Identifying Gaps in Grid Middleware on Fast Networks with The Advanced Networking Initiative

D. Dykstra, G. Garzoglio, H. Kim, P. Mhashilkar
Scientific Computing Division, Fermi National Accelerator Laboratory

CHEP2012 Poster ID 214

Motivation

Goal of the High Throughput Data Program(HTDP) at the Fermilab Computing Sector is to support Fermilab and its stakeholders in the adoption of a 100GE networking infrastructure.

Focus

- compile a list of key services used by relevant research communities/facilities
- identify gaps in current infrastructure and tools that interface with 100GE networks

We are conducting a series of tests with key tools on a test bed 100 GE network which is operated by US DoE ESnet's Advanced Networking Initiative(ANI)

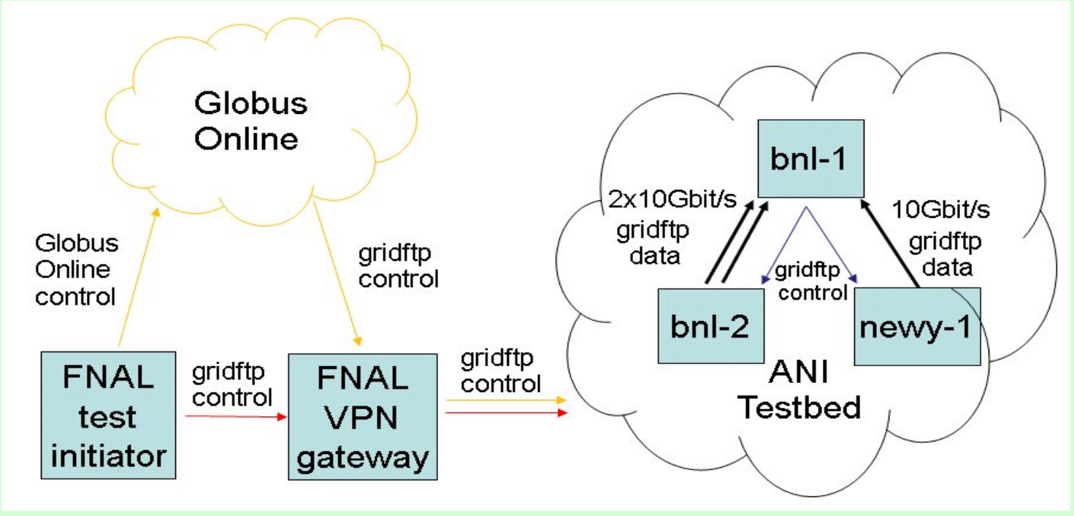
Conclusion

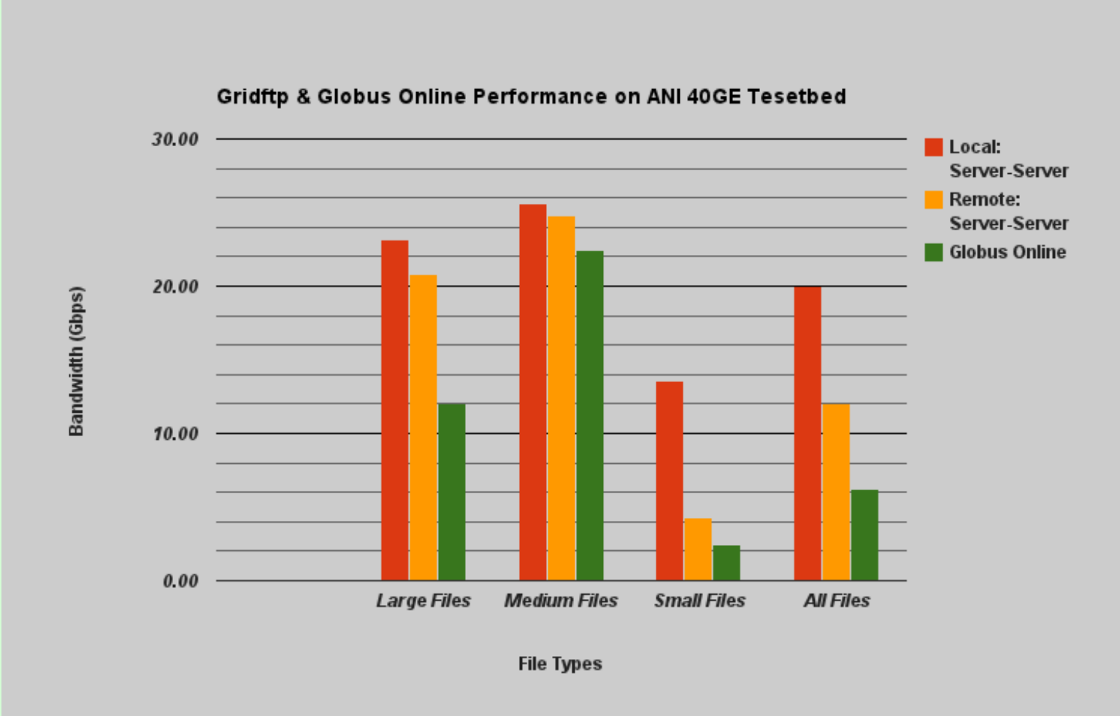
- Basic network capacity test is close to 100 GE
- Can saturate the bandwidth capacity by increasing streams
- **GridFTP:** suffers from protocol overhead for small files
- **Globus Online:** working with GO to improve performance
- **XrootD:** test at initial stage but gives throughput comparable to GridFTP. Not many performance-tuning options are available

May – October 2011

LIMAN Testbed : 40 GE

- Main tools tested on the Long Island Metropolitan Area Network: Globus Online(GO) and GridFTP(GF)
- Compare 3 transfer mechanisms (to see overheads from GO and control channels)
  - ✓Local GF transfer (server to server)
  - ✓FNAL-controlled GridFTP transfer
  - ✓GO-controlled GridFTP transfer
- Compare 3 sets of files with different sizes (to see the effects of transfer protocol overhead on small files)
- Result: Overheads observed in the use of Globus-Online and small files
- RTT between FNAL & BNL (ctrl) : 36 ms
- RTT among testbed nodes (data): 2 ms

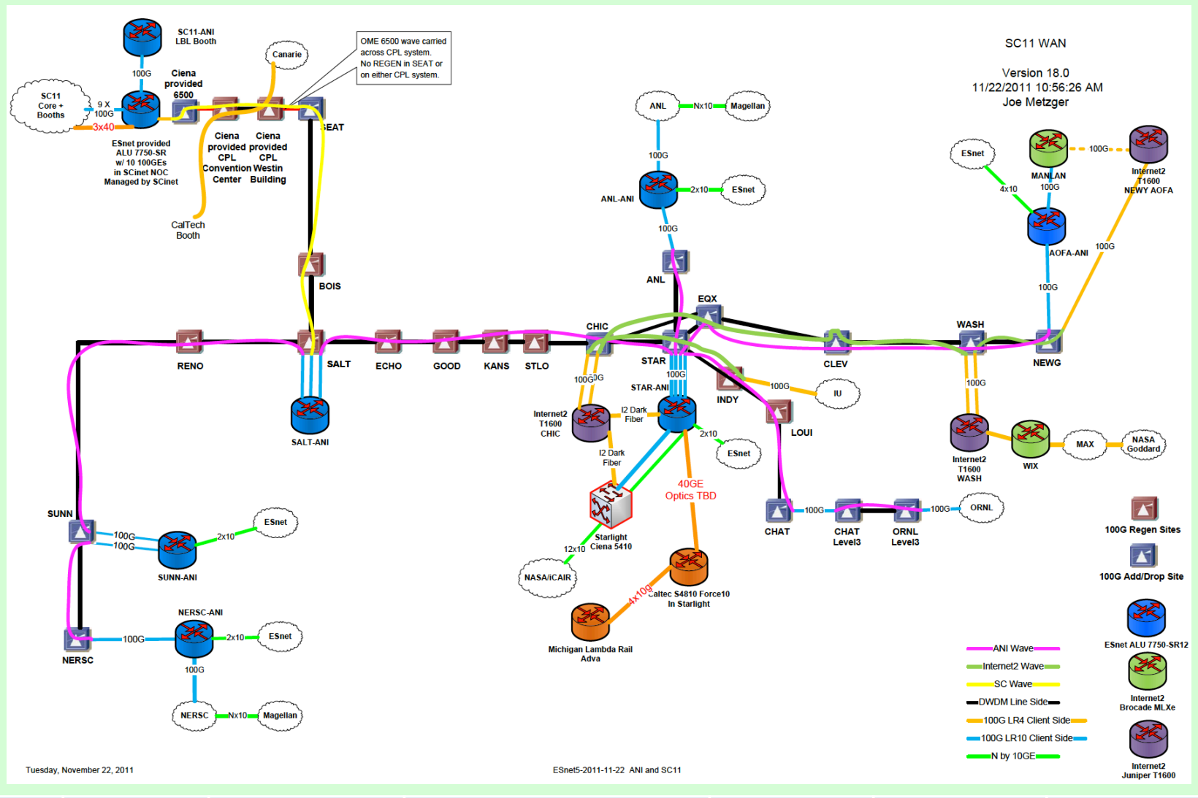




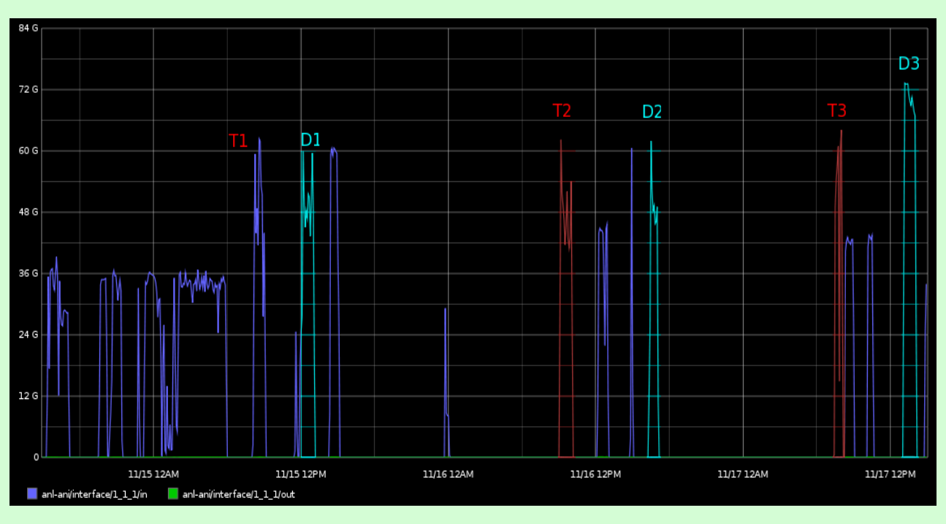
November 2011

SC2011 Demo : shared 100 GE

- The Grid & Cloud Computing Dept. of Fermilab demonstrated the use of 100 GE network to move CMS data with GridFTP
- Test Characteristics
  - ✓15 NERSC & 26 ANL nodes w/ 10 GE NIC
  - ✓10 CMS files of 2 GB (RAM to RAM only)
  - ✓Total 30 TB transferred in one hour
- Result: data transfer rate at ~70 Gbps sustained with peaks at 75 Gbps



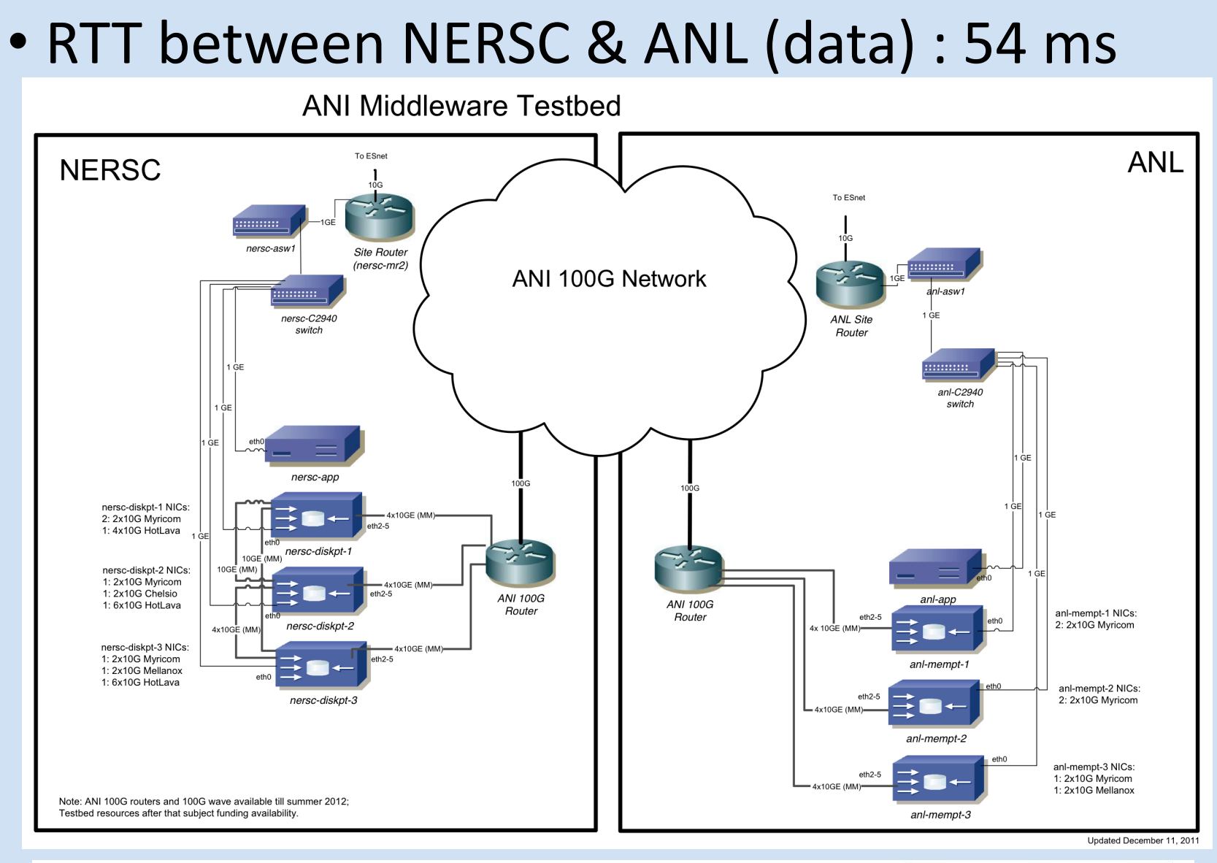
	GUC/ core	GUC streams	GUC TCP Window Size	Files/ GUC	MAX BW	Sustain BW
T1	-	-	-	-	-	-
D1	1	2	Default	60	65	50
T2	1	2	2MB	1	65	52
D2	1	2	2MB	1	65	52
T3	4	2	2MB	1	73	70
D3	4	2	2MB	1	75	70

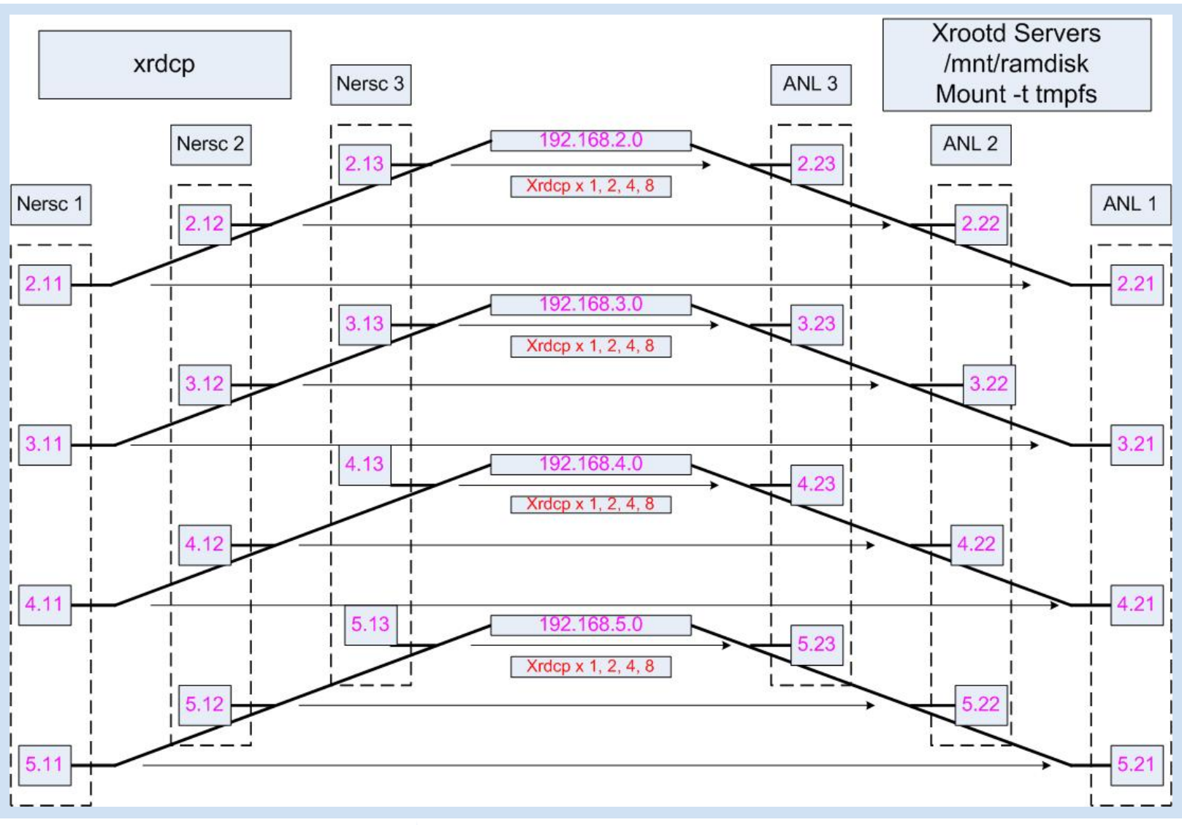


January 2012 - Present

Current Testbed : 100 GE

- 2 sites (NERSC, ANL) with 3 nodes each. Each node with 4 x 10 GE NICs
- Measure various overheads from protocols and file sizes
  - ✓Basic network capacity using nuttcp
  - ✓GridFTP and Globus Online
  - ✓XrootD
- RTT between FNAL & NERSC (control) : 55 ms
- RTT between FNAL & ANL (control) : 108 ms
- RTT between NERSC & ANL (data) : 54 ms


ANI Middleware Testbed
NERSC
ANL
ANI 100G Network
To ESnet
10G
1GE
nersc-asw1
Site Router (nersc-mr2)
nersc-C2940 switch
1 GE
eth0
nersc-app
nersc-diskpt-1 NICs: 2: 2x10G Myricom 1: 4x10G HotLava
nersc-diskpt-1
nersc-diskpt-2 NICs: 1: 2x10G Myricom 1: 2x10G Chelsio 1: 6x10G HotLava
nersc-diskpt-2
nersc-diskpt-3 NICs: 1: 2x10G Myricom 1: 2x10G Mellanox 1: 6x10G HotLava
nersc-diskpt-3
4x10GE (MM)
10GE (MM)
eth2-5
100G
ANI 100G Router
ANL Site Router
anl-asw1
anl-C2940 switch
anl-app
anl-mempt-1
anl-mempt-1 NICs: 2: 2x10G Myricom
anl-mempt-2
anl-mempt-2 NICs: 2: 2x10G Myricom
anl-mempt-3
anl-mempt-3 NICs: 1: 2x10G Myricom 1: 2x10G Mellanox
Note: ANI 100G routers and 100G wave available till summer 2012; Testbed resources after that subject funding availability.
Updated December 11, 2011



Tests on the Current ANI 100 GE Testbed

Basic Network Throughput Test with nuttcp

- Motivation: Confirm basic performance of network with parameters to tune and Compare with baseline provided by ANI team
- Results
  - ✓ NIC to NIC : 9.89 Gbps (as expected from 10 GE NIC)
  - ✓ 4 NICs to 4 NICs between 2 nodes : 39 Gbps (as expected from 4 NICs)
  - ✓ Aggregate throughput using 10 TCP streams (10 pairs of NIC-NIC) : 99 Gbps

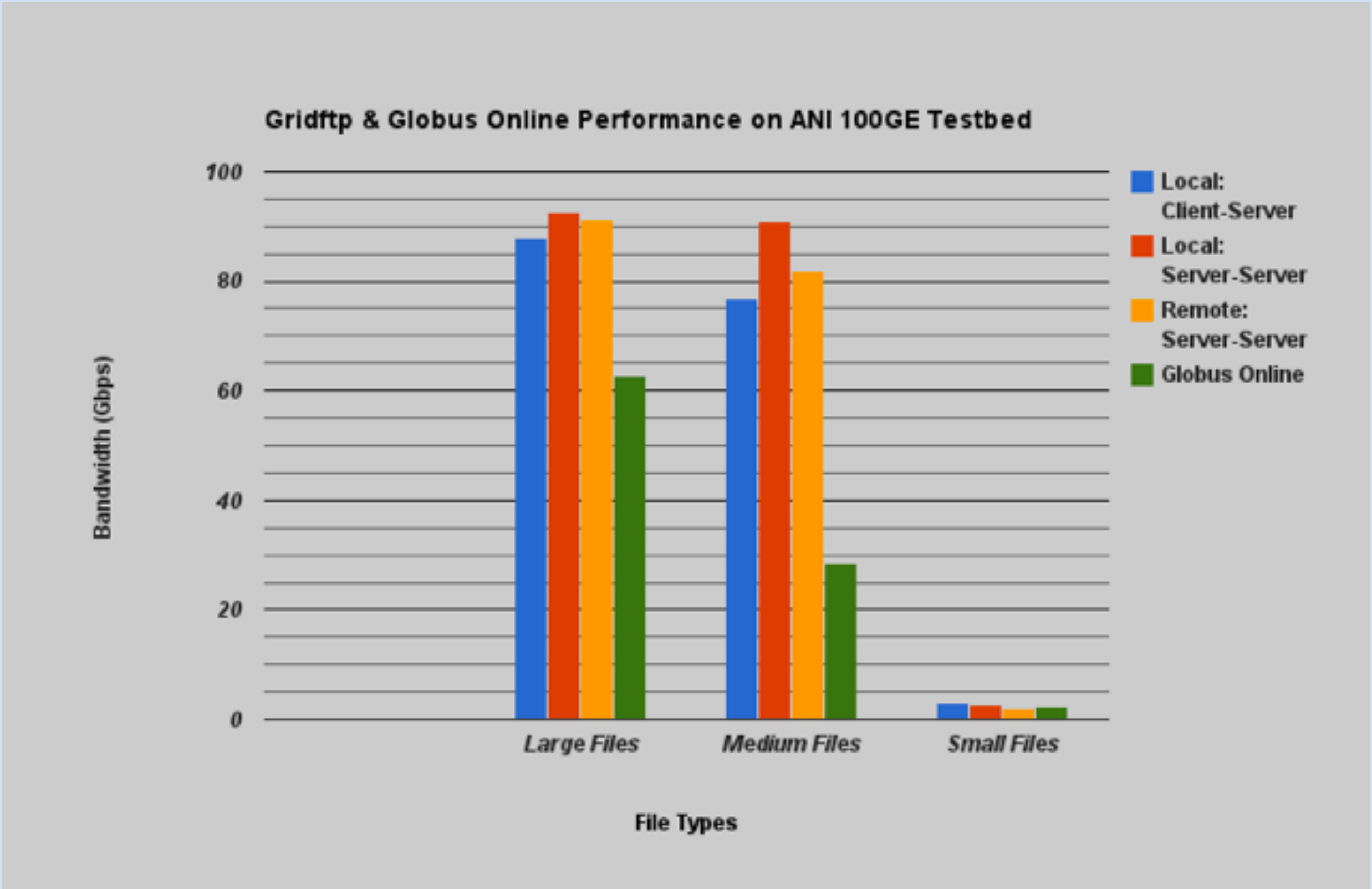
GridFTP and Globus Online Test

- **Motivation 1** : Using a single instance of GridFTP client/server is not efficient
- What is the efficient way to increase the throughput via each NIC?
- What is the efficient way to transfer a single file?
  - ✓Answer: use multiple parallel streams for each file transfer, globus-url-copy –p N
- What is the efficient way to transfer a set of files?
  - ✓Answer: use multiple concurrent globus-gridftp-servers , globus-url-copy –cc M
- We launch multiple clients and servers with multiple streams opened between them

- **Motivation 2** : we expect protocol overheads to be different across various file sizes
- Files of various sizes are transferred from client disk to server memory
  - ✓ Dataset split into 3 sets: Small(8KB - 4MB), Medium(8MB -1G), Large(2, 4, 8 GB)

- **Motivation 3** : In addition to locally-controlled GridFTP, we tested 2 remotely-controlled configurations
  1. Use port-forwarding to access GridFTP clients/servers (labeled: "Remote")
  2. Use Globus-Online
- We also compare server-server transfer with client-server transfer

- Results
  - ✓GridFTP does not suffer from protocol overhead for large & medium size files
  - ✓Observe significant overhead in the case of small size files
  - ✓Remote use of GridFTP via Globus Online suffers from protocol overhead
  - ✓We think RTT affects results for small files



	Local: Client-Server	Local: Server-Server	Remote: Server-Server	Globus Online
Large	87.92 Gbps	92.74 Gbps	91.19 Gbps	62.90 Gbps
Medium	76.90 Gbps	90.94 Gbps	81.79 Gbps	28.49 Gbps
Small	2.99 Gbps	2.57 Gbps	2.11 Gbps	2.36 Gbps

XrootD Test

	1 client	2 clients	4 clients	8 clients
8 GB, 1-NIC	3 Gbps	5 Gbps	7.9 Gbps	N/A
Large, 1-NIC (2 / 4 GB)	2.3 / 2.7 Gbps	3.5 / 4.4 Gbps	5.6 / 6.9 Gbps	7.7 / 8.7 Gbps
Medium (64M / 256M)	2.9 / 8.8 Gbps	5.7 / 14.7 Gbps	11.2 / 23.9 Gbps	22 / 39 Gbps
Small (256K / 4M)	0.03 / 0.19 Gbps	0.07 / 0.38 Gbps	0.11 / 0.76 Gbps	0.1 / 1.4 Gbps

- Motivation: What is the efficient way to increase the throughput via each NIC?
- We are focusing only on tuning transfer parameters of xrootd
- Test begins with single instance of xrdcp and xrootd
- ✓ Server side: one xrootd writing to RAMdisk or HDD
- Are multiple concurrent transfers possible in xrootd?
- ✓ The equivalent of the "GridFTP –cc" option is not available but we can emulate it by launching multiple xrdcp. xrootd server accepts multiple connections by using multithreading. How efficient is it?
- Are multiple parallel transfers possible in xrootd? Not practical for our test
- Results : Limited by RAMdisk, we estimate the aggregate throughput by scaling one-NIC result for files of over 2 GB
- ✓ 2GB, 4GB and 8GB file transfer results are estimated to be 77 Gbps, 87 Gbps and 80 Gbps respectively.(Assume 10 NICs. 8GB uses maximum 4 clients)

U.S. DEPARTMENT OF ENERGY

Fermilab is Operated by the Fermi Research Alliance, LLC
under Contract No. DE-AC02-07CH11359 with the United States Department of Energy