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

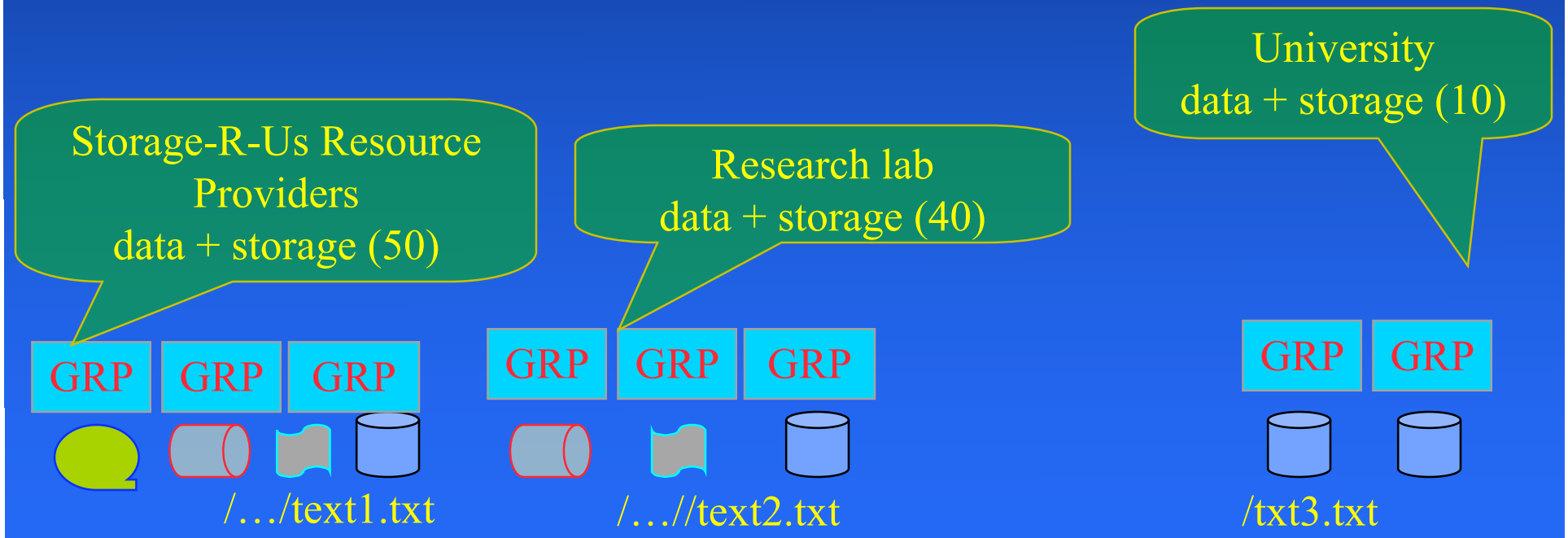
Asilomar, California, 28 October 2009

Agenda

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

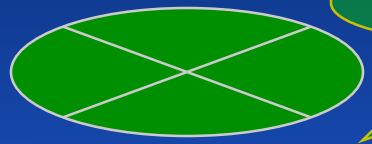
Slides from HPTS 2005

Data Grid Administrative domains



Data Grid: Logical view of data & resources

/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)

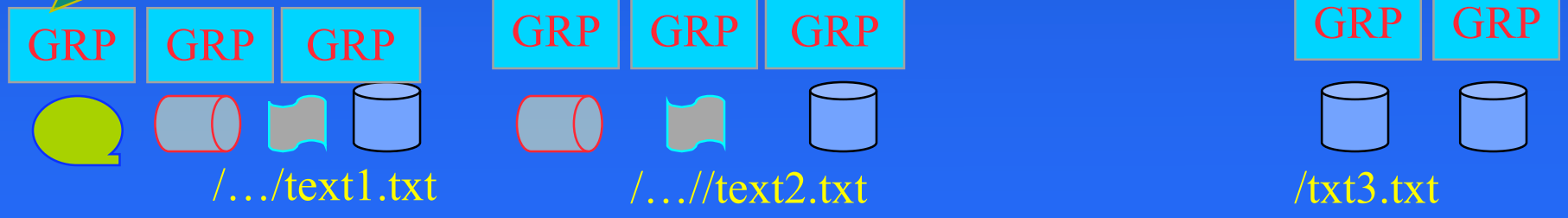


Logical Namespace (Need not be same as physical view of resources)

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

Research Lab
data + storage (40)

University
data + storage (10)



/…/text1.txt

/…//text2.txt

/txt3.txt

Data Grid Language

- **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

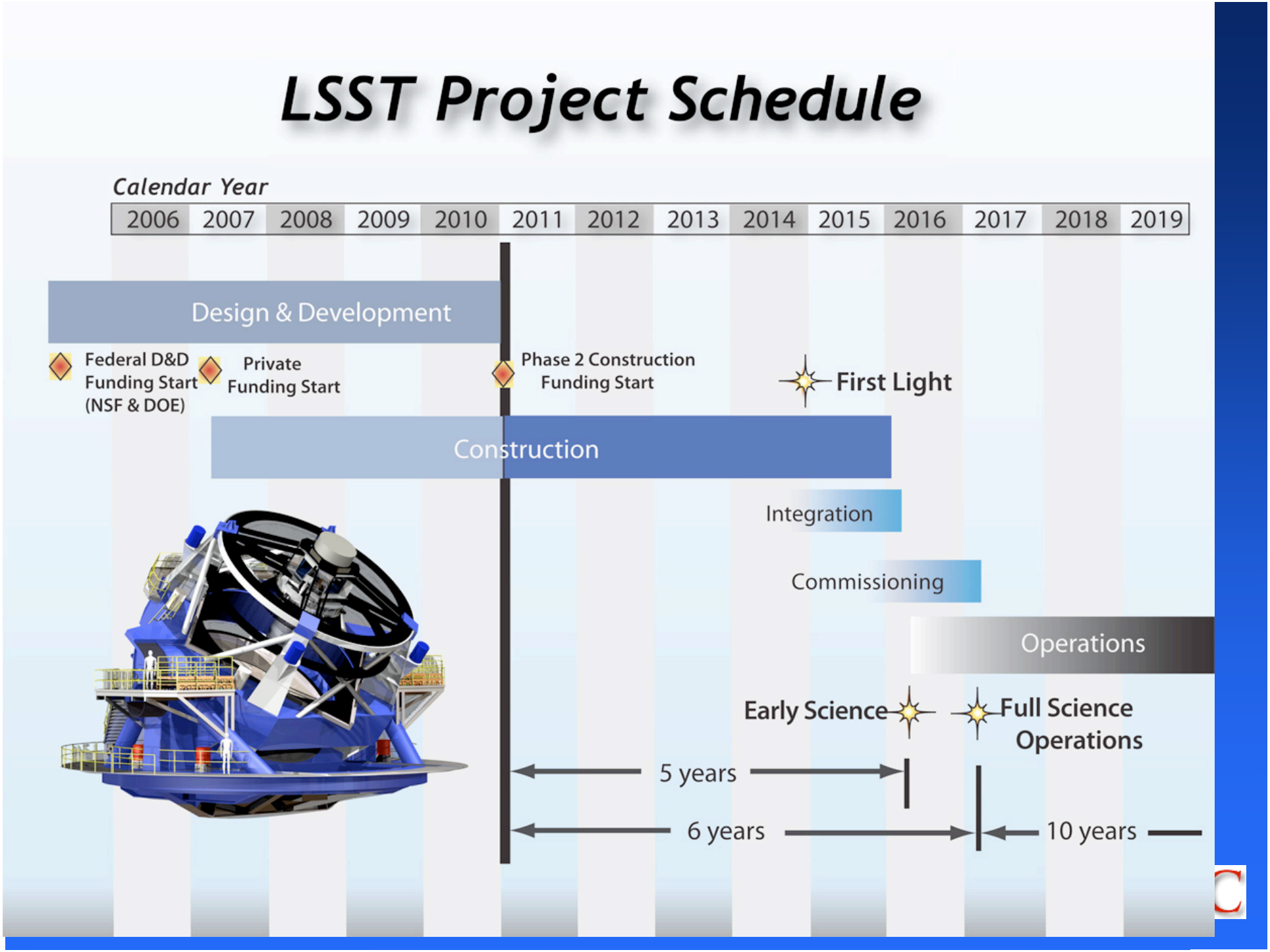
Agenda

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

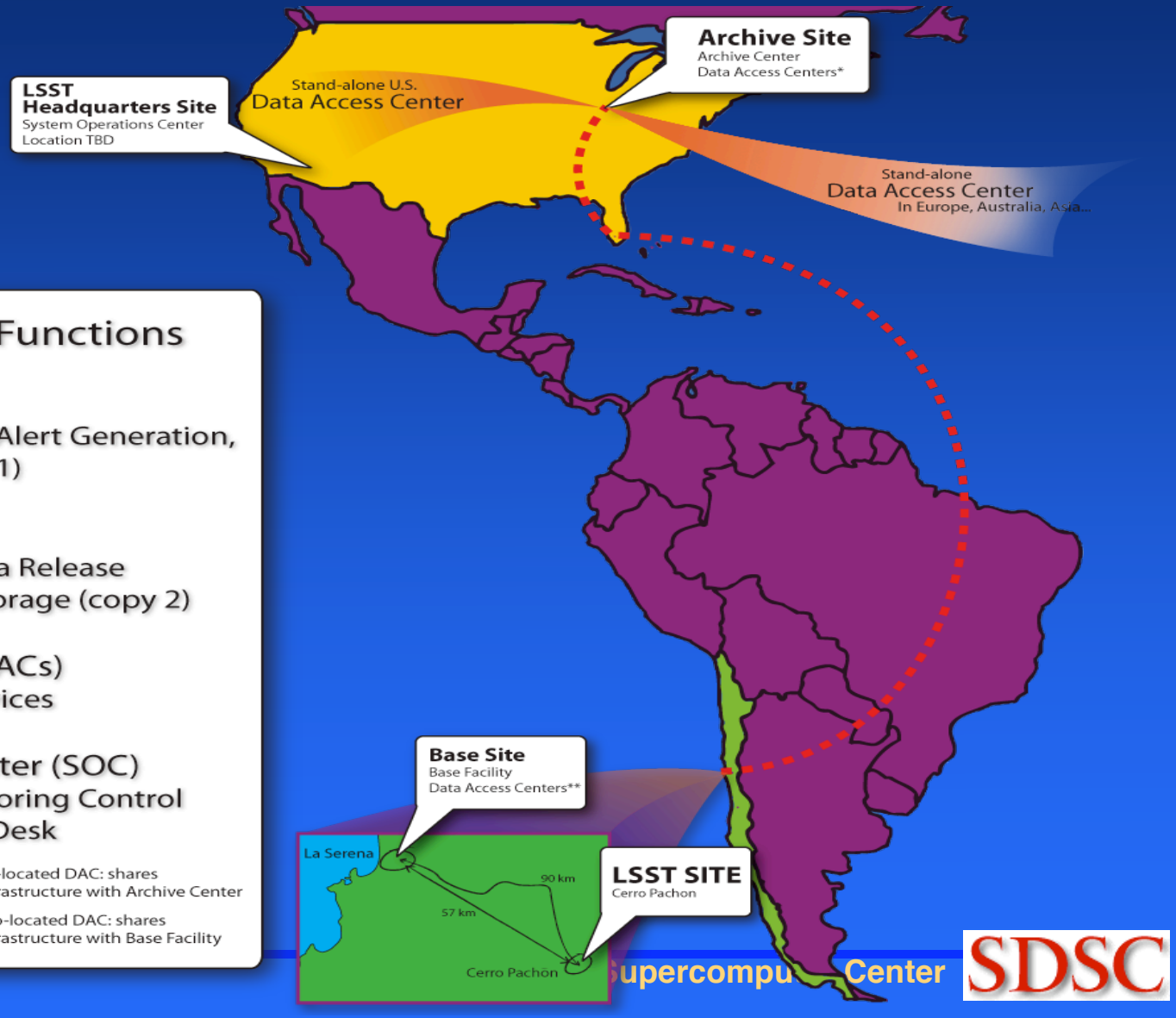
LSST

- **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

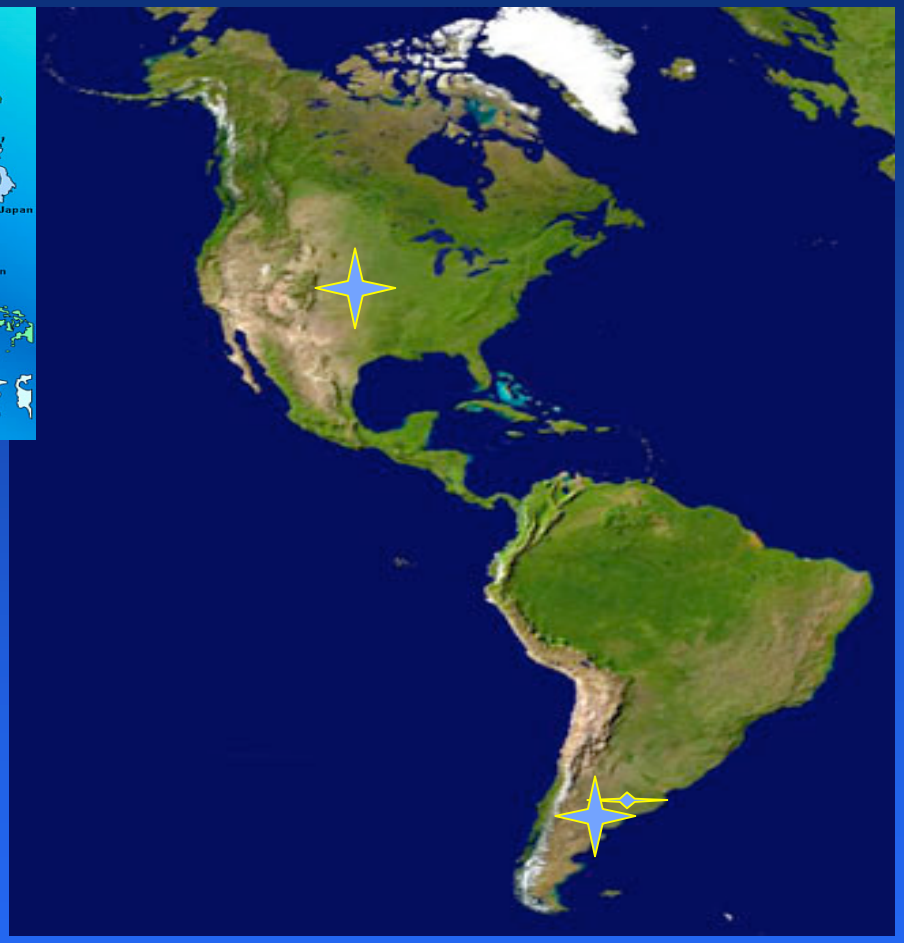
** Co-located DAC: shares infrastructure with Base Facility

iRODS.org

Supercompu Center SDSC

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

LSST sites growing…



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

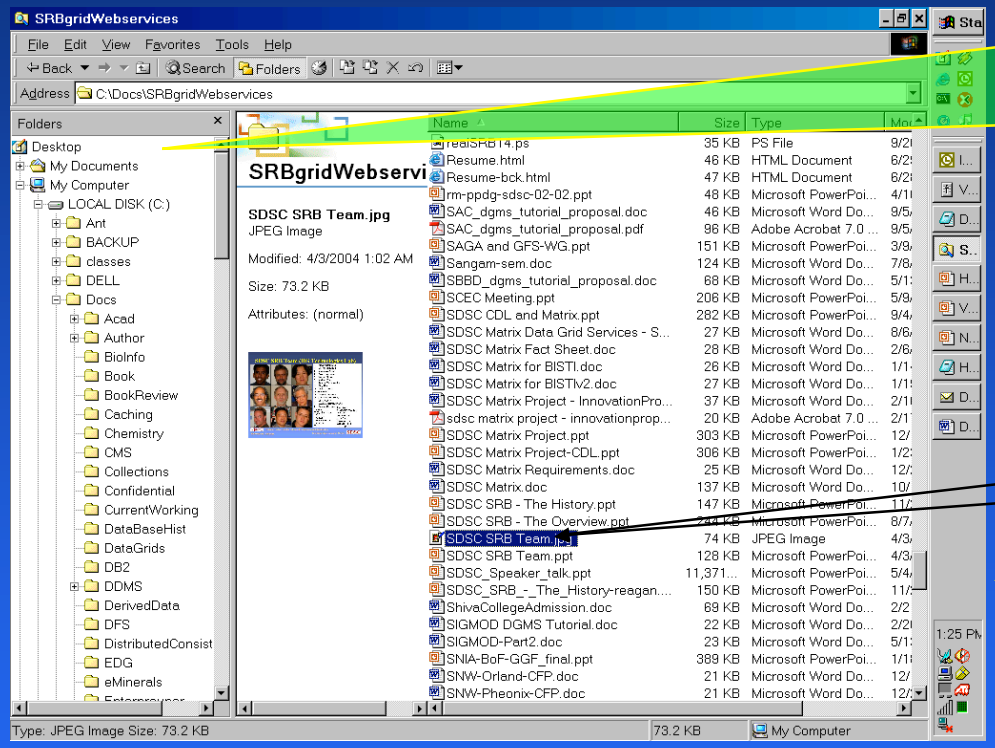
/exp/file1.fits
/exp/file2.fits

Disadvantages of separate file systems

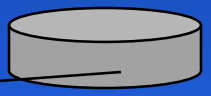
- — Increased cost of operation
  - — Storage cost for data backups (in petabytes)
- — No load sharing
- — No inter-site failover
- — Need for scripts at each site to coordinate with each other while mirroring the data
- — Lots of telecons, emails and frustrated sys-admins
- \+ Autonomous operation of 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

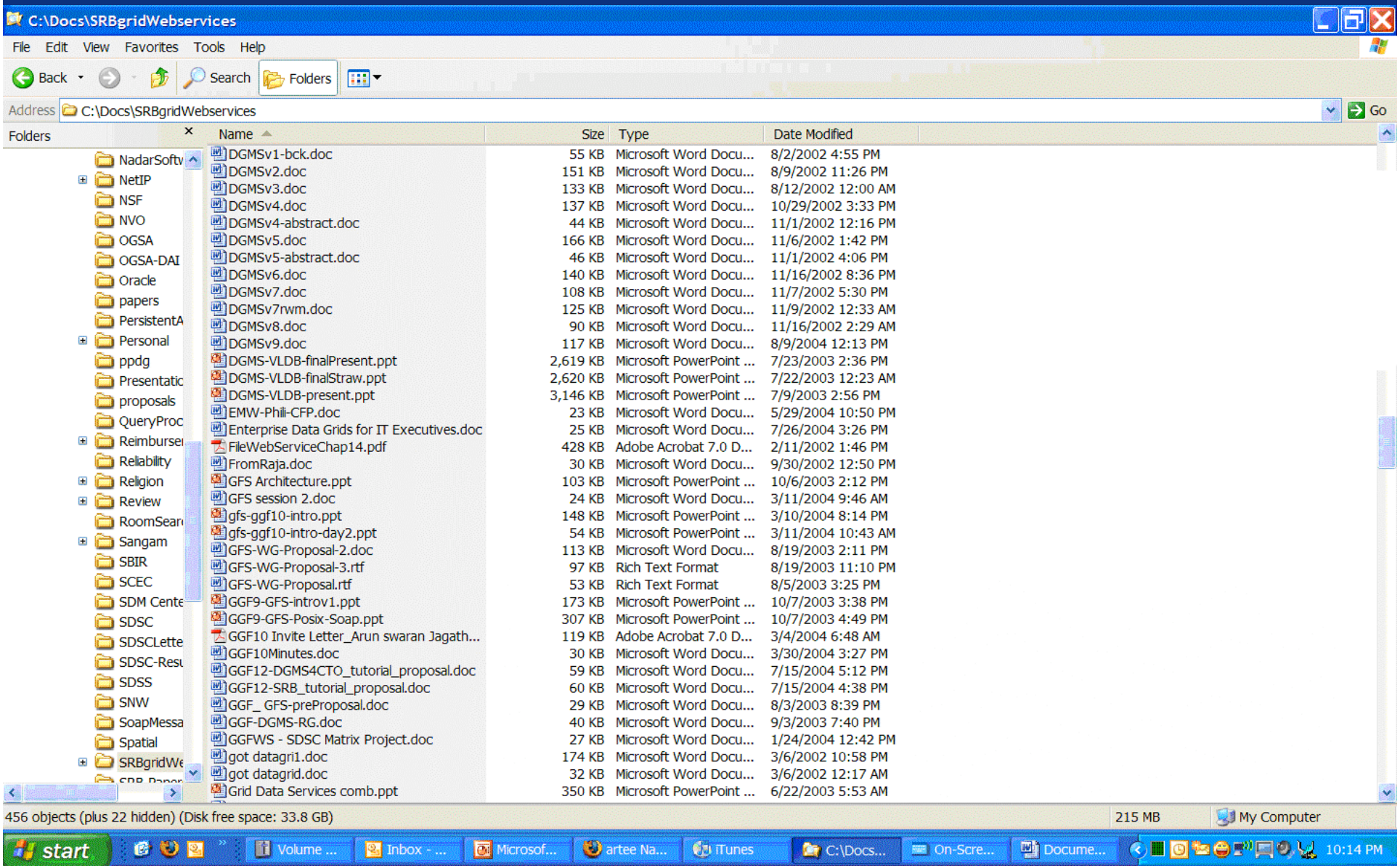
Is logical data namespace sufficient?

- 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

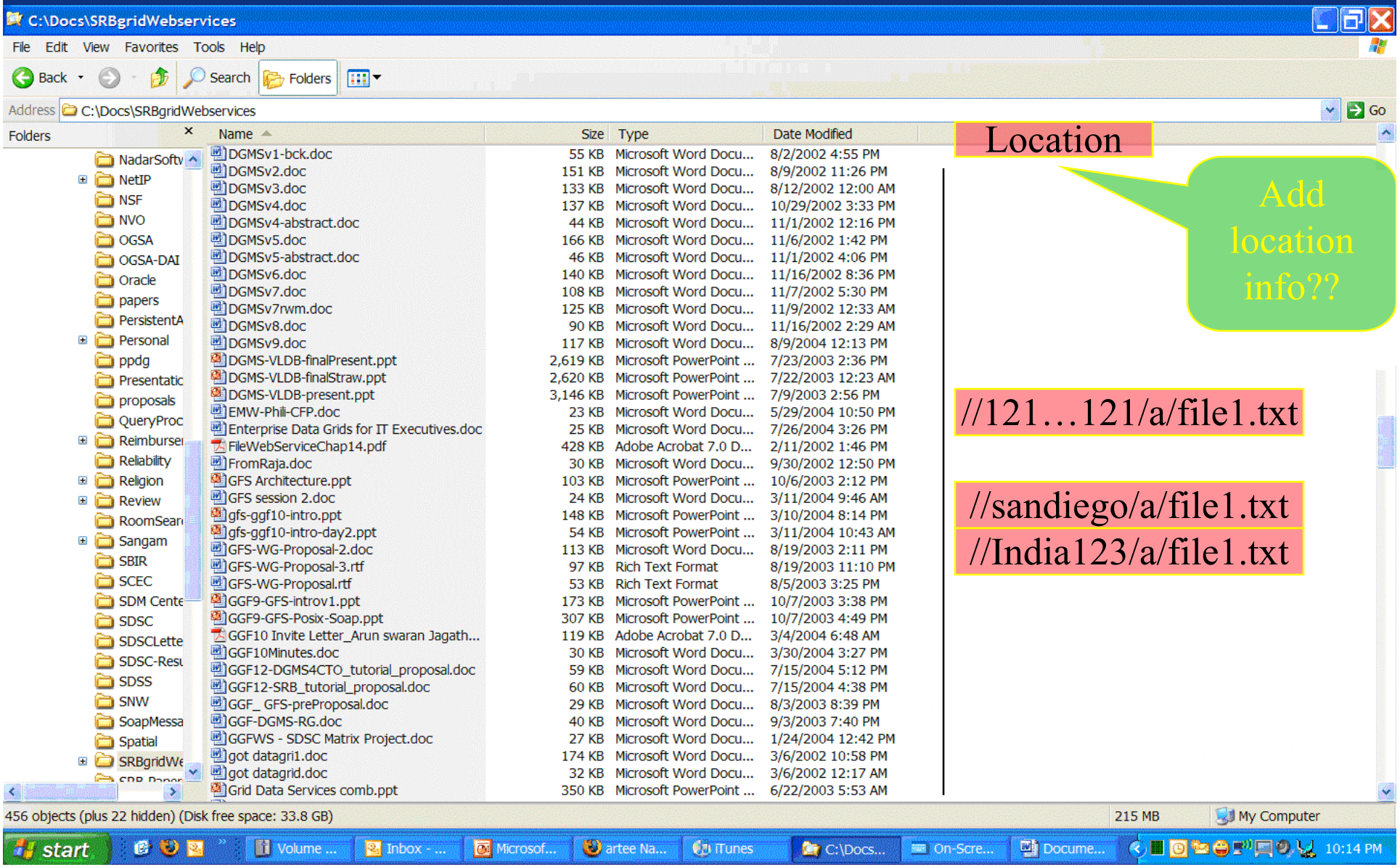
Possible solution (1)

- 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



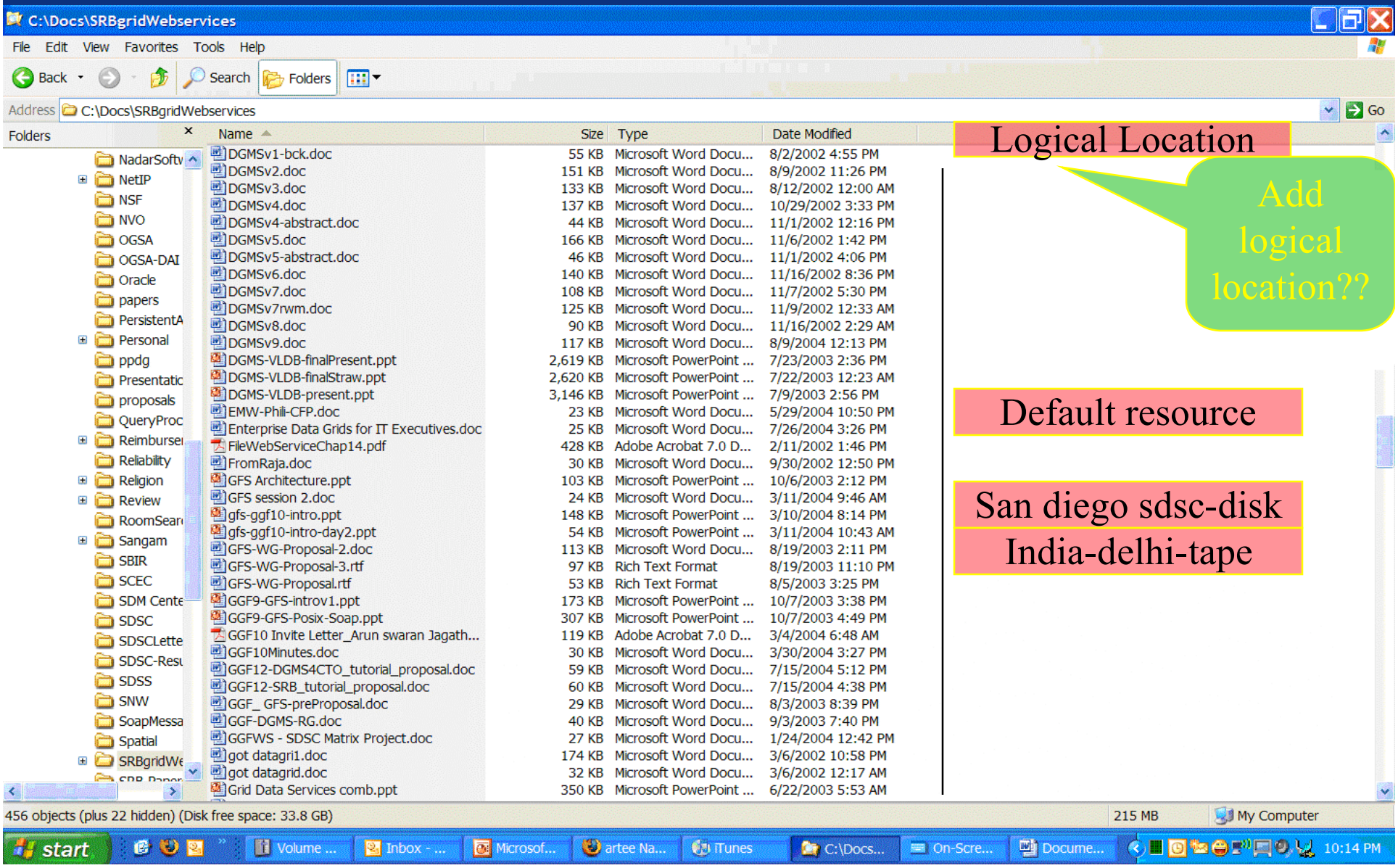
Logical namespace + location



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

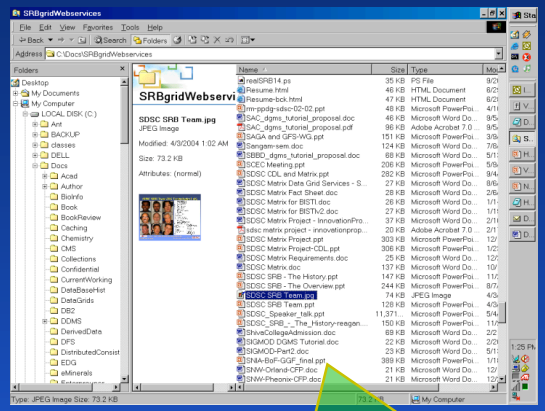


Possible Solution (3)

- *"Houston, We have a solution!"*

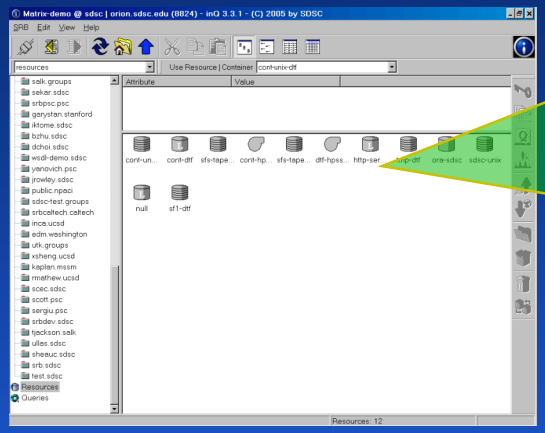
- **We will create a separate (human readable) logical resource namespace of all data resources in the data grid**
- **The well-known logical file namespace can be "joined" with this logical resource namespace to create this new file-name space for our Wide-Area Collaborative FileSystem**

The solution (or a part of the solution)



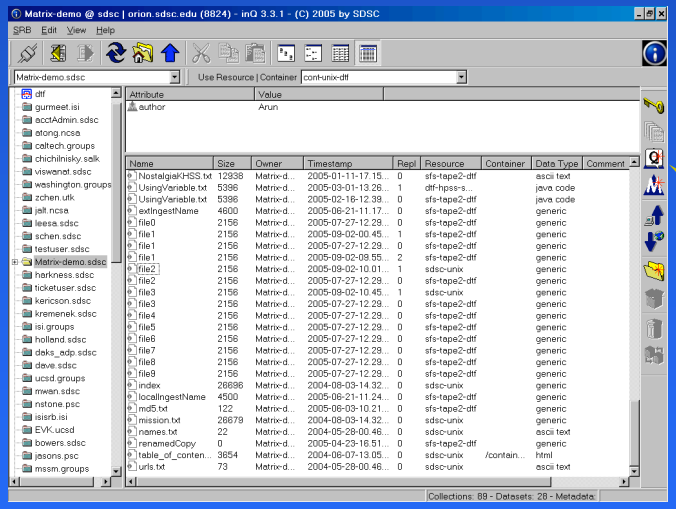
Traditional Logical Namespace

+



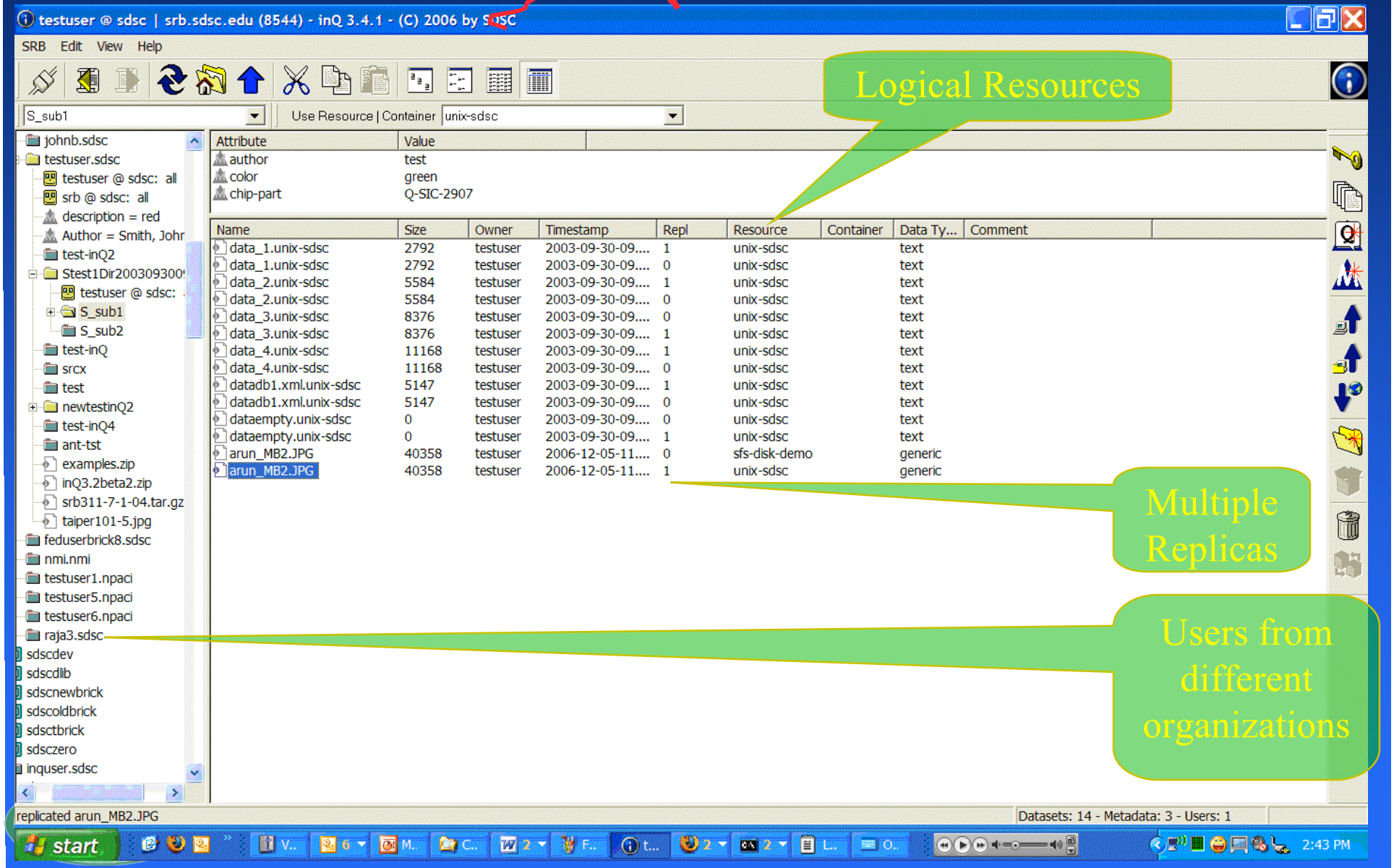
Logical Resource Namespace (Each logical resource is a combination of one or more physical resources)

=



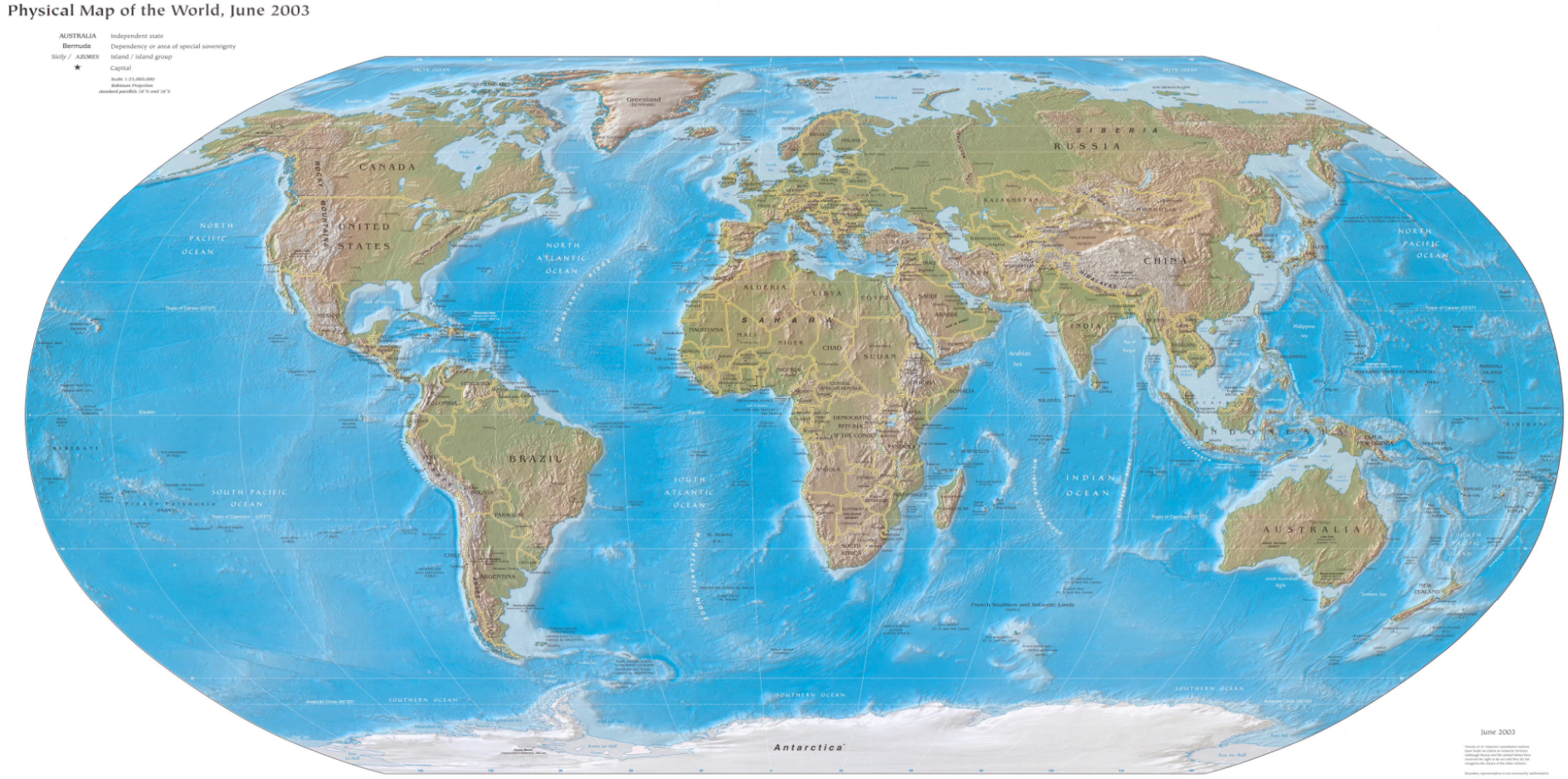
Data Grid Namespace

Cloud Data Grid Namespace



myGlobal500.com
(single business entity)

myMovieStudio.com (collaboration of businesses)



myFreeBackup.com
(1 man IT-army)

Agenda

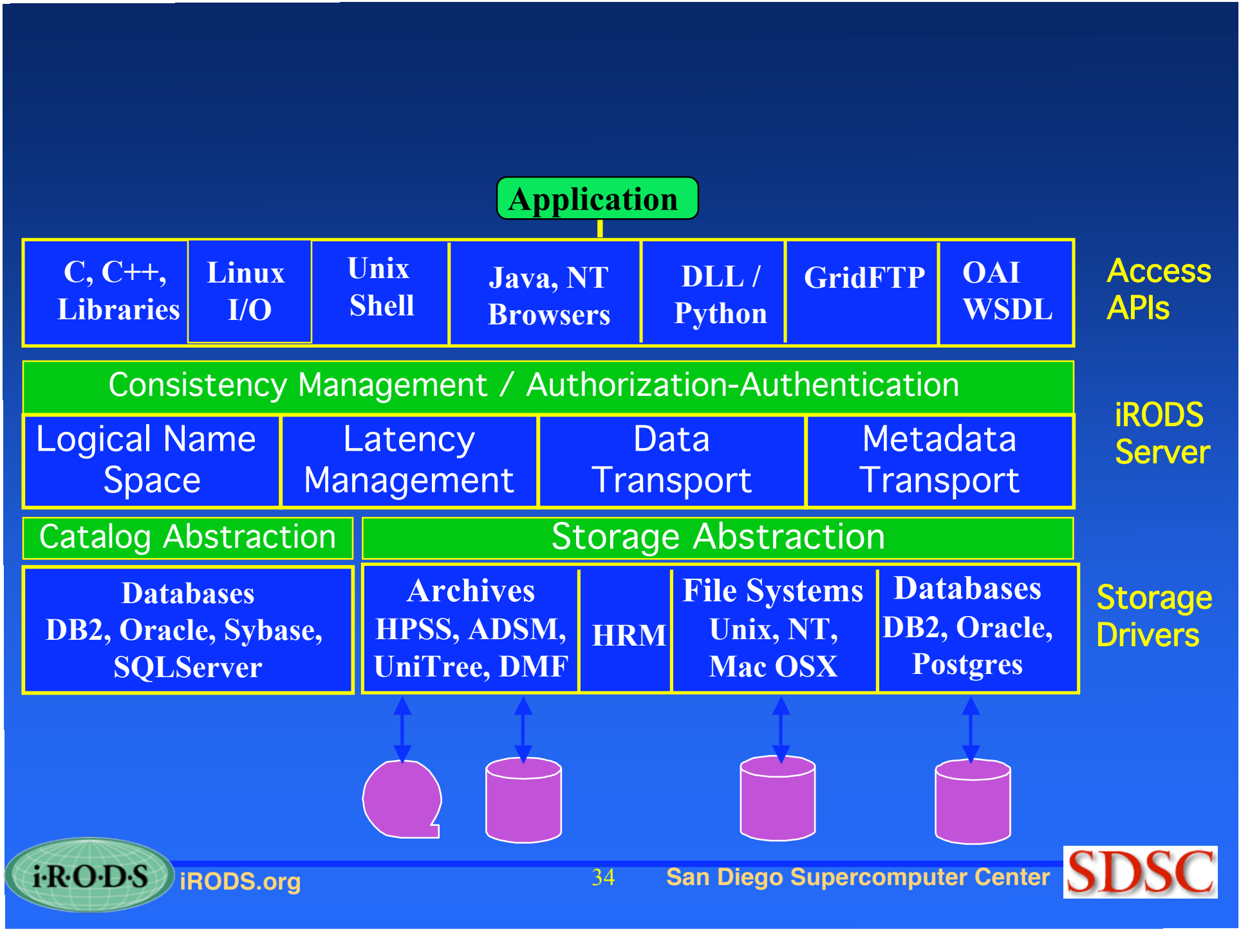
- Intro
- Problem Motivation: LSST and myGlobal500.com
- Solution Path
- Implementation: iRODS
- Improvements Possible / Future Plans

iRODS

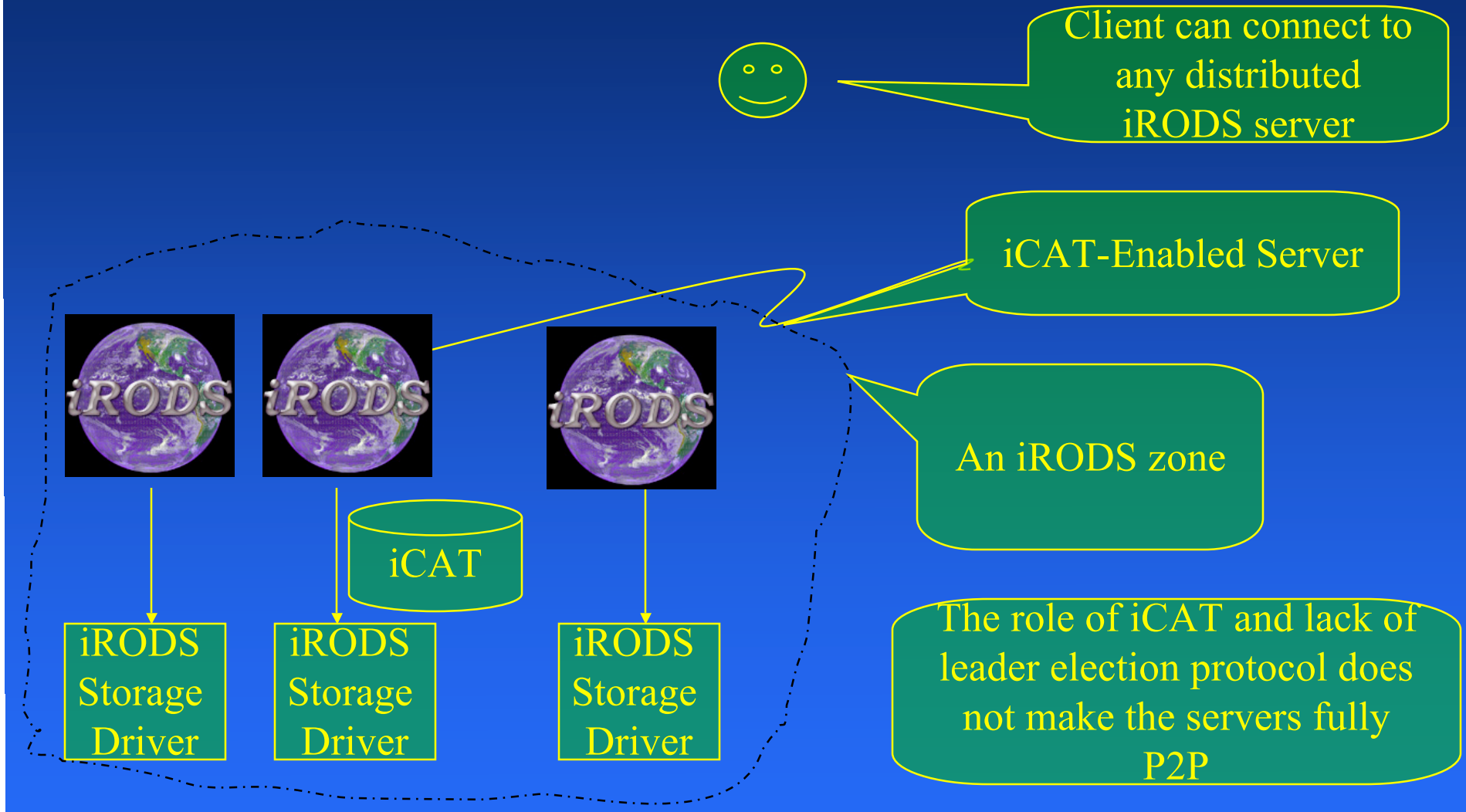
- **Client-server architecture**
  - Clients implement our protocol in your favorite programming language
- **Servers**
  - Use an hybrid Model of P2P and centralized concepts
  - Servers have storage associated with them (in the form of directories on host/physical file system)
  - One (or more) servers connect to a database that stores the file metadata
  -

Three Tier Architecture

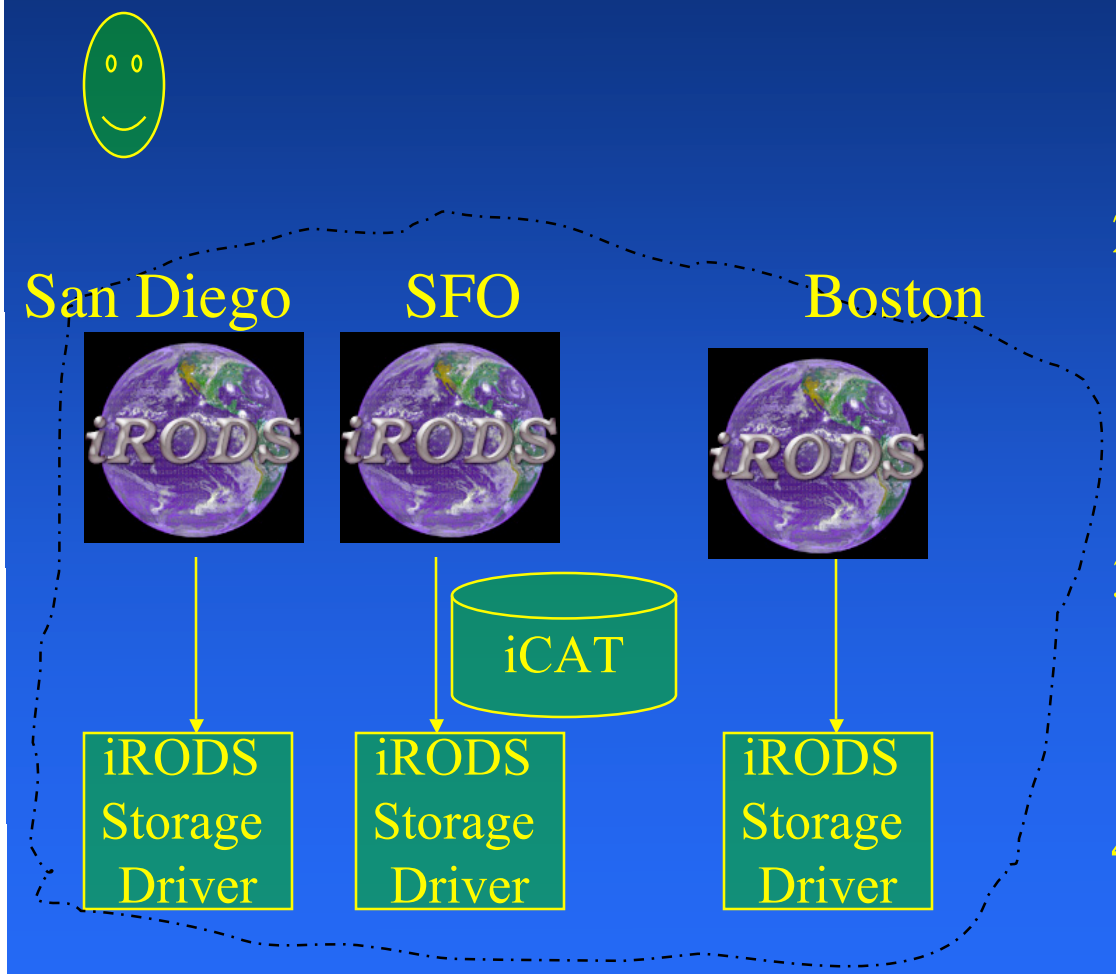
- **Client Implementations (Any interface/API)**
  - Your preferred access mechanism
- **Servers (iRODS Server)**
    - Federated to support direct interactions between servers
  - Metadata catalog (iCAT)
    - Separation of metadata management from data storage
    - State persistence using a database
- **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, ADSM, UniTree, DMF
HRM
File Systems Unix, NT, Mac OSX
Databases DB2, Oracle, Postgres
Storage Drivers

Peer-2-peer (like) iRODS servers

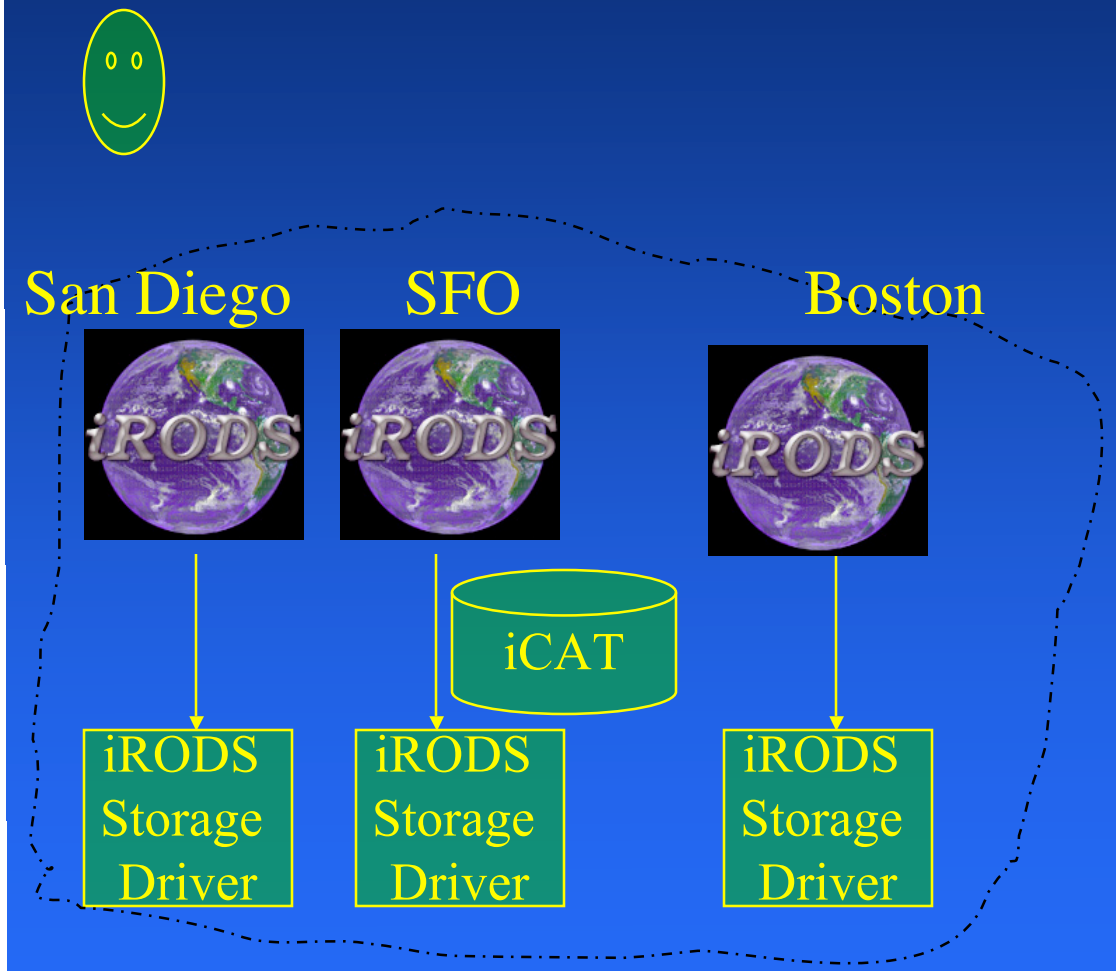


Basic get explained



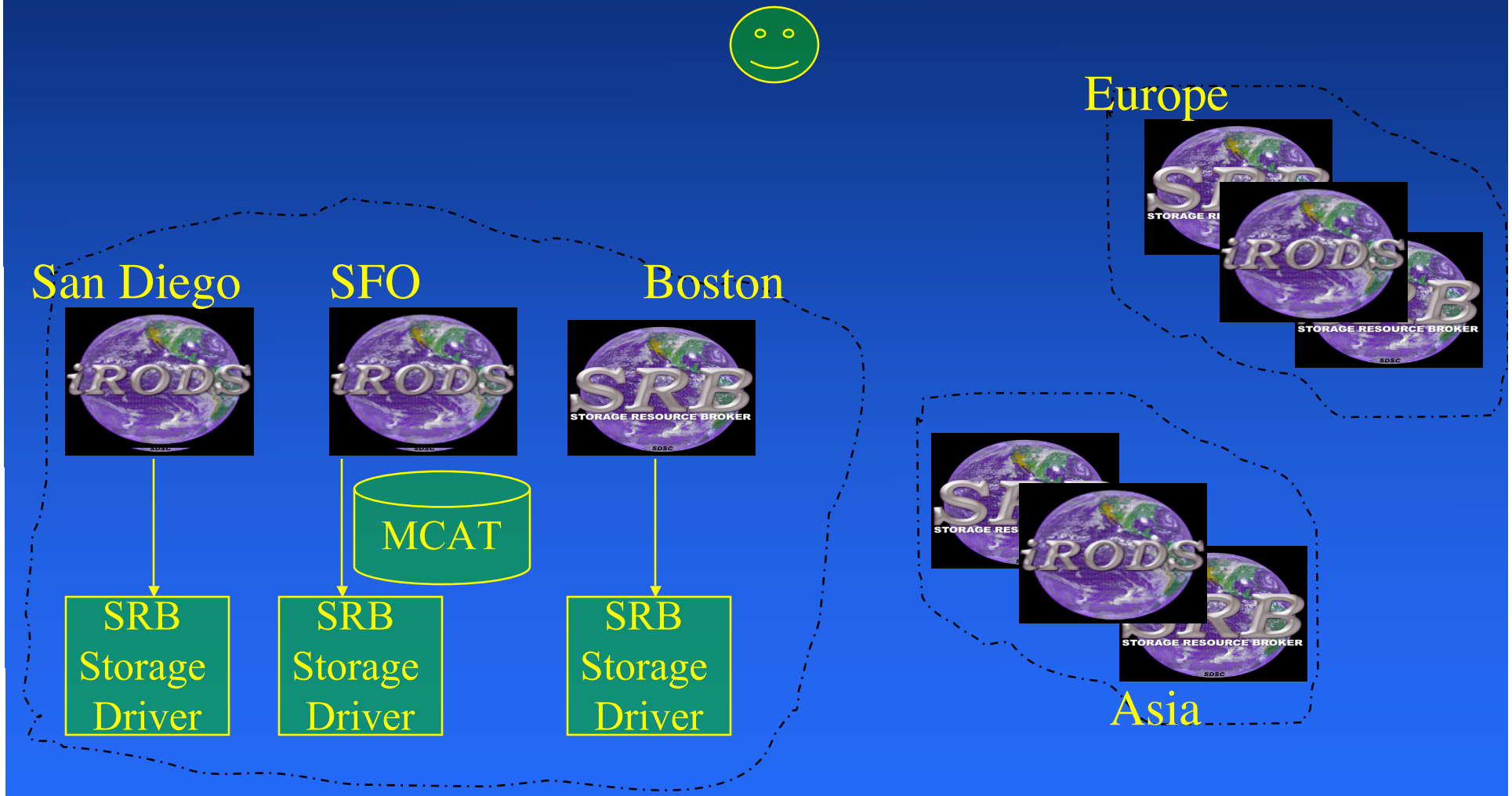
1. Check Auth (Logon-server connects to iCAT server)
2. Find optimal copy of the file for that particular client request (uses simple heuristics)
3. Decide on a data path option, number of threads, bandwidth etc
4. Send the data (failover to replica automagically

Basic put explained (with iRules - Trigger_like)



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 automagically]
5. <post_process>
6. <error_recovery>

Peer-2-peer SRB Zones



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
  - RENCI/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