100G R&D at Fermilab

Gabriele Garzoglio
(for the High Throughput Data Program team)
Grid and Cloud Computing Department
Computing Sector, Fermilab

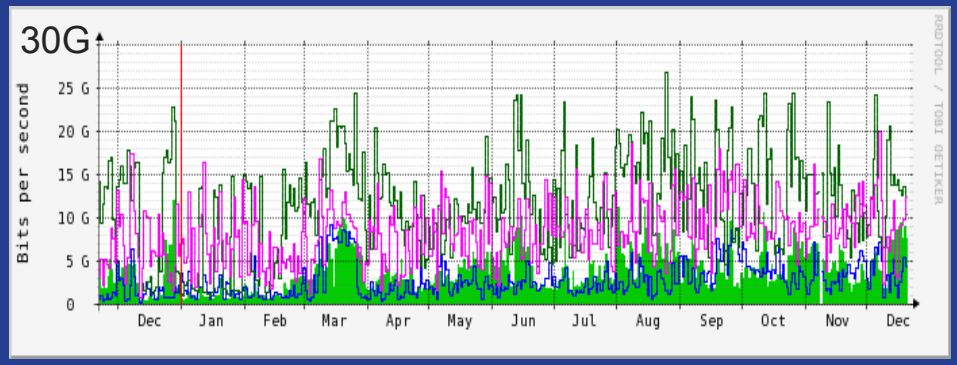
Overview

- Fermilab's interest in 100 GE
- Results from the ESNet 100GE testbed
- 100 GE infrastructure at Fermilab

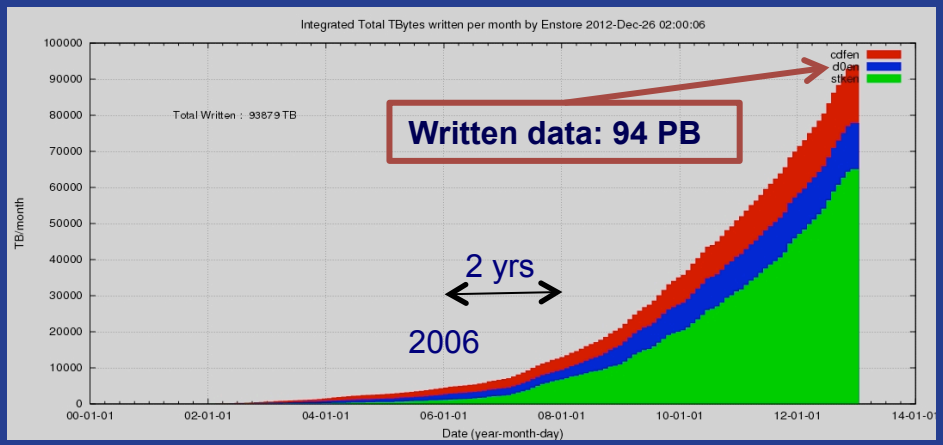
U.S. DEPARTMENT OF ENERGY
Fermilab

Fermilab Users and 100 GE

- Decades long dependence on 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

U.S. DEPARTMENT OF ENERGY

Fermilab

Goals of 100 GE Program at Fermilab

- End-to-end experiment analysis systems include a deep stack of software layers and services.

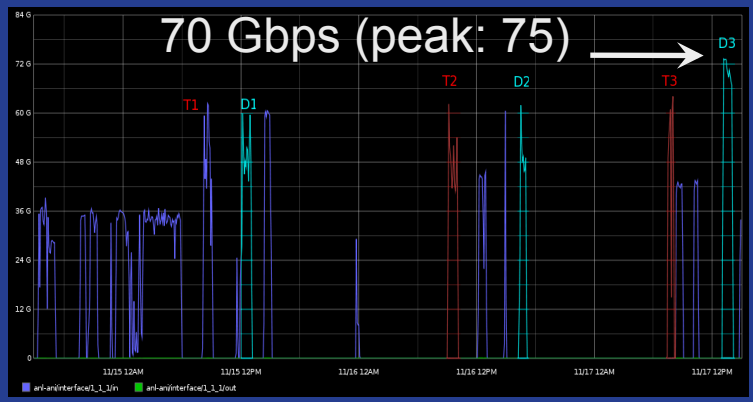
- **Need to ensure these are functional and effective at the 100 GE scale.**
  - Determine and tune the configuration to ensure full throughput in and across each layer/service.
  - Measure and determine efficiency of the end-to-end solutions.
  - Monitor, identify and mitigate error conditions.

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

Ongoing Program of Work

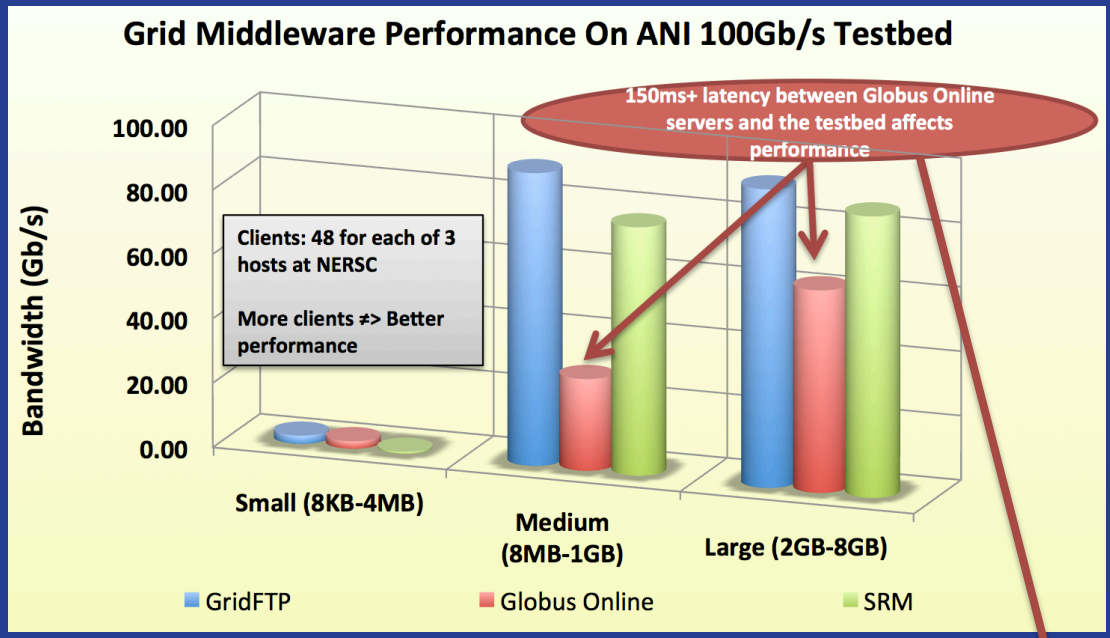
- 2011: ANI Long Island MAN (LIMAN) testbed.
  - GO / GridFTP over 3x10GE.
- 2011-2012: Super Computing '11
  - Fast access to ~30TB of CMS data in 1h from NERSC to ANL using GridFTP.
  - 15 srv / 28 clnt – 4 gFTP / core; 2 strms; TCP Win. 2MB



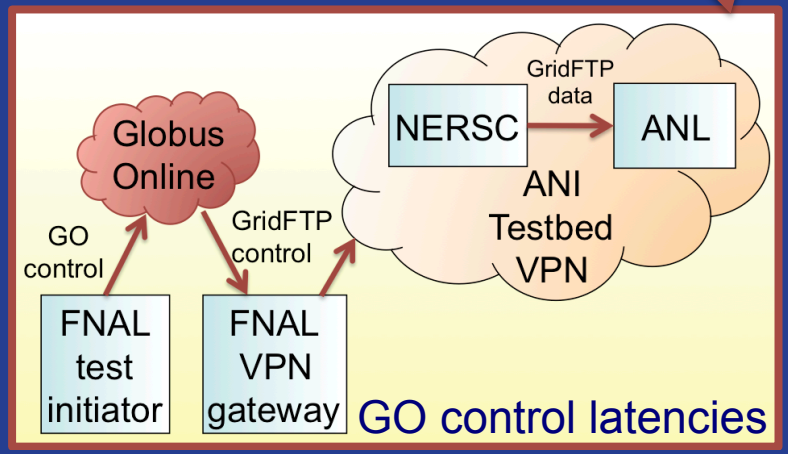
- **Currently: 100GE testbed (focus of this talk)**
  - **Tuning parameters of middleware for data movement: xrootd, GridFTP, SRM, Globus Online, Squid**
  - **Achieved ~97Gbps**
- Spring 2013: 100GE Endpoint at Fermilab
  - Plan to repeat and extend tests using CMS current datasets.

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 100G Testbed:
  - GridFTP Pipelining needs to be fixed on Globus implementation



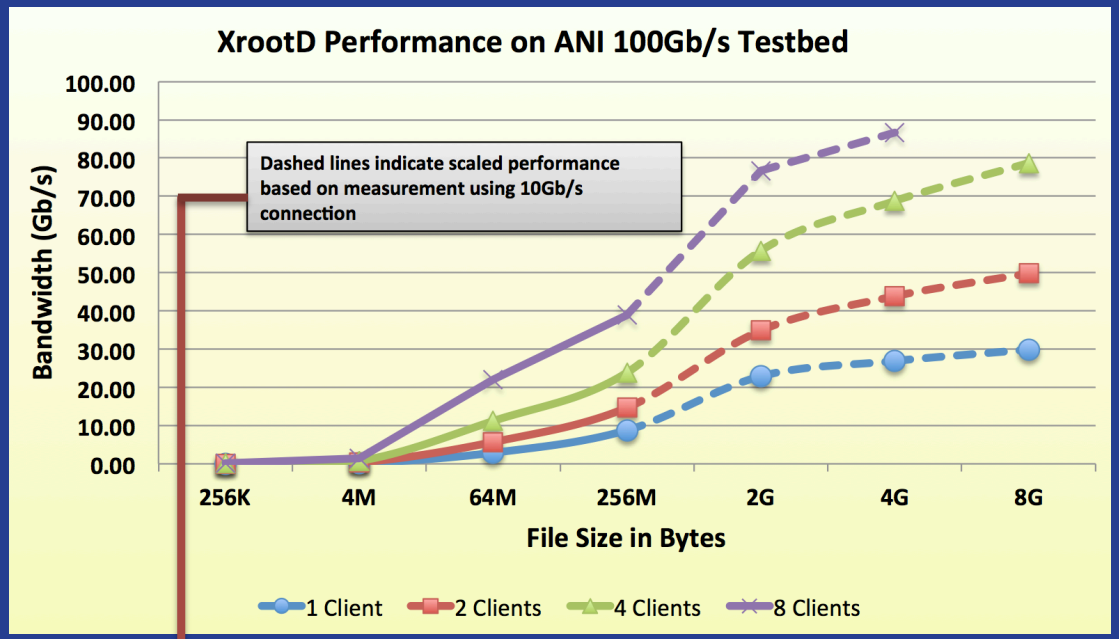
Optimal performance: 97 Gbps w/ GridFTP
2 GB files – 3 nodes x 16 streams / node



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

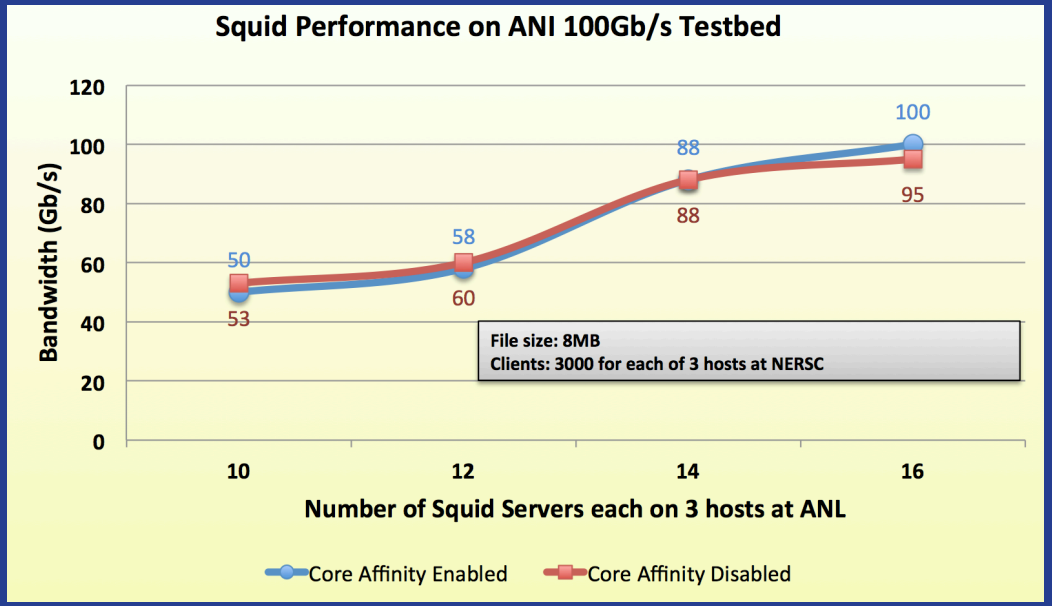


Dataset (GB)	1 NIC measurements (Gb/s)	Aggregate Measurements (12 NIC) (Gb/s)	Scale Factor per NIC	Aggregate estimate (12 NIC) (Gb/s)
0.512	4.5	46.9	0.87	–
1	6.2	62.4	0.83	–
4	8.7 (8 clients)	–	0.83	86.7
8	7.9 (4 clients)	–	0.83	78.7

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 servers per node (4 NIC per node)
  - Testing core affinity on/off: pin Squid to core i.e. to L2 cache
  - Testing all clnt noded vs. all servers AND aggregate one node vs. only one server



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

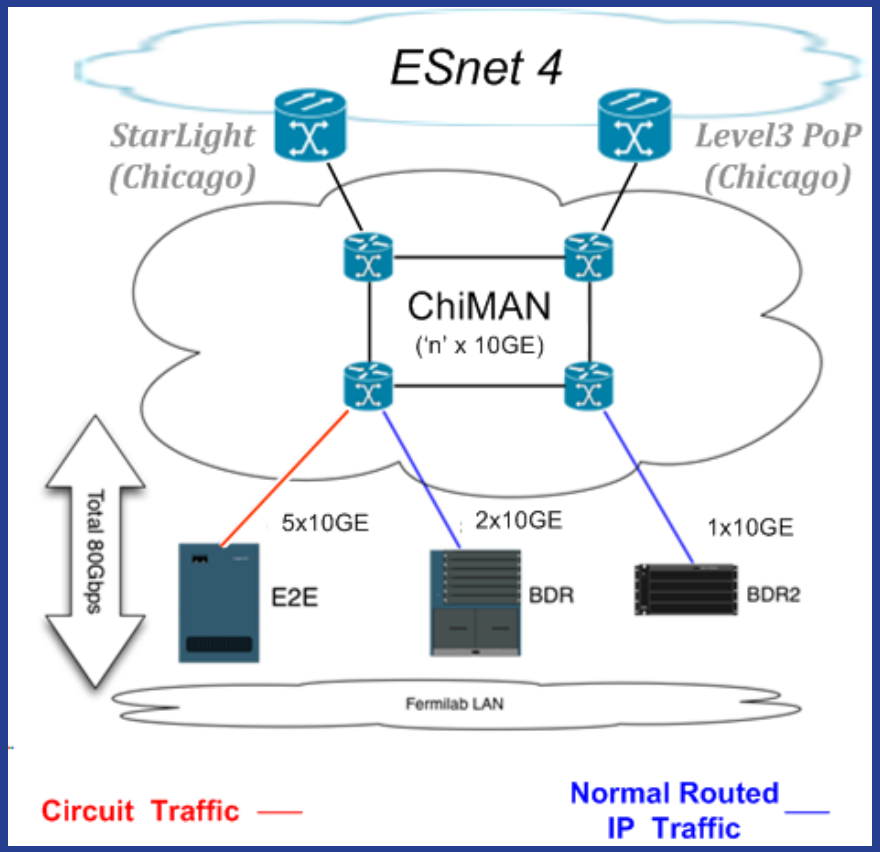
U.S. DEPARTMENT OF ENERGY

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

Current Fermilab WAN Capabilities

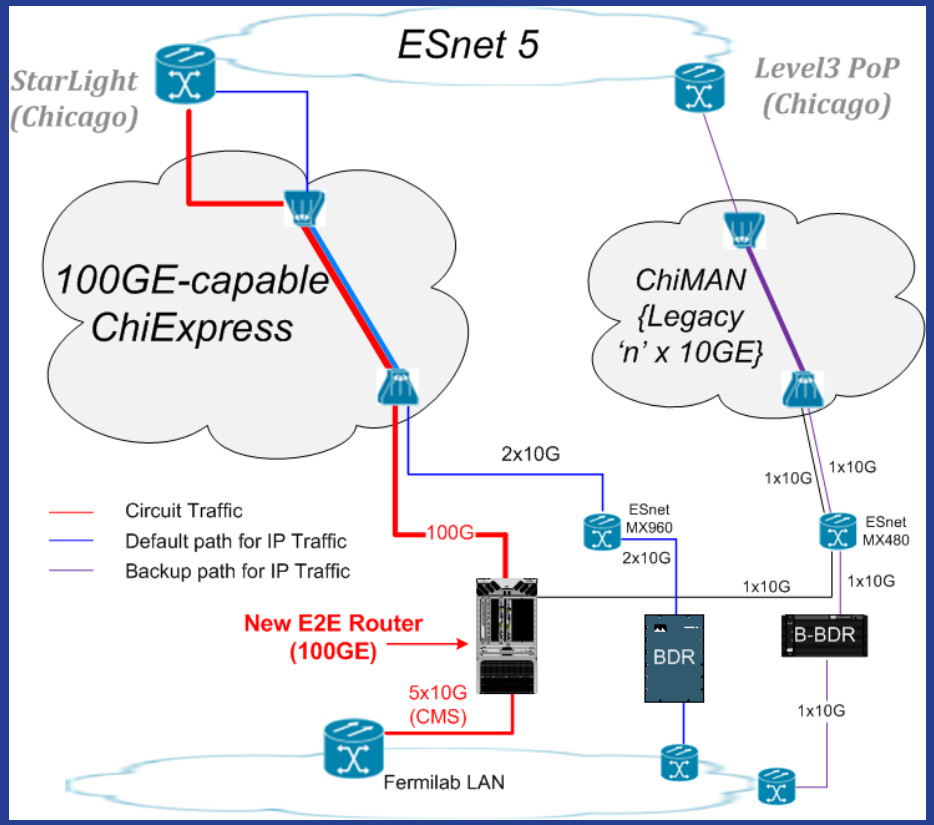
- Metropolitan Area Network provides 10GE channels:
  - Currently 8 deployed
- Five channels used for circuit traffic
  - Supports CMS WAN traffic
- Two used for normal routed IP traffic
  - Backup 10GE for redundancy
  - Circuits fail over to routed IP paths



* Diagram courtesy of Phil Demar

Upcoming 100GE WAN capability

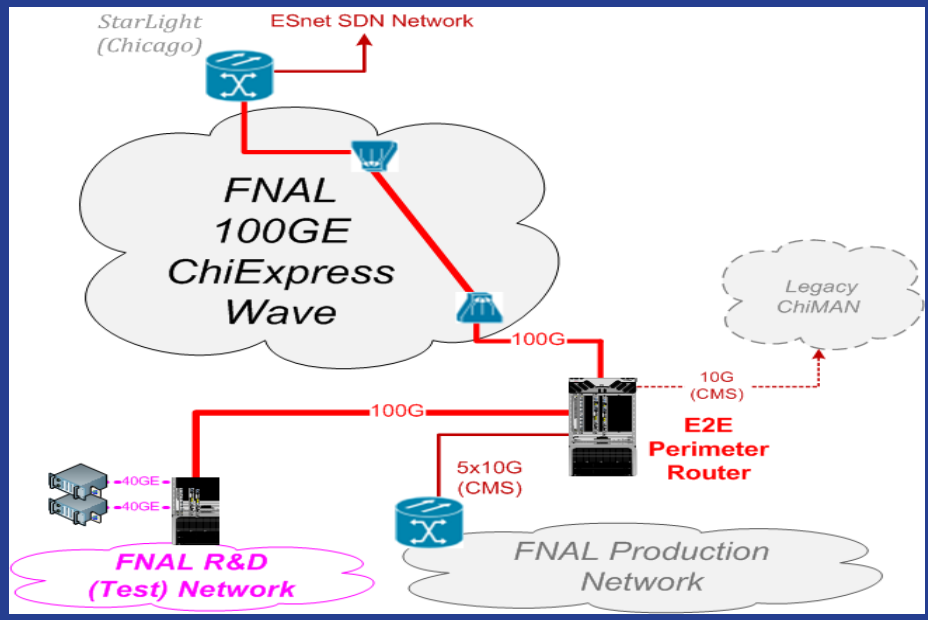
- ESnet deploying 100GE MAN as part of ESnet5
  - One 100GE wave for FNAL
  - Also 2x10GE channels for routed IP traffic
- 100GE wave will be used to support circuit traffic
- Legacy 10GE MAN will remain for diversity
  - Backup routed IP path
  - One 10GE circuit path



* Diagram courtesy of Phil Demar

Use of 100GE Wave for FNAL R&D

- 100GE wave will support 4x10GE circuit paths
  - Excess capacity available for WAN R&D activities
- Planning ~6 x 10GE link to FNAL R&D network
  - Network will host 10GE test/development systems
  - Possibly 40GE systems later
- Anticipate WAN circuit into ESnet ANI test bed



* Diagram courtesy of Phil Demar

Summary

- The High Throughput Data program at Fermilab is testing 100GE networks for its scientific stakeholders
- The collaboration with ANI and ESNet has been central to this program
- Fermilab will have 100GE capability in the Spring 2013 – planning for involvement with 100G ESNet testbed