

100G R&D at Fermilab

Gabriele Garzoglio
Grid and Cloud Computing Department
Computing Sector, Fermilab

Overview

- Fermilab Network R&D
- 100G Infrastructure at Fermilab
- Results from the ESnet 100G testbed

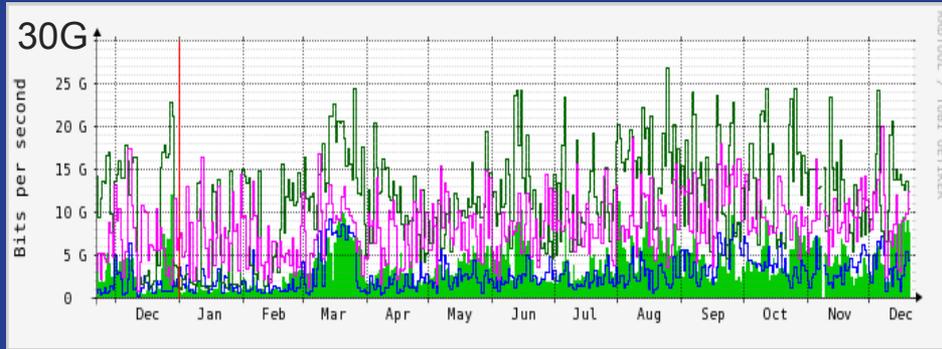
Fermilab Users and 100 GE

- Using the network for decades in the process of scientific discovery for sustained, high speed, large and wide-scale distribution of and access to data

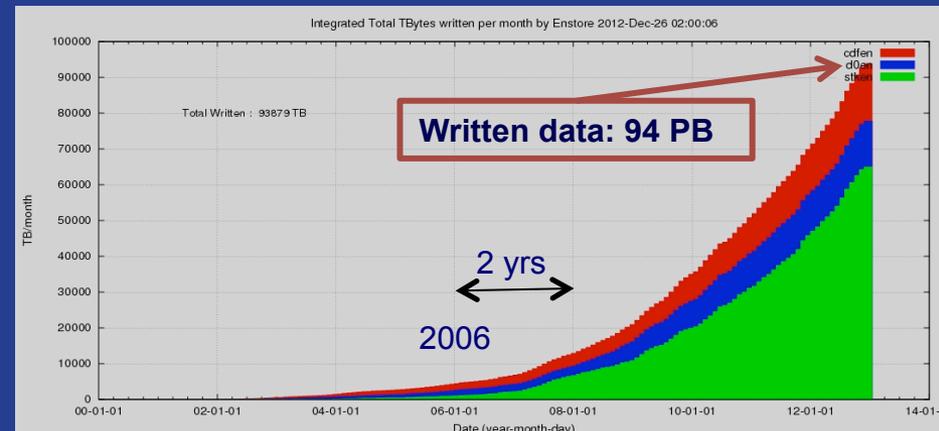
- High Energy Physics community
- Multi-disciplinary communities using grids (OSG, XSEDE)

- Figures of merit

- 40 Petabytes on tape, today mostly coming from offsite
- 140Gbps LAN traffic from archive to local processing farms
- LHC peak WAN usage at 20-30 Gbps

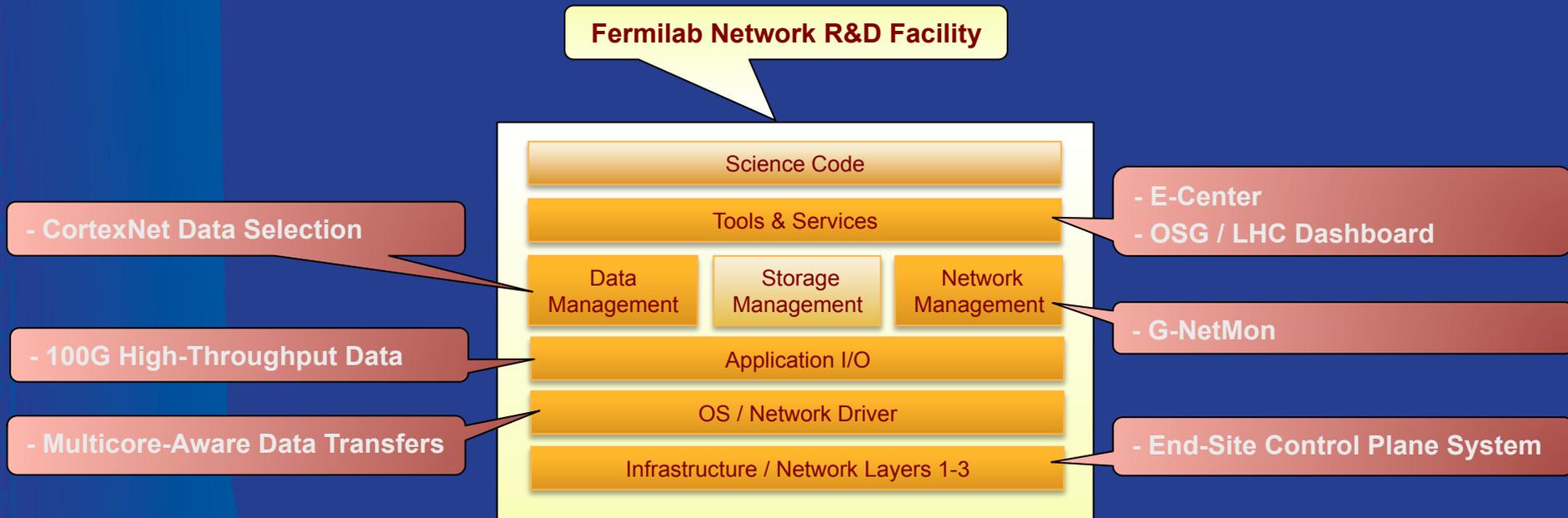


Compact Muon Solenoid (CMS) routinely peaks at 20-30 Gbps.



94 PB of data ever written to the Enstore tape archive – 54 PB available for retrieval

Network R&D at Fermilab

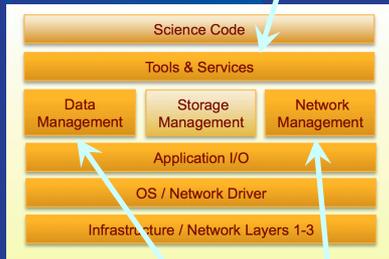


- A diverse program of work that spans all layers of computing for scientific discovery
- A collaborative process benefitting from the effort of multiple research organizations
- A broad range of activities internally and externally funded

Pulling all R&D effort together from the top layers...

Providing tools & services to enable users / applications to optimize use of the network

- Collaborating with the OSG Network Area for the deployment of perfSONAR at 100 OSG facilities
- Aggregating and displaying data through E-Center and the OSG Dashboard for end-to-end hop-by-hop paths across network domains



Developing tools to monitor real-time 100G network traffic through multi-core architectures

Proposed integration with Data Management through network-aware data source selection – CortexNET

- Seeking collaborators for network forecast module

Pulling all R&D effort together from the bottom layers...

Application-level R&D through the High Throughput Data Program

- R&D on 100GE for production use by CMS & FNAL high-capacity high-throughput Storage facility
- Identifying gaps in data movement middleware for the applications common to our stakeholders – GridFTP, SRM, Globus Online, XRootD, Frontier / Squid, NFS v4



OS-level R&D on multicore-aware data transfer middleware

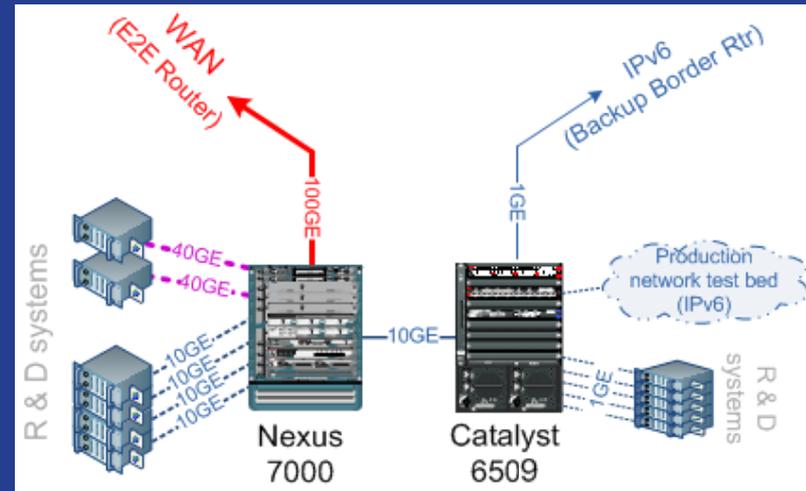
- Optimizing network I/O for 40/100GE environments

Integrating local network infrastructure with WAN circuit technologies through policy-driven configuration (ESCPS)

A dedicated R&D Network facility

- 100GE R&D
- Production-like env for tech eval
- Testing of firmware upgrades

- Nexus 7000 w/ 2-port 100GE module / 6-port 40GE module / 10GE copper module
- 12 nodes w/ 10GE Intel X540-AT2 (PCIe) / 8 cores / 16 GB RAM
- 2 nodes w/ 40GE Mellanox ConnectX®-2 (PCIe-3) / 8 cores w/ Nvidia M2070 GPU

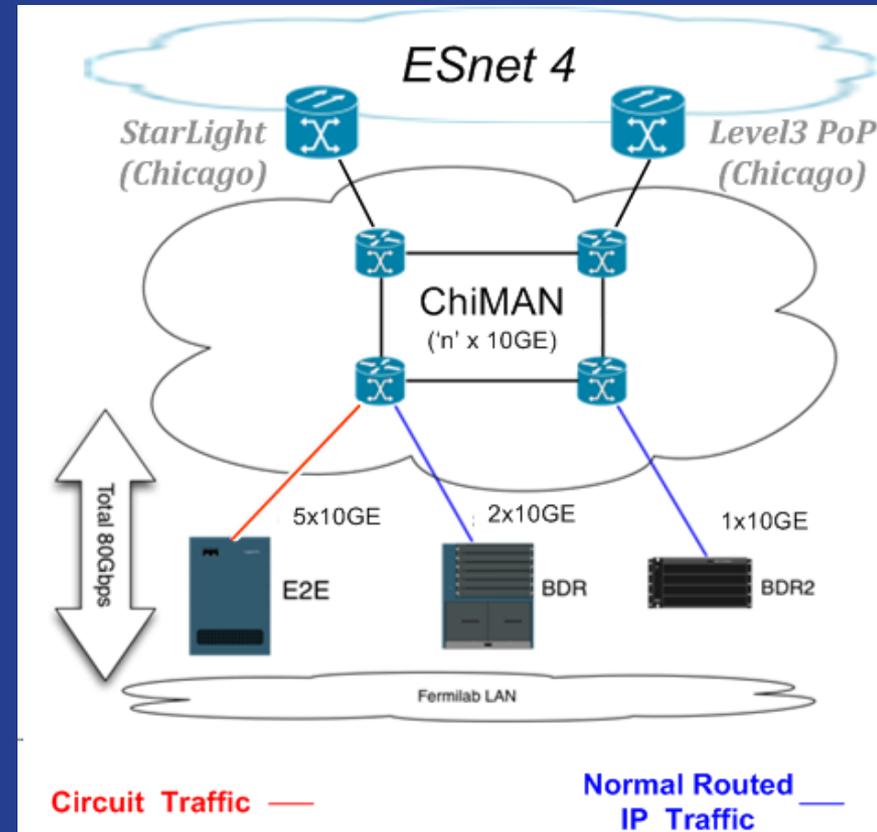


* Diagram courtesy of Phil Demar

- Catalyst 6509E for 1GE systems
 - IPv6 tests / F5 load balancer / Infoblox DNS, Palo Alto firewall

Current Fermilab WAN Capabilities

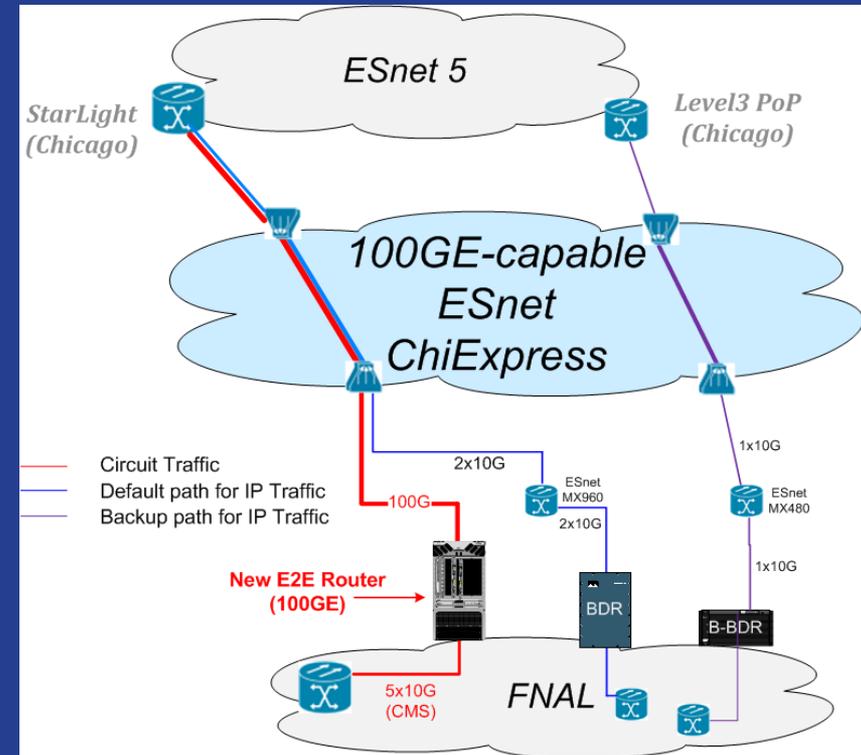
- Metropolitan Area Network provides 10GE channels:
 - Currently 8 deployed
- Five channels used for circuit traffic
 - Supports CMS WAN traffic
- Two used for normal routed IP traffic
 - Backup 10GE for redundancy
 - Circuits fail over to routed IP paths



* Diagram courtesy of Phil Demar

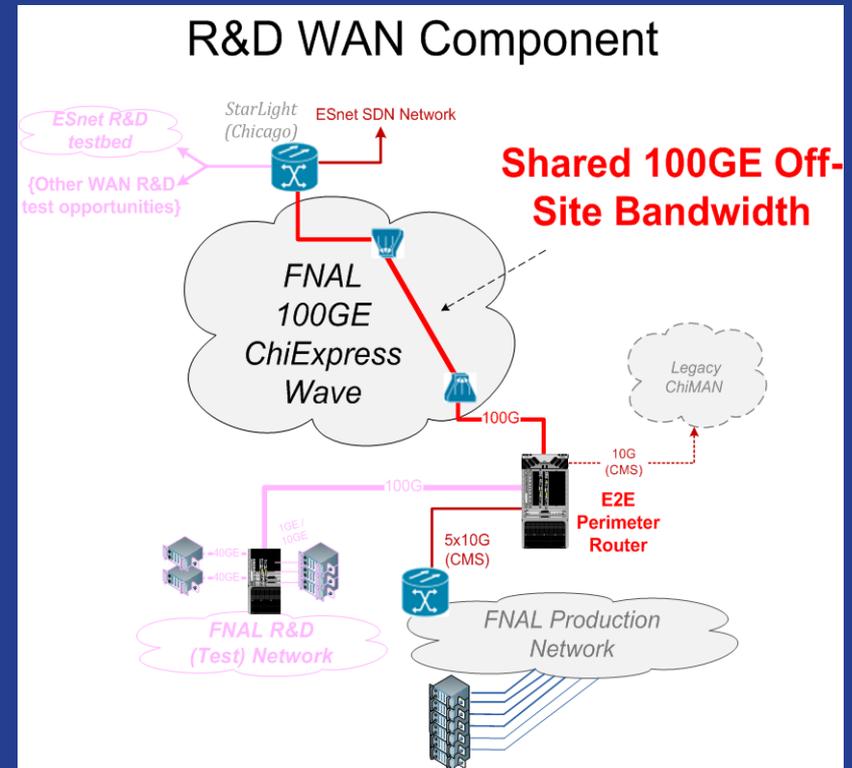
Near-Future Fermilab WAN Capabilities

- ESnet ChiExpress MAN:
 - One 100GE channel
 - Circuit-based high impact science data traffic
 - Network R&D activities
 - Three 10GE channels
 - For default routed IP traffic
 - Full geographic diversity within MAN
 - Production deployment in spring of 2013



Use of 100GE Wave for FNAL R&D Test Bed

- 100GE wave will support 50GE of CMS traffic
- Remaining ~50GE for FNAL R&D network
 - Potentially higher when CMS traffic levels are low
- Planning WAN circuit into ESnet 100G testbed
 - Potential for circuits to other R&D collaborations



Goals of 100 GE Program at Fermilab

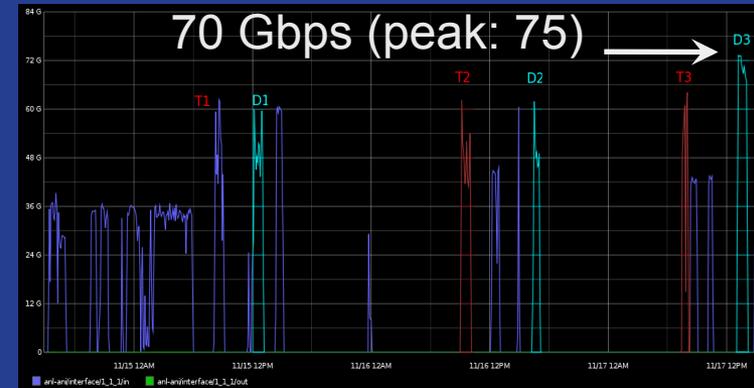
- End-to-end experiment analysis systems include a deep stack of software layers and services.
- **Need to ensure these are functional and effective at the 100 GE scale.**
 - Determine and tune the configuration to ensure full throughput in and across each layer/service.
 - Measure and determine efficiency of the end-to-end solutions.
 - Monitor, identify and mitigate error conditions.

Fermilab Network R&D Facility



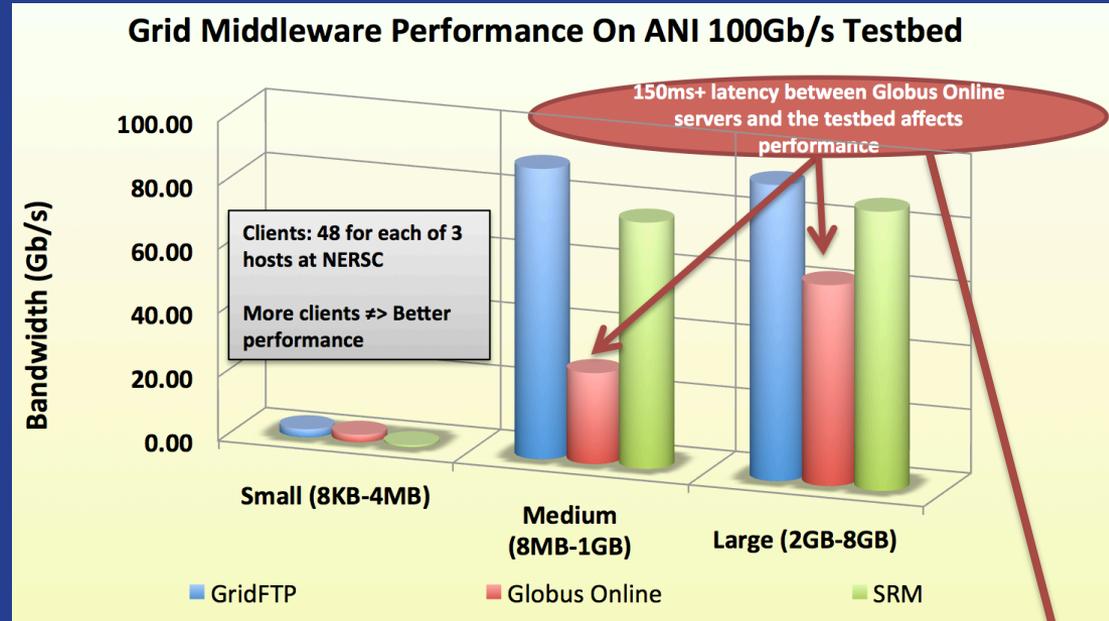
100G High Throughput Data Program

- 2011: Advanced Network Initiative (ANI) Long Island MAN (LIMAN) testbed.
 - GO / GridFTP over 3x10GE.
- 2011-2012: Super Computing '11
 - Fast access to ~30TB of CMS data in 1h from NERSC to ANL using GridFTP.
 - 15 srv / 28 clnt – 4 gFTP / core; 2 strms; TCP Win. 2MB
- 2012-2013: ESnet 100G testbed
 - Tuning parameters of middleware for data movement: xrootd, GridFTP, SRM, Globus Online, Squid. Achieved ~97Gbps
 - Commissioning Fermilab Network R&D facility: 8.5 Gbps per 10G node
- Spring 2013: 100GE Endpoint at Fermilab
 - Validate hardware link w/ transfer apps for CMS current datasets
 - Test NFS v4 over 100G for dCache and gpfs (collab. w/ IBM research)

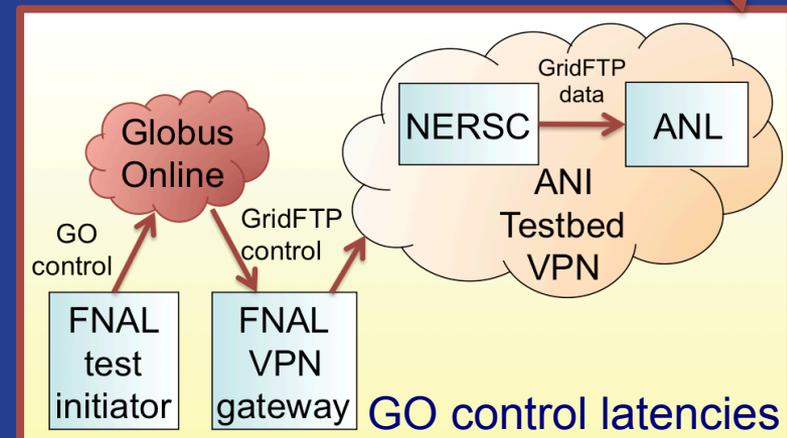


GridFTP / SRM / GlobusOnline Tests

- Data Movement using GridFTP
 - 3rd party Srv to Srv trans.: Src at NERSC / Dest at ANL
 - Dataset split into 3 size sets
- Large files transfer performance ~ 92Gbps
- Small files transfer performance - abysmally low
- Issues uncovered on Esnet 100G Testbed:
 - GridFTP Pipelining needs to be fixed on Globus implementation



Optimal performance: 97 Gbps w/ GridFTP
2 GB files – 3 nodes x 16 streams / node



GO control channel sent to the VPN through port forwarding

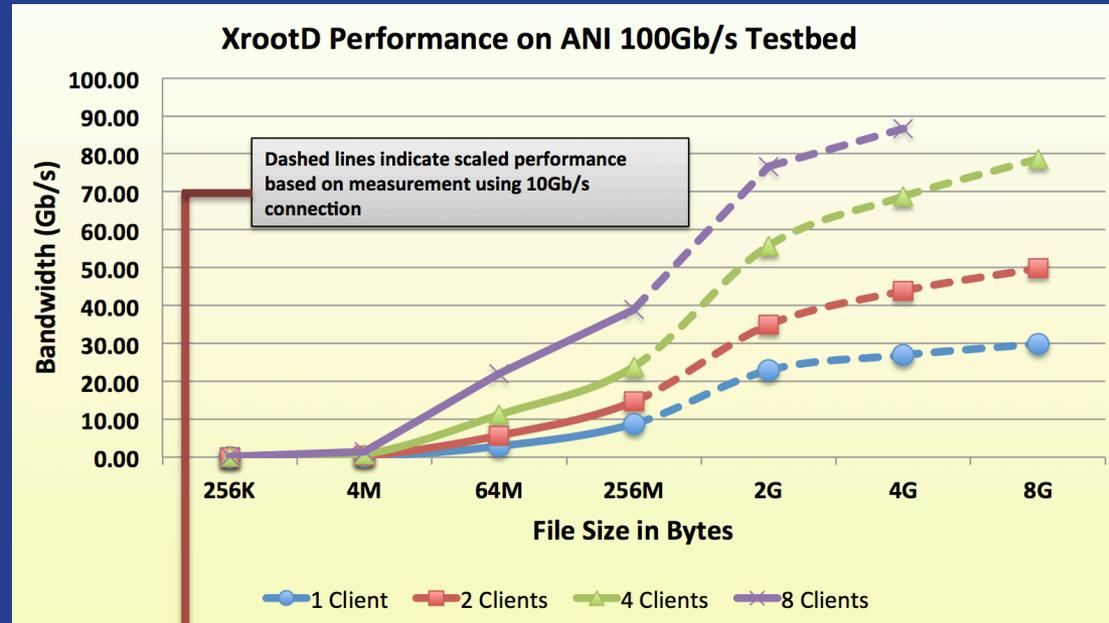
XRootD Tests

- Data Movement over XRootD, testing LHC experiment (CMS / Atlas) analysis use cases.

- Clients at NERSC / Servers at ANL
- Using RAMDisk as storage area on the server side

Challenges

- Tests limited by the size of RAMDisk
- Little control over xrootd client / server tuning parameters

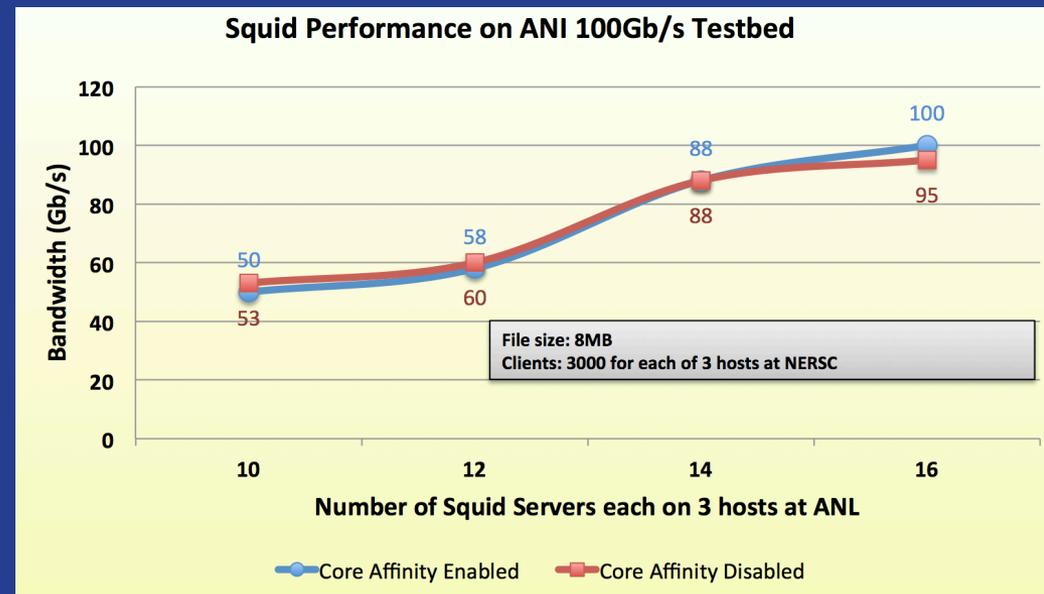


Dataset (GB)	1 NIC measurements (Gb/s)	Aggregate Measurements (12 NIC) (Gb/s)	Scale Factor per NIC	Aggregate estimate (12 NIC) (Gb/s)
0.512	4.5	46.9	0.87	—
1	6.2	62.4	0.83	—
4	8.7 (8 clients)	—	0.83	86.7
8	7.9 (4 clients)	—	0.83	78.7

Calculation of the scaling factor between 1 NIC and an aggregated 12 NIC for datasets too large to fit on the RAM disk

Squid / Frontier Tests

- Data transfers
 - Cache 8 MB file on Squid – This size mimics LHC use case for large calib. data
 - Clients (wget) at NERSC / Servers at ANL
 - Data always on RAM
- Setup
 - Using Squid2: single threaded
 - Multiple squid processes per node (4 NIC per node)
 - Testing core affinity on/off: pin Squid to core i.e. to L2 cache
 - Testing all clnt nodes vs. all servers AND aggregate one node vs. only one server



- Results
 - Core-affinity improves performance by 21% in some tests
 - Increasing the number of squid processes improves performance
 - Best performance w/ 9000 clients: ~100 Gbps

Summary

- The Network R&D at Fermilab spans all layers of the communication stack
- Fermilab is deploying a Network R&D facility with 100G capability
- ESnet 100G Testbed has been fundamental for our middleware validation program
- Fermilab will have 100GE capability in the Spring 2013
 - Planning to participate in the ESnet 100G Testbed

High Throughput Data Program (HTDP) at Fermilab

- **Mission:** prepare the Computing Sector and its stakeholders for the 100GE infrastructure and put Fermilab in a strategic position of leadership.
- Establish collaborations with stakeholders, computing facilities, scientific communities, and institutions, to coordinate a synergistic program of work on 100GE.
- The program includes technological investigations, prototype development, and participation on funding agency solicitations.
- Close collaboration with the OSG network area to instrument the cyber-infrastructure (PerfSONAR) and provide nation-wide network metrics.
- The ANI has been the major testbed used since 2011 in close partnership with ESNet

Plans

- 100GE production endpoint coming to Fermilab (see next slides)
 - Expecting 100 GE capabilities by Spring 2013
 - Connecting local cluster (FermiCloud Integration Testbed) to ESNet testbed
 - Validating 100GE link at Fermilab running measurements of middleware already tested on ANI.
- Continue testing of middleware technologies defined by stakeholders
 - Now planning measurements on NFS v4 for dCache on ESNet testbed

