

Experience and proposal for 100 GE R&D at Fermilab

Interactomes – May 22, 2012

Gabriele Garzoglio
Grid and Cloud Computing Department
Computing Sector, Fermilab

Overview

- The Hight Throughput Data Program
- Results from the ANI testbed
- Future program

Goals of 100 GE Program @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 the participation to funding agency solicitations.
- The ANI has been the major testbed used since last year in close partnership with ESNNet

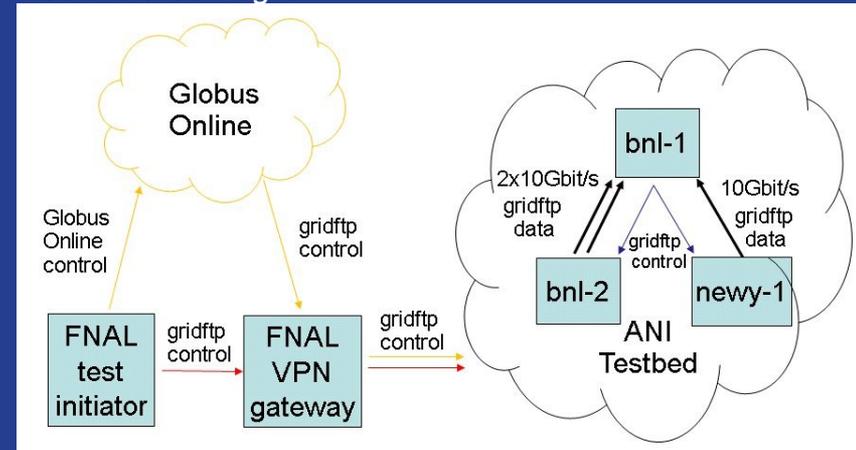
Ongoing Program of Work

- 2011: ANI Long Island MAN (LIMAN) testbed.
 - Tested GridFTP and Globus Online for the data movement use cases of HEP over 3x10GE.
- 2011-2012: Super Computing 2011.
 - Demonstration of fast access to ~30TB of CMS data from NERSC to ANL using GridFTP.
 - Achieved 70 Gbps
- Currently: ANI 100GE testbed.
 - Tuning parameters of middleware for data movement: xrootd, GridFTP and Globus Online.
 - Achieved ~97Gbps w/ simple GridFTP tests; more work needed for small files.
- Summer 2012: 100GE Endpoint at Fermilab
 - Plan to repeat and extend tests.

Experience on the ANI LIMAN Testbed

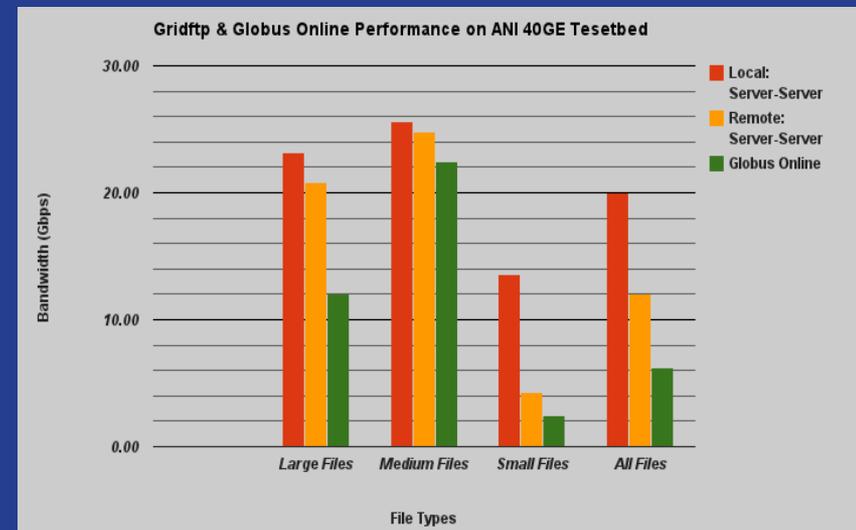
Work by Dave Dykstra w/ contrib. by Raman Verma & Gabriele Garzoglio

- Testing with GridFTP using 3x10GE in preparation for 100GE on ANI Testbed.



- Characteristics:
 - 300GB of data split into 42,432 files (8KB – 8GB; varied sizes).
 - Aggregated 3 x 10Gbit/s link to Long Island test end-point.

- Results:
 - Almost equal throughput for Globus Online (green) as for direct GridFTP (red) for medium-size files.
 - Increased throughput by 30% through increasing concurrency and pipelining on small files.
 - Auto-tuning in Globus Online works better for medium sized files than for large files.

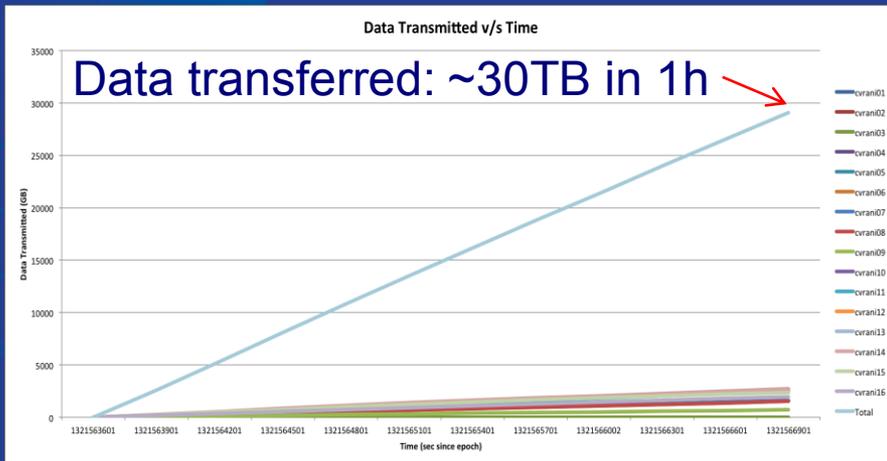


Super Computing 2011

- Test transfer of CMS experiment data between NERSC and ANL over 100 GE network.
- Characteristics:
 - 15 server / 28 client nodes (multi-cores, 48 GB RAM, 10Gbps)
 - 2 globus-url-copy (GUC) clients / server

Work by Parag Mhashilkar, Gabriele Garzoglio (Fermilab) and Haifeng Pi (UCSD)

	GUC/core	GUC streams	GUC TCP Window Size	Files/GUC	MAX BW	Sustain BW
T1	-	-	-	-	-	-
D1	1	2	Default	60	65	50
T2	1	2	2MB	1	65	52
D2	1	2	2MB	1	65	52
T3	4	2	2MB	1	73	70
D3	4	2	2MB	1	75	70



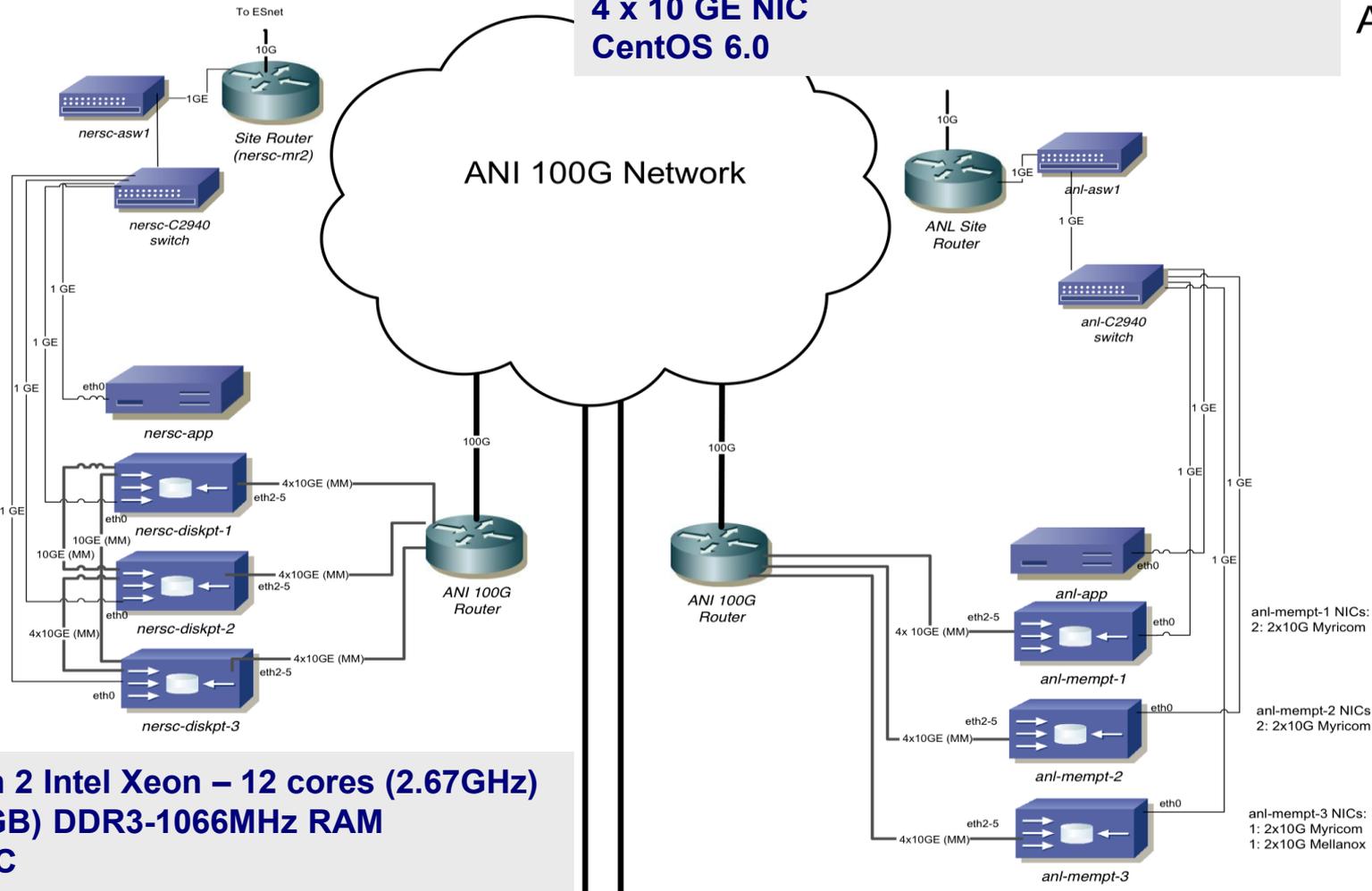
100 GE ANI Testbed

ANI Middleware Testbed

**3 Nodes with 2 AMD 6140 – 8 cores (2.6GHz)
64 GB (8x8GB) DDR3-1333MHz
4 x 10 GE NIC
CentOS 6.0**

NERSC

ANL

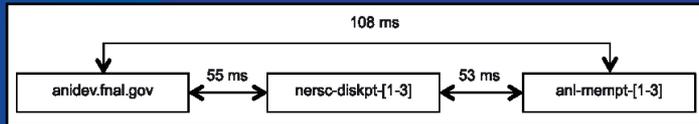


**3 Nodes with 2 Intel Xeon – 12 cores (2.67GHz)
48GB (12x4GB) DDR3-1066MHz RAM
4 x 10 GE NIC
CentOS 5.7**

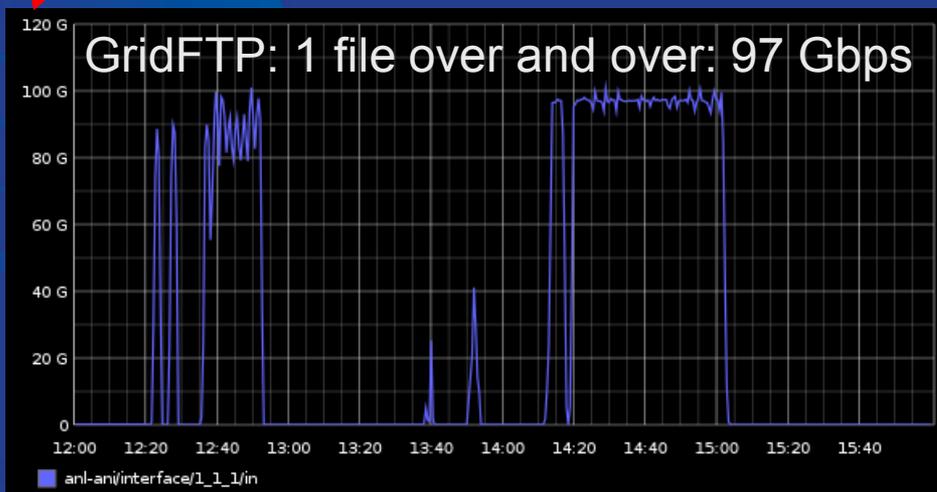
Updated December 11, 2011

GridFTP and GO on the ANI 100G Testbed

- 3 tests w/ GridFTP
 - Local Client-Server
 - Local Server-Server
 - Remote Server-Server (VPN port fwd'ing)



- GO Tests w/ port-fwd'ing
- Challenges
 - Simple tests can saturate the network
 - Need more work for realistic use cases

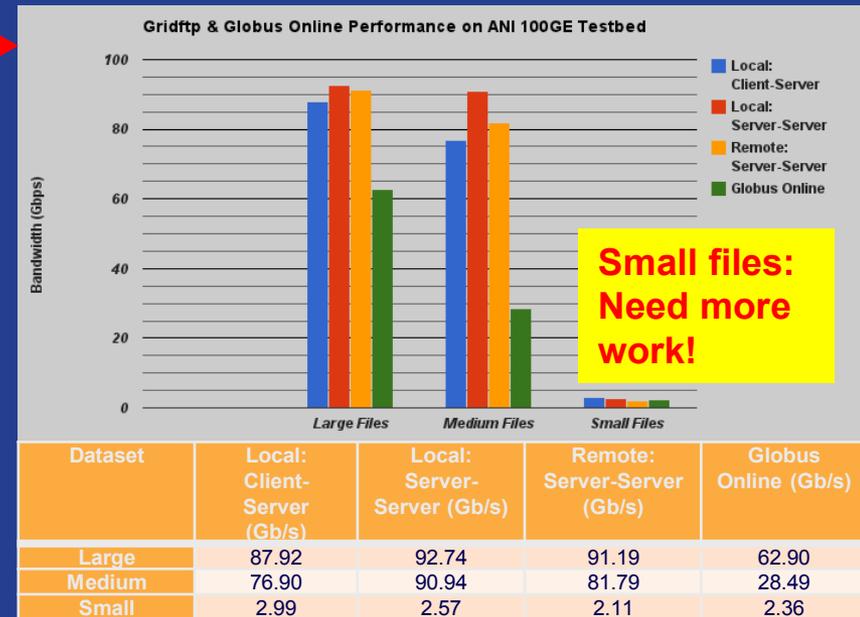


GridFTP Parameter Tuning *Work by Parag Mhashilkar*

Test Type	Dataset	GUC -p	GUC -cc	GUC -pp	GUCs per host	Times dataset transferred	Transfer Time (sec)
Local: Client-Server	Large	4	4	No	12	3	422.67
	Medium	-	4	Yes	12	5	114.67
	Small	-	-	Yes	16	250	787.33
Local: Server-Server	Large	4	4	No	48	3	1602.67
	Medium	-	4	Yes	48	5	387.67
	Small	-	-	Yes	16	250	917.00
Remote: Server-Server	Large	4	4	No	48	3	1630.00
	Medium	-	4	Yes	48	5	431.00
	Small	-	-	Yes	48	250	3349.00

GO Parameter Tuning

Test Type	Dataset	--perf-p	--perf-cc	--perf-pp	Number of transfer requests	Times dataset transferred	Transfer Time (sec)
Globus Online	Large	4	4	32	12	3	262.61
	Medium	-	4	32	12	5	137.58
	Small	-	-	32	16	250	333.83



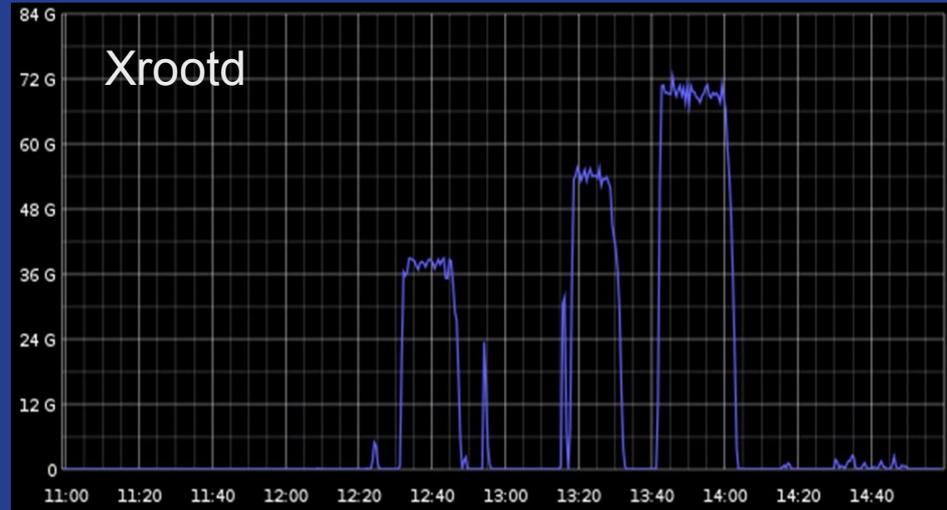
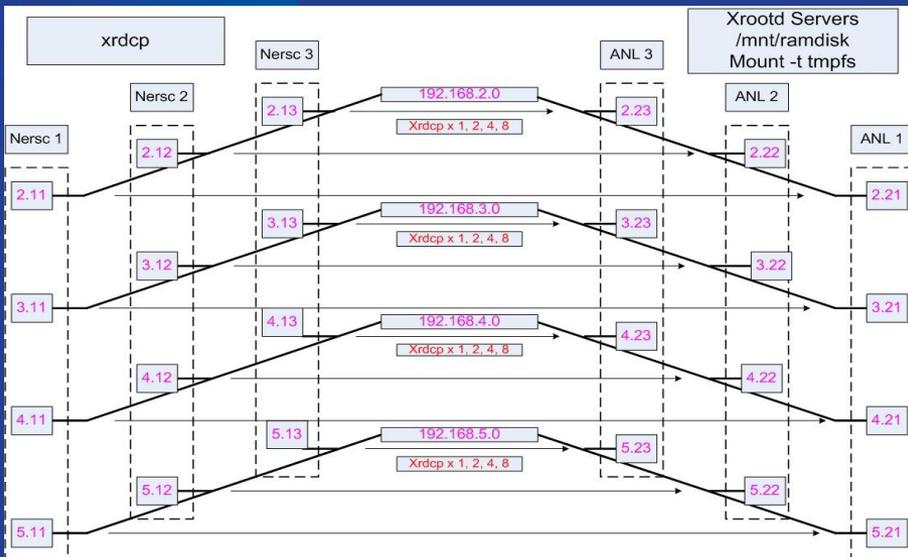
Xrootd on the ANI 100G Testbed

- 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

Work by Hyunwoo Kim (Fermilab)

# Clients / NIC	Input File 512 MB	Input File 1 GB	Input File 2 GB	Input File 4 GB
1	~12 Gbps	~18 Gbps	~26 Gbps	~32 Gbps
2	~22 Gbps	~32 Gbps	~40 Gbps	~56 Gbps
4	~42 Gbps	~56 Gbps	~70 Gbps	~77 Gbps
8	~60 Gbps	~75 Gbps	~80 Gbps	-

Increased Throughput



Current Plans & Constraints

- ANI 100G testbed
 - Current time window: until Aug 2012
 - Complete tests of Xrootd, GridFTP, and Globus Online
 - Test Squid for condition data access (preliminary: 7 Gpbs / 1 server)
- Planning to test more technologies. Priorities agreed w/ Stakeholders:
 - CVMFS, dCache, IRODS, Luster.
 - Risk to the stakeholders: extension to ANI not available.
- 100GE production endpoint coming to Fermilab
 - Expecting 100 GE capabilities in summer 2012.
 - Creating a local testbed connecting to ANI.
 - Continue testing of middleware technologies defined by stakeholders.

Summary

- Fermilab has a program of work to test 100GE network for its scientific stakeholders
- The collaboration with ANI and ESNNet has been central to this program
- The current timeline for ANI is not sufficient to evaluate all technologies of interest to the Fermilab stakeholders
- It is important to plan for an RnD 100GE network at Fermilab to...
 - ... hedge the risk of ANI closing down
 - ... bootstrap knowledge of 100 GE technologies across the sector.