Report on the 100 Gigabit Networking Expansion

Shawn McKee/University of Michigan

XRootD Workshop @ UCSD

Univ. of CA, San Diego

January 29th, 2015

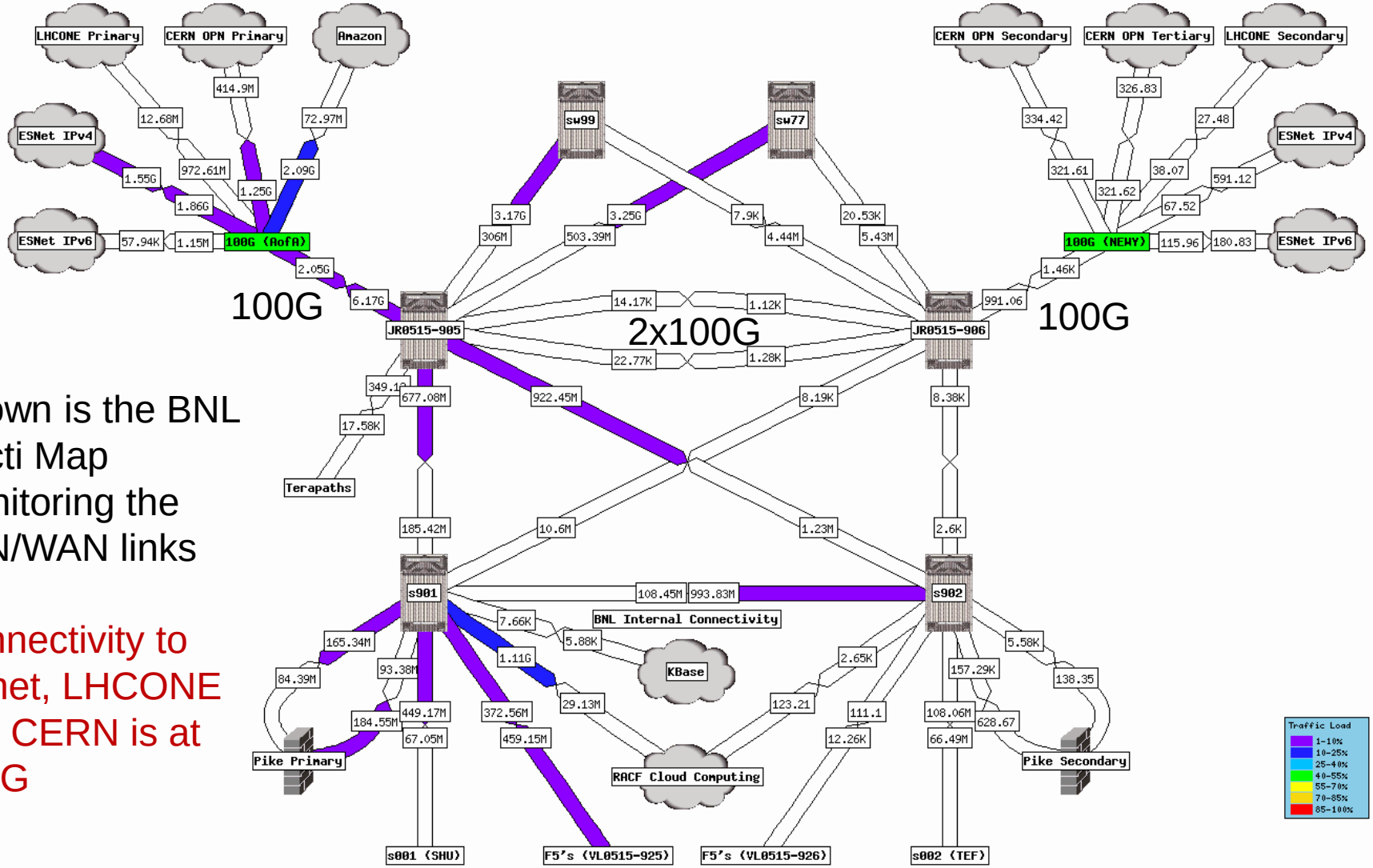
Acknowledgements

- Thanks to everyone who contributed with text, diagrams and experiences. In order of response ☺
- **Horst Severini, Wei Yan, Frank Wuerthwein, Azher Mughal, Garhan Attebury, Ken Bloom, Samir Cury, Brian Bockelman, Gabriele Garzoglio, Hiro Ito, Harvey Newman, Saul Youssef, Sarah Williams, Dave Lesny, Antonio Ceseracciu, Kaushik De, Valerie Polichar, Joshua Alexander, Dale Carder, Manoj Jha, John Bigrow**

Introduction

- I was tasked with summarizing the state of 100G networking at our sites, especially as relates to WAN data access. Globally, this is a large task
- I simplified a little and focused on what has been happening at our US LHC Tier-N sites and solicited input using mailing lists for USATLAS and USCMS.
- I got responses from the Tier-1 sites: BNL and FNAL and a large fraction of the US Tier-2 sites: AGLT2, Caltech, MWT2(UC,IU,UIUC), Nebraska, NET2, Purdue, SWT2, UCSD, Wisconsin, WT2(SLAC)
- I will show a slide or two per site covering the information I was able to gather. Starting with the Tier-1s and going alphabetically…

BNL

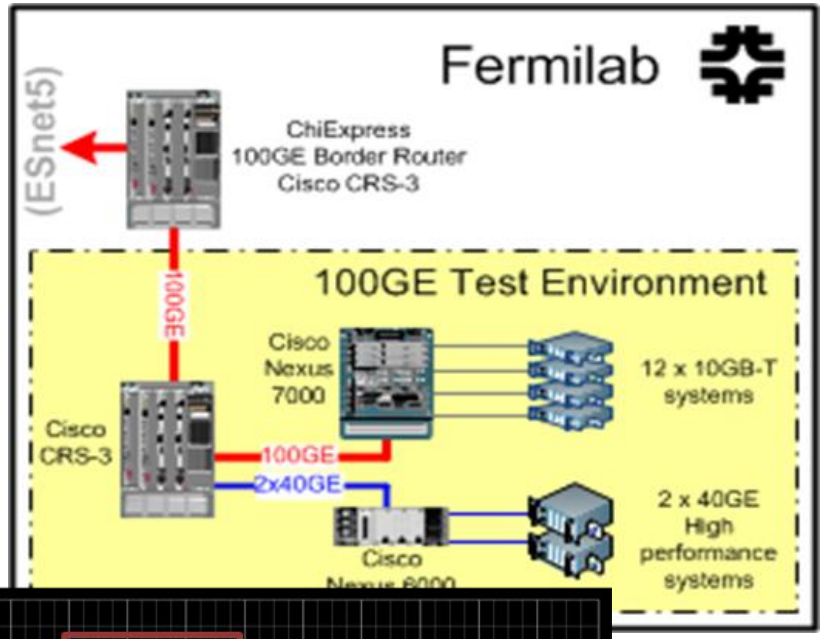


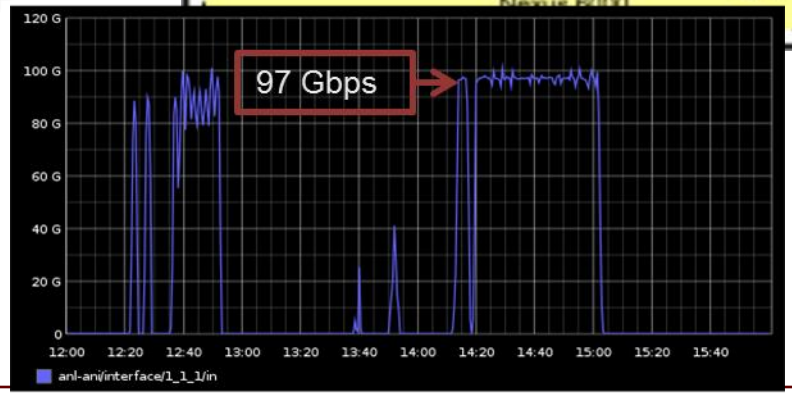
Shown is the BNL Cacti Map monitoring the LAN/WAN links

Connectivity to ESnet, LHCONE and CERN is at 100G

FNAL

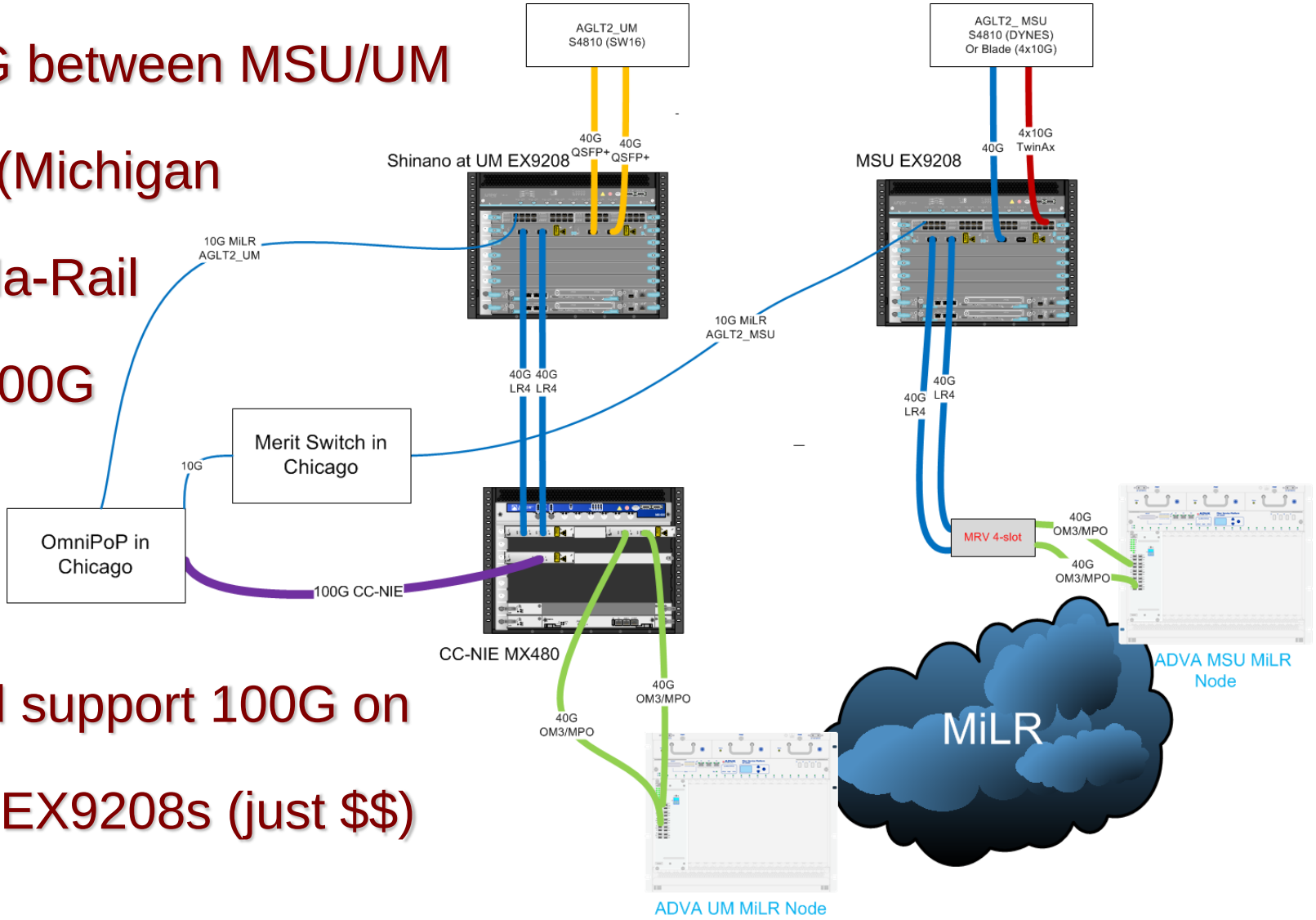
- Fermilab has been investigating 100G networking since December 2011.
- They have some interesting results for GridFTP, Globus Online, squid and xrootd summarized at: http://cd-docdb.fnal.gov/cgi-bin/ShowDocument?docid=5063
- 2012-2013: ESnet 100G testbed
- Tuned parameters of middleware for data movement: xrootd, GridFTP, SRM, Globus Online, Squid.
- Rapid turn around on the testbed thanks to custom boot images
- Optimal performance: **97 Gbps** w/GridFTP
  2 GB files – 3 nodes x 16 streams / node
- Tested NFS v4 over 100G using dCache (collab. w/ IBM research)





AGLT2

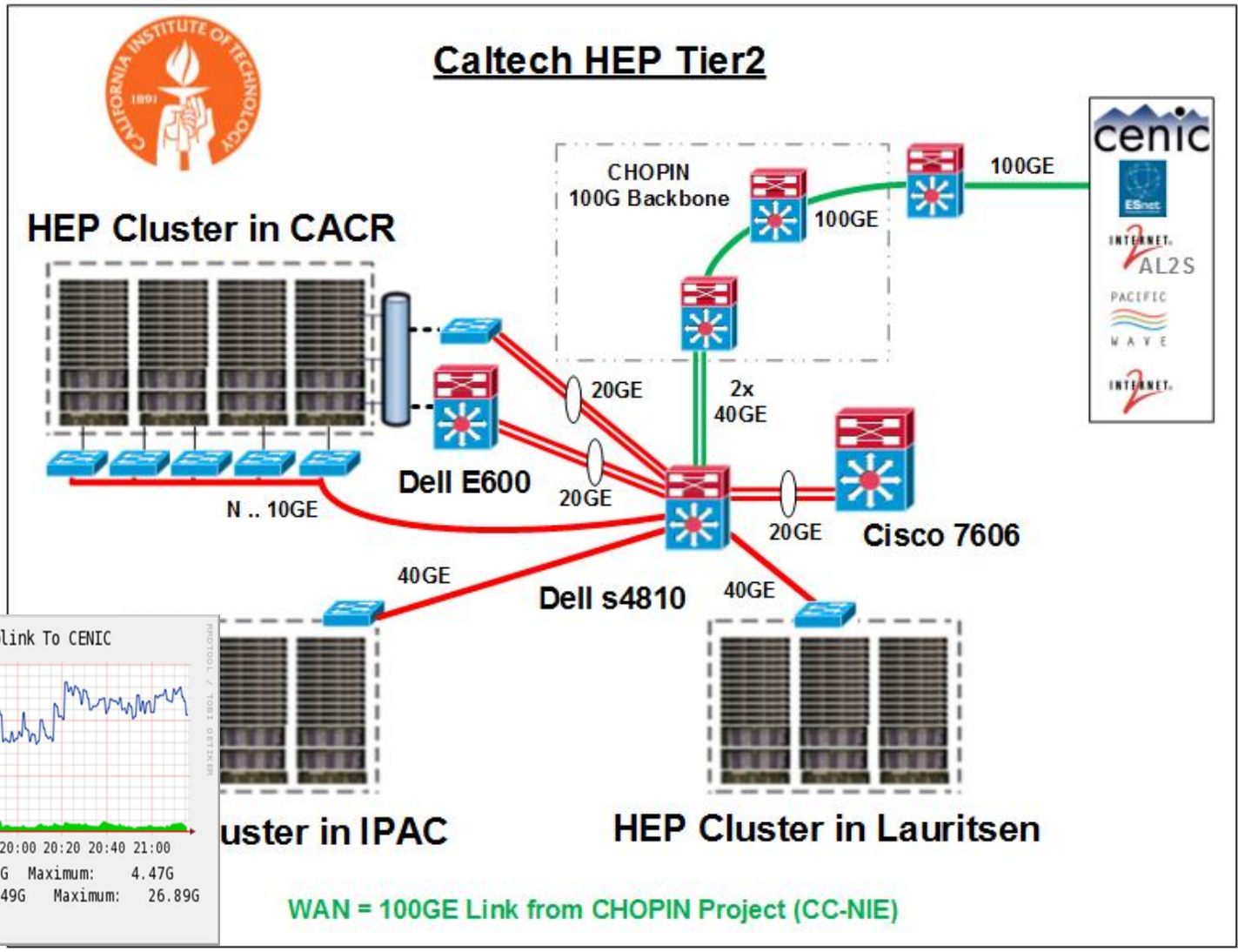
- 2x40G between MSU/UM
- MiLR (Michigan Lambda-Rail hosts 100G
- Would support 100G on Juniper EX9208s (just $$)

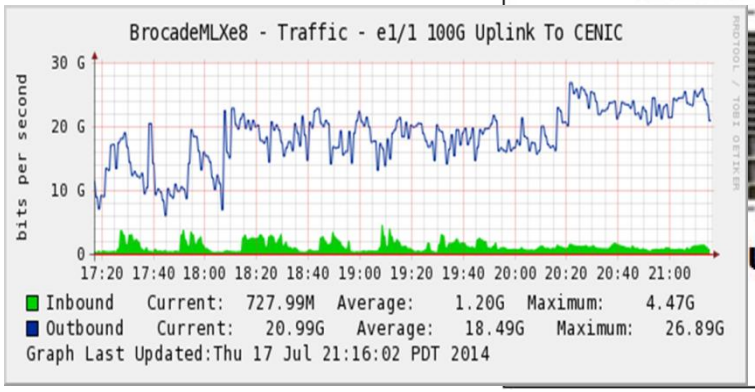


Caltech

Has 100G into CENIC from campus Tier-2 has 2x40G

Link was validated with required USCMS 20 Gbps PhEDEx stress test



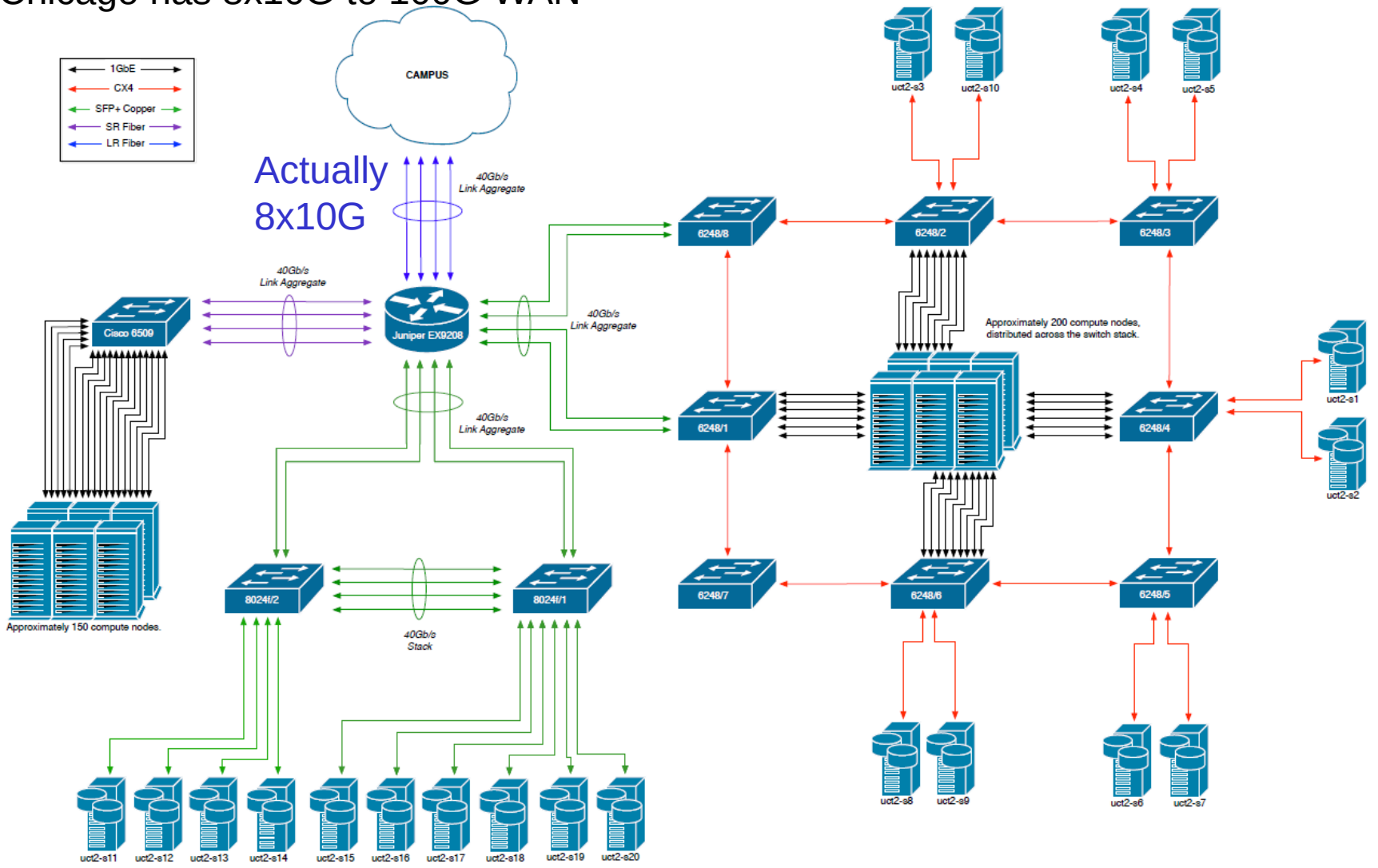


MWT2

- The Midwest Tier-2 (MWT2) comprises three sites, all of which have 100G connections to the wide-area network.
  - Connections between Tier-2 locations and the 100G are not yet at 100G though (but all close; each site 8x10G)
- Individual network diagrams for each site follow.
- Storage nodes at each site is typically connected at 10G.

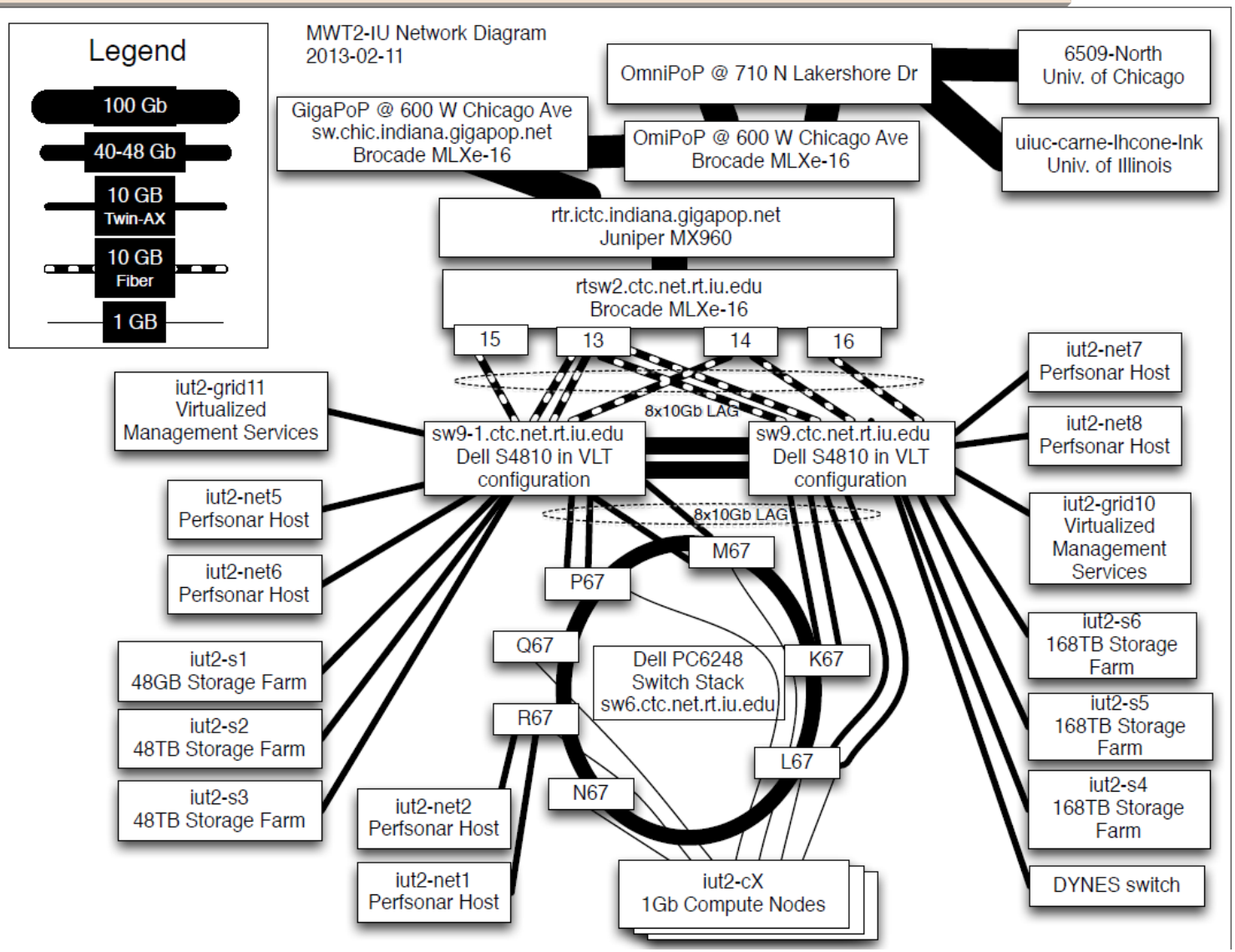
MWT2_UC

U Chicago has 8x10G to 100G WAN



MWT2_IU

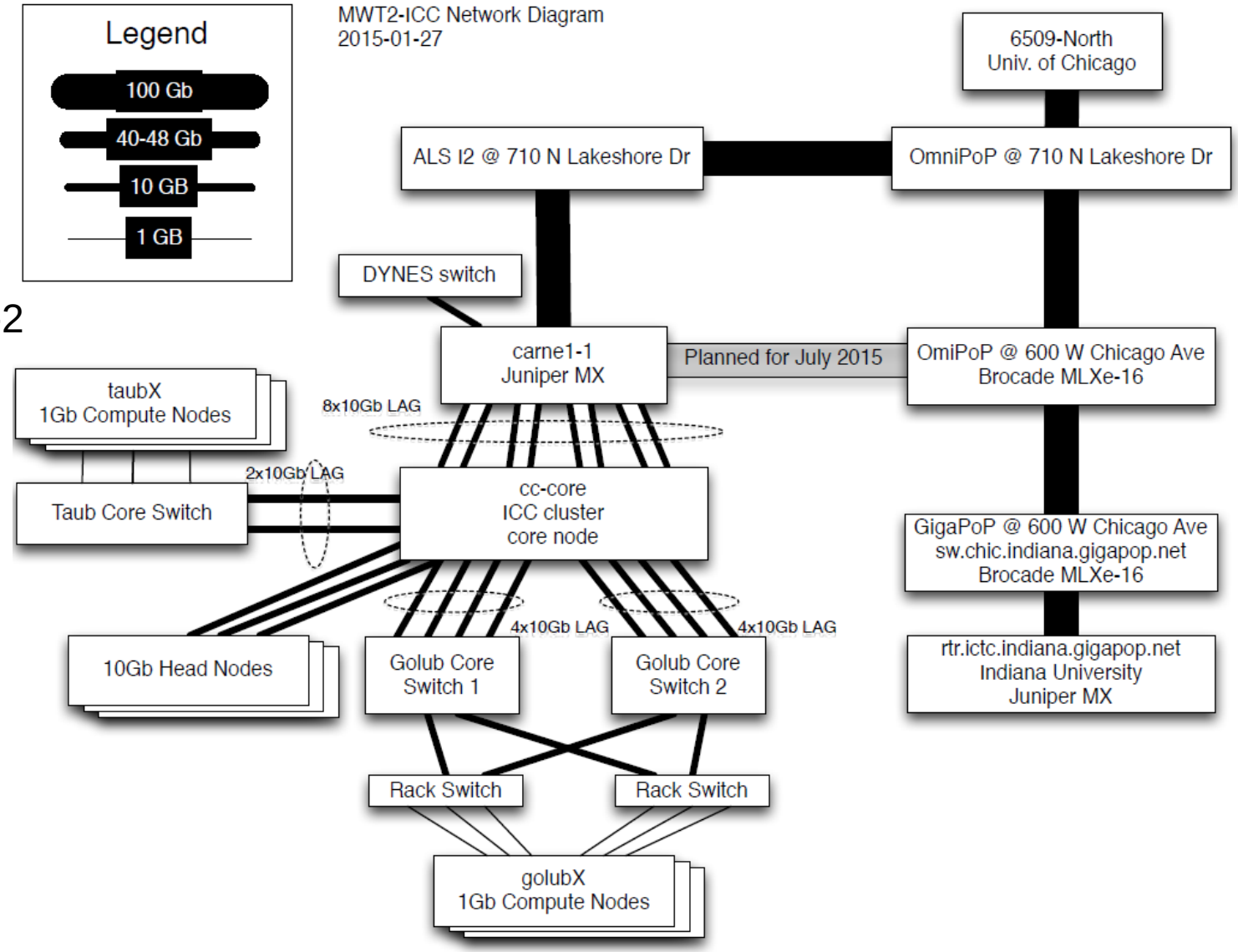
The Indiana site has 8x10G to the 100G path



MWT2_UIUC

Univ Illinois Urbana Champaign

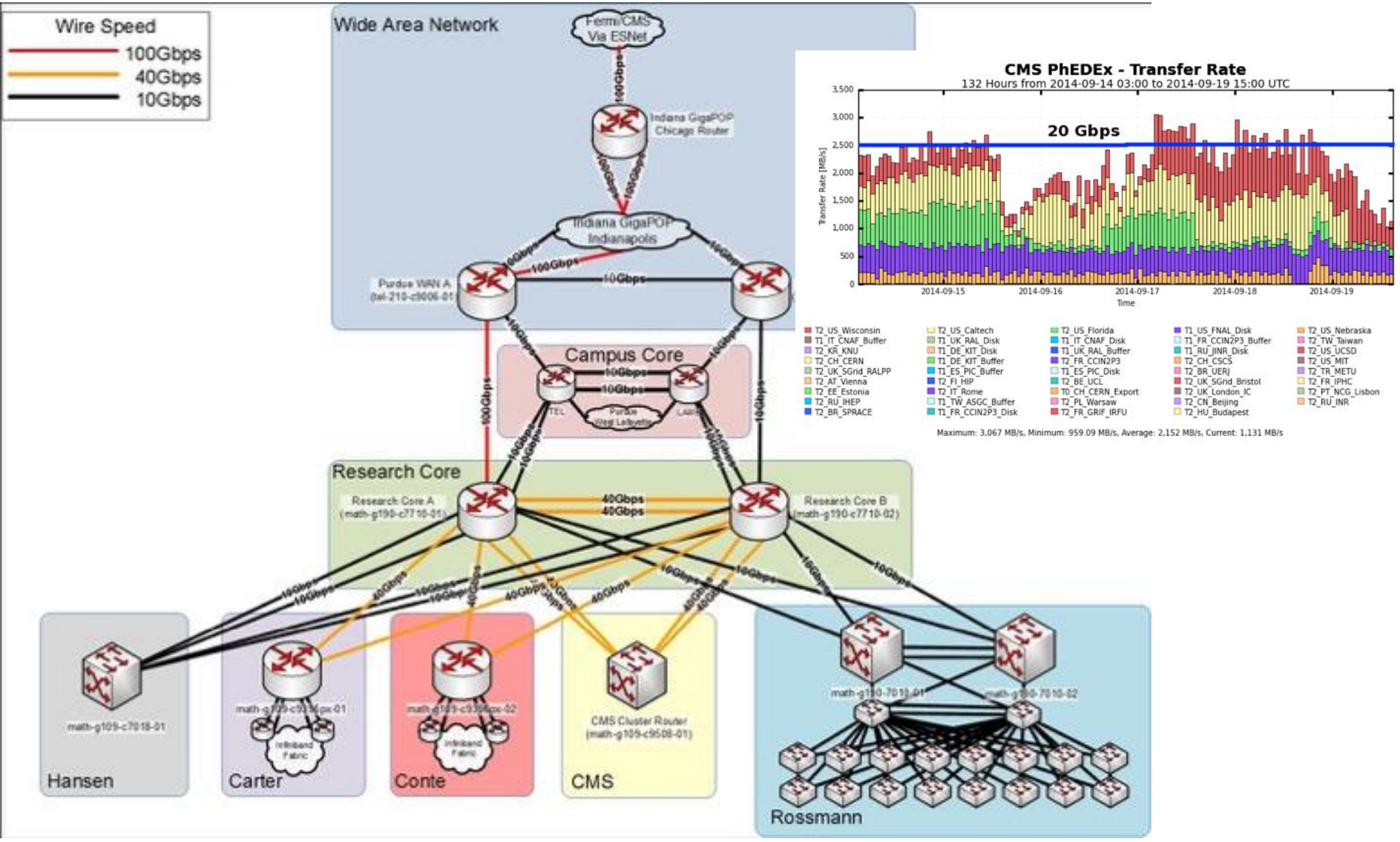
Has 8x10G to/from the Tier-2



Purdue

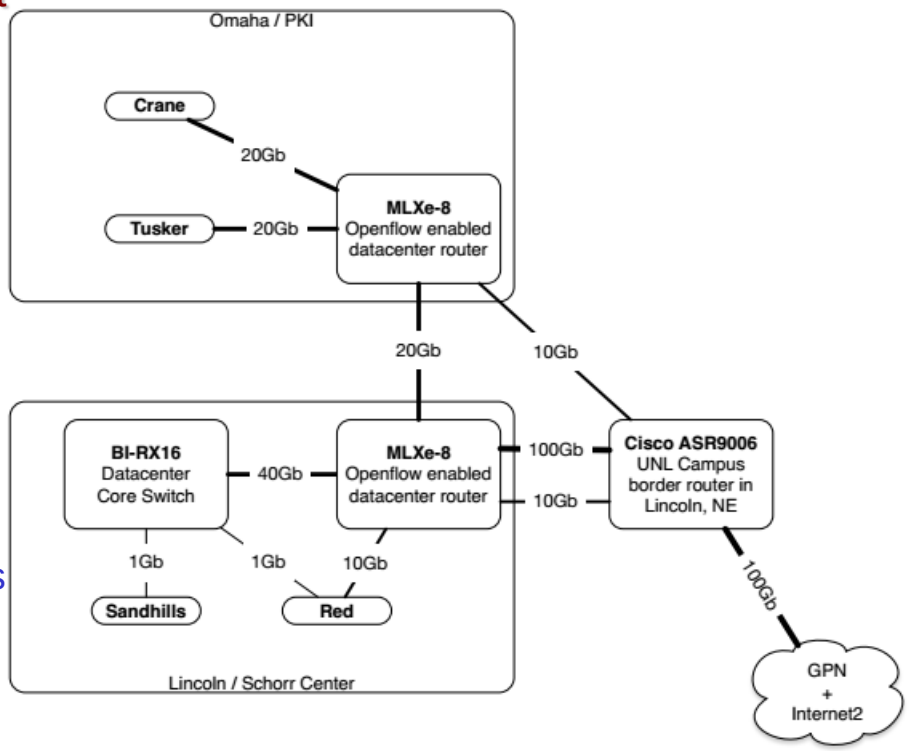
- In spring 2014, Purdue deployed 100G wide area connection.
- To utilize 100G WAN connection, CMS dedicated cluster at Purdue upgraded local LAN to 160 Gbps (4 x 40 Gbps) link. Local and wide area network of computing resources at Purdue are shown in the following slide
- After upgrading WAN to 100G, site passed the 20 Gbps throughput test which was conducted among US Tier-1 and several Tier-2's.
- Experienced two network outages at the site which appear to be related to optic or switch failure. Observed 100G optic can partially fail (25%) without generating errors logged by Cisco router.
  - Opened a support ticket with Cisco concerning this issue.

Purdue Network and Test Result



Nebraska

- Upgrade to 100G uneventful. Took effort and time to 1) get the NSF grant allowing the purchase of 100Gb line card(s) and optical equipment and 2) for campus to acquire and put into production a new 100Gb capable border router.
- UNL's border router peers with GPN and Internet2 directly over a 100Gb link, and in addition we at HCC receive a handful of VLANs from GPN where we peer with LHCONE
- Networking within the data-center is the next challenge
  - Right now UNL has ~10x 10Gb 'paths' from worker nodes in the Tier2 cluster to the MLXe
  - Lots of room for 10Gb and/or 40Gb improvements in both datacenters in the near future.
- Have yet to stress the 100Gb.
  - Have attempted ~40+Gbps before via lots of manual xrootd transfers / gridftp transfers / attempted annihilations of ESNet's transfer test servers
  - To date haven't actually broken 37Gbps sustained over the WAN



NET2

- NET2 (Boston University, Harvard) are still in the planning phase but hope to have 100G connectivity soon.
- Currently connected with multiple 10G paths to the wide area
- Storage servers typically connected to LAN at 10G

SLAC

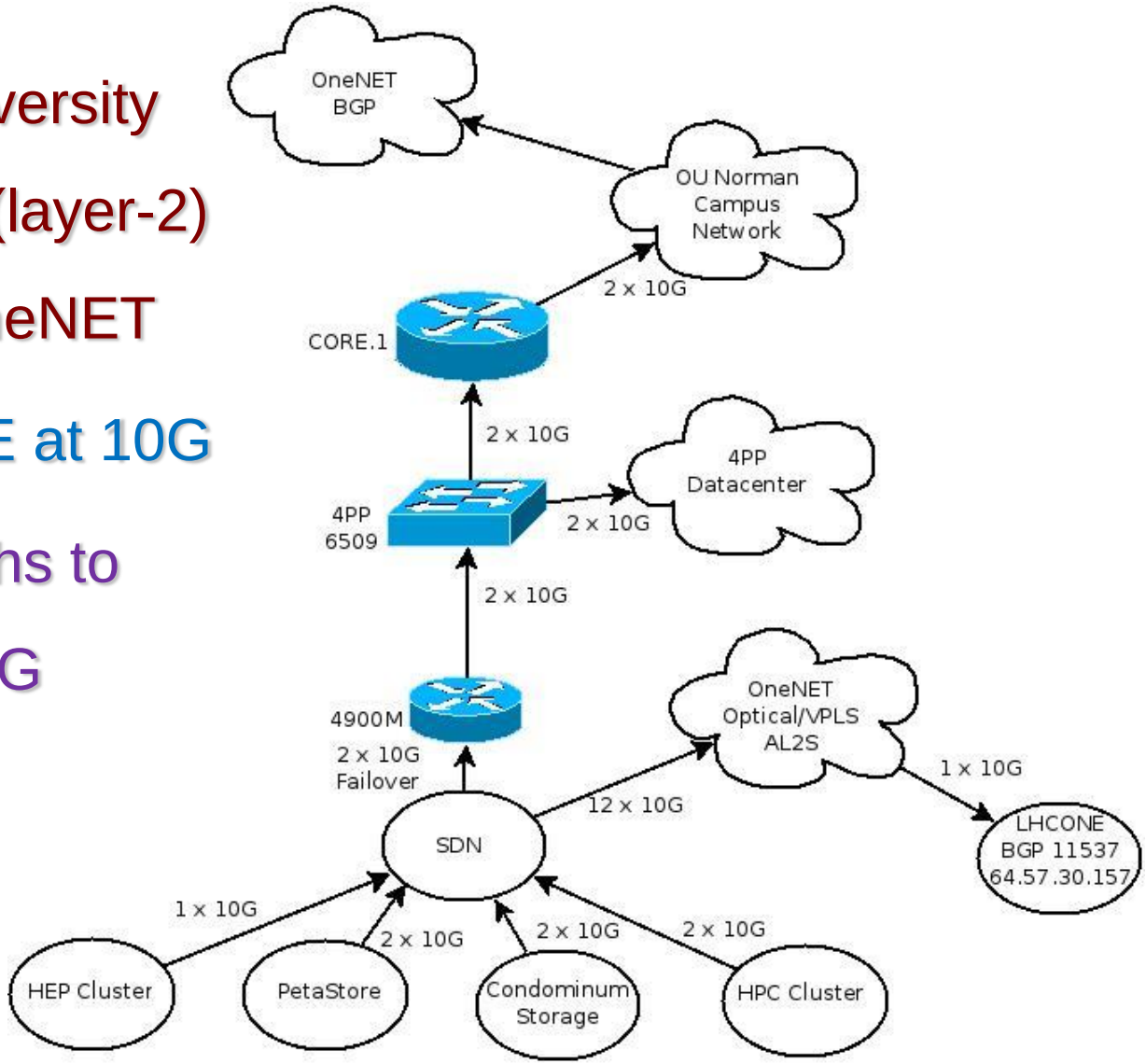
- Earlier this month, SLAC has established a 100G connection to ESnet which is used for general IP connectivity.
- In addition their LHCONE migrated to that connection on January 20th
- Along with the 100G link to ESNet, a dedicated 100G link has been established between the SLAC Scientific Computing network (which includes all ATLAS computing), and the SLAC 100G border router.
- The ATLAS Tier-2 now has multiple 10G's of usable bandwidth to LHCONE, across multiple DTNs (disruption tolerant networks).
  - This dedicated link bypasses the 10G SLAC core network. It is intended to be a temporary solution, to be undone once the SLAC core network gets its 100G upgrade, possibly later this year.
  - So for ATLAS, today, there are two main DTNs, with a single 10G link each. I know that there are short term plans to double that to 2x10G links, and later to acquire more DTNs.

SWT2

- The South West Tier-2 (SWT2), comprised of the University of Texas Arlington (UTA) and Oklahoma University, doesn't yet have 100G, at least in a single network connection
- For UTA, still in the planning stages. Nothing concrete yet, but a LEARN will go to 40G soon, and eventually up to 100G.
- SWT2 UTA is working on 10G->20G (available)->40G->100G plans gradually, in collaboration with LEARN

SWT2 OU

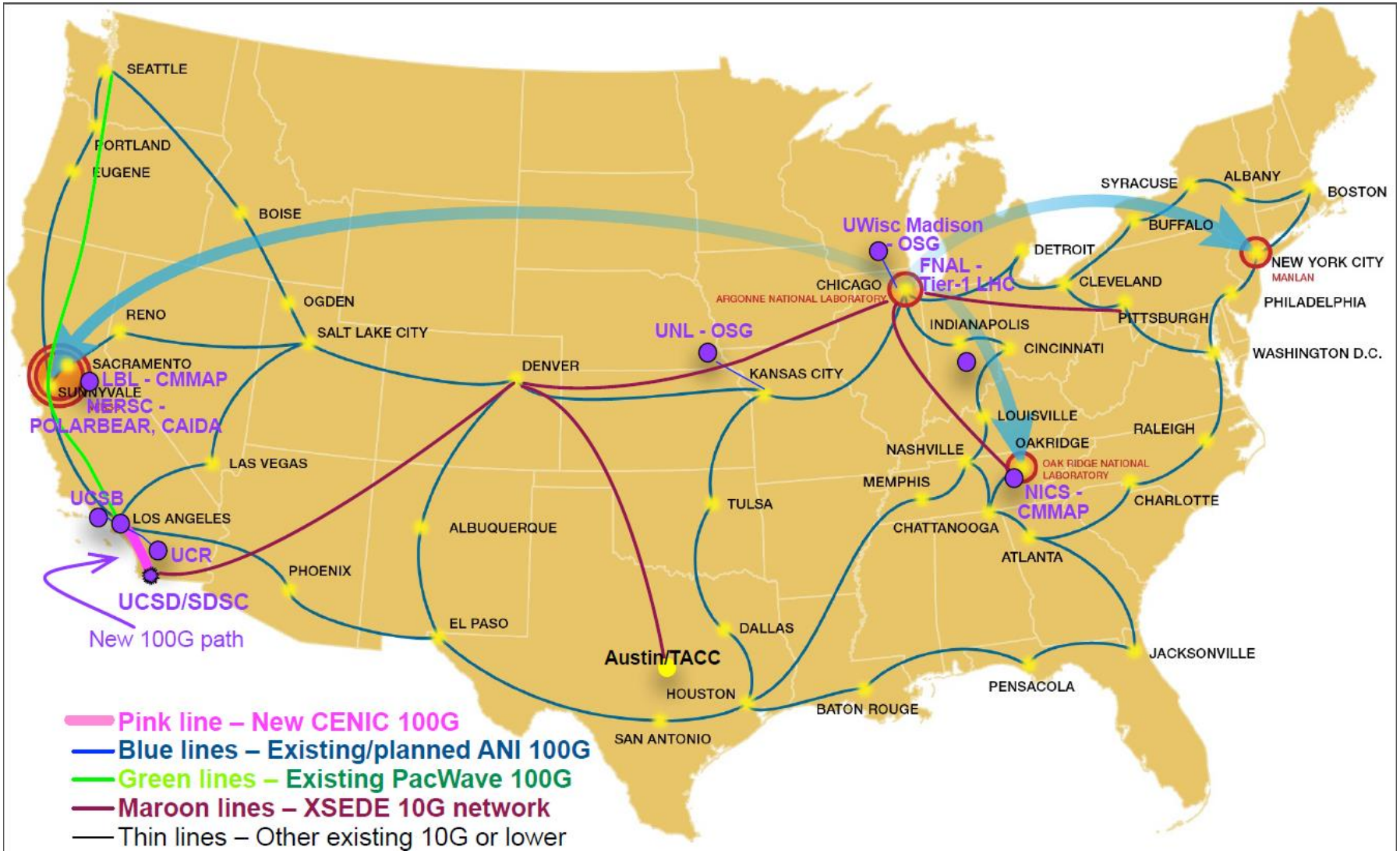
- For Oklahoma University they have 12x10G(layer-2) connectivity via OneNET
- Peer with LHCONE at 10G
- Have alternate paths to the WAN via 2x10G



UCSD

- Next slide shows a diagram showing UCSD's 100G connection in relation to the US networks. It employs an ANI map, where ESnet & I2 are confounded.
- VLANs are in place to support the Tier-2 site; once the switches are in place and configured, routing will be switched to the new path
- See details at 100g.ucsd.edu which is updated to show the status of the connection and its configuration.
  - There's been one proper test — see the news entry for 9/2014.
  - The most commonly encountered problems during testing were with flawed or dirty optics.
  - Small problems with optics led to huge throughput loss, so this is an area that warrants particular care.
- UCSD attempted to test 100G to New Zealand, but because the NZ end was configured as 10x10G, we were unable to do so.
  - Bonded channels don't handle single large flows, so this can be considered a case where architecture choices at a remote site can create network bottlenecks even when the aggregate bandwidth is high.
- CENIC wants to enable L3 on UCSD's connection and to use these connections for production purposes. That is expected 6-18 months from now.

UCSD


SEATTLE
PORTLAND
EUGENE
BOISE
RENO
OGDEN
SALT LAKE CITY
SACRAMENTO
LBL - CMMAP
SUNNYVALE
MERSC -
POLARBEAR, CAIDA
LAS VEGAS
UCSB
LOS ANGELES
UCR
UCSD/SDSC
New 100G path
PHOENIX
ALBUQUERQUE
EL PASO
DENVER
UNL - OSG
KANSAS CITY
TULSA
DALLAS
Austin/TACC
HOUSTON
SAN ANTONIO
UWisc Madison
- OSG
CHICAGO
ARGONNE NATIONAL LABORATORY
FNAL -
Tier-1 LHC
DETROIT
CLEVELAND
INDIANAPOLIS
CINCINNATI
LOUISVILLE
NASHVILLE
MEMPHIS
CHATTANOOGA
ATLANTA
OAKRIDGE
OAK RIDGE NATIONAL
LABORATORY
NICS -
CMMAP
CHARLOTTE
RALEIGH
JACKSONVILLE
PENSACOLA
BATON ROUGE
SYRACUSE
ALBANY
BOSTON
BUFFALO
NEW YORK CITY
MANLAN
PHILADELPHIA
PITTSBURGH
WASHINGTON D.C.
Pink line – New CENIC 100G
Blue lines – Existing/planned ANI 100G
Green lines – Existing PacWave 100G
Maroon lines – XSEDE 10G network
Thin lines – Other existing 10G or lower

Wisconsin

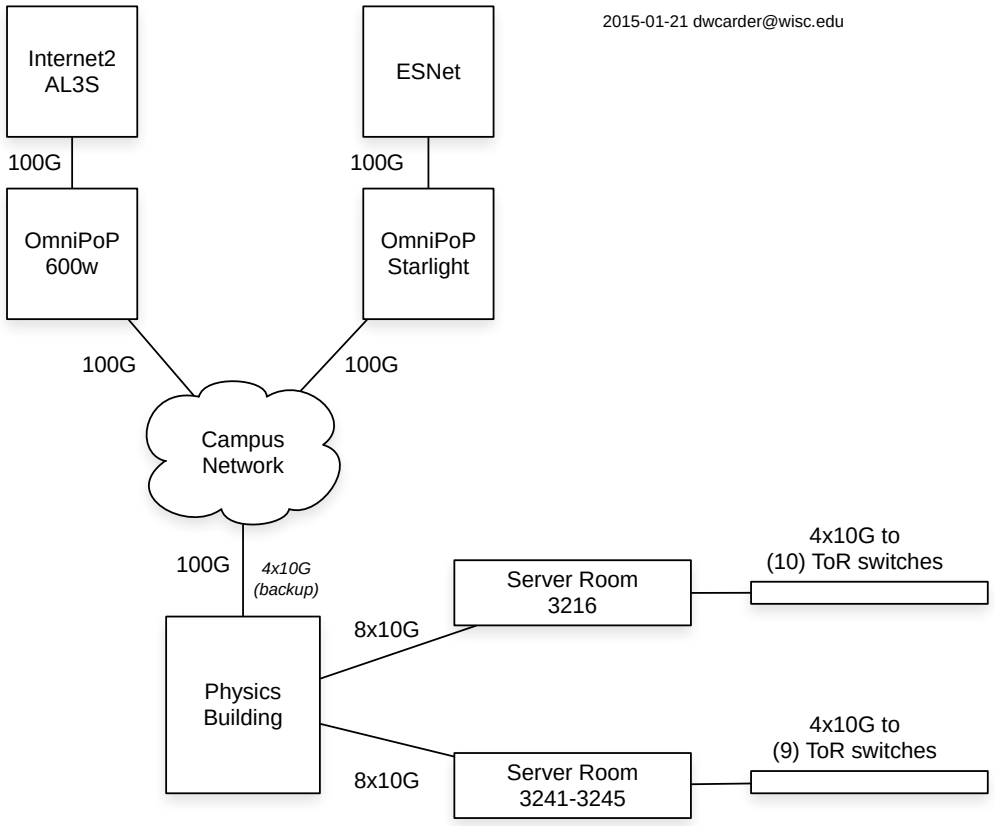
Shown is the logical network diagram for Wisconsin's Tier-2 connectivity.

Connected to both ESnet and Internet2 at 100G

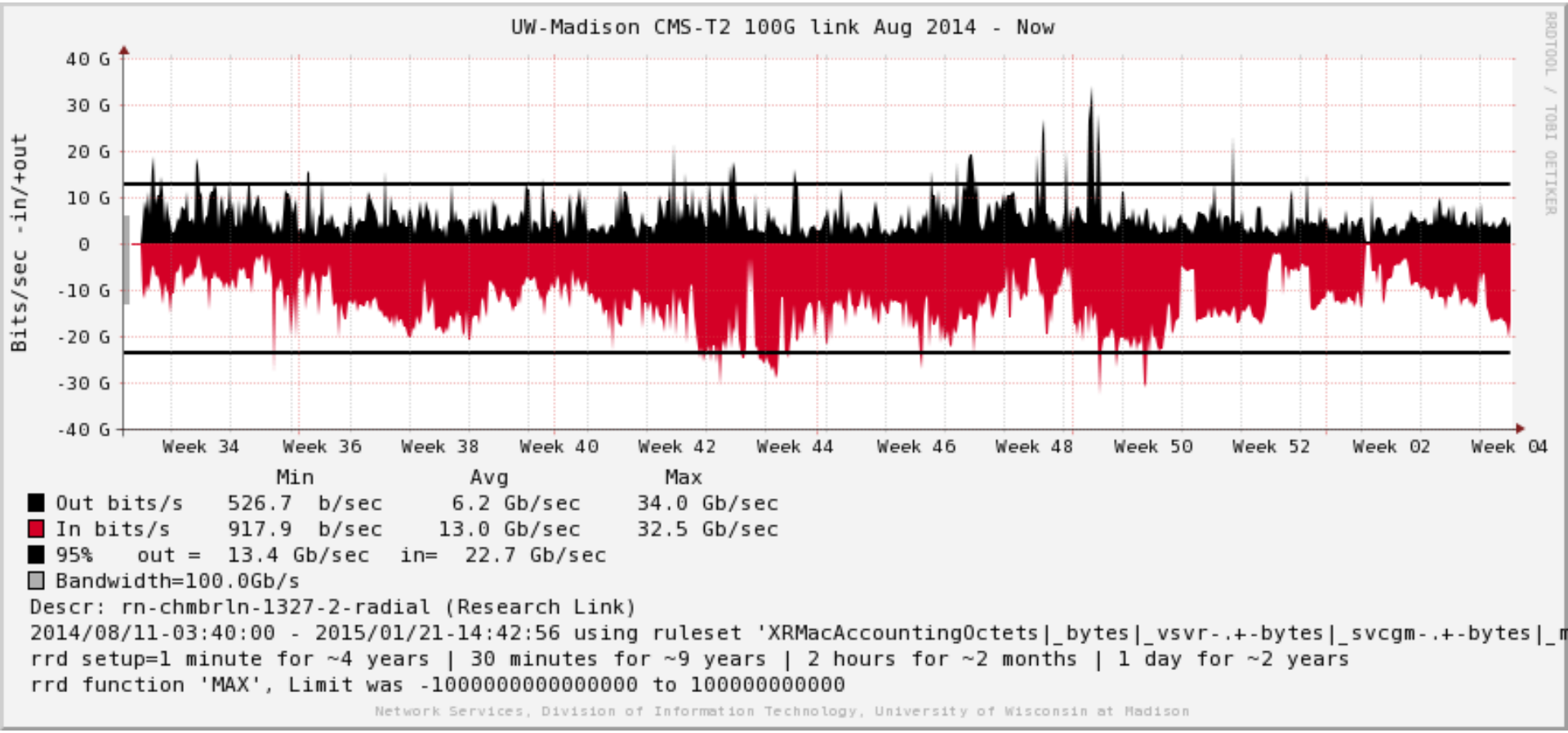
Some slides follow showing use of this network

University of Wisconsin - Madison
High Energy Physics

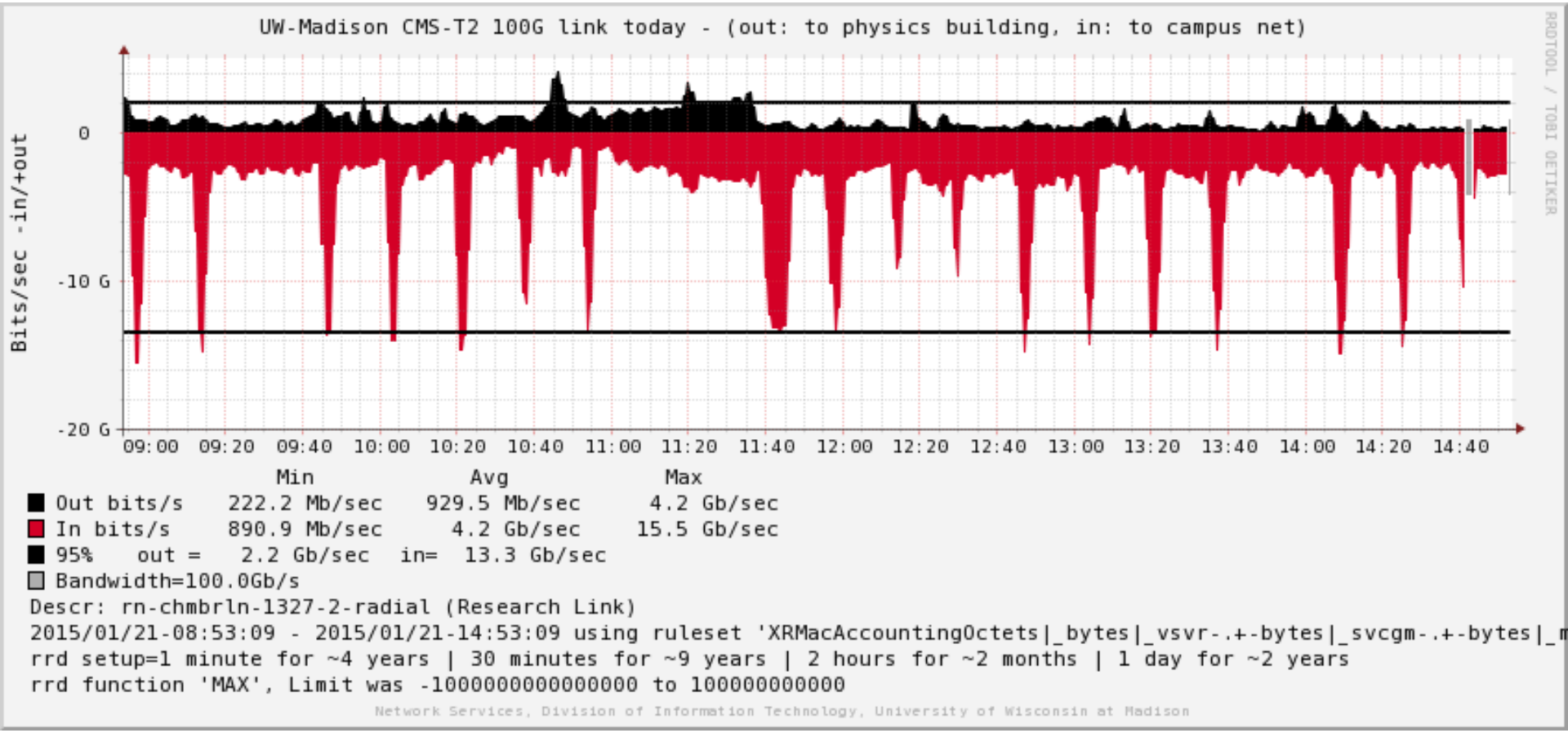
2015-01-21 dwcarder@wisc.edu



Wisconsin CMS T2 100G link since commissioning



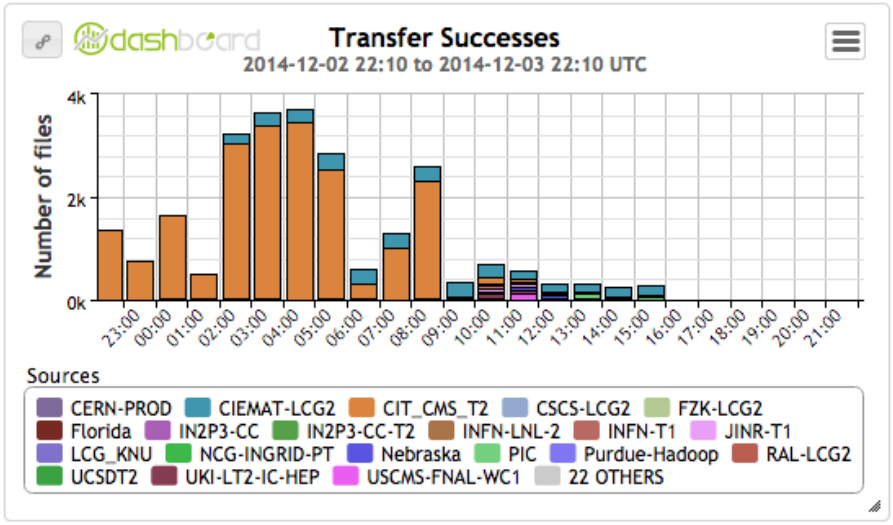
Wisconsin CMS T2 100G link, a 6hr snapshot

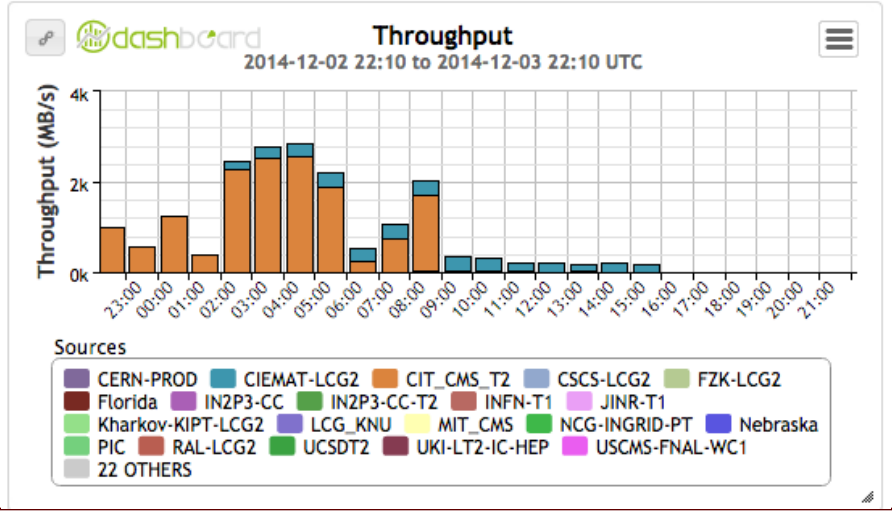


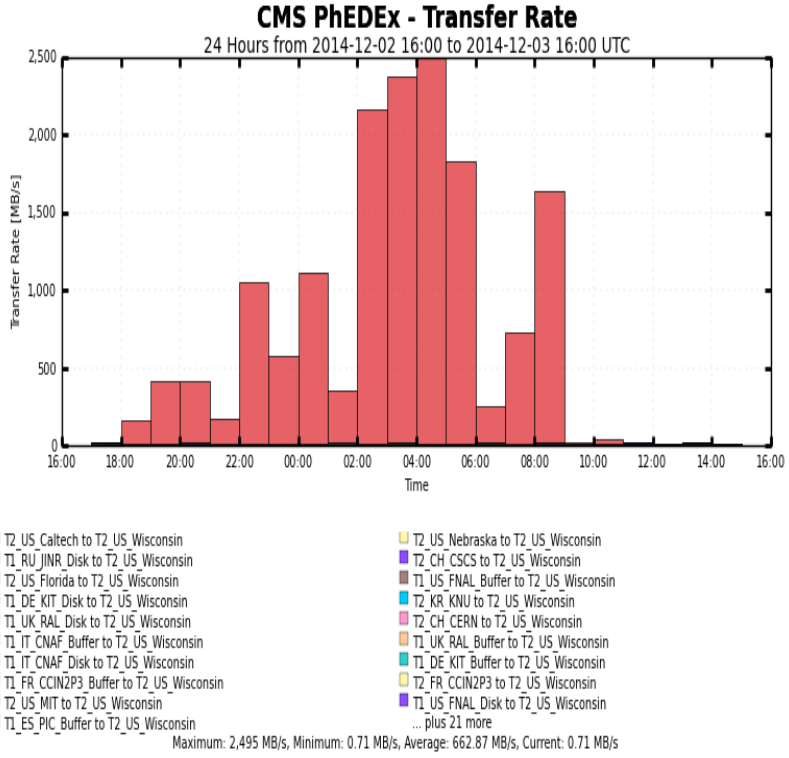
This appeared to mostly be traffic destined to Nebraska

CMS PhEDEx Transfers To T2_US_Wisconsin

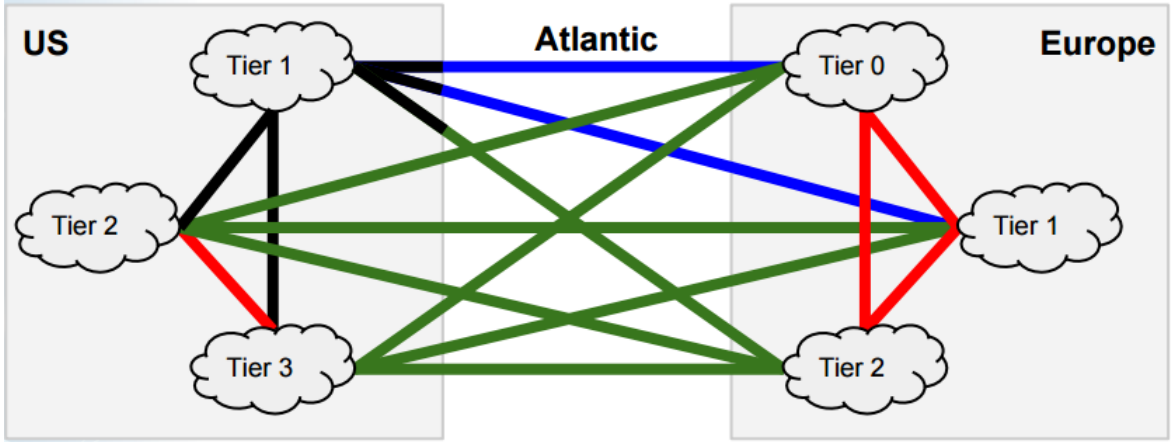
Rate reached 20Gbps







Trans-Atlantic Networking



ESnet has now taken over the trans-Atlantic networking for LHC

<= Original and new scope

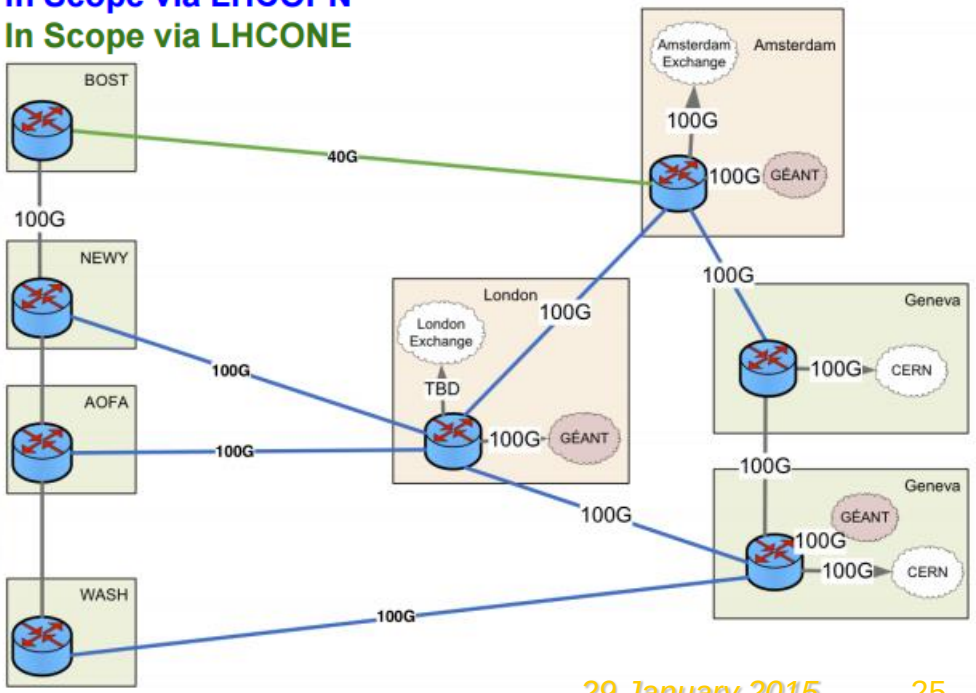
BLACK = Historical ESnet Scope
RED = Out of Scope

BLUE = In Scope via LHCOPN
GREEN = In Scope via LHCONE

Resilient multiple high bandwidth paths across the Atlantic

Normal operations provide 3x100G and 1x40G links on diverse fibers

Now working on connecting US Tier-2 sites



Summary of Facilities

- Overall all our facilities are well connected with sufficient bandwidth and resiliency
- Most of our Tier-2s either have 100G connections already or will soon
- For Run-2 and beyond we anticipate new levels of network use. Having 100G (or at least beyond 10G) connectivity will be vital for things like our XRootD federations.
  - High-speed networks can enable new modes of operation and should allow us to optimize our use of storage and compute

Projects Active in HEP Networking

- LHCOPN/LHCONE Working Group -- R&E network providers, network engineers and physicists.
- Energy Sciences Network – One of the primary developers of perfSONAR and related supporting technologies
- ANSE Project -- NSF project integrating "networking" into ATLAS and CMS
  - FTS3, Rucio Developers – Tracking ANSE and WLCG net monitoring for possible use in FTS
- WLCG Network and Transfer Metrics Working Group -- Ensure functioning and maintenance of net/transfer metrics
- OSG Networking Area – New area in OSG since June 2012 that is creating and hosting WLCG networking service.
- Federated ATLAS Xrootd (FAX) -- Measuring inter-site xrootd performance to create path/site-pair cost metrics

Using Networks Beyond 10G

- There are a number of challenges for networking beyond 10G after high-speed physical links are in place.
- A couple people on my email thread requesting details about 100G networking raised a very important point.
  - Most sites didn't have a problem on getting 10 Gbps use of the WAN
  - GridFTP and xrootd servers perform close to line rate, BUT we still didn't hear of sites running production transfers with 40+ Gbps rates out of these servers
- There is consensus that use of 100G (or 40G or 8x10G, etc) paths to WAN will happen, at least for the near term, by via lots of servers connected at 10G (or 2x10G) and sourcing and sinking data 4-6 Gbps for each 10G NIC (storage systems are often the bottleneck)
  - This is a cost effective way to benefit from improved WAN BW
- The challenge we now face is how best to manage and tune-up our data movers so the can effectively benefit from big WAN pipes

Discussion/Questions

WLCG
Worldwide LHC Computing Grid

Questions or Comments?

Some URLs

- FAX in ATLAS: https://twiki.cern.ch/twiki/bin/view/AtlasComputing/AtlasXrootdSystems
  http://dashb-atlas-xrootd-transfers.cern.ch/ui/#m.content=(active,throughput,volume)&tab=matrix
- OSG networking pages https://www.opensciencegrid.org/bin/view/Documentation/NetworkingInOSG
- WLCG Network and Transfer Metrics Working Group: https://twiki.cern.ch/twiki/bin/view/LCG/NetworkTransferMetrics
- WLCG perfSONAR installation information https://twiki.opensciencegrid.org/bin/view/Documentation/DeployperfSONAR
- Esmond (network datastore) in GitHub https://github.com/esnet/esmond

Reference

ADDITIONAL SLIDES

WLCG Network and Transfer Metrics WG

- With the current challenges in mind, we proposed to form a new WG in May:
  - **Network and Transfer Metrics WG**
- Mandate
  - Ensure all relevant network and transfer metrics are identified, collected and published
  - Ensure sites and experiments can better understand and fix networking issues
  - Enable use of network-aware tools to improve transfer efficiency and optimize experiment workflows
- Objectives
  - Identify and continuously make available relevant transfer and network metrics
    - Ensure we can consistently publish all the relevant metrics
    - Common metric attributes semantics needed for analytics/correlations
    - Improve our understanding on what metrics are needed and how we can get them
  - Document metrics and their use
  - Facilitate their integration in the middleware and/or experiment tool chain
    - Work with experiments on their use cases
  - Coordinate commissioning and maintenance of WLCG network monitoring
    - Ensure all links continue to be monitored and sites stay correctly configured
    - Verify coverage and optimize test parameters

Advance **N**etwork **S**ervices for **E**xperiments (**ANSE**) Project Overview

- ANSE is a project funded by NSF's CC-NIE program
  - Two years funding, started in January 2013, ~3 FTEs
- Collaboration of 4 institutes:
  - **Caltech (CMS)**
  - **University of Michigan (ATLAS)**
  - **Vanderbilt University (CMS)**
  - **University of Texas at Arlington (ATLAS)**
- Goal: Enable strategic workflow planning including network capacity as well as CPU and storage as a co-scheduled resource
- Path Forward: Integrate advanced network-aware tools with the mainstream production workflows of ATLAS and CMS
- Network provisioning and in-depth monitoring
- Complex workflows: a natural match and a challenge for SDN
- **Exploit state of the art progress in high throughput long distance data transport, network monitoring and control**

ANSE Objectives

- Deterministic, optimized workflows
  - Use network resource allocation along with storage and CPU resource allocation in planning data and job placement
  - Use accurate (as much as possible) information about the network to optimize workflows
  - Improve overall throughput and task times to completion
- Integrate advanced network-aware tools in the mainstream production workflows of ATLAS and CMS
  - Use tools and deployed installations where they exist
  - Extend functionality of the tools to match experiments' needs
  - Identify and develop tools and interfaces where they are missing
- Build on several years of invested manpower, tools and ideas
- Details about getting perfSONAR metrics into ANSE to follow

Beyond Monitoring

- The consensus is that good monitoring information from the network will help improve our ability to use our resources more effectively but what about negotiating with the network to further improve things?
  - Networks have moved beyond black boxes that transmit bits with some delay and variable bandwidth.
  - Users have the option to negotiate for the service(s) they require.
- Various networking services have been (and are being) developed to better optimize both network resource use and end-user experience:
  - SDN: Software Defined Networks; OpenFlow
  - NSI: Network Service Interface
  - Dynamic circuits via DYNES/AutoBahn/ION/OSCARS, etc
- **We want to make sure LHC experiments can utilize and benefit from these developments**
- **ANSE is providing "hooks" for PANDA (and PheDEx) to use SDN but it's still too early for production level end-to-end SDN (but it is coming).**

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 OSG 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?)*

OSG Networking Service

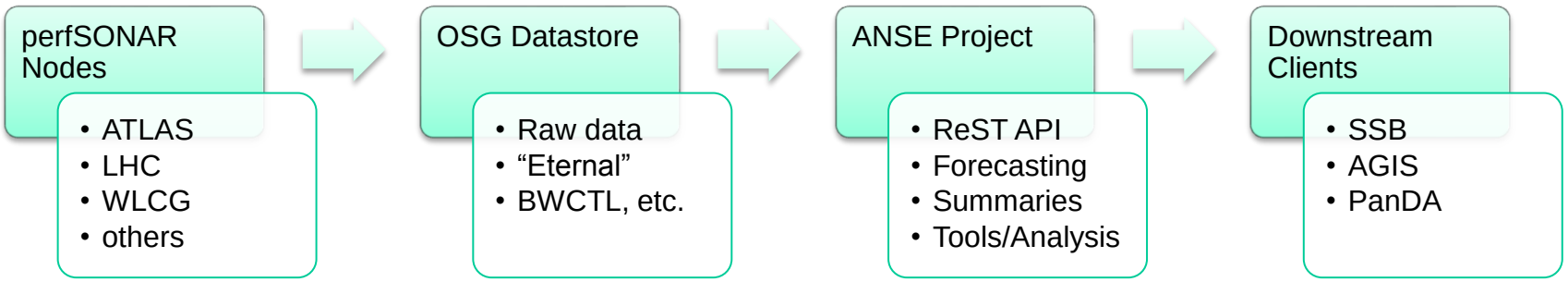
- OSG is building a centralized service for gathering, viewing and providing network information to users and applications.
- OSG is testing/deploying Esmond (Casandra backend) to organize and store the network metrics and associated metadata (Esmond is part of perfSONAR 3.4 from ESnet)
  - perfSONAR-PS stores data in a MA (Measurement Archive)
    - Each host stores **its** measurements (locally)
  - OSG (via MaDDash) is gathering relevant metrics from the complete set of OSG and WLCG perfSONAR-PS instances
  - This data must be available via an API, must be visualized and must be organized to provide the **"OSG Networking Service"**
- **This service then feeds downstream clients like ANSE, WLCG, and higher level services needing network info**
  - **Experiment frameworks, network researchers, alarming services, GUIs, etc.**

Finding/Debugging Network Problems

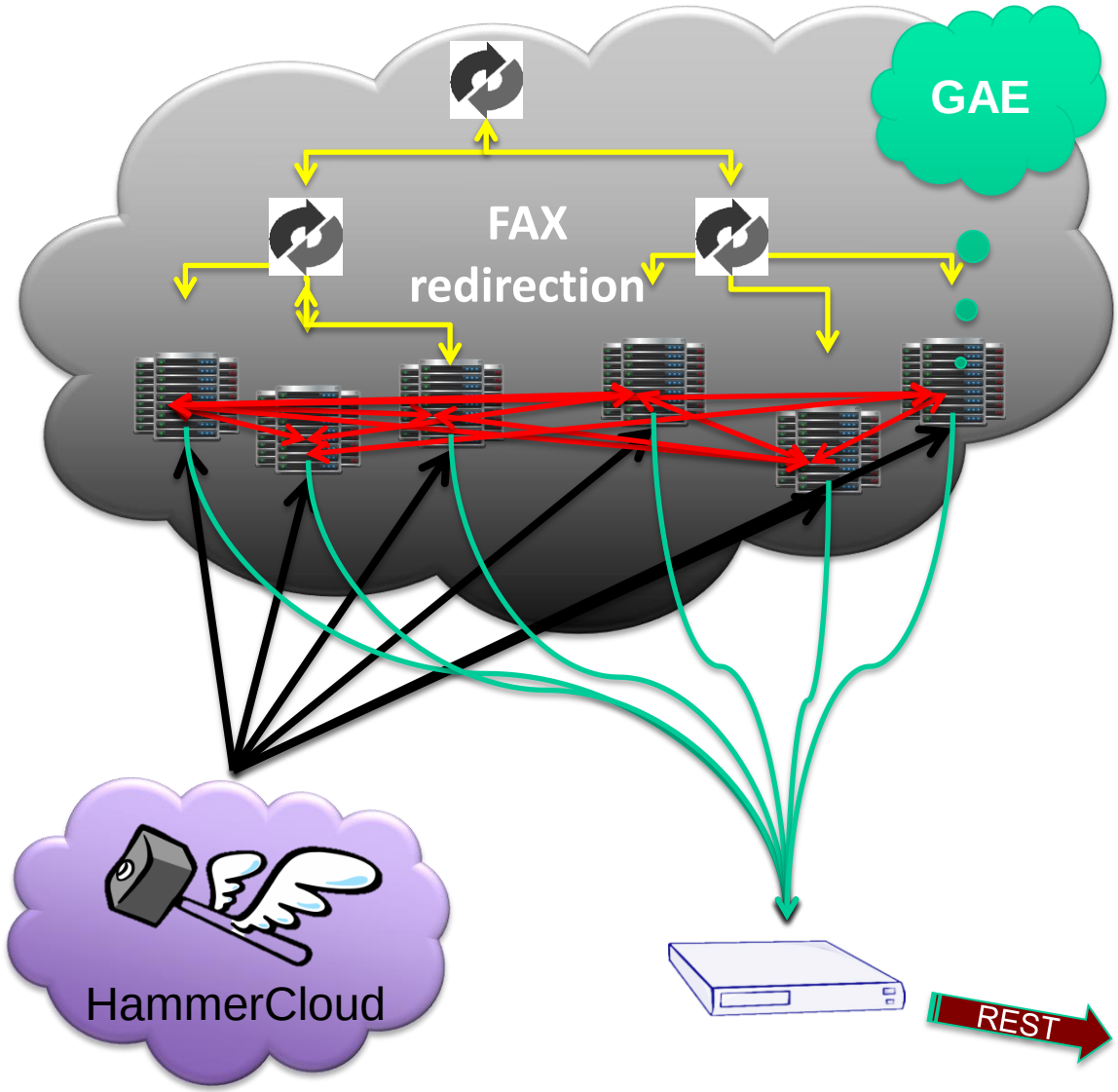
- One of the primary use-cases for LHC is to be able to quickly find network problems when they arise.
  - Often this is very difficult and time-consuming for Wide-Area Network (WAN) problems
- Scheduled perfSONAR bandwidth and latency metrics monitor WLCG network paths
  - Significant **packet-loss** or **consistent large deviation from baseline bandwidth** indicate a potential network problem (see in GUI or via alarms).
  - On-demand tests to perfSONAR instances can verify the problem exists. Different test points along the path can help pin-point the location.
  - Correlation with other paths sharing common segments can be used to localize the issue.
  - The **time** things change is also very useful to find the root causes. Scheduled tests provide this.

perfSONAR to ANSE Dataflow

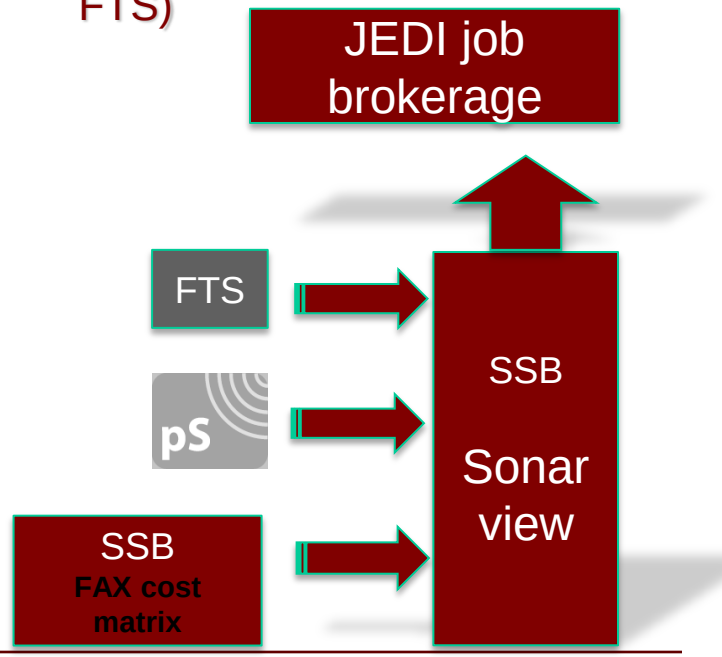
- OSG will provide the network datastore for indefinite storage of network metrics,including perfSONAR data
- ANSE provides a ReST API for access of raw data, data summaries, and generated forecasts
- Downstream clients may include
  - SSB (WLCG) – for raw historical data
  - AGIS (ATLAS) – for recent data
  - PanDA – for a forecast matrix to use in generating weights for PanDA site selections. Predictor “smooths” variations, creating better estimator for our use-cases.



FAX Cost Matrix Generation



- Data collected between 20 ANALY queues (compute sites) and 58 FAX endpoints
- Jobs submitted by HammerCloud
- Results to ActiveMQ, consumed by SSB with network & throughput measurements (perfSONAR and FTS)



FAX cost matrix