PSC
PITTSBURGH SUPERCOMPUTING CENTER

Lustre WAN

OSG Storage Forum

September 21, 2010

J. Ray Scott

Project Summary

- Evaluation of a Global Widearea File System for:
  - Performance
  - Robustness
- Leverage Work from Teragrid
- Software Support
  - PSC
  - Josephine Palencia, Brian Johanson
- Hardware Support
  - UF
- Testing
  - UF, FSI, FIU, PSC, others

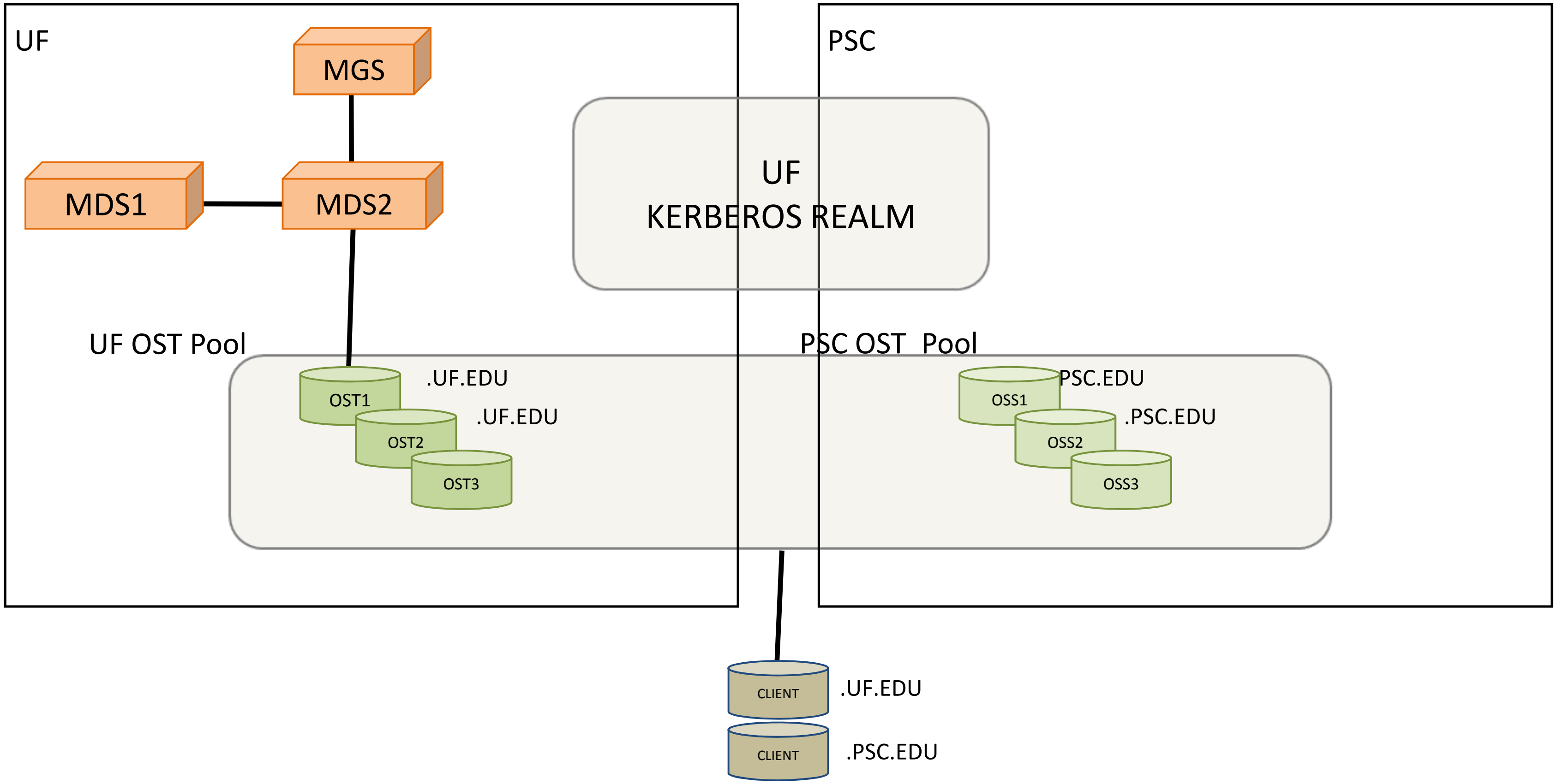
Project Approach

- Secure Infrastructure
- Installation Support
- Authentication Mapping
- Network Performance Measurement
- Application Integration
- Assessment and Project Support

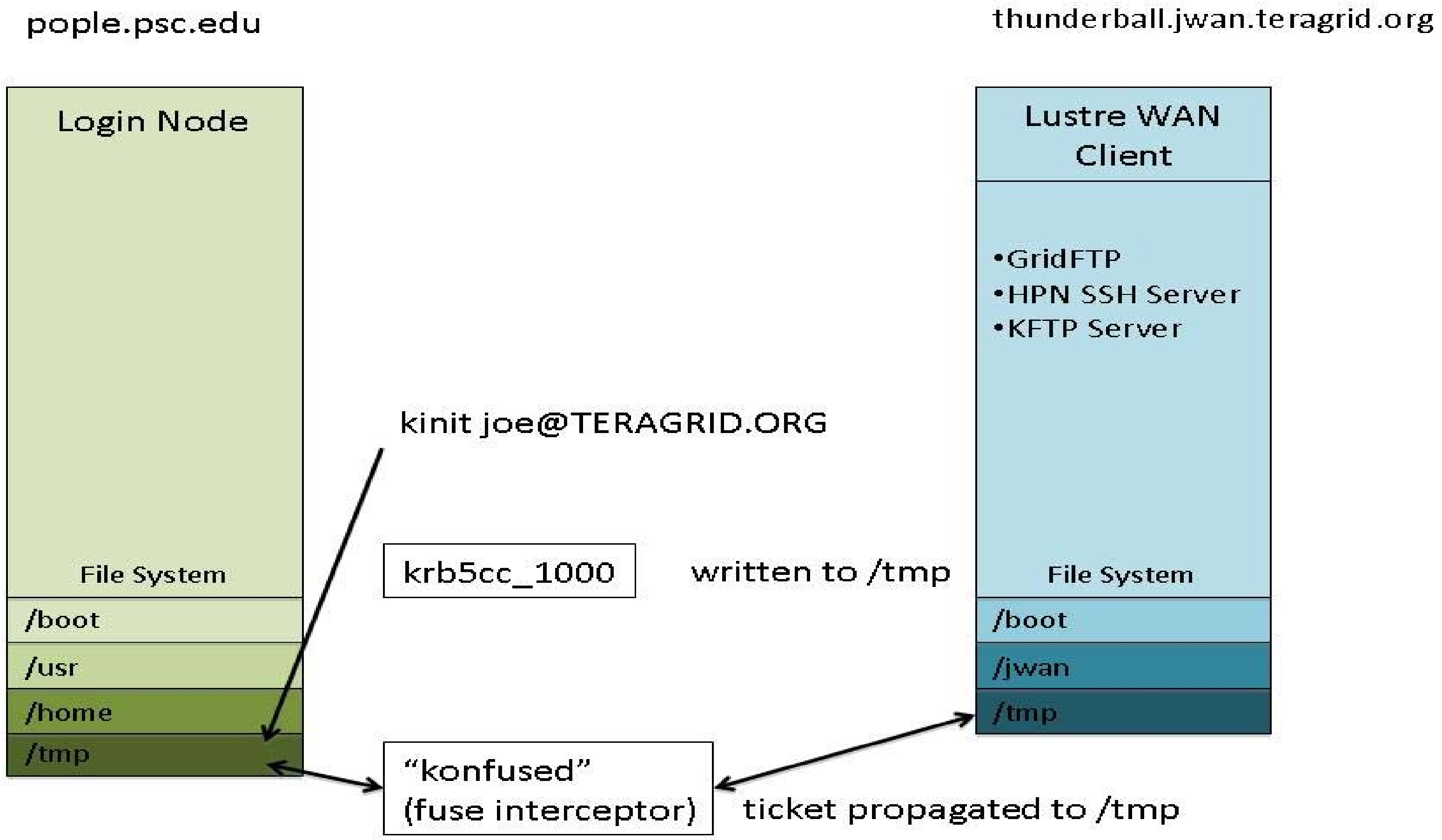
Secure Infrastructure

- Kerberos security infrastructure
- Lustre 2.0
- Installation Packages
  - Ease the software installation
  - Hide Kerberos from site administration

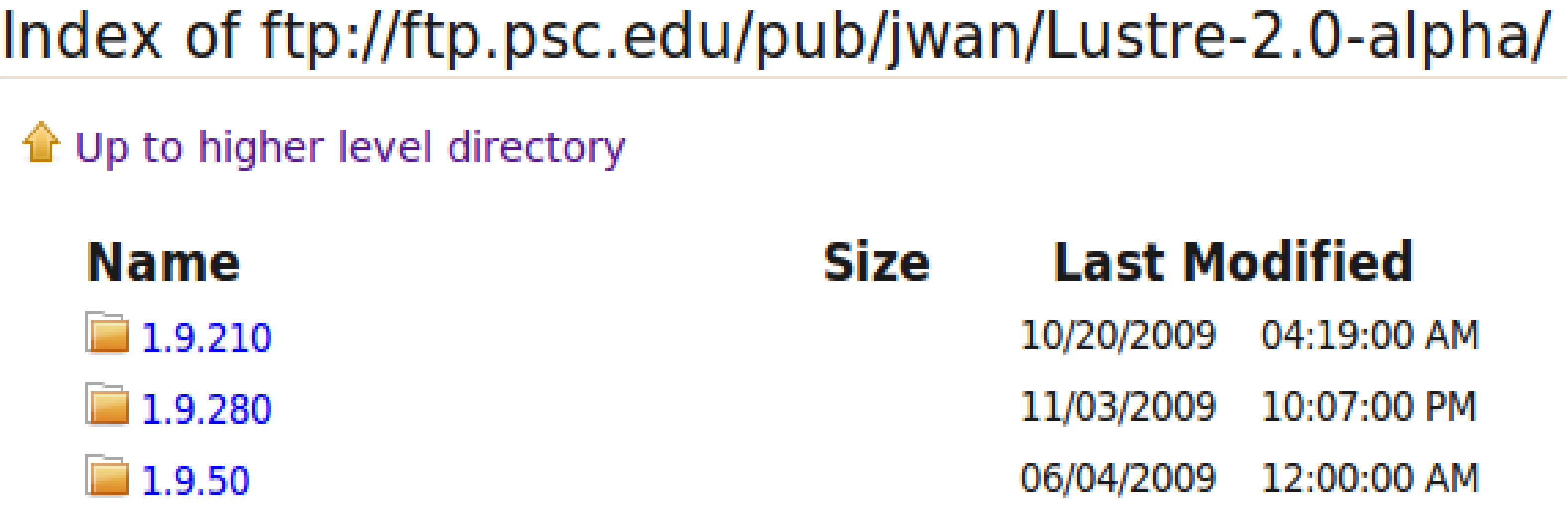
Example Site Configuration



Kerberized scp/kftp/gridftp: konFUSEd



Installation Support – RPM Packaging



Index of ftp://ftp.psc.edu/pub/jwan/Lustre-2.0-alpha/

⇧ Up to higher level directory

Name	Size	Last Modified
1.9.210		10/20/2009 04:19:00 AM
1.9.280		11/03/2009 10:07:00 PM
1.9.50		06/04/2009 12:00:00 AM

- **1.10.0.40 Lustre 2.0 Beta 1**
- **VM client/server rpms**

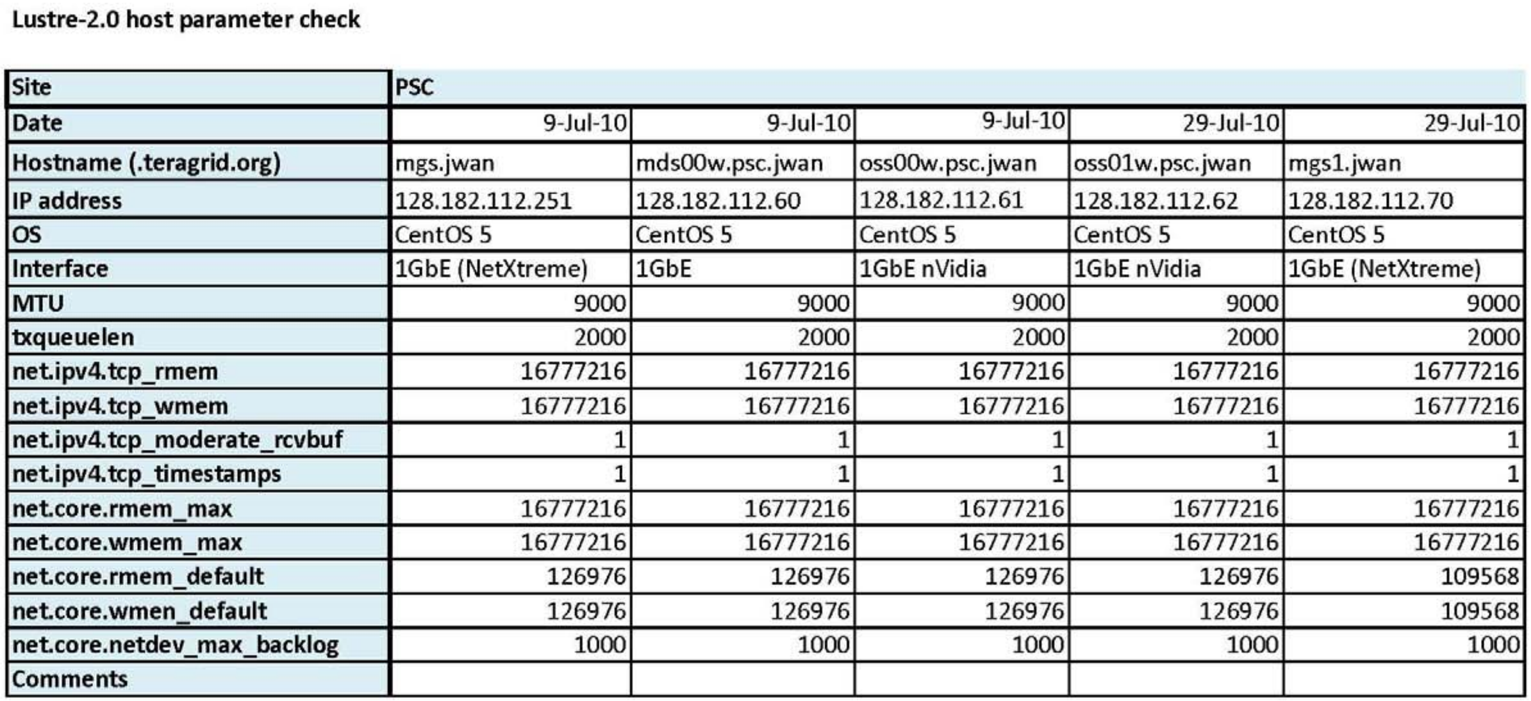
Authentication Mapping

- UID Mapping Using IU Developed Code
- Only Necessary Across Administrative Domains
  - Without UID synchronization

Network Performance Testing

- Pre-Production Baseline Testing
- Ongoing Production Testing

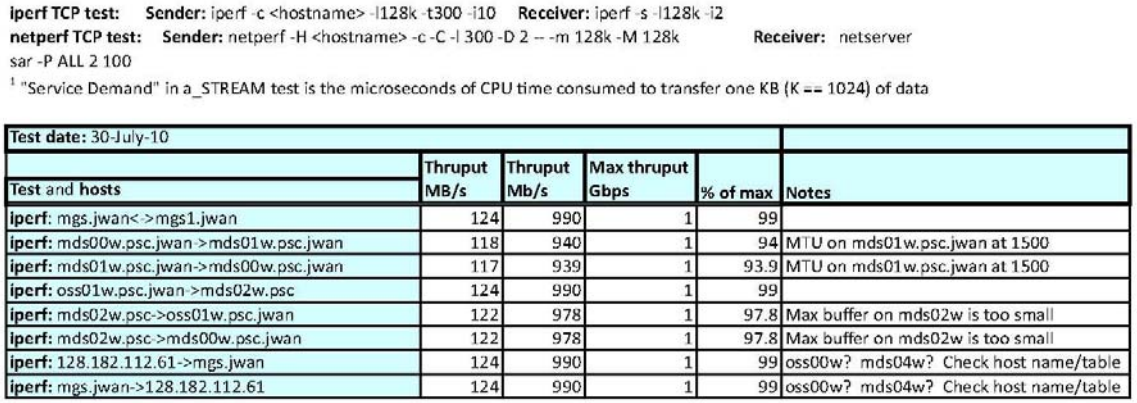
Performance Measurement –Internal Testing



Lustre-2.0 host parameter check

Site	PSC				
Date	9-Jul-10	9-Jul-10	9-Jul-10	29-Jul-10	29-Jul-10
Hostname (.teragrid.org)	mgs.jwan	mds00w.psc.jwan	oss00w.psc.jwan	oss01w.psc.jwan	mgs1.jwan
IP address	128.182.112.251	128.182.112.60	128.182.112.61	128.182.112.62	128.182.112.70
OS	CentOS 5	CentOS 5	CentOS 5	CentOS 5	CentOS 5
Interface	1GbE (NetXtreme)	1GbE	1GbE nVidia	1GbE nVidia	1GbE (NetXtreme)
MTU	9000	9000	9000	9000	9000
txqueuelen	2000	2000	2000	2000	2000
net.ipv4.tcp_rmem	16777216	16777216	16777216	16777216	16777216
net.ipv4.tcp_wmem	16777216	16777216	16777216	16777216	16777216
net.ipv4.tcp_moderate_rcvbuf	1	1	1	1	1
net.ipv4.tcp_timestamps	1	1	1	1	1
net.core.rmem_max	16777216	16777216	16777216	16777216	16777216
net.core.wmem_max	16777216	16777216	16777216	16777216	16777216
net.core.rmem_default	126976	126976	126976	126976	109568
net.core.wmen_default	126976	126976	126976	126976	109568
net.core.netdev_max_backlog	1000	1000	1000	1000	1000
Comments					

Performance Measurement – Internal Testing



iperf TCP test: **Sender:** iperf -c <hostname> -l128k -t300 -i10 **Receiver:** iperf -s -l128k -i2

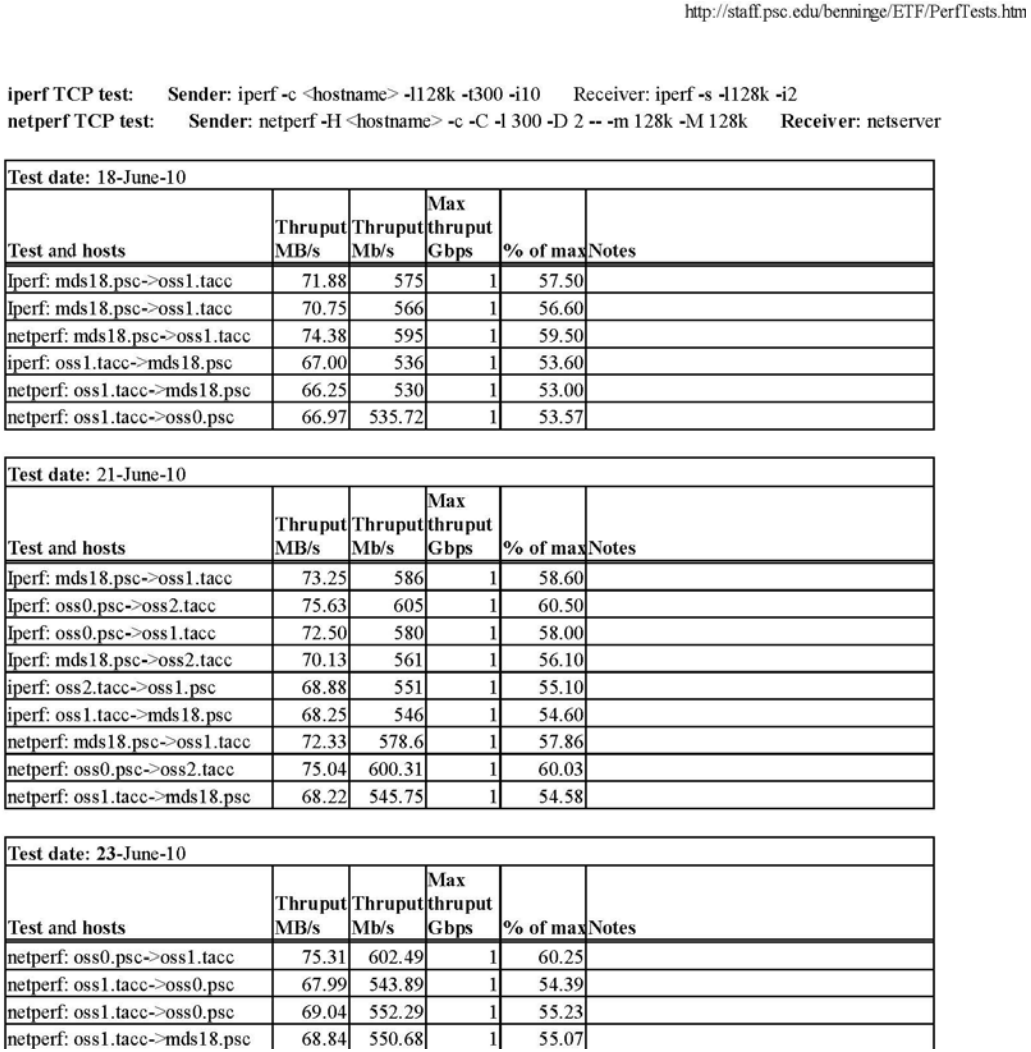
netperf TCP test: **Sender:** netperf -H <hostname> -c -C -l 300 -D 2 -- -m 128k -M 128k **Receiver:** netserver

sar -P ALL 2 100

[1] "Service Demand" in a_STREAM test is the microseconds of CPU time consumed to transfer one KB (K == 1024) of data

Test date: 30-July-10					
	Thruput	**Thruput**	**Max thruput**		
Test and **hosts**	**MB/s**	**Mb/s**	**Gbps**	**% of max**	**Notes**
iperf: mgs.jwan<->mgs1.jwan	124	990	1	99	
iperf: mds00w.psc.jwan->mds01w.psc.jwan	118	940	1	94	MTU on mds01w.psc.jwan at 1500
iperf: mds01w.psc.jwan->mds00w.psc.jwan	117	939	1	93.9	MTU on mds01w.psc.jwan at 1500
iperf: oss01w.psc.jwan->mds02w.psc	124	990	1	99	
iperf: mds02w.psc->oss01w.psc.jwan	122	978	1	97.8	Max buffer on mds02w is too small
iperf: mds02w.psc->mds00w.psc.jwan	122	978	1	97.8	Max buffer on mds02w is too small
iperf: 128.182.112.61->mgs.jwan	124	990	1	99	oss00w? mds04w? Check host name/table
iperf: mgs.jwan->128.182.112.61	124	990	1	99	oss00w? mds04w? Check host name/table

Performance Measurement –TeraGrid



http://staff.psc.edu/benninge/ETF/PerfTests.htm

iperf TCP test: **Sender:** iperf -c <hostname> -l128k -t300 -i10 Receiver: iperf -s -l128k -i2

netperf TCP test: **Sender:** netperf -H <hostname> -c -C -l 300 -D 2 -- -m 128k -M 128k **Receiver:** netserver

Test date: 18-June-10

Test and hosts	Thruput MB/s	Thruput Mb/s	Max thruput Gbps	% of max	Notes
Iperf: mds18.psc->oss1.tacc	71.88	575	1	57.50	
Iperf: mds18.psc->oss1.tacc	70.75	566	1	56.60	
netperf: mds18.psc->oss1.tacc	74.38	595	1	59.50	
iperf: oss1.tacc->mds18.psc	67.00	536	1	53.60	
netperf: oss1.tacc->mds18.psc	66.25	530	1	53.00	
netperf: oss1.tacc->oss0.psc	66.97	535.72	1	53.57	

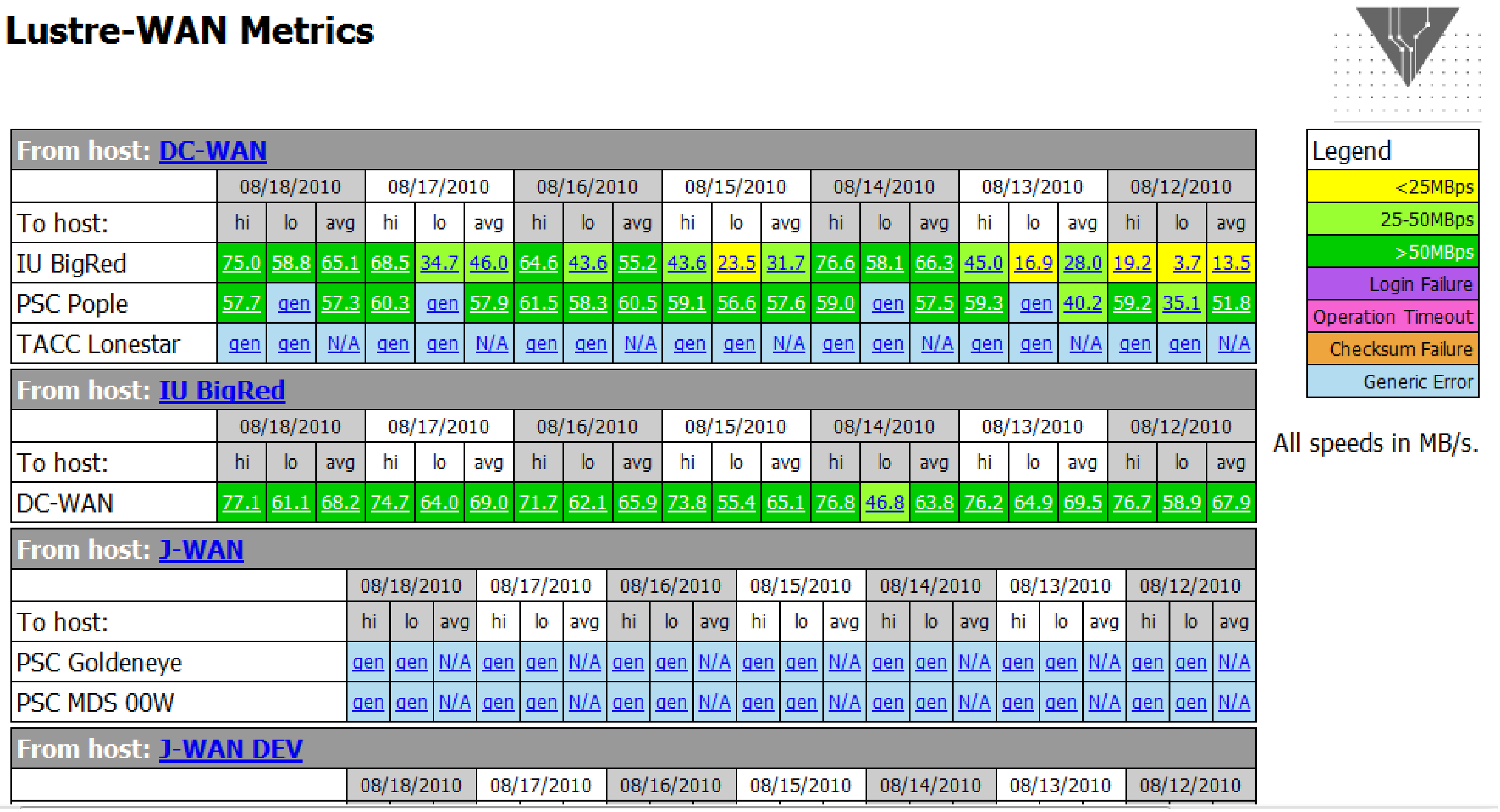
Test date: 21-June-10

Test and hosts	Thruput MB/s	Thruput Mb/s	Max thruput Gbps	% of max	Notes
Iperf: mds18.psc->oss1.tacc	73.25	586	1	58.60	
Iperf: oss0.psc->oss2.tacc	75.63	605	1	60.50	
Iperf: oss0.psc->oss1.tacc	72.50	580	1	58.00	
Iperf: mds18.psc->oss2.tacc	70.13	561	1	56.10	
iperf: oss2.tacc->oss1.psc	68.88	551	1	55.10	
iperf: oss1.tacc->mds18.psc	68.25	546	1	54.60	
netperf: mds18.psc->oss1.tacc	72.33	578.6	1	57.86	
netperf: oss0.psc->oss2.tacc	75.04	600.31	1	60.03	
netperf: oss1.tacc->mds18.psc	68.22	545.75	1	54.58	

Test date: 23-June-10

Test and hosts	Thruput MB/s	Thruput Mb/s	Max thruput Gbps	% of max	Notes
netperf: oss0.psc->oss1.tacc	75.31	602.49	1	60.25	
netperf: oss1.tacc->oss0.psc	67.99	543.89	1	54.39	
netperf: oss1.tacc->oss0.psc	69.04	552.29	1	55.23	
netperf: oss1.tacc->mds18.psc	68.84	550.68	1	55.07	

On Going Network Performance Testing



Lustre-WAN Metrics

From host: DC-WAN

	08/18/2010			08/17/2010			08/16/2010			08/15/2010			08/14/2010			08/13/2010			08/12/2010		
To host:	hi	lo	avg	hi	lo	avg	hi	lo	avg	hi	lo	avg	hi	lo	avg	hi	lo	avg	hi	lo	avg
IU BigRed	75.0	58.8	65.1	68.5	34.7	46.0	64.6	43.6	55.2	43.6	23.5	31.7	76.6	58.1	66.3	45.0	16.9	28.0	19.2	3.7	13.5
PSC Pople	57.7	gen	57.3	60.3	gen	57.9	61.5	58.3	60.5	59.1	56.6	57.6	59.0	gen	57.5	59.3	gen	40.2	59.2	35.1	51.8
TACC Lonestar	gen	gen	N/A	gen	gen	N/A	gen	gen	N/A	gen	gen	N/A	gen	gen	N/A	gen	gen	N/A	gen	gen	N/A

From host: IU BigRed

	08/18/2010			08/17/2010			08/16/2010			08/15/2010			08/14/2010			08/13/2010			08/12/2010		
To host:	hi	lo	avg	hi	lo	avg	hi	lo	avg	hi	lo	avg	hi	lo	avg	hi	lo	avg	hi	lo	avg
DC-WAN	77.1	61.1	68.2	74.7	64.0	69.0	71.7	62.1	65.9	73.8	55.4	65.1	76.8	46.8	63.8	76.2	64.9	69.5	76.7	58.9	67.9

From host: J-WAN

	08/18/2010			08/17/2010			08/16/2010			08/15/2010			08/14/2010			08/13/2010			08/12/2010		
To host:	hi	lo	avg	hi	lo	avg	hi	lo	avg	hi	lo	avg	hi	lo	avg	hi	lo	avg	hi	lo	avg
PSC Goldeneye	gen	gen	N/A	gen	gen	N/A	gen	gen	N/A	gen	gen	N/A	gen	gen	N/A	gen	gen	N/A	gen	gen	N/A
PSC MDS 00W	gen	gen	N/A	gen	gen	N/A	gen	gen	N/A	gen	gen	N/A	gen	gen	N/A	gen	gen	N/A	gen	gen	N/A

From host: J-WAN DEV

	08/18/2010	08/17/2010	08/16/2010	08/15/2010	08/14/2010	08/13/2010	08/12/2010

Legend
<25MBps
25-50MBps
>50MBps
Login Failure
Operation Timeout
Checksum Failure
Generic Error

All speeds in MB/s.

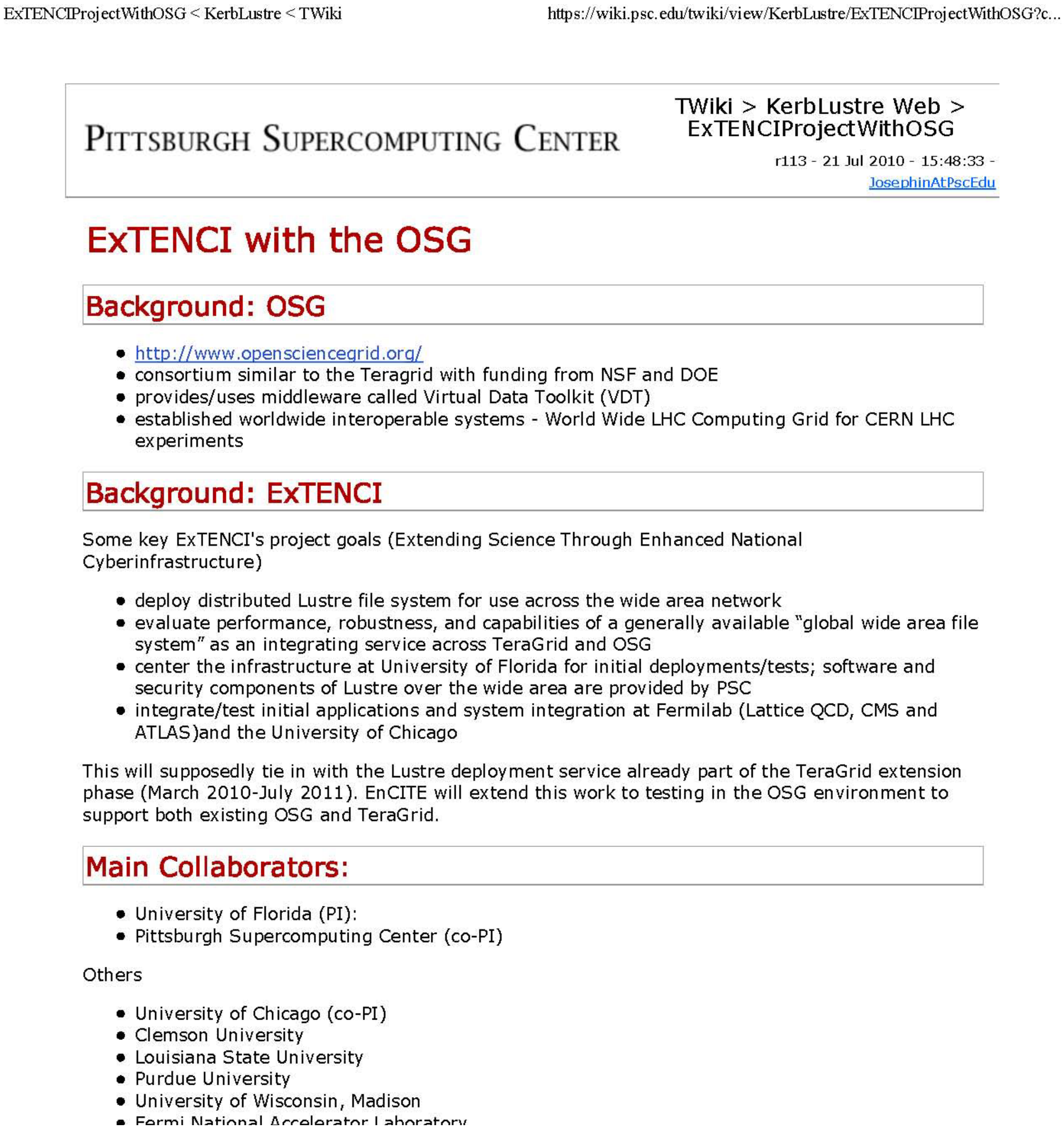
Application Integration

- Largely Invisible to Application
- Performance
  - Large Metadata Operations
  - Data Locality
- Independent Assessment for LQCD, CMS services to include:
  - data integrity
  - accessibility
  - usability

Application Integration, cont.

- maintainability
- ability to troubleshoot/isolate problems
- namespace
- IO performance
- Metrics and Assessment evaluate acceptability as production storage for LHC physics
- compare with Hadoop20 implementation
- test with SCEC and Protein Structure applications

Project WIKI



ExTENCIProjectWithOSG < KerbLustre < TWiki https://wiki.psc.edu/twiki/view/KerbLustre/ExTENCIProjectWithOSG?c...

PITTSBURGH SUPERCOMPUTING CENTER

TWiki > KerbLustre Web > ExTENCIProjectWithOSG

r113 - 21 Jul 2010 - 15:48:33 - JosephinAtPscEdu

ExTENCI with the OSG

Background: OSG

- http://www.opensciencegrid.org/
- consortium similar to the Teragrid with funding from NSF and DOE
- provides/uses middleware called Virtual Data Toolkit (VDT)
- established worldwide interoperable systems - World Wide LHC Computing Grid for CERN LHC experiments

Background: ExTENCI

Some key ExTENCI's project goals (Extending Science Through Enhanced National Cyberinfrastructure)

- deploy distributed Lustre file system for use across the wide area network
- evaluate performance, robustness, and capabilities of a generally available "global wide area file system" as an integrating service across TeraGrid and OSG
- center the infrastructure at University of Florida for initial deployments/tests; software and security components of Lustre over the wide area are provided by PSC
- integrate/test initial applications and system integration at Fermilab (Lattice QCD, CMS and ATLAS)and the University of Chicago

This will supposedly tie in with the Lustre deployment service already part of the TeraGrid extension phase (March 2010-July 2011). EnCITE will extend this work to testing in the OSG environment to support both existing OSG and TeraGrid.

Main Collaborators:

- University of Florida (PI):
- Pittsburgh Supercomputing Center (co-PI)

Others

- University of Chicago (co-PI)
- Clemson University
- Louisiana State University
- Purdue University
- University of Wisconsin, Madison
- Fermi National Accelerator Laboratory

Teragrid WIKI

- http://teragridforum.org/mediawiki/index.php?title=JWAN:_lustre-wan_advanced_features_testing

Thank You

- Josephine Palencia – Josephin@psc.edu
- J. Ray Scott – Scott@psc.edu