Status and Trends in Networking at LHC Tier1 Facilities

Fermi National Accelerator Laboratory, U.S.A.
Brookhaven National Laboratory, U.S.A,
Karlsruhe Institute of Technology, Germany

Presented by Andrey Bobyshev (Fermilab)

CHEP 2012, New York, U.S.A. May 21-25, 2012

Motivations

- Over decade of preparations for LHC working on all aspects of LHC computing, almost 3 years of operation
- Good cooperation between LHC centers on Wide-Area networking (LHCOPN, USLHCNET, ESNet, Internet2, GEANT) to support data movement
- Not so much on LAN issues
- Each site might have its own specifics but can we determine any commonalities ?
- Would it be useful to exchange our experiences, ideas, expectations, are we on the same track with regard to general data center networking ?

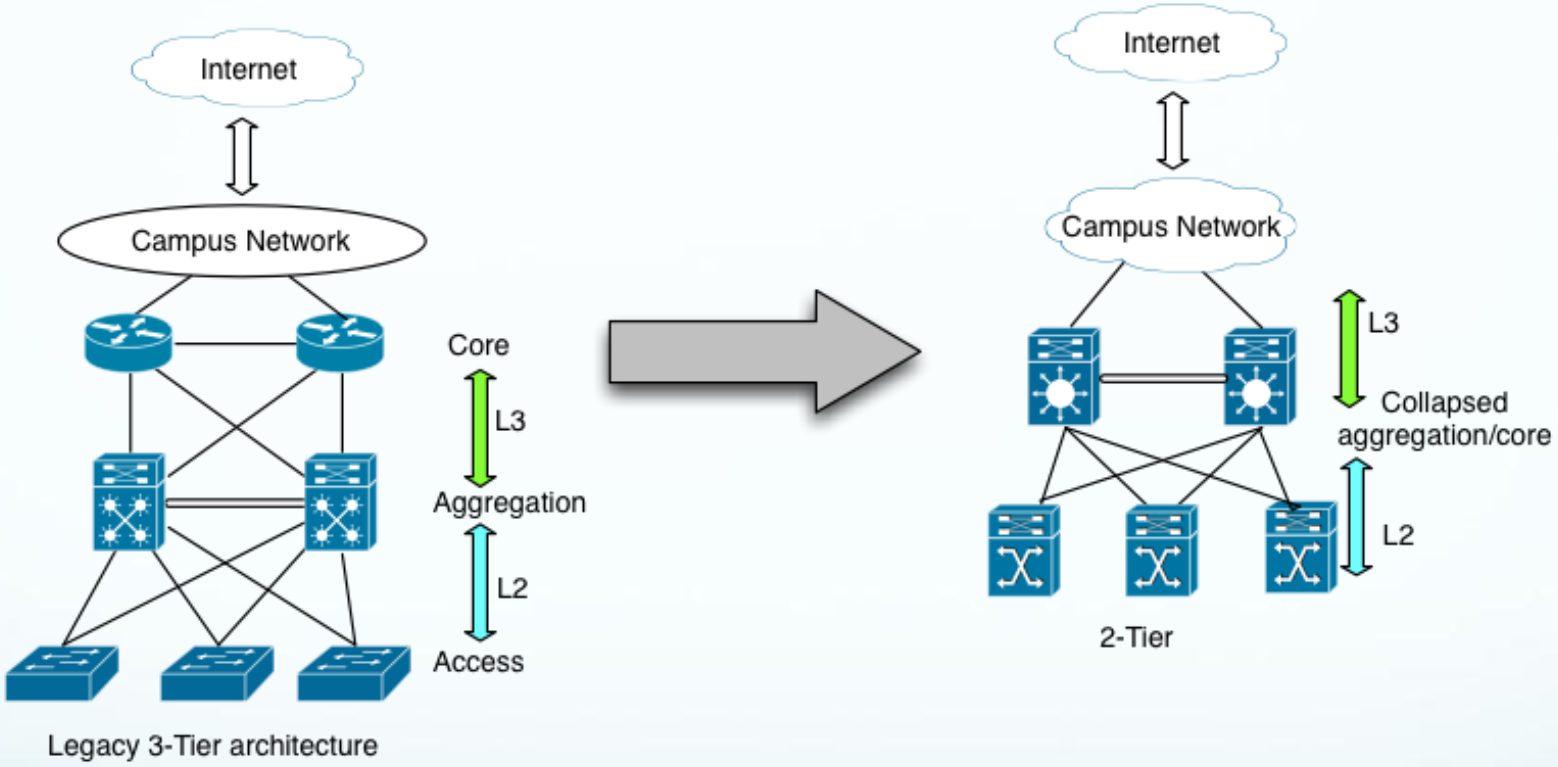
Authors:

- US-BNL-T1 (ATLAS), Brookhaven National Laboratory,U.S.A.
  - John Bigrow, big@bnl.gov
- DE-KIT Tier1 (multi-LHC), Karlsruhe Institute of Technology, Germany:
  - Bruno Hoeft, Aurelie Reymund, {bruno.hoeft, aurelie.reymund}@kit.edu
- USCMS-Tier1, Fermi National Accelerator Laboratory, U.S.A.
  - Andrey Bobyshev, Phil DeMar, Vyto Grigaliunas, {bobyshev,demar,vyto}@fnal.gov

Our objectives:

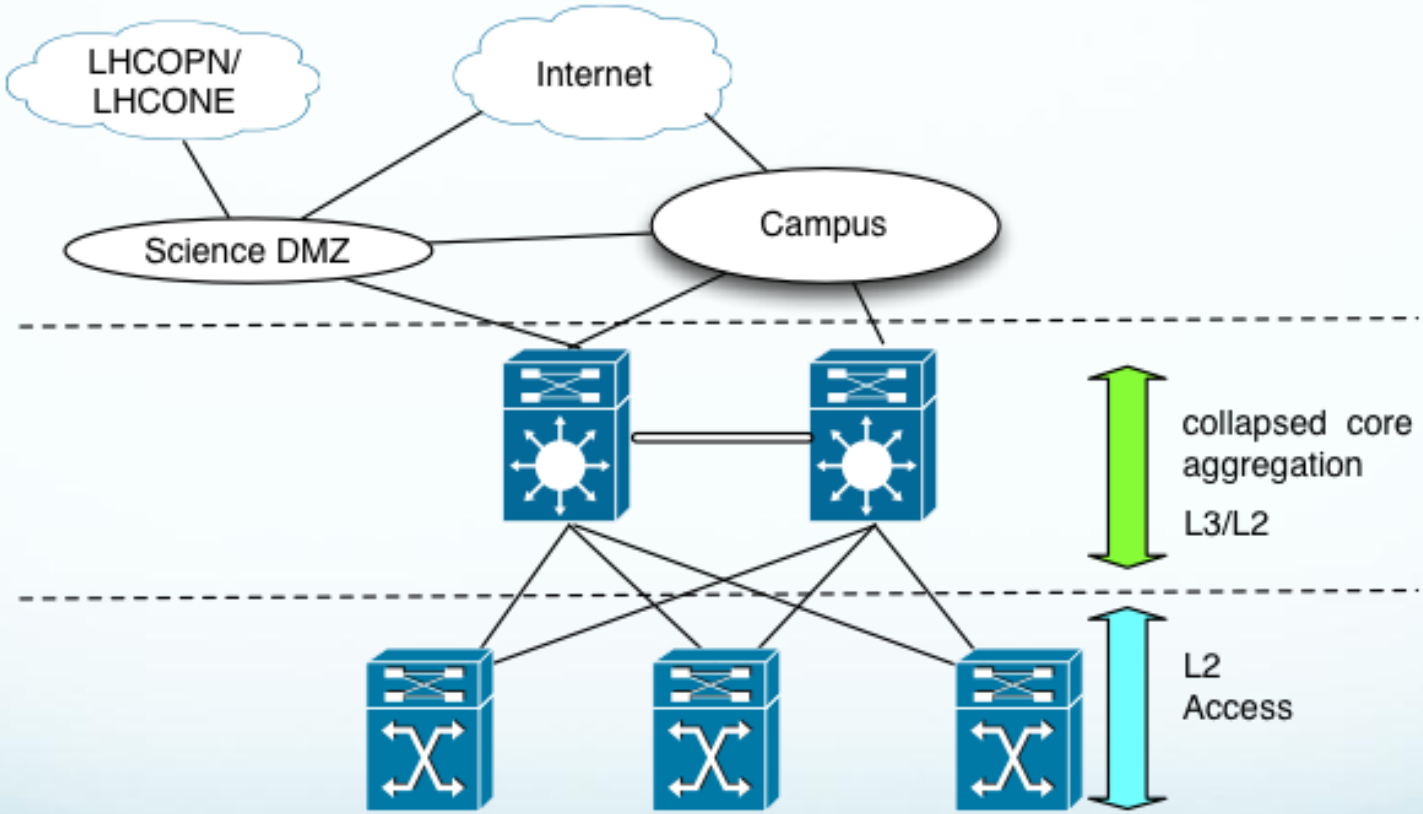
- Review Status of
  - Network architectures
  - Access solutions
- Analyze trends in :
  - 10G End systems, 40/100G inter-switch aggregation
  - Network Virtualization/sharing resources
  - Unified fabrics, Ethernet Fabrics, new architectures
  - Software-Defined Networks
  - IPv6
  - In our analysis we tried to be generic and not about any particular institution
  - Initially we planned to involve more Tier1 facilities. Due to daily routine we had to lower our ambitions and have a smaller team of volunteers

Transition of Network Architecture

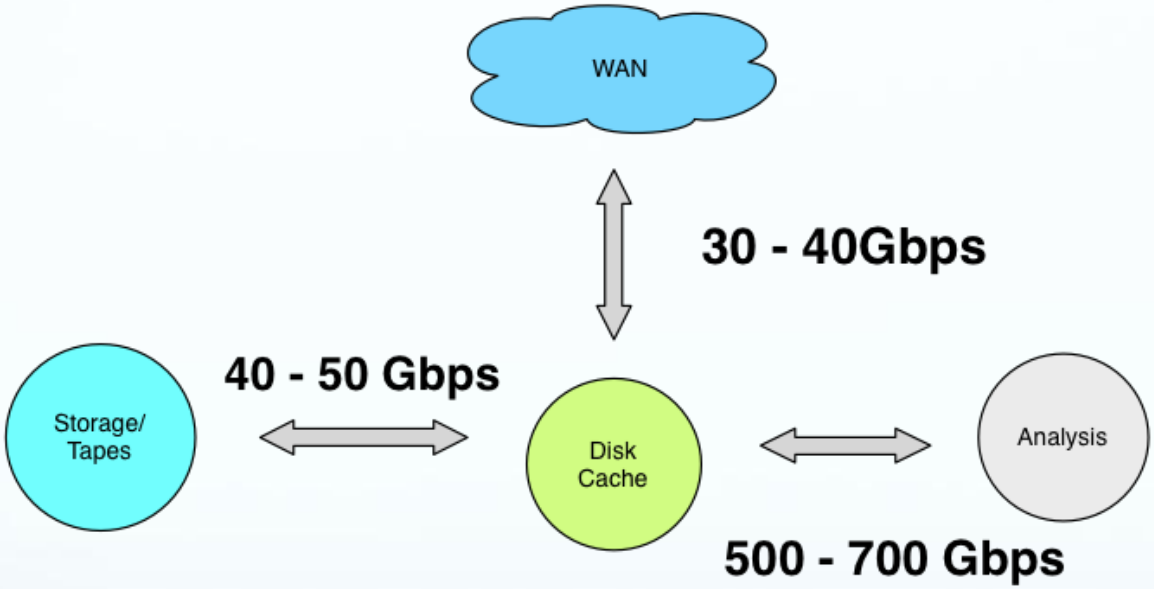


Jashree Ullal (CEO at Arista Networking) "….. *we are witnessing a shift from multi-tier enterprises with north south traffic patterns using tens/hundreds of servers, to two-tier cloud networks with east- west traffic patterns scaling across thousands of servers.. "*

Typical architecture of an LHC Tier1



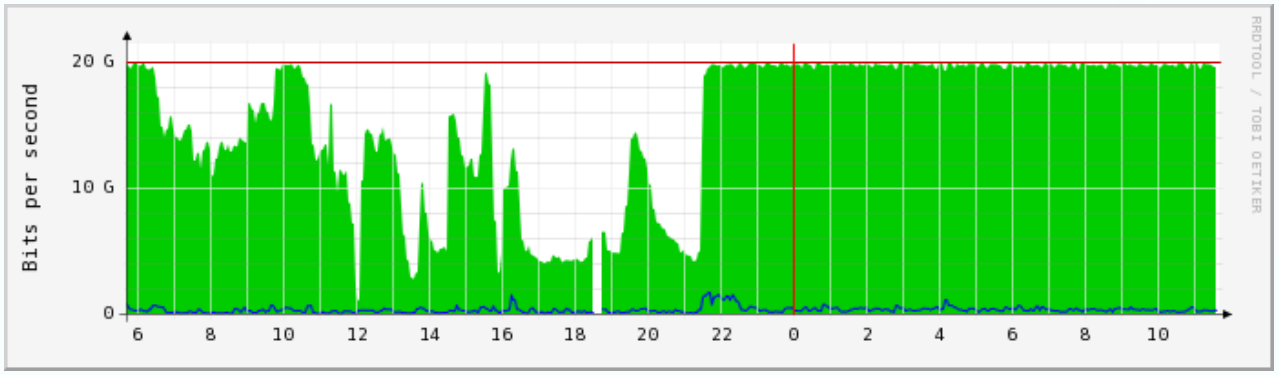
Typical bandwidth provisioned between elements of Tier1's data model



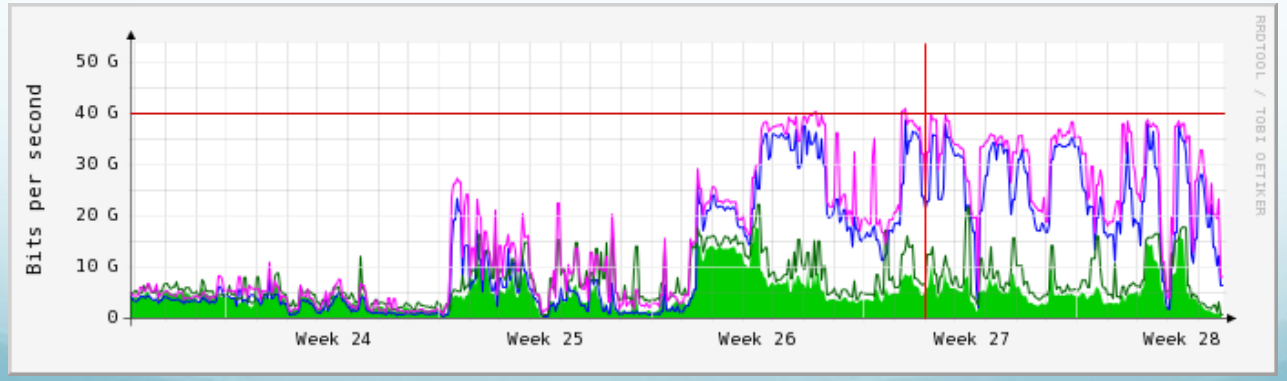
- Very low oversubscription (3:1 or 2:1) towards servers,
- No over-subscription at the aggregation layer
- QoS to ensure special treatment of selected traffic (dCache, tapes)
- Preferable access solution – "big" switches (C6509, BigIron, MLX) rather than ToR switches

Typical utilization of a LAN channel during data movement in 2008 - 2009

Channel 1 hourly



Channel 2 Weekly

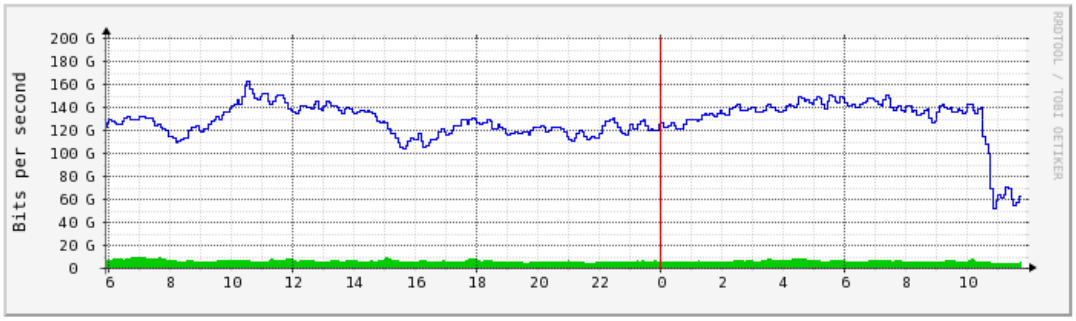


Typical utilization of a LAN channel during data movement in 2011 - 2012

Max Speed: 290 Gbits/s

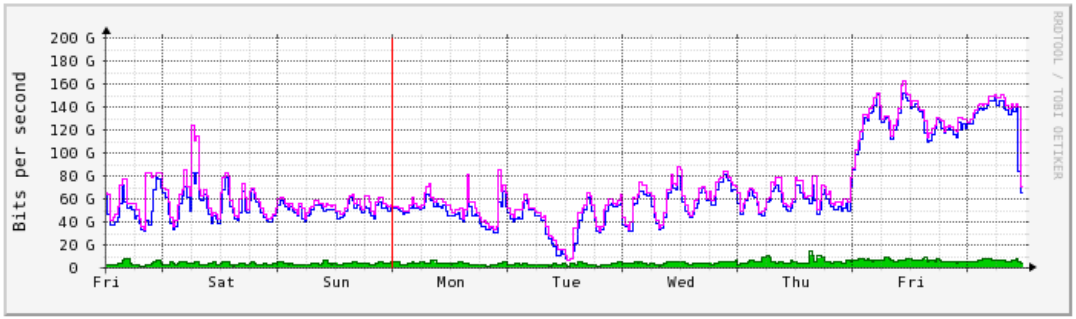
The statistics were last updated **Saturday, 12 May, 11:47:03 CDT**

`Daily' Graph (5 Minute Average)



Max In: 9731.4 Mb/s (3.4%) Average In: 5751.4 Mb/s (2.0%) Current In: 4951.3 Mb/s (1.7%)
Max Out: 162.3 Gb/s (56.0%) Average Out: 128.6 Gb/s (44.3%) Current Out: 63.0 Gb/s (21.7%)

`Weekly' Graph (30 Minute Average)



Max In: 10.3 Gb/s (3.5%) Average In: 3926.7 Mb/s (1.4%) Current In: 3851.7 Mb/s (1.3%)
Max Out: 152.2 Gb/s (52.5%) Average Out: 65.5 Gb/s (22.6%) Current Out: 64.9 Gb/s (22.4%)

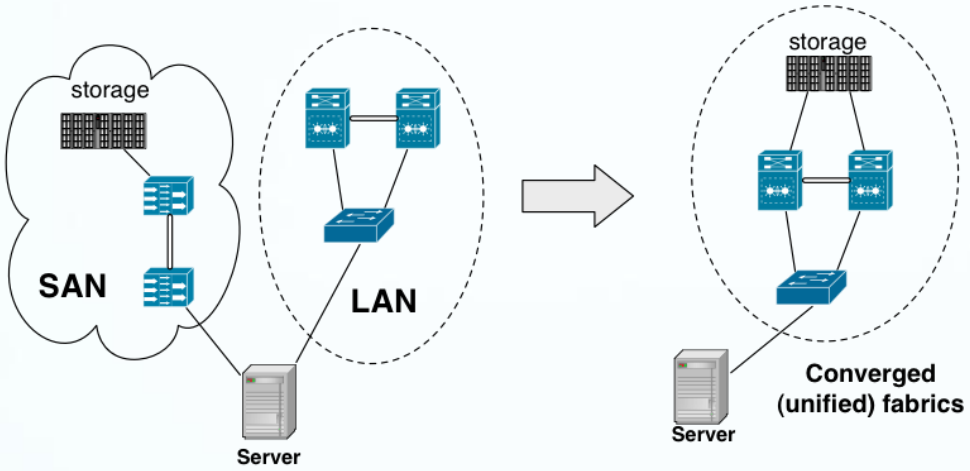
Connecting 10G servers

Cost connecting 10G servers goes down. A good situation with bandwidth provisioned now could change quickly

Different scenarios :

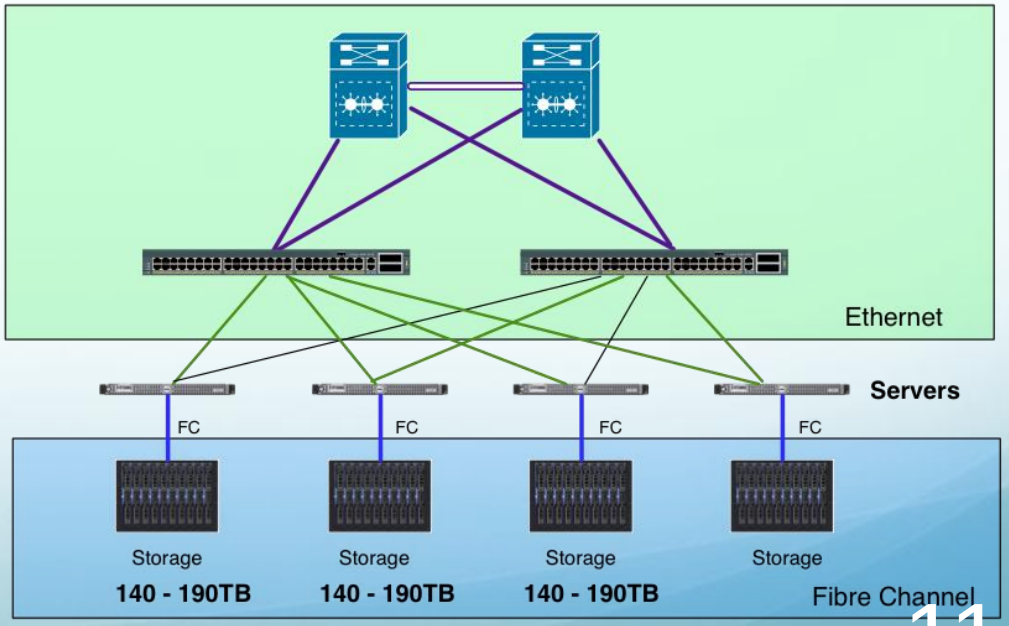
- Small deployment, 10G ToR switches
- Large deployment: Aggregation switches with high capacity switching fabrics (220+ Gbps/slot)

Converged /Unified fabrics



Storage in a generic DC

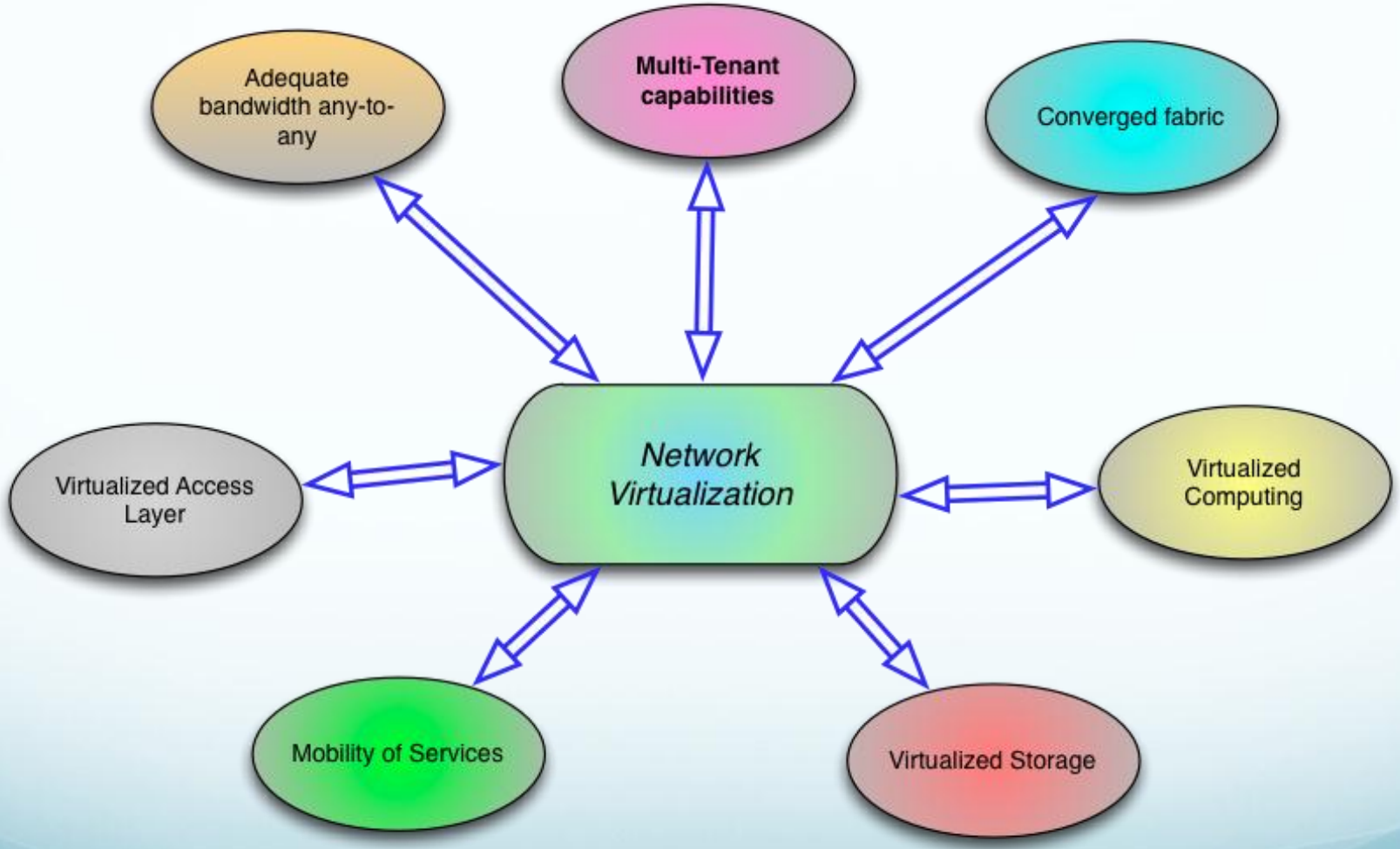
Disk Storage at LHC Tier1s



Network Virtualization Technologies

- VLANs and VPNs (a long time existing)
- Virtual Device Context (VDC)
- Virtual Routing and Forwarding (VRFs)
- Ethernet Fabrics (new L2 technology, not exactly virtualization but enables new virtualization capabilities, e.g. multiple L2 topologies)

Network Virtualization



Why Tier1s are interested in virtualization ?

- Apply security, routing, other policies per logical network, globally rather than per interfaces, devices and etc..
- Segmentation and management of resources
- Creating different levels of services
- Traffic isolation per application, user group, services

Software Defined Network

- LHC Tier1 facilities participate in early deployments and R&D
- Usually at the perimeter to access WAN circuits via Web-GUI (ESnet OSCARS, GEANT AutoBAHN, Internet2 ION)
- Not mature enough for deployment in production LAN

IPv6

- Experiments do not express any strong requirements for IPv6 yet
- In each organization there is a IPv6 deployment
- In US, OMB mandates to support IPv6 natively for all public-facing services by end FY2012. Scientific computing is not within scope
- Anticipate appearance of IPv6 only Tiers in 2-3 years, that means IPv6 deployment needs to be accelerated

Site News

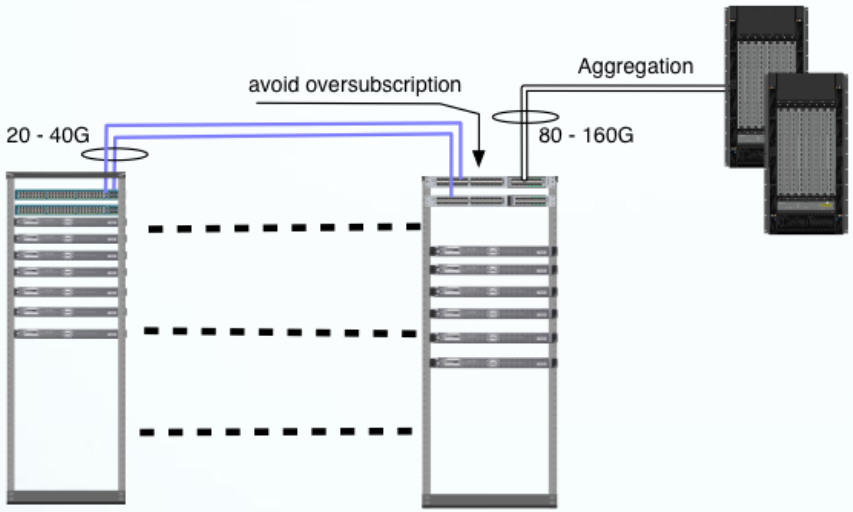
- BNL completed upgrade of the Perimeter Defense Network from 2 6509' s to a pair of Nexus 7010 to give a path to 100G from the perimeter inward
- DE-KIT has deployed at the Core two Nexus 7010's. A border router upgrade is planned for provisioning of the 100G capability for the external link to LHCOPN
- DE-KIT has deployed 50 additional 10G fileservers for matching the experiment expectations of storage accessibility (intern /extern)
- Fermilab has reached an agreement with ESnet to establish a dedicated 100G link along with several 10G waves to a new 100G ESnet5 network. This project is planned to be completed this summer
- Fermilab deployed one hundred ten 10G servers, primarily dCache nodes and tape movers

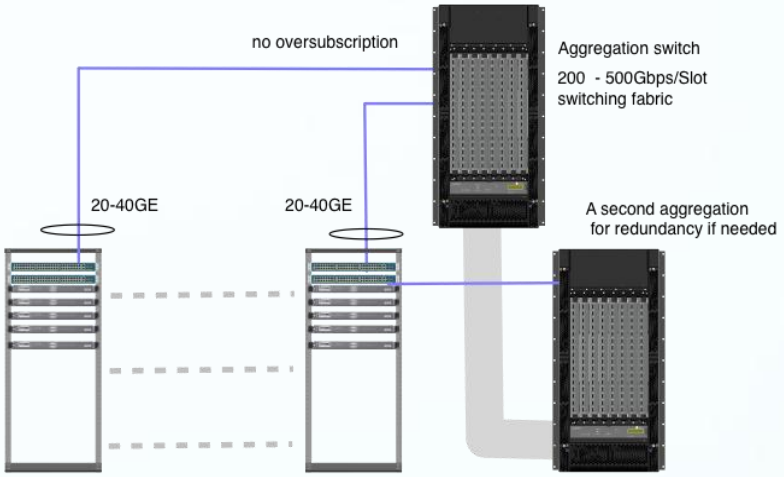
To conclude:

- We would like to be more proactive in planning network resources in LAN. For this we need to understand where we are at any particular time, what we might need to anticipate on requirements, technology progress and so on. Exchange of information, ideas, solutions between LHC centers might be useful
- If folks from other Tier1/2/3 are interested to join this informal forum feel free to contact any person listed at the beginning of this presentation

END

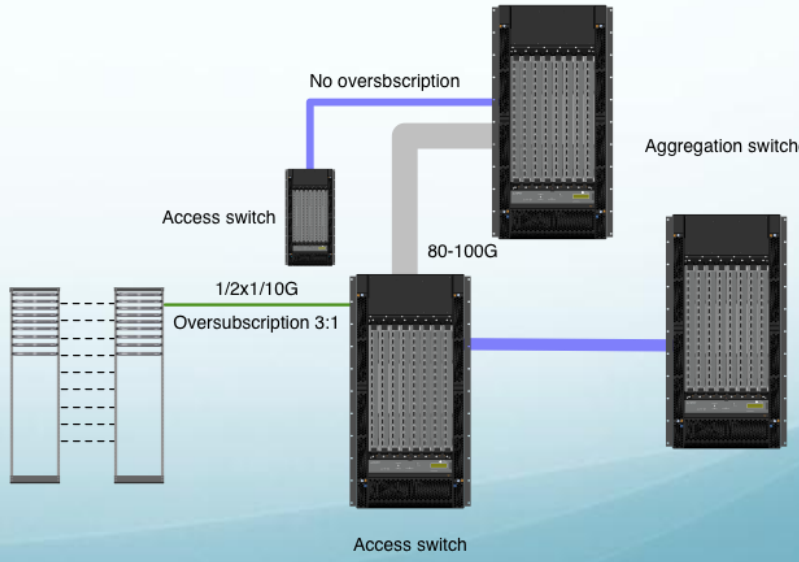
Access solutions at LHC Tier1s





A generic Data Center

An LHC Data Center



Graphs -> Weathermap -> sw99 - Traffic - Port-channel 177 - Mozilla Firefox

File Edit View History Bookmarks Tools Help

BNL Continuous Monitoring | Console -> Weathermap - Wea... | Console -> Devices -> (Edit) | Graphs -> Weathermap -> mut... | Console -> Graph Management | Graphs -> Weathermap -> sw9...

https://network.racf.bnl.gov/cacti/graph.php?rra_id=all&local_graph_id=93

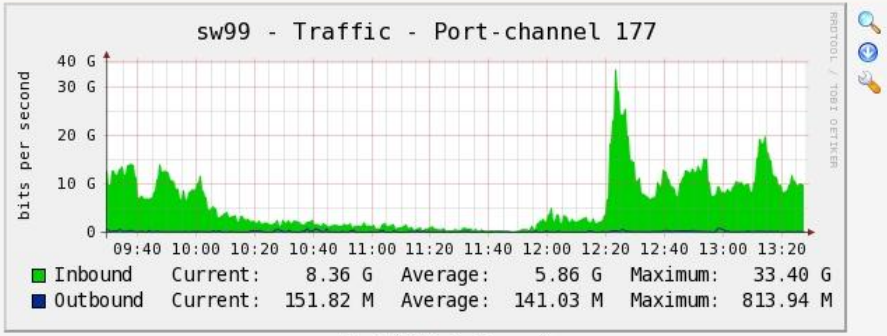
Most Visited | 32BitsOnline.com - Be... | Best of the Web | Channel Guide | Customize Links | Free Hotmail | Internet Explorer News | Internet Start | Troublshooting Guide | Windows Marketplace | Windows Media

graphs | weathermap | settings

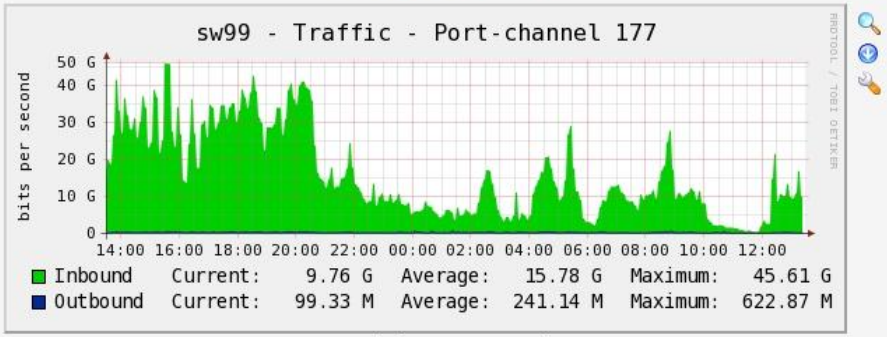
Graphs -> Weathermap -> sw99 - Traffic - Port-channel 177

Logged in as **bigrow** (Logout)

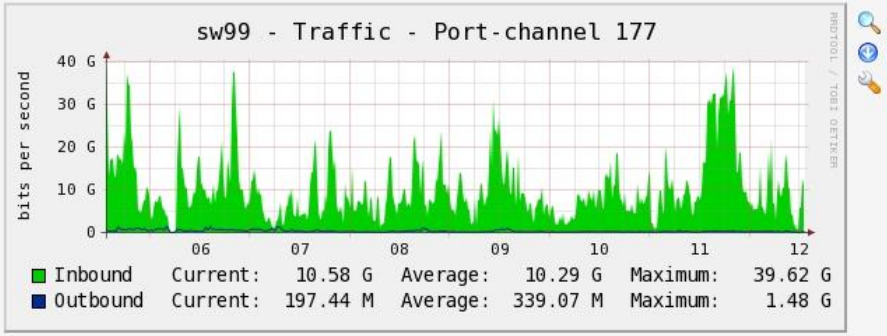
Viewing Graph 'sw99 - Traffic - Port-channel 177'



Hourly (1 Minute Average)



Daily (5 Minute Average)



Weekly (30 Minute Average)

Multi-plane Virtualized Architecture