initially pursuing asymmetric routes and link congestion in US
- I2 opened ticket with GEANT Feb 27
- GEANT brought up LHCONE perfSONAR in Frankfurt
  - Tests to SARA(close) and AGLT2(far) showed 3x better bandwidth to AGLT2. Problem close to SARA
  - Isolated link between SARA and GEANT March 4

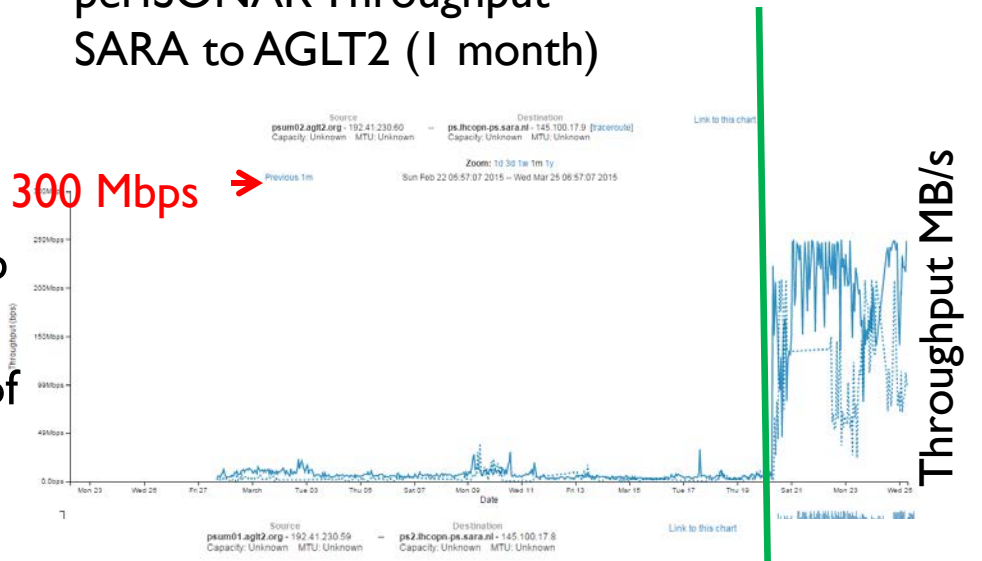


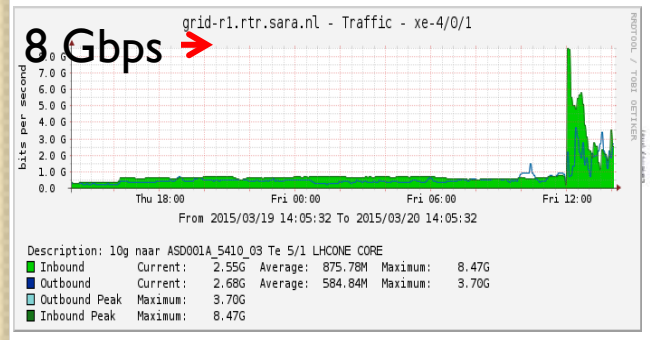
SARA noticed problems once we highlighted the link

Problem Found/Fixed – Mar 20

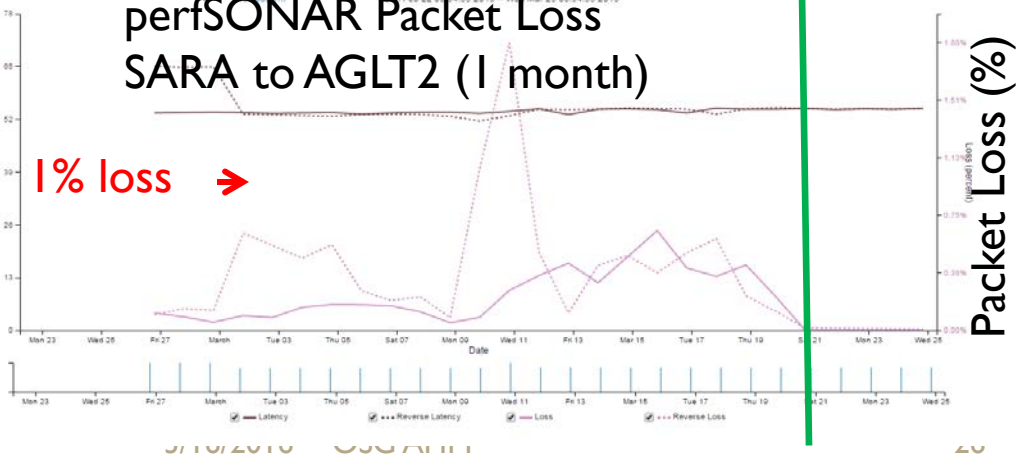
- **2 weeks of link-debug**
- Problem identified and fixed Mar 20
  - Bandwidth "policing" for LHCONE. More than a year ago, LHCONE setup 1 Gbps. *Never enforced.*
  - Turned on policing start of February.
  - **Changed BW 10Gbps March 20…fixed**

perfSONAR Throughput
SARA to AGLT2 (1 month)





perfSONAR Packet Loss
SARA to AGLT2 (1 month)



Open Science Grid