

Xtremely Large File Systems for the small collaborative world

Arun Jagatheesan

**San Diego Supercomputer Center
University of California at San Diego**

&

**DiceResearch.org
(arun@sdsc.edu)**

HPTS Workshop

Asilomar, California, 28 October 2009



iRODS.org

San Diego Supercomputer Center



Agenda

- Intro
- Problem Motivation: LSST and myGlobal500.com
- Solution Path
- Implementation: iRODS
- Improvements Possible / Future Plans



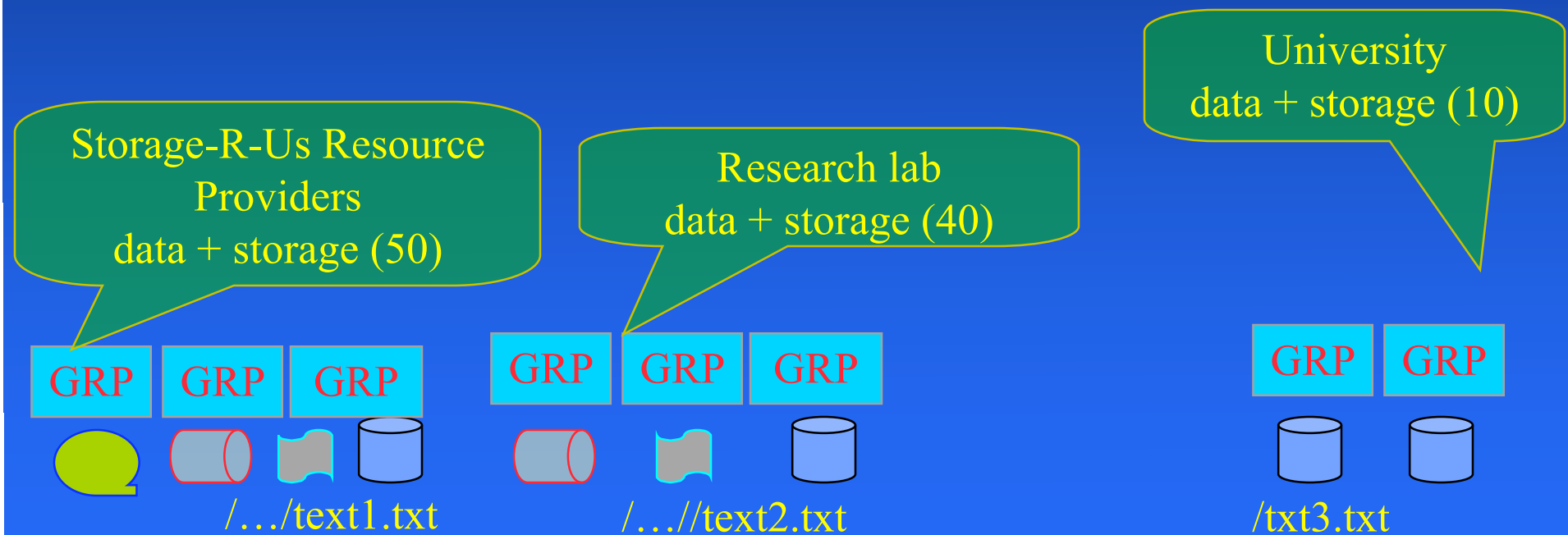
Slides from HPTS 2005



2005

Data Grid Administrative domains

2005



iRODS.org

2005

Data Grid: Logical view of data & resources

2005

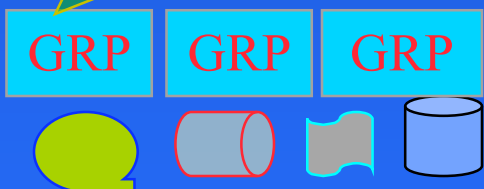
/home/arun.sdsc/exp1
/home/arun.sdsc/exp1/text1.txt
/home/arun.sdsc/exp1/text2.txt
/home/arun.sdsc/exp1/text3.txt
data + storage (100)

Logical Namespace (Need not be same as physical view of resources)

University
data + storage (10)

Storage-R-Us Resource
Providers
data + storage (50)

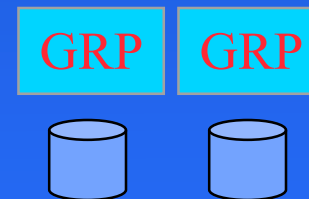
Research Lab
data + storage (40)



/.../text1.txt



/.../text2.txt



/txt3.txt



iRODS.org

2005

Data Grid Language

2005

- **Requirement**
 - Data Grid ILM process
 - The long run process that has to be run is described in DGL
 - Data Grid Triggers
 - Action part of the ECA (Event-Condition-Action) logic
 - Data Gridflows
 - Step by step execution of long run process on Data Grid
- **Analogy of SQL in relational databases**
 - Long-run procedures stored and executed in Data Grid it self
 - Captures the “Infrastructure Execution Logic”



Since HPTS 2005

- SRB -> iRODS (open source)
- .org
- DGL (triggers etc) -> iRODS rules



Agenda

- Intro
- Problem Motivation: LSST and myGlobal500.com
- Solution Path
- Implementation: iRODS
- Improvements Possible / Future Plans

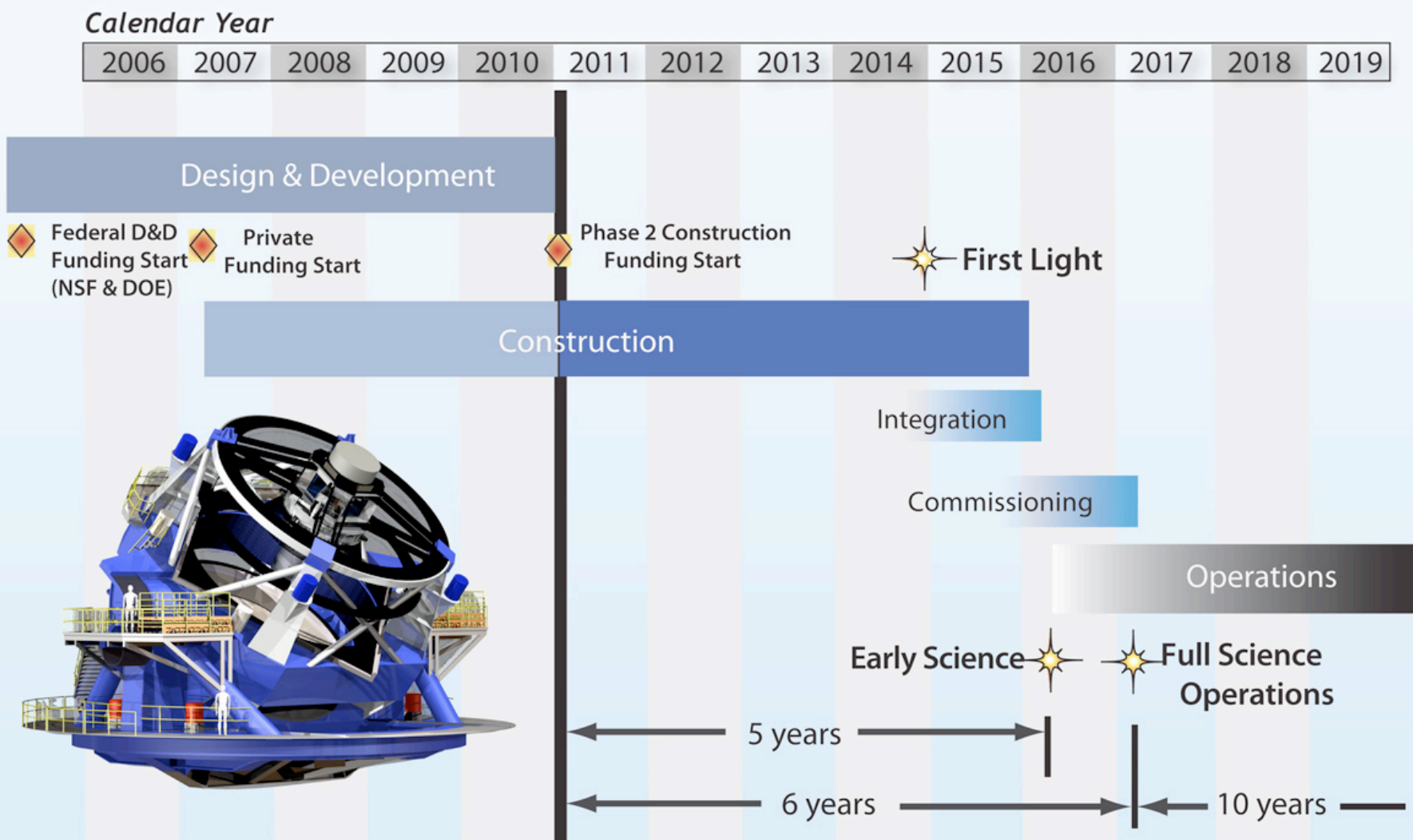


LSST

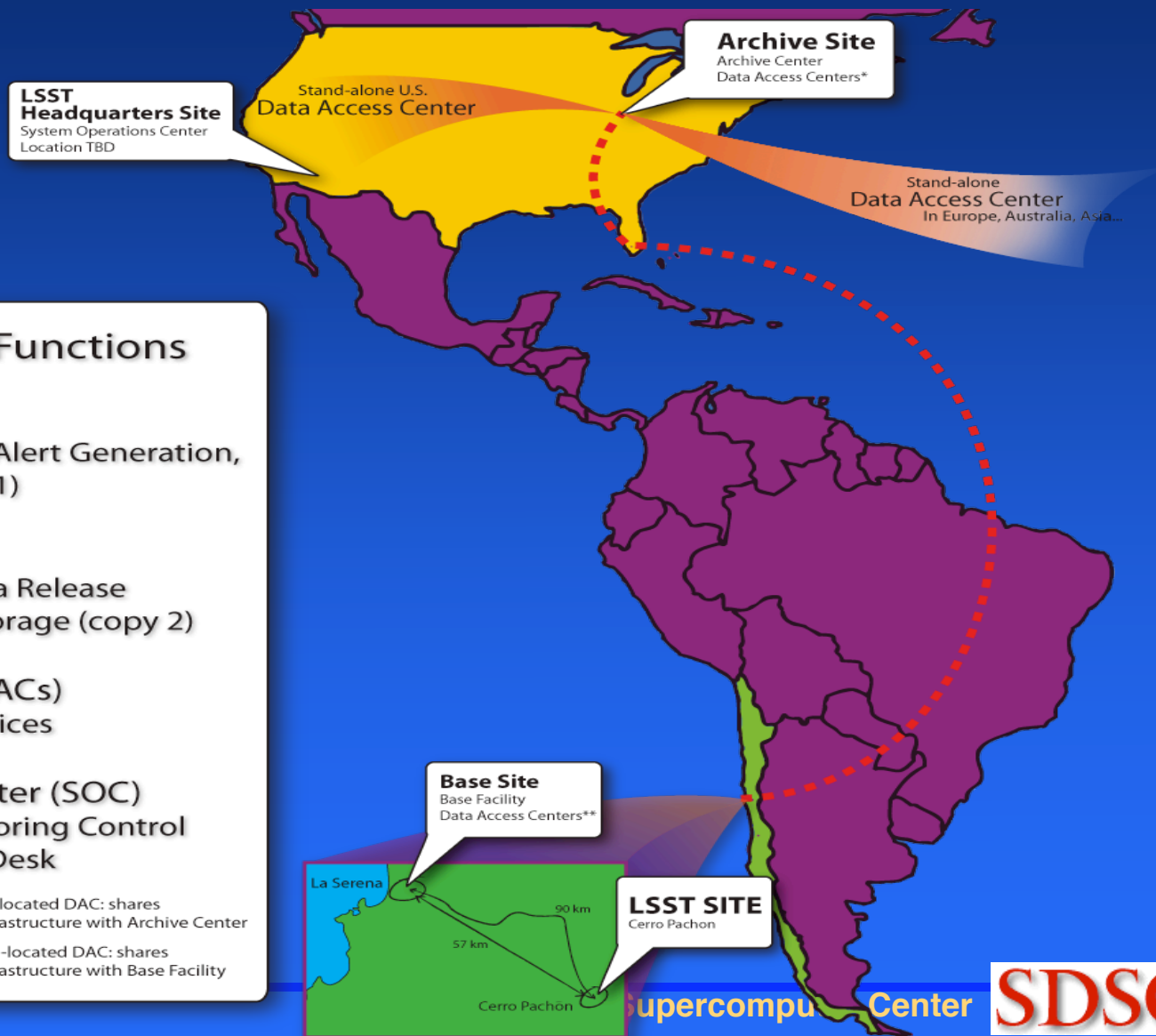
- **Large Synoptic Survey Telescope (LSST)**
 - Survey entire sky every 3 nights
 - Dark Energy, Dark Matter, Near Earth Asteroids, and more
 - Largest digital camera in the world (3 billion pixels)
 - Images 3000 times wider than Hubble
 - http://www.youtube.com/watch?v=LtMJ_WwvBb8
 - Public data
- **LSST Data Management**
 - Data from Chile to US and rest of the world
 - 15 TB/night, over hundred(s) petabytes
 - Multiple data centers around the world
 - Hundreds of millions of files (more)



LSST Project Schedule



LSST current sites



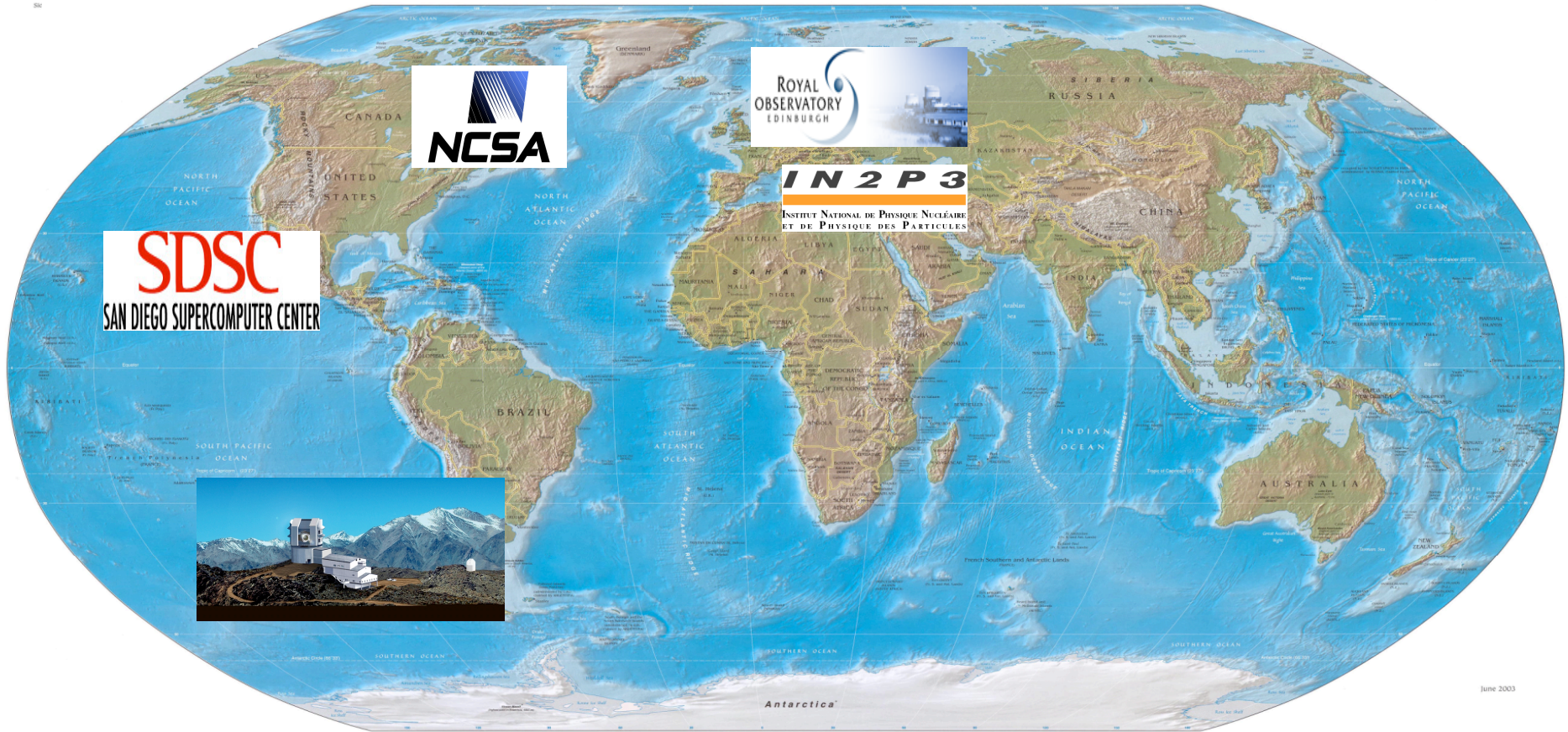
Site Roles and their Functions

- **Base Facility**
Real-time Processing and Alert Generation,
Long-term storage (copy 1)
- **Archive Center**
Nightly Reprocessing, Data Release
Processing, Long-term Storage (copy 2)
- **Data Access Centers (DACs)**
Data Access and User Services
- **System Operations Center (SOC)**
System Supervisory Monitoring Control
& End User Support/Help Desk

* Co-located DAC: shares infrastructure with Archive Center

** Co-located DAC: shares infrastructure with Base Facility

If it were live now ... (Optimist's simulation)



iRODS.org

12

San Diego Supercomputer Center



LSST sites growing...



iRODS.org

Assuming files from 1st light (1st dir)

- /LSST/exp/
 - file1.fits
 - file2.fits
 - file3.fits



Separate file systems...



`\\i\exp\file1.fits`
`\\i\exp\file3.fits`



`/u/exp/file1.fits`
`/u/exp/file2.fits`

`/res/chile/exp/file1.fits`



`/euro/exp/file2.fits`

`/exp/file1.fits`
`/exp/file2.fits`



iRODS.org

Disadvantages of separate file systems

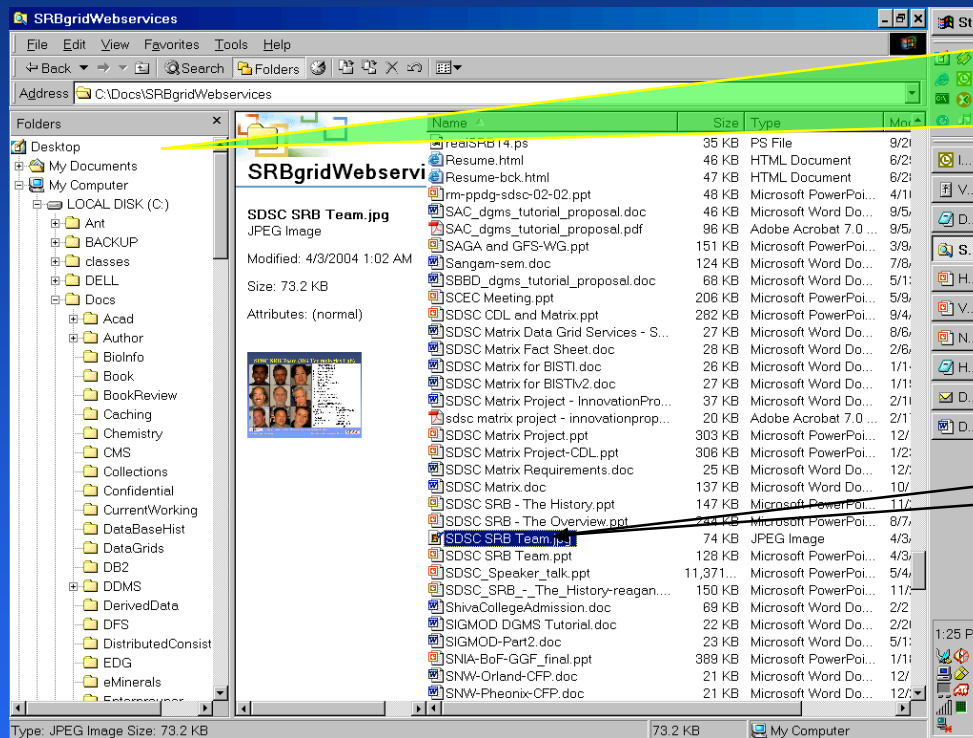
- **Increased cost of operation**
 - Storage cost for data backups (in petabytes)
- **No load sharing**
- **No inter-site failover**
- **Need for scripts at each site to coordinate with each other while mirroring the data**
- **Lots of telecons, emails and frustrated sys-admins**
- + **Autonomous operation of data centers within each funding agency (or country) to satisfy their tax payer's dollars / euros / ...**



**If we used a traditional single file
system namespace (WAFS)...**



Mapping physical data to logical view



Hierarchical view,
independent of
network, disk, sector,
track, fragments

Rule : Storage
Abstraction – Hide
storage resources



Is logical data namespace sufficient?

- Logical Namespace hides all the heterogeneities and all data seem to be on a single resource.
- However, users are hidden from the advantages of internet, distributed data management etc
- How do we handle replication of data between storage resources? (There is no concept of “resource” or “storage space” in logical data namespace)
- Makes the file system a black-box as though only FS admins know the internet and replication



Possible solution (1)

- Within the logical data namespace, we need a way to indicate the presence of distributed resources (from different organizations)
- Why not add an attribute to every data item to specify where the data is physically present



Logical namespace

C:\Docs\SRBgridWebservices

File Edit View Favorites Tools Help

Back Forward Stop Search Folders

Address C:\Docs\SRBgridWebservices Go

Folders	Name	Size	Type	Date Modified
	DGMSv1-bck.doc	55 KB	Microsoft Word Docu...	8/2/2002 4:55 PM
	DGMSv2.doc	151 KB	Microsoft Word Docu...	8/9/2002 11:26 PM
	DGMSv3.doc	133 KB	Microsoft Word Docu...	8/12/2002 12:00 AM
	DGMSv4.doc	137 KB	Microsoft Word Docu...	10/29/2002 3:33 PM
	DGMSv4-abstract.doc	44 KB	Microsoft Word Docu...	11/1/2002 12:16 PM
	DGMSv5.doc	166 KB	Microsoft Word Docu...	11/6/2002 1:42 PM
	DGMSv5-abstract.doc	46 KB	Microsoft Word Docu...	11/1/2002 4:06 PM
	DGMSv6.doc	140 KB	Microsoft Word Docu...	11/16/2002 8:36 PM
	DGMSv7.doc	108 KB	Microsoft Word Docu...	11/7/2002 5:30 PM
	DGMSv7nwm.doc	125 KB	Microsoft Word Docu...	11/9/2002 12:33 AM
	DGMSv8.doc	90 KB	Microsoft Word Docu...	11/16/2002 2:29 AM
	DGMSv9.doc	117 KB	Microsoft Word Docu...	8/9/2004 12:13 PM
	DGMS-VLDB-finalPresent.ppt	2,619 KB	Microsoft PowerPoint ...	7/23/2003 2:36 PM
	DGMS-VLDB-finalStraw.ppt	2,620 KB	Microsoft PowerPoint ...	7/22/2003 12:23 AM
	DGMS-VLDB-present.ppt	3,146 KB	Microsoft PowerPoint ...	7/9/2003 2:56 PM
	EMW-Phil-CFP.doc	23 KB	Microsoft Word Docu...	5/29/2004 10:50 PM
	Enterprise Data Grids for IT Executives.doc	25 KB	Microsoft Word Docu...	7/26/2004 3:26 PM
	FileWebServiceChap14.pdf	428 KB	Adobe Acrobat 7.0 D...	2/11/2002 1:46 PM
	FromRaja.doc	30 KB	Microsoft Word Docu...	9/30/2002 12:50 PM
	GFS Architecture.ppt	103 KB	Microsoft PowerPoint ...	10/6/2003 2:12 PM
	GFS session 2.doc	24 KB	Microsoft Word Docu...	3/11/2004 9:46 AM
	gfs-ggf10-intro.ppt	148 KB	Microsoft PowerPoint ...	3/10/2004 8:14 PM
	gfs-ggf10-intro-day2.ppt	54 KB	Microsoft PowerPoint ...	3/11/2004 10:43 AM
	GFS-WG-Proposal-2.doc	113 KB	Microsoft Word Docu...	8/19/2003 2:11 PM
	GFS-WG-Proposal-3.rtf	97 KB	Rich Text Format	8/19/2003 11:10 PM
	GFS-WG-Proposal.rtf	53 KB	Rich Text Format	8/5/2003 3:25 PM
	GGF9-GFS-introv1.ppt	173 KB	Microsoft PowerPoint ...	10/7/2003 3:38 PM
	GGF9-GFS-Posix-Soap.ppt	307 KB	Microsoft PowerPoint ...	10/7/2003 4:49 PM
	GGF10 Invite Letter_Arun swaran Jagath...	119 KB	Adobe Acrobat 7.0 D...	3/4/2004 6:48 AM
	GGF10Minutes.doc	30 KB	Microsoft Word Docu...	3/30/2004 3:27 PM
	GGF12-DGMS4CTO_tutorial_proposal.doc	59 KB	Microsoft Word Docu...	7/15/2004 5:12 PM
	GGF12-SRB_tutorial_proposal.doc	60 KB	Microsoft Word Docu...	7/15/2004 4:38 PM
	GGF_GFS-preProposal.doc	29 KB	Microsoft Word Docu...	8/3/2003 8:39 PM
	GGF-DGMS-RG.doc	40 KB	Microsoft Word Docu...	9/3/2003 7:40 PM
	GGFWS - SDSC Matrix Project.doc	27 KB	Microsoft Word Docu...	1/24/2004 12:42 PM
	got datagrid1.doc	174 KB	Microsoft Word Docu...	3/6/2002 10:58 PM
	got datagrid.doc	32 KB	Microsoft Word Docu...	3/6/2002 12:17 AM
	Grid Data Services comb.ppt	350 KB	Microsoft PowerPoint ...	6/22/2003 5:53 AM

456 objects (plus 22 hidden) (Disk free space: 33.8 GB) 215 MB My Computer

start Volume ... Inbox - ... Microsof... artee Na... iTunes C:\Docs... On-Scre... Docume... 10:14 PM

Logical namespace + location

C:\Docs\SRBgridWebservices

File Edit View Favorites Tools Help

Back Forward Stop Search Folders

Address C:\Docs\SRBgridWebservices Go

Folders	Name	Size	Type	Date Modified
NadarSoftv	DGMSv1-bck.doc	55 KB	Microsoft Word Docu...	8/2/2002 4:55 PM
NetIP	DGMSv2.doc	151 KB	Microsoft Word Docu...	8/9/2002 11:26 PM
NSF	DGMSv3.doc	133 KB	Microsoft Word Docu...	8/12/2002 12:00 AM
NVO	DGMSv4.doc	137 KB	Microsoft Word Docu...	10/29/2002 3:33 PM
OGSA	DGMSv4-abstract.doc	44 KB	Microsoft Word Docu...	11/1/2002 12:16 PM
OGSA-DAI	DGMSv5.doc	166 KB	Microsoft Word Docu...	11/6/2002 1:42 PM
Oracle	DGMSv5-abstract.doc	46 KB	Microsoft Word Docu...	11/1/2002 4:06 PM
papers	DGMSv6.doc	140 KB	Microsoft Word Docu...	11/16/2002 8:36 PM
PersistentA	DGMSv7.doc	108 KB	Microsoft Word Docu...	11/7/2002 5:30 PM
Personal	DGMSv7nwm.doc	125 KB	Microsoft Word Docu...	11/9/2002 12:33 AM
ppdg	DGMSv8.doc	90 KB	Microsoft Word Docu...	11/16/2002 2:29 AM
Presentatic	DGMSv9.doc	117 KB	Microsoft Word Docu...	8/9/2004 12:13 PM
proposals	DGMS-VLDB-finalPresent.ppt	2,619 KB	Microsoft PowerPoint ...	7/23/2003 2:36 PM
QueryProc	DGMS-VLDB-finalStraw.ppt	2,620 KB	Microsoft PowerPoint ...	7/22/2003 12:23 AM
Reimburse	DGMS-VLDB-present.ppt	3,146 KB	Microsoft PowerPoint ...	7/9/2003 2:56 PM
Reliability	EMW-Phil-CFP.doc	23 KB	Microsoft Word Docu...	5/29/2004 10:50 PM
Religion	Enterprise Data Grids for IT Executives.doc	25 KB	Microsoft Word Docu...	7/26/2004 3:26 PM
Review	FileWebServiceChap14.pdf	428 KB	Adobe Acrobat 7.0 D...	2/11/2002 1:46 PM
RoomSear	FromRaja.doc	30 KB	Microsoft Word Docu...	9/30/2002 12:50 PM
Sangam	GFS Architecture.ppt	103 KB	Microsoft PowerPoint ...	10/6/2003 2:12 PM
SBIR	GFS session 2.doc	24 KB	Microsoft Word Docu...	3/11/2004 9:46 AM
SCEC	gfs-ggf10-intro.ppt	148 KB	Microsoft PowerPoint ...	3/10/2004 8:14 PM
SDM Cente	gfs-ggf10-intro-day2.ppt	54 KB	Microsoft PowerPoint ...	3/11/2004 10:43 AM
SDSC	GFS-WG-Proposal-2.doc	113 KB	Microsoft Word Docu...	8/19/2003 2:11 PM
SDSCLette	GFS-WG-Proposal-3.rtf	97 KB	Rich Text Format	8/19/2003 11:10 PM
SDSC-Resu	GFS-WG-Proposal.rtf	53 KB	Rich Text Format	8/5/2003 3:25 PM
SDSS	GGF9-GFS-introv1.ppt	173 KB	Microsoft PowerPoint ...	10/7/2003 3:38 PM
SNW	GGF9-GFS-Posix-Soap.ppt	307 KB	Microsoft PowerPoint ...	10/7/2003 4:49 PM
SoapMessa	GGF10 Invite Letter_Arun swaran Jagath...	119 KB	Adobe Acrobat 7.0 D...	3/4/2004 6:48 AM
Spatial	GGF10Minutes.doc	30 KB	Microsoft Word Docu...	3/30/2004 3:27 PM
SRBgridWe	GGF12-DGMS4CTO_tutorial_proposal.doc	59 KB	Microsoft Word Docu...	7/15/2004 5:12 PM
SRBgridWe	GGF12-SRB_tutorial_proposal.doc	60 KB	Microsoft Word Docu...	7/15/2004 4:38 PM
	GGF_GFS-preProposal.doc	29 KB	Microsoft Word Docu...	8/3/2003 8:39 PM
	GGF-DGMS-RG.doc	40 KB	Microsoft Word Docu...	9/3/2003 7:40 PM
	GGFWS - SDSC Matrix Project.doc	27 KB	Microsoft Word Docu...	1/24/2004 12:42 PM
	got datagrid1.doc	174 KB	Microsoft Word Docu...	3/6/2002 10:58 PM
	got datagrid.doc	32 KB	Microsoft Word Docu...	3/6/2002 12:17 AM
	Grid Data Services comb.ppt	350 KB	Microsoft PowerPoint ...	6/22/2003 5:53 AM

456 objects (plus 22 hidden) (Disk free space: 33.8 GB) 215 MB My Computer

start Volume ... Inbox - ... Microsof... artee Na... iTunes C:\Docs... On-Scre... Docume...

10:14 PM

Location

Add location info??

//121...121/a/file1.txt

//sandiego/a/file1.txt

//India123/a/file1.txt

Possible Solution (2)

- But, providing physical information in logical data namespace might not be a good solution. All advantages due to “infrastructure-independence” will be lost.
- Imagine the Internet without “www” and only physical IP addresses.
- Also, users need a way to specify “human readable” resource names (example: `SandiegoDisk`, `FastestIO`, `capacityStorage3` etc)



Logical namespace + location

C:\Docs\SRBgridWebservices

File Edit View Favorites Tools Help

Back Search Folders

Address C:\Docs\SRBgridWebservices Go

Folders	Name	Size	Type	Date Modified
	DGMSv1-bck.doc	55 KB	Microsoft Word Docu...	8/2/2002 4:55 PM
	DGMSv2.doc	151 KB	Microsoft Word Docu...	8/9/2002 11:26 PM
	DGMSv3.doc	133 KB	Microsoft Word Docu...	8/12/2002 12:00 AM
	DGMSv4.doc	137 KB	Microsoft Word Docu...	10/29/2002 3:33 PM
	DGMSv4-abstract.doc	44 KB	Microsoft Word Docu...	11/1/2002 12:16 PM
	DGMSv5.doc	166 KB	Microsoft Word Docu...	11/6/2002 1:42 PM
	DGMSv5-abstract.doc	46 KB	Microsoft Word Docu...	11/1/2002 4:06 PM
	DGMSv6.doc	140 KB	Microsoft Word Docu...	11/16/2002 8:36 PM
	DGMSv7.doc	108 KB	Microsoft Word Docu...	11/7/2002 5:30 PM
	DGMSv7nwm.doc	125 KB	Microsoft Word Docu...	11/9/2002 12:33 AM
	DGMSv8.doc	90 KB	Microsoft Word Docu...	11/16/2002 2:29 AM
	DGMSv9.doc	117 KB	Microsoft Word Docu...	8/9/2004 12:13 PM
	DGMS-VLDB-finalPresent.ppt	2,619 KB	Microsoft PowerPoint ...	7/23/2003 2:36 PM
	DGMS-VLDB-finalStraw.ppt	2,620 KB	Microsoft PowerPoint ...	7/22/2003 12:23 AM
	DGMS-VLDB-present.ppt	3,146 KB	Microsoft PowerPoint ...	7/9/2003 2:56 PM
	EMW-Phil-CFP.doc	23 KB	Microsoft Word Docu...	5/29/2004 10:50 PM
	Enterprise Data Grids for IT Executives.doc	25 KB	Microsoft Word Docu...	7/26/2004 3:26 PM
	FileWebServiceChap14.pdf	428 KB	Adobe Acrobat 7.0 D...	2/11/2002 1:46 PM
	FromRaja.doc	30 KB	Microsoft Word Docu...	9/30/2002 12:50 PM
	GFS Architecture.ppt	103 KB	Microsoft PowerPoint ...	10/6/2003 2:12 PM
	GFS session 2.doc	24 KB	Microsoft Word Docu...	3/11/2004 9:46 AM
	gfs-ggf10-intro.ppt	148 KB	Microsoft PowerPoint ...	3/10/2004 8:14 PM
	gfs-ggf10-intro-day2.ppt	54 KB	Microsoft PowerPoint ...	3/11/2004 10:43 AM
	GFS-WG-Proposal-2.doc	113 KB	Microsoft Word Docu...	8/19/2003 2:11 PM
	GFS-WG-Proposal-3.rtf	97 KB	Rich Text Format	8/19/2003 11:10 PM
	GFS-WG-Proposal.rtf	53 KB	Rich Text Format	8/5/2003 3:25 PM
	GGF9-GFS-introv1.ppt	173 KB	Microsoft PowerPoint ...	10/7/2003 3:38 PM
	GGF9-GFS-Posix-Soap.ppt	307 KB	Microsoft PowerPoint ...	10/7/2003 4:49 PM
	GGF10 Invite Letter_Arun swaran Jagath...	119 KB	Adobe Acrobat 7.0 D...	3/4/2004 6:48 AM
	GGF10Minutes.doc	30 KB	Microsoft Word Docu...	3/30/2004 3:27 PM
	GGF12-DGMS4CTO_tutorial_proposal.doc	59 KB	Microsoft Word Docu...	7/15/2004 5:12 PM
	GGF12-SRB_tutorial_proposal.doc	60 KB	Microsoft Word Docu...	7/15/2004 4:38 PM
	GGF_GFS-preProposal.doc	29 KB	Microsoft Word Docu...	8/3/2003 8:39 PM
	GGF-DGMS-RG.doc	40 KB	Microsoft Word Docu...	9/3/2003 7:40 PM
	GGFWS - SDSC Matrix Project.doc	27 KB	Microsoft Word Docu...	1/24/2004 12:42 PM
	got datagrid1.doc	174 KB	Microsoft Word Docu...	3/6/2002 10:58 PM
	got datagrid.doc	32 KB	Microsoft Word Docu...	3/6/2002 12:17 AM
	Grid Data Services comb.ppt	350 KB	Microsoft PowerPoint ...	6/22/2003 5:53 AM

456 objects (plus 22 hidden) (Disk free space: 33.8 GB) 215 MB My Computer

start Volume ... Inbox - ... Microsof... artee Na... iTunes C:\Docs... On-Scre... Docume...

10:14 PM

Logical Location

Add logical location??

Default resource

San diego sdsc-disk

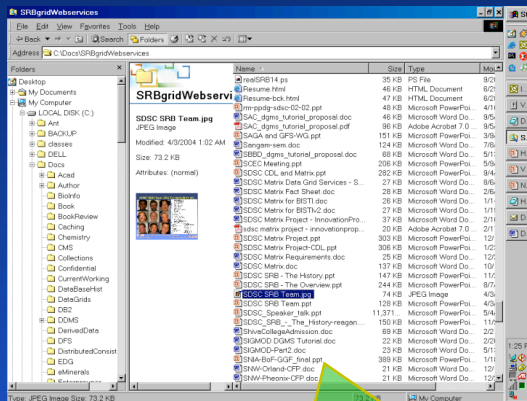
India-delhi-tape

Possible Solution (3)

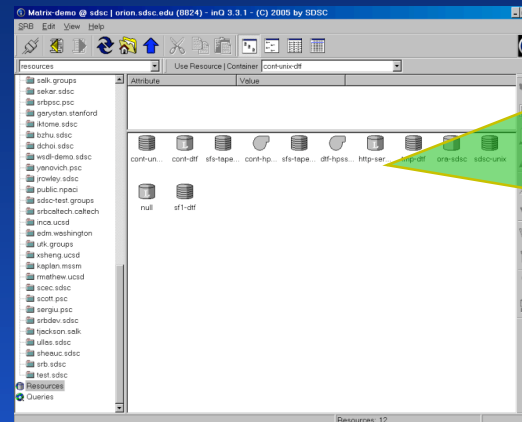
- *“Houston, We have a solution!”*
- We will create a separate (human readable) logical resource namespace of all data resources in the data grid
- The well-known logical file namespace can be “joined” with this logical resource namespace to create this new file-name space for our Wide-Area Collaborative FileSystem



The solution (or a part of the solution)



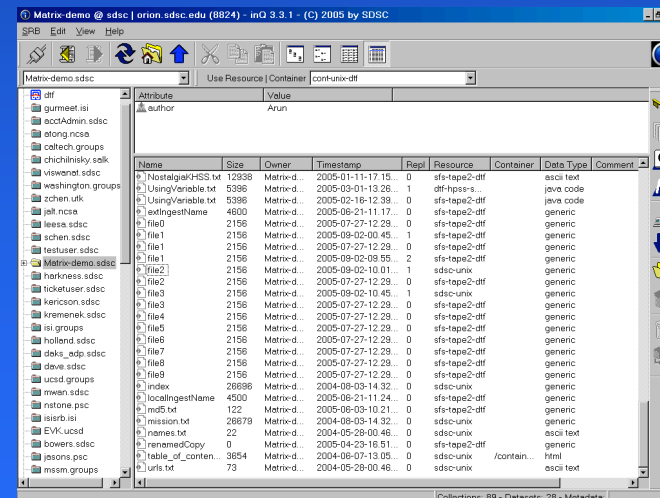
+



Logical Resource Namespace (Each logical resource is a combination of one or more physical resources)

Traditional Logical Namespace

=



Data Grid Namespace



iRODS.org

26

San Diego Supercomputer Center



~~Cloud~~ Data Grid Namespace

testuser @ sdsc | srb.sdsc.edu (8544) - inQ 3.4.1 - (C) 2006 by SDSC

SRB Edit View Help

Use Resource | Container unix-sdsc

Logical Resources

Attribute	Value
author	test
color	green
chip-part	Q-SIC-2907

Name	Size	Owner	Timestamp	Repl	Resource	Container	Data Ty...	Comment
data_1.unix-sdsc	2792	testuser	2003-09-30-09....	1	unix-sdsc		text	
data_1.unix-sdsc	2792	testuser	2003-09-30-09....	0	unix-sdsc		text	
data_2.unix-sdsc	5584	testuser	2003-09-30-09....	1	unix-sdsc		text	
data_2.unix-sdsc	5584	testuser	2003-09-30-09....	0	unix-sdsc		text	
data_3.unix-sdsc	8376	testuser	2003-09-30-09....	0	unix-sdsc		text	
data_3.unix-sdsc	8376	testuser	2003-09-30-09....	1	unix-sdsc		text	
data_4.unix-sdsc	11168	testuser	2003-09-30-09....	1	unix-sdsc		text	
data_4.unix-sdsc	11168	testuser	2003-09-30-09....	0	unix-sdsc		text	
datadb1.xml.unix-sdsc	5147	testuser	2003-09-30-09....	1	unix-sdsc		text	
datadb1.xml.unix-sdsc	5147	testuser	2003-09-30-09....	0	unix-sdsc		text	
dataempty.unix-sdsc	0	testuser	2003-09-30-09....	0	unix-sdsc		text	
dataempty.unix-sdsc	0	testuser	2003-09-30-09....	1	unix-sdsc		text	
arun_MB2.JPG	40358	testuser	2006-12-05-11....	0	sfs-disk-demo		generic	
arun_MB2.JPG	40358	testuser	2006-12-05-11....	1	unix-sdsc		generic	

Multiple Replicas

Users from different organizations

replicated arun_MB2.JPG

Datasets: 14 - Metadata: 3 - Users: 1

start

2:43 PM

myGlobal500.com (single business entity)

Physical Map of the World, June 2003



myMovieStudio.com (collaboration of businesses)

Physical Map of the World, June 2003



iRODS.org

29

San Diego Supercomputer Center



myFreeBackup.com (1 man IT-army)



Agenda

- Intro
- Problem Motivation: LSST and myGlobal500.com
- Solution Path
- Implementation: iRODS
- Improvements Possible / Future Plans



iRODS

- **Client-server architecture**

- Clients implement our protocol in your favorite programming language

- **Servers**

- Use an hybrid Model of P2P and centralized concepts
 - Servers have storage associated with them (in the form of directories on host/physical file system)
 - One (or more) servers connect to a database that stores the file metadata

-



Three Tier Architecture

- **Client Implementations (Any interface/API)**
 - Your preferred access mechanism
- **Servers (iRODS Server)**
 - Federated to support direct interactions between servers
 - Metadata catalog (iCAT)
 - Separation of metadata management from data storage
 - State persistence using a database
- **Storage Resources (iRODS drivers)**
 - Your storage device with our iRODS driver
 - Just some basic 16 POSIX-like operations
 - Can also work with your favorite storage vendor



Application

C, C++, Libraries	Linux I/O	Unix Shell	Java, NT Browsers	DLL / Python	GridFTP	OAI WSDL
----------------------	--------------	---------------	----------------------	-----------------	---------	-------------

Access
APIs

Consistency Management / Authorization-Authentication

Logical Name Space	Latency Management	Data Transport	Metadata Transport
-----------------------	-----------------------	-------------------	-----------------------

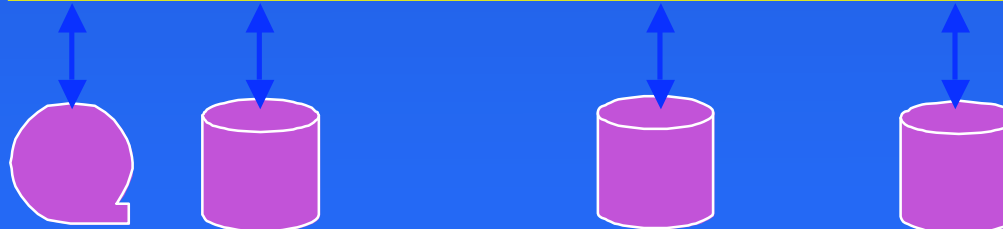
iRODS
Server

Catalog Abstraction

Storage Abstraction

Databases DB2, Oracle, Sybase, SQLServer	Archives HPSS, ADMS, UniTree, DMF	HRM	File Systems Unix, NT, Mac OSX	Databases DB2, Oracle, Postgres
--	---	-----	--------------------------------------	---------------------------------------

Storage
Drivers



iRODS.org

34

San Diego Supercomputer Center

SDSC

Peer-2-peer (like) iRODS servers

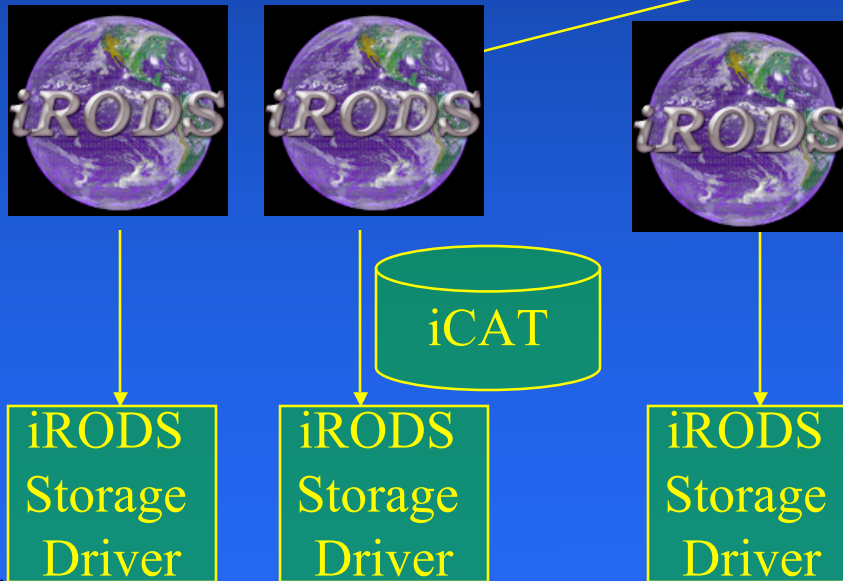


Client can connect to
any distributed
iRODS server

iCAT-Enabled Server

An iRODS zone

The role of iCAT and lack of
leader election protocol does
not make the servers fully
P2P



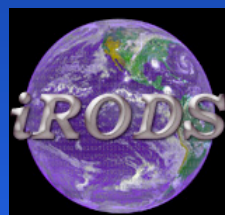
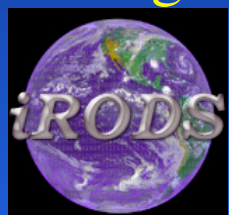
Basic get explained



San Diego

SFO

Boston



iRODS
Storage
Driver

iRODS
Storage
Driver



iRODS
Storage
Driver

1. Check Auth (Logon-server connects to iCAT server)
2. Find optimal copy of the file for that particular client request (uses simple heuristics)
3. Decide on a data path option, number of threads, bandwidth etc
4. Send the data (failover to replica automatically)



iRODS.org

36

San Diego Supercomputer Center

SDSC

Basic put explained (with iRules - Trigger_like)



San Diego



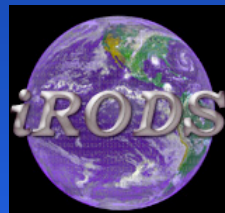
iRODS
Storage
Driver

SFO



iRODS
Storage
Driver

Boston



iRODS
Storage
Driver



1. Check Auth
2. <pre_process>
3. Decide on a data path option, number of threads, bandwidth etc
4. [sink the data (failover to replica resource automatically)]
5. <post_process>
6. <error_recovery>



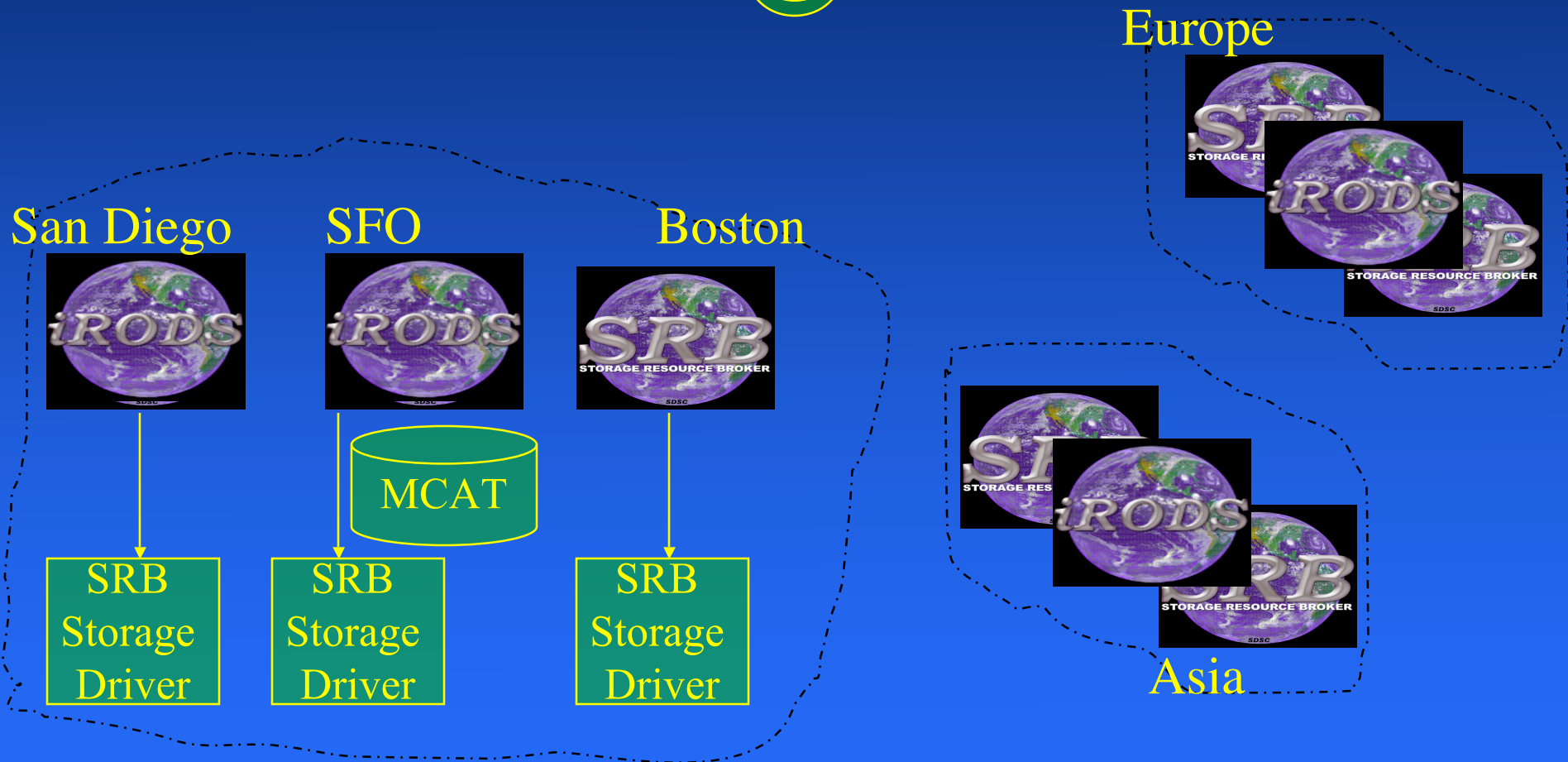
iRODS.org

37

San Diego Supercomputer Center

SDSC

Peer-2-peer SRB Zones



iRODS.org

Supercomputing 2008 Storage Challenge

- **Sites and hardware**

- IN2P3 (France), ROE/U.Edinburgh (UK), NCSA, SDSC, Chile and (RENCI).
- Multiple hardware and heterogeneous environments
- Different capabilities and roles at each site.
 - Telescope, Base, Data Archival Center, Data Access Centers, User Centers



Simulation of the pipeline

- **Pipeline processing of images**
 - Data from telescope (IN2P3) ingested into iRODS resource
 - Images automatically replicated into Base at UK (iRules)
 - ImageSubtract Pipeline process started by iRODS software itself at Base after each Image exposure is replicated
 - Data again replicated to NCSA - Archival center
 - More detailed ImageSubtract pipeline
 - Selective replication of data to data access center
- **Data-lifecycle in Action (remember its Infrastructure Software)**
 - Rules or policies managing data pipelines, replication
 - Files have the same file name everywhere on this single confluence of systems spanning HPC, data delivery, archives
 - All files at different sites and stages of lifecycle available through the LSST infrastructure software



Performance & Scalability (GREEN TESTS)

- **MAX number of files**
 - 9.2 quintillion (billion times billion)
 - LSST will have to have an ingest rate of little more than 30 billion files/second to reach MAX count in our infrastructure software
- **MAX File Size for one file (NOT TESTED)**
 - 1 Exabyte (if you have a file system that can store it and bandwidth to transfer it)
- **MAX File System size for WHOLE system (NOT TESTED)**
 - 9.2 quintillion exabytes (10^{36}) or undecillion bytes
 - Considering replicas also it will be just over one hundredth of quindecillion bytes (10^{47}) bytes (way smaller than a googol)
 - Considering replicas and also their versions, its Infinite (sorry could not calculate permutations of a 250 char array)
- **MAX number of files in a directory (collection)**
 - 9.2 quintillion
- **MAX number of storage resources and sites**
 - 9.2 quintillion



The QUINTILLION MARK (GREEN WAY)

```
srbbrick15:/data1/LSST-SC08/V4-stressTest/iRODS % iput -R quintillion+ Makefile  
countown6
```

```
srbbrick15:/data1/LSST-SC08/V4-stressTest/iRODS % iput -R quintillion+ Makefile  
countown5
```

```
srbbrick15:/data1/LSST-SC08/V4-stressTest/iRODS % iput -R quintillion+ Makefile  
countown4
```

```
srbbrick15:/data1/LSST-SC08/V4-stressTest/iRODS % iput -R quintillion+ Makefile  
countown3
```

```
srbbrick15:/data1/LSST-SC08/V4-stressTest/iRODS % iput -R quintillion+ Makefile  
countown2
```

```
srbbrick15:/data1/LSST-SC08/V4-stressTest/iRODS % iput -R quintillion+ Makefile  
countown1
```

```
srbbrick15:/data1/LSST-SC08/V4-stressTest/iRODS % iput -R quintillion+ Makefile  
countown0
```

```
ERROR: putUtil: put error for /LSSTzone/home/rods/quintillion/countown0, status  
= -806000 status = -806000 CAT_SQL_ERR
```



The QUINTILLIONth storage resource

srbbrick15:/data1/LSST-SC08/V4-stressTest/iRODS/% **ilsresc -l**

resource name: **quintillion+**

resc id: **9223372036854775800**

zone: LSSTzone

type: **unix file system**

class: **cache**

location: **srbbrick15.sdsc.edu**

vault: /data1/LSST-SC08/V4-stressTest/iRODS/vault2

free space:

info:

comment:

create time: 01226098441: 2008-11-07.14:54:01

modify time: 01226098441: 2008-11-07.14:54:01





Performance (1000 file ingests)

- **End-to-end (put,client-read, transfer, server-write, ack recv)**
 - Ranges from 5 to 8 seconds
- **Just raw file system speed on same hardware configuration (used copy)**
 - Exactly 3 seconds
- **Is it ok to conclude that the infrastructure software alone (on a local system)**
 - 2 to 5 seconds

No bulk operations or asynchronous operations were used to speed up the response of the infrastructure software

Hardware used is not “performance-oriented”



Agenda

- Intro
- Problem Motivation: LSST and myGlobal500.com
- Solution Path
- Implementation: iRODS
- Improvements Possible / Future Plans



Problems and Improvements

- **Metadata lookup (our implementation uses DB)**
 - In Memory
 - No SQL?
- **Sites going down or delay**
 - Coda Concepts for data grids
 - Can a large file system network operate even when one or more sites are down
 - Can the telescope handle its own failure when it can not connect to any other site
- **iRODS.org for non-science users**
 - Open source software
 - Big companies : alternative to CDN? Alternative to expensive h/w ?



Acknowledgements

- **iRODS**
 - DICE team at UCSD
 - DICE team at UNC
- **LSST**
- **SC2009**
 - CC-IN2P3, France
 - iRODS.org
 - LSST.org
 - NCSA/UIUC
 - NOAO, Chile
 - RENC/UNC
 - SDSC/UCSD
 - University of Edinburgh (EPCC, ROE), UK



Summary

- File systems for collaboration (not just large-scale data)
- Logical file namespace or file-tree along with logical human-readable resource names
- iRODS.org
- Near Future: More work and collaborations



Xtremely Large File Systems for the small collaborative world

Arun Jagatheesan

**San Diego Supercomputer Center
University of California at San Diego**

&

**DiceResearch.org
(arun@sdsc.edu)**

HPTS Workshop

Asilomar, California, 28 October 2009



iRODS.org

San Diego Supercomputer Center

