the globus alliance
www.globus.org

National
e-Science
Centre

Grid Monitoring and Information Services:

Globus Toolkit MDS4
&
TeraGrid Inca

Jennifer M. Schopf

Argonne National Lab

UK National eScience Center (NeSC)

Overview

- Brief overview of what I mean by “Grid monitoring”
- Tool for Monitoring/Discovery:
  - Globus Toolkit MDS 4
- Tool for Monitoring/Status Tracking
  - Inca from the TeraGrid project
- Just added: GLUE schema in a nutshell

What do I mean by monitoring?

- Discovery and expression of data
- Discovery:
  - Registry service
  - Contains descriptions of data that is available
  - Sometimes also where last value of data is kept (caching)
- Expression of data
  - Access to sensors, archives, etc.
  - Producer (in consumer producer model)

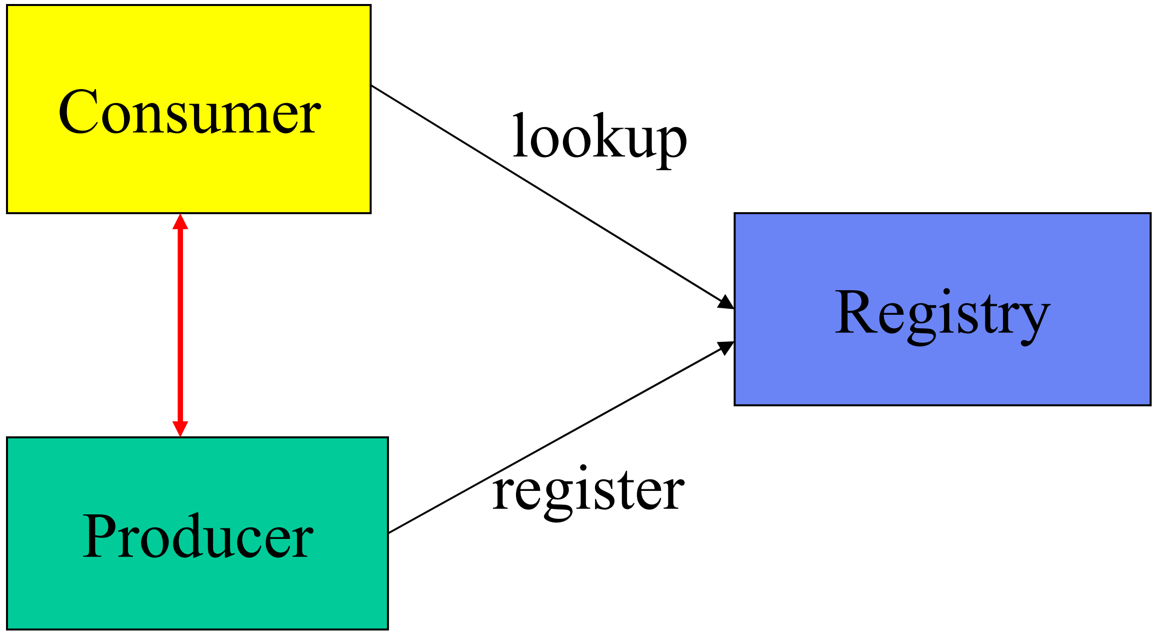
What do I mean by Grid monitoring?

- Grid level monitoring concerns data that is:
  - Shared between administrative domains
  - For use by multiple people
  - Often summarized
  - (think scalability)
- Different levels of monitoring needed:
  - Application specific
  - Node level
  - Cluster/site Level
  - Grid level
- Grid monitoring may contain summaries of lower level monitoring

Grid Monitoring Does Not Include…

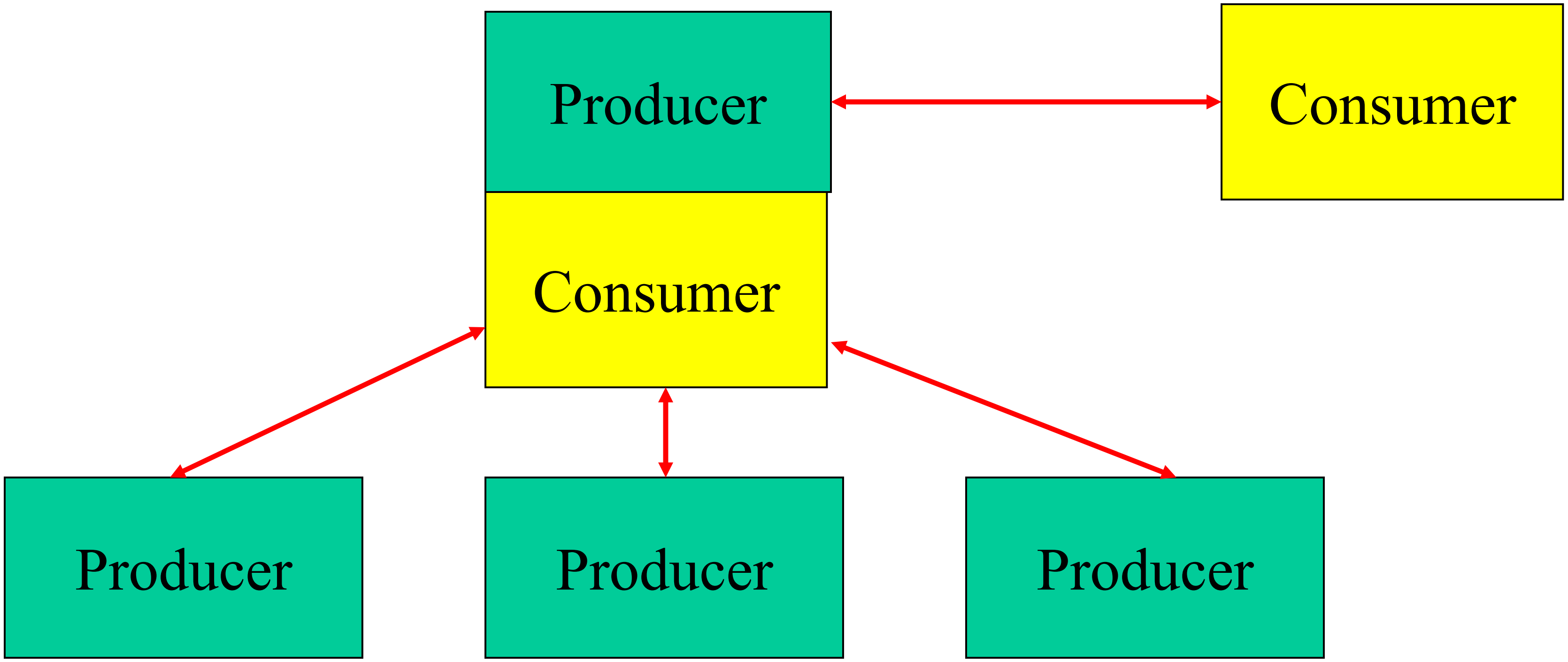
- All the data about every node of every site
- Years of utilization logs to use for planning next hardware purchase
- Low-level application progress details for a single user
- Application debugging data (except perhaps notification of a failure of a heartbeat)
- Point-to-point sharing of all data over all sites

What monitoring systems look like GMA architecture



Compound Producer-Consumers

- In order to have more than just data sources and simple sinks approaches combine these



Pieces of a Grid Monitoring System

- Producer
  - Any component that publishes monitoring data (also called a sensor, data source, information provider, etc)
- Consumer
  - Any component the requests data from a producer
- Registry or directory service
  - A construct (database?) containing information on what producer publishes what events, and what the event schemas are for those events
  - Some approaches cache data (last value) as well
- Higher-Level services
  - Aggregation, Trigger Services, Archiving
- Client Tools
  - APIs, Viz services, etc

PGI Monitoring Defined Usecases

- Joint PPDG, GriPhyN and iVDGL effort to define monitoring requirements
- http://www.mcs.anl.gov/~jms/pg-monitoring
- 19 use cases from ~9 groups
- Roughly 4 categories
  - Health of system (NW, servers, cpus, etc)
  - System upgrade evaluation
  - Resource selection
  - Application-specific progress tracking

Why So Many Monitoring Systems?

- There is no ONE tool for this job
  - Nor would you ever get agreement between sites to all deploy it if there was
- Best you can hope for is
  - An understanding of overlap
  - Standard-defined interactions when possible

Things to Think About When Comparing Systems

- What is the main use case your system addresses?
- What are the base set of sensors given with a system?
- How does that set get extended?
- What are you doing for discovery/registry?
- What schema are you using (do you interact with)?
- Is this system meant to monitor a machine, a cluster, or send data between sites, or some combination of the above?
- What kind of testing has been done in terms of scalability (several pieces to this - how often is data updated, how many users, how many data sources, how many sites, etc)

Two Systems To Consider

- Globus Toolkit Monitoring and Discovery System 4 (MDS4)
  - WSRF-compatible
  - Resource Discovery
  - Service Status
- Inca test harness and reporting framework
  - TeraGrid project
  - Service agreement monitoring – software stack, service up/down, performance

Monitoring and Discovery Service in GT4 (MDS4)

- WS-RF compatible
- Monitoring of basic service data
- Primary use case is discovery of services
- Starting to be used for up/down statistics

MDS4 Producers: Information Providers

- Code that generates resource property information
  - Were called service data providers in GT3
- XML Based – not LDAP
- Basic cluster data
  - Interface to Ganglia
  - GLUE schema
- Some service data from GT4 services
  - Start, timeout, etc
- Soft-state registration
- Push and pull data models

MDS4 Registry: Aggregator

- Aggregator is both registry and cache
- Subscribes to information providers
  - Data, datatype, data provider information
- Caches last value of all data
- In memory default approach

MDS4 Trigger Service

- Compound consumer-producer service
- Subscribe to a set of resource properties
- Set of tests on incoming data streams to evaluate trigger conditions
- When a condition matches, email is sent to pre-defined address

- GT3 tech-preview version in use by ESG
- GT4 version alpha is in GT4 alpha release currently available

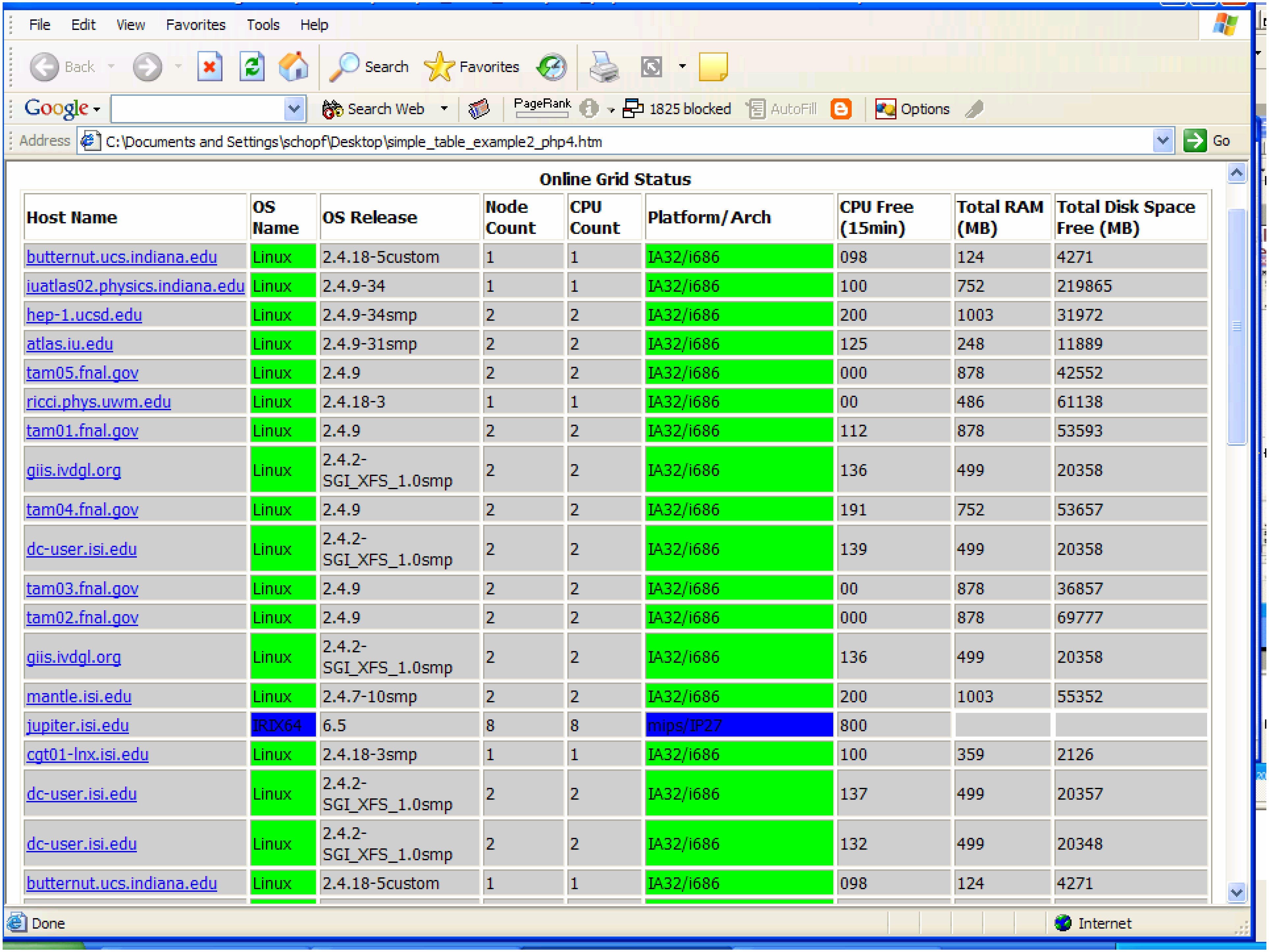
MDS4 Archive Service

- Compound consumer-producer service
- Subscribe to a set of resource properties
- Data put into database (Xindice)
- Other consumers can contact database archive interface

- Will be in GT4 beta release

MDS4 Clients

- Command line, Java and C APIs
- MDSWeb Viz service
  - Tech preview in current alpha (3.9.3 last week)


File Edit View Favorites Tools Help
Back
Search
Favorites
Google
Search Web
PageRank
1825 blocked
AutoFill
Options
Address
C:\Documents and Settings\schopf\Desktop\simple_table_example2_php4.htm
Go
Done
Internet

Online Grid Status

Host Name	OS Name	OS Release	Node Count	CPU Count	Platform/Arch	CPU Free (15min)	Total RAM (MB)	Total Disk Space Free (MB)
butternut.ucs.indiana.edu	Linux	2.4.18-5custom	1	1	IA32/i686	098	124	4271
iuatlas02.physics.indiana.edu	Linux	2.4.9-34	1	1	IA32/i686	100	752	219865
hep-1.ucsd.edu	Linux	2.4.9-34smp	2	2	IA32/i686	200	1003	31972
atlas.iu.edu	Linux	2.4.9-31smp	2	2	IA32/i686	125	248	11889
tam05.fnal.gov	Linux	2.4.9	2	2	IA32/i686	000	878	42552
ricci.phys.uwm.edu	Linux	2.4.18-3	1	1	IA32/i686	00	486	61138
tam01.fnal.gov	Linux	2.4.9	2	2	IA32/i686	112	878	53593
giis.ivdgl.org	Linux	2.4.2-SGI_XFS_1.0smp	2	2	IA32/i686	136	499	20358
tam04.fnal.gov	Linux	2.4.9	2	2	IA32/i686	191	752	53657
dc-user.isi.edu	Linux	2.4.2-SGI_XFS_1.0smp	2	2	IA32/i686	139	499	20358
tam03.fnal.gov	Linux	2.4.9	2	2	IA32/i686	00	878	36857
tam02.fnal.gov	Linux	2.4.9	2	2	IA32/i686	000	878	69777
giis.ivdgl.org	Linux	2.4.2-SGI_XFS_1.0smp	2	2	IA32/i686	136	499	20358
mantle.isi.edu	Linux	2.4.7-10smp	2	2	IA32/i686	200	1003	55352
jupiter.isi.edu	IRIX64	6.5	8	8	mips/IP27	800		
cgt01-lnx.isi.edu	Linux	2.4.18-3smp	1	1	IA32/i686	100	359	2126
dc-user.isi.edu	Linux	2.4.2-SGI_XFS_1.0smp	2	2	IA32/i686	137	499	20357
dc-user.isi.edu	Linux	2.4.2-SGI_XFS_1.0smp	2	2	IA32/i686	132	499	20348
butternut.ucs.indiana.edu	Linux	2.4.18-5custom	1	1	IA32/i686	098	124	4271

Coming Up Soon…

- Extend MDS4 information providers
  - More data from GT4 services (GRAM, RFT, RLS)
  - Interface to other tests (Inca, GRASP)
  - Interface to archiver (PinGER, Ganglia, others)
- Scalability testing and development
- Additional clients
- If tracking job stats is of interest this is something we can talk about

TeraGrid Inca

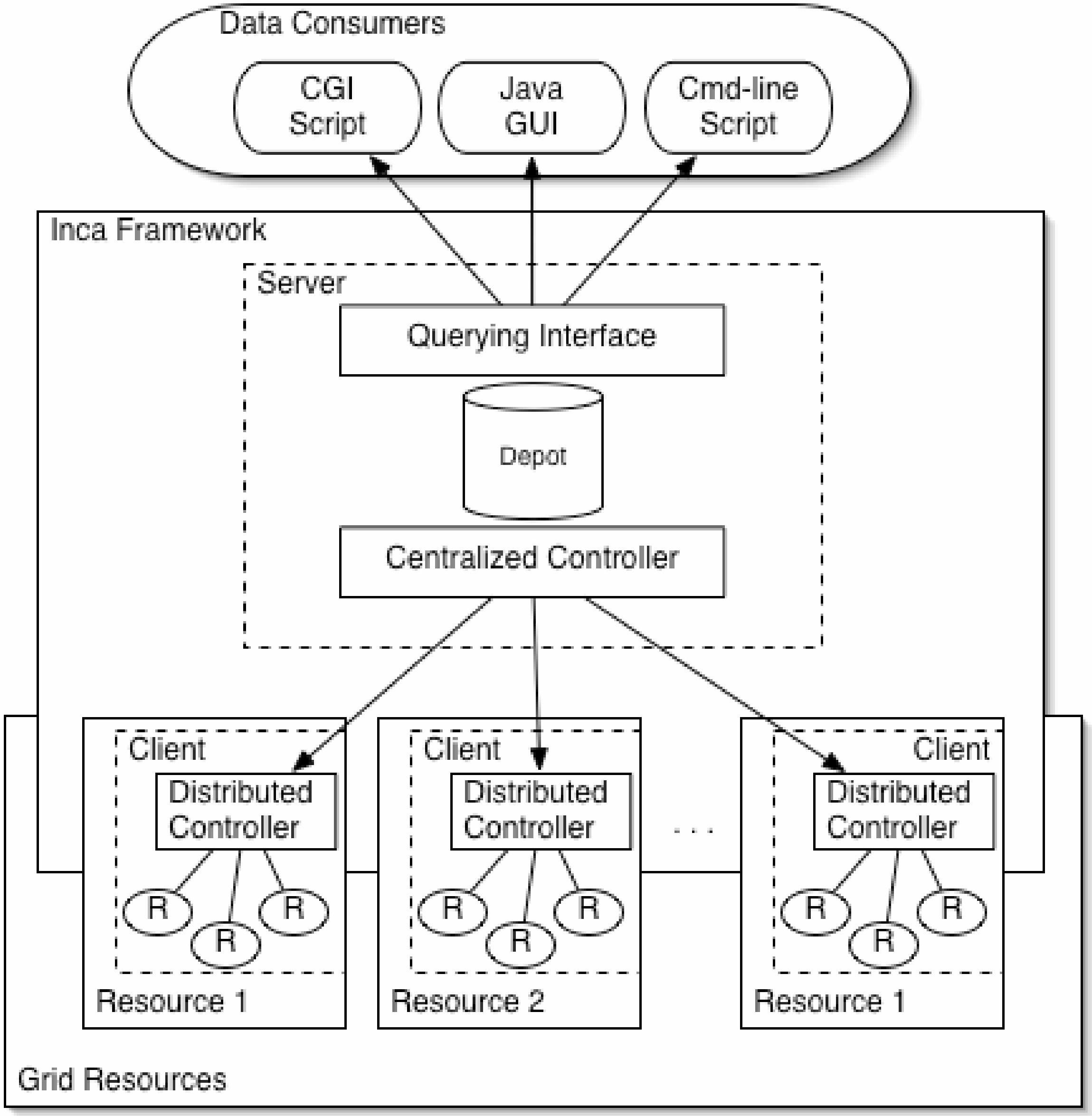
- Originally developed for the TeraGrid project to verify its software stack
- Now part of the NMI GRIDS center software
- Now performs automated verification of service-level agreements
  - Software versions
  - Basic software and service tests – local and cross-site
  - Performance benchmarks
- Best use: CERTIFICATION
  - Is this site Project Compliant?
  - Have upgrades taken place in a timely fashion?

Inca Producers: Reporters

- Over 100 tests deployed on each TG resource (9 sites)
  - Load on host systems less than 0.05% overall
- Primarily specific software versions and functionality tests
  - Versions not functionality because functionality is an open question
  - Grid service capabilities cross-site
  - GT 2.4.3 GRAM jobs submission & GridFTP
  - OpenSSH
  - MyProxy
- Soon to be deployed: SRB, VMI, BONNIE benchmarks, LAPACK Benchmarks

Support Services

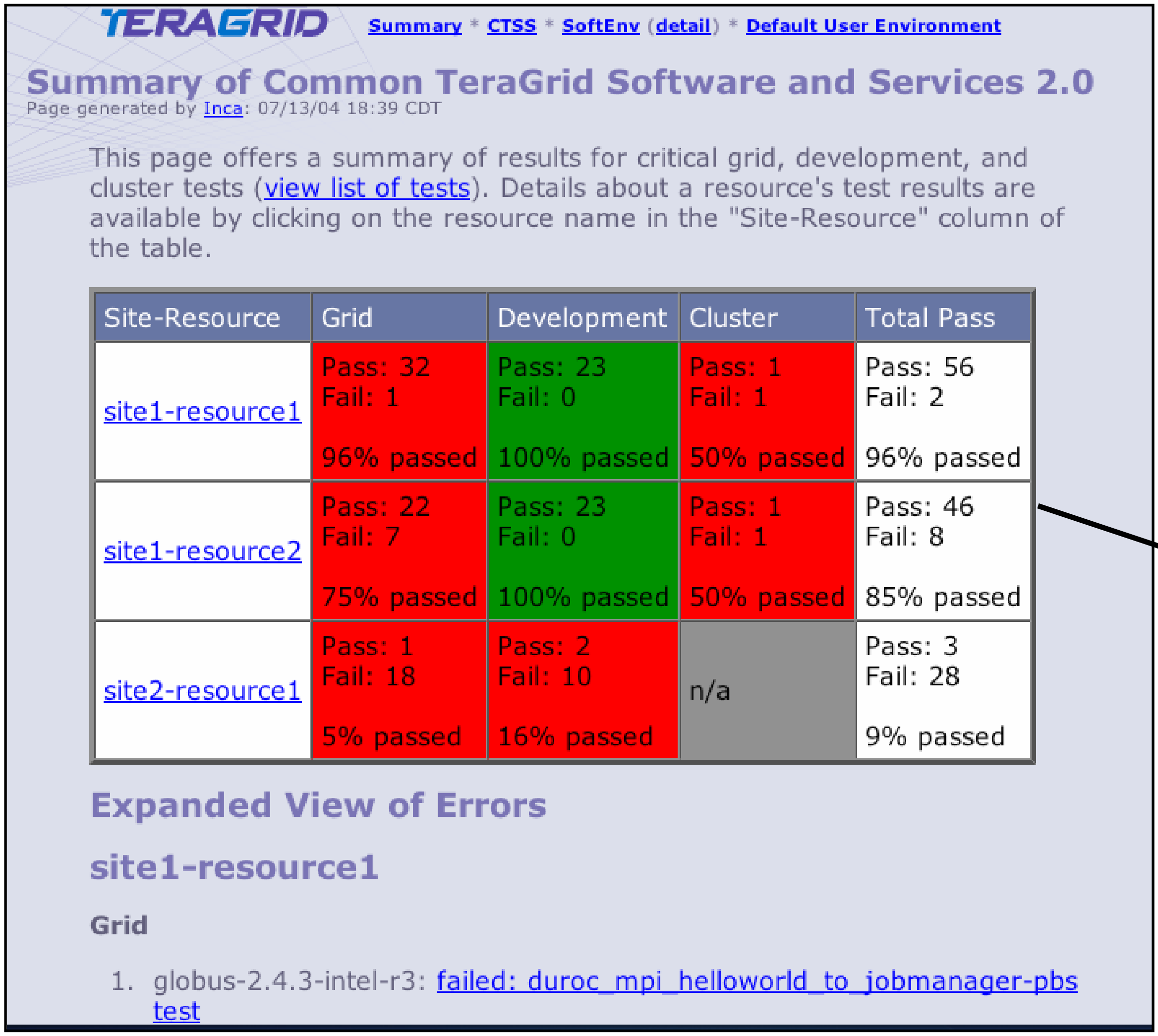
- Distributed controller
  - runs on each client resource
  - controls the local data collection through the reporters
- Centralized controller
  - system administrators can change data collection rates and deployment of the reporters
- Archive system (depot)
  - collects all the reporter data using a round-robin database scheme.


Data Consumers
CGI Script
Java GUI
Cmd-line Script
Inca Framework
Server
Querying Interface
Depot
Centralized Controller
Client
Distributed Controller
R
R
R
Resource 1
Client
Distributed Controller
R
R
R
Resource 2
. . .
Client
Distributed Controller
R
R
R
Resource 1
Grid Resources

Interfaces

- Command line, C, and Perl APIs
- Several GUI clients
- Executive view
  - http://tech.teragrid.org/inca/TG/html/execView.html
- Overall Status
  - http://tech.teragrid.org/inca/TG/html/stackStatus.html

Example Summary View Snapshot



TERAGRID **Summary** * **CTSS** * **SoftEnv** (**detail**) * **Default User Environment**

Summary of Common TeraGrid Software and Services 2.0

Page generated by Inca: 07/13/04 18:39 CDT

This page offers a summary of results for critical grid, development, and cluster tests (view list of tests). Details about a resource's test results are available by clicking on the resource name in the "Site-Resource" column of the table.

Site-Resource	Grid	Development	Cluster	Total Pass
site1-resource1	Pass: 32 Fail: 1 96% passed	Pass: 23 Fail: 0 100% passed	Pass: 1 Fail: 1 50% passed	Pass: 56 Fail: 2 96% passed
site1-resource2	Pass: 22 Fail: 7 75% passed	Pass: 23 Fail: 0 100% passed	Pass: 1 Fail: 1 50% passed	Pass: 46 Fail: 8 85% passed
site2-resource1	Pass: 1 Fail: 18 5% passed	Pass: 2 Fail: 10 16% passed	n/a	Pass: 3 Fail: 28 9% passed

Expanded View of Errors

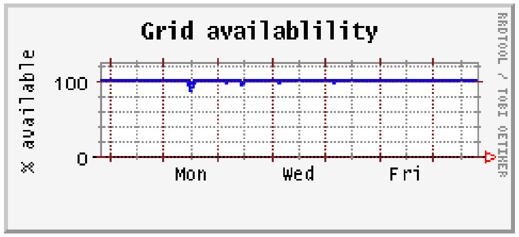
site1-resource1

Grid

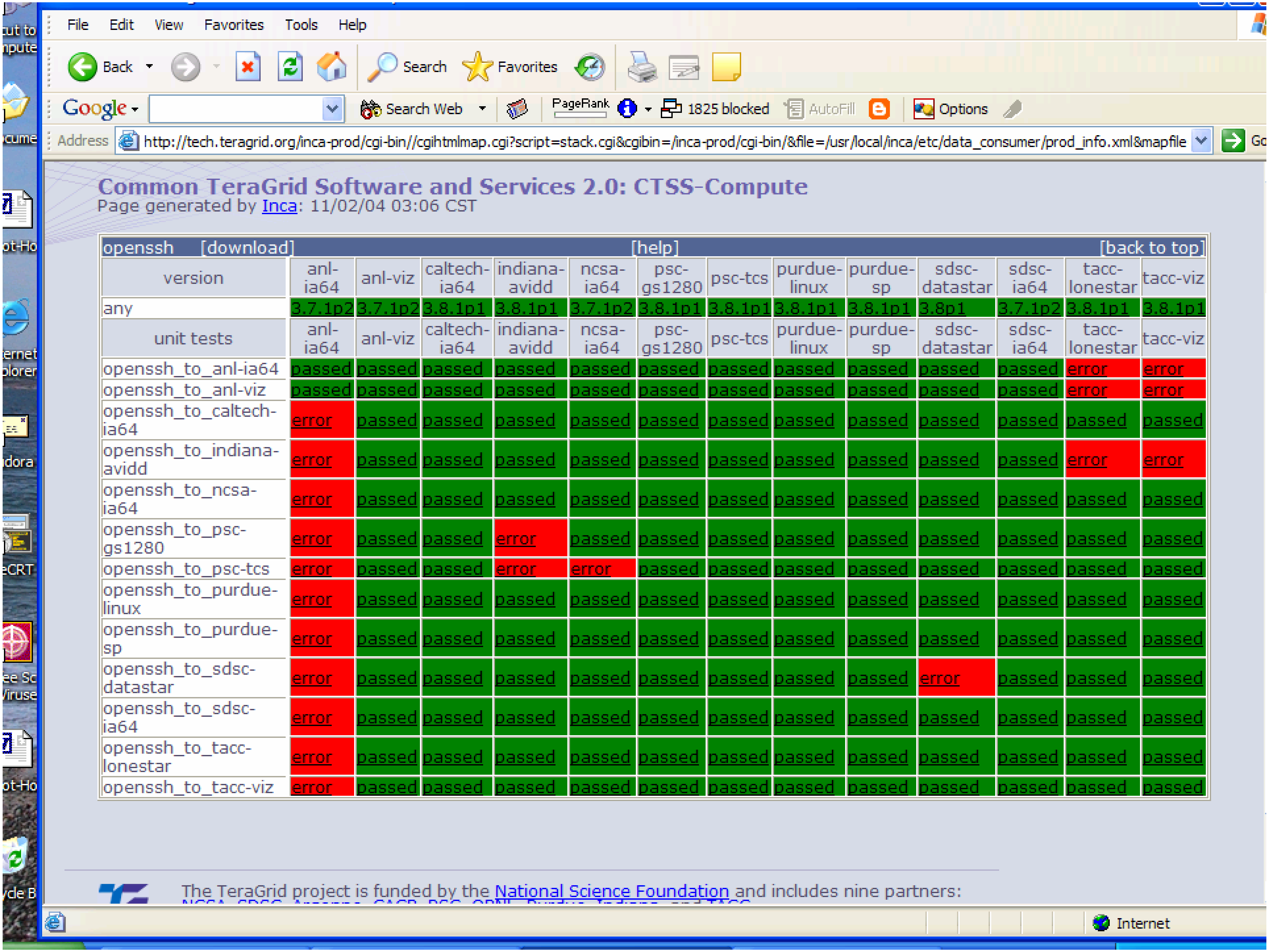
1. globus-2.4.3-intel-r3: failed: duroc_mpi_helloworld_to_jobmanager-pbs test

🟩	All tests passed: 100%
🟥	One or more tests failed: < 100%
▨	Tests not applicable to machine or have not yet been ported

Key



History of percentage of tests passed in “Grid” category for a one week period



http://tech.teragrid.org/inca-prod/cgi-bin//cgihtmlmap.cgi?script=stack.cgi&cgibin=/inca-prod/cgi-bin/&file=/usr/local/inca/etc/data_consumer/prod_info.xml&mapfile

Common TeraGrid Software and Services 2.0: CTSS-Compute

Page generated by Inca: 11/02/04 03:06 CST

openssh [download] [help] [back to top]

version	anl-ia64	anl-viz	caltech-ia64	indiana-avidd	ncsa-ia64	psc-gs1280	psc-tcs	purdue-linux	purdue-sp	sdsc-datastar	sdsc-ia64	tacc-lonestar	tacc-viz
any	3.7.1p2	3.7.1p2	3.8.1p1	3.8.1p1	3.7.1p2	3.8.1p1	3.8.1p1	3.8.1p1	3.8.1p1	3.8p1	3.7.1p2	3.8.1p1	3.8.1p1
unit tests	anl-ia64	anl-viz	caltech-ia64	indiana-avidd	ncsa-ia64	psc-gs1280	psc-tcs	purdue-linux	purdue-sp	sdsc-datastar	sdsc-ia64	tacc-lonestar	tacc-viz
openssh_to_anl-ia64	passed	passed	passed	passed	passed	passed	passed	passed	passed	passed	passed	error	error
openssh_to_anl-viz	passed	passed	passed	passed	passed	passed	passed	passed	passed	passed	passed	error	error
openssh_to_caltech-ia64	error	passed	passed	passed	passed	passed	passed	passed	passed	passed	passed	passed	passed
openssh_to_indiana-avidd	error	passed	passed	passed	passed	passed	passed	passed	passed	passed	passed	error	error
openssh_to_ncsa-ia64	error	passed	passed	passed	passed	passed	passed	passed	passed	passed	passed	passed	passed
openssh_to_psc-gs1280	error	passed	passed	error	passed	passed	passed	passed	passed	passed	passed	passed	passed
openssh_to_psc-tcs	error	passed	passed	error	error	passed	passed	passed	passed	passed	passed	passed	passed
openssh_to_purdue-linux	error	passed	passed	passed	passed	passed	passed	passed	passed	passed	passed	passed	passed
openssh_to_purdue-sp	error	passed	passed	passed	passed	passed	passed	passed	passed	passed	passed	passed	passed
openssh_to_sdsc-datastar	error	passed	passed	passed	passed	passed	passed	passed	passed	error	passed	passed	passed
openssh_to_sdsc-ia64	error	passed	passed	passed	passed	passed	passed	passed	passed	passed	passed	passed	passed
openssh_to_tacc-lonestar	error	passed	passed	passed	passed	passed	passed	passed	passed	passed	passed	passed	passed
openssh_to_tacc-viz	error	passed	passed	passed	passed	passed	passed	passed	passed	passed	passed	passed	passed

The TeraGrid project is funded by the National Science Foundation and includes nine partners:

Inca Future Plans

- Paper being presented at SC04
  - Scalability results (soon to be posted here)
  - www.mcs.anl.gov/~jms/Pubs/jmspubs.html
- Extending information and sites
- Restructuring depot (archiving) for added scalability (RRDB won't meet future needs)
- Cascading reporters – trigger more info on failure
- Discussions with several groups to consider adoption/certification programs
  - NEES, GEON, UK NGS, others

GLUE Schema

- Why do we need a fixed schema?
  - Communication between projects
- Condor doesn't have one – why do we need one?
  - Condor has a defacto schema
  - OS won't match to OpSys – major problem when matchmaking between sites
- What about doing updates?
  - Schema updates should NOT be done on the fly if you want to maintain compatibility
  - On the other hand, they don't need to be since by definition they include deploying new sensors to gather data
  - Whether or not sw has to be re-started after a deployment is an implementation issue, not a schema issue

Glue Schema

- Does a schema have to define everything?
  - No – GLUE schema v1 was in use and by plan did NOT define everything
  - It had extendable pieces so we could get more hands on use
  - This is what projects have been doing since it was defined 18 months ago

Extending the GLUE Schema

- Sergio Andreozzi proposed extending the GLUE schema to take into account project-specific details
  - We now have hands on experience
  - Every project has added their own extension
  - We need to unify them
- Mailman list
  - www.hicb.org/mailman/listinfo/glue-schema
- Bugzilla-like system for tracking the proposed changes
  - infnforge.cnaf.infn.it/projects/glueinfomodel/
  - Currently only used by Sergio :)
- Mail this morning suggesting better requirement gathering and phone call/meeting to move forward

Ways Forward

- Sharing of tests between infrastructures
- Help contribute to GLUE schema
- Share use cases and scalability requirements

- Hardest thing in Grid computing isn’t technical, it’s socio-political and communication

For More Information

- Jennifer Schopf
  - jms@mcs.anl.gov
  - http://www.mcs.anl.gov/~jms
- Globus Toolkit MDS4
  - http://www.globus.org/mds
- Inca
  - http://tech.teragrid.org/inca
- Scalability comparison of MDS2, Hawkeye, R-GMA

www.mcs.anl.gov/~jms/Pubs/xuehaijeff-hpdc2003.pdf

- Monitoring Clusters, Monitoring the Grid – ClusterWorld
  - http://www.grids-center.org/news/clusterworld/