Storage Challenges at the San Diego Supercomputer Center

Richard Marciano

P.O. Box 85608

San Diego, CA 92186

Ph: (619) 534-8345 **Fax:** (619 822-0906

E-mail: marciano@sdsc.edu

Presented at the THIC meeting at the Amberley Suite Hotel in Albuquerque, NM on April 21, 1998

SAN DIEGO SUPERCOMPUTER CENTER

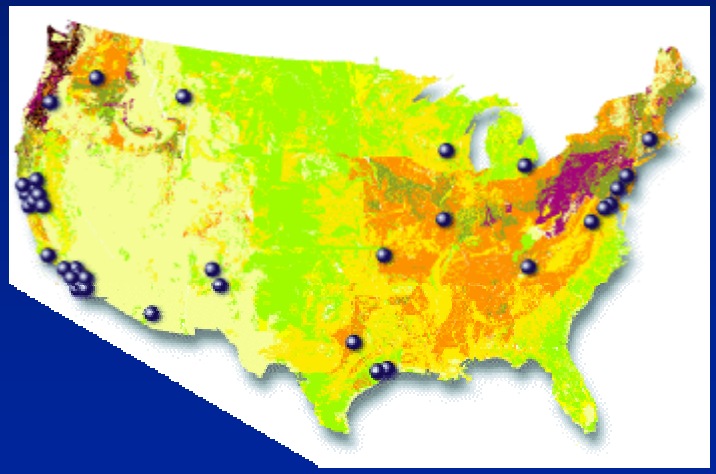
A National Laboratory for Computational Science & Engineering

Data Intensive & Mass Storage

- ***Systems Group***
  - Phil **Andrews**
- ***Systems Storage Group***
  - Joe **Lopez**
  - Mike **Gleicher**
- ***Data Intensive Group***
  - Reagan **Moore**
  - Chaitan **Baru**
  - Amarnath **Gupta**
  - Richard **Marciano**
  - Arcot **Rajasekar**
  - Wayne **Schroeder**
  - Michael **Wan**

NPACI Program

- **WHO: 37 partners in 18 states led by UCSD**
  resource / research & education / associate / industrial partners
- **WHAT:**
  - **1.** Create a HPC infrastructure with teraflops performance, ***petabyte archives, and discipline-specific caches***
  - **2.** Increase focus on ***data-intensive computing*** to analyze ***terabyte*** data sets
  - **3.** Education, outreach, and training
  - **4.** Promote industrial partnerships



Partner Sites

- UCSD/SDSC
- UC Berkeley
- UCLA
- U Maryland
- CSU/SDSU
- Scripps
- Montana State
- U Kansas
- Salk Institute
- EPSCoR Foundation
- Los Alamos NL
- PNNL / EMSL
- U Massachusetts
- Caltech
- U Michigan
- UC Santa Barbara
- Washington U
- UC Santa Cruz
- U Virginia
- USC
- LTER / U New Mexico
- Stanford
- JPL
- LBNL / NERSC
- UC Irvine
- U Pennsylvania
- U Texas
- UC Davis
- U Houston/Keck Center
- CRPC
- Oregon State
- U Wisconsin
- U Tennessee
- Rutgers
- CARB
- Kitt Peak Nat. Obs.
- Lawrence Livermore NL
- UC San Francisco

Planned NPACI Data Caches (FY'98)

Partner Site	***Cache Size (GB)***	***Hardware (OS)***	***Network***
UCSD/SDSC	2000	IBM SP, Wildcat (AIX)	vBNS, CalREN-2, Esnet, AAInet
Caltech	500	IBM SP (AIX)	vBNS
UC Berkeley	50	Sun (Solaris)	CalREN-2
U Michigan	200	IBM SP (AIX)	vBNS
U Texas	100	IBM SP (AIX)	vBNS
UC Davis	250	Origin 2000 (SGI)	CalREN-2
UC Los Angeles	200	SGI (SGI)	CalREN-2
UC Santa Barbara	100	Digital (OSF)	CalREN-2
Washington U	400	Sun (SunOS)	vBNS
U Maryland	200	IBM SP (AIX)	vBNS
U Houston	200	IBM SP (AIX)	vBNS

Data Storage Vision

- **Dramatically improve the ability of science researchers to build upon prior knowledge by:**
  - establishing ***domain specific*** data repositories
  - enabling ***publication*** of scientific data sets and associated methods
  - providing ***information discovery*** and ***data handling*** APIs
- **Support integration of digital library, metacomputing, scalable tools, interaction environments**

Data Cache Testbed Projects

- **Molecular Science**
  - U Houston (molecular dynamics trajectory DB), U Houston Keck Center for Computational Biology (enhanced molecule image DB)
- **Neuroscience**
  - Sharing brain images between data repositories at Washington U, UCLA, UCSD/SDSC
- **Earth System Science**
  - UCSD (climate DB), UCLA (earth sys. DB), UC Davis (California natural resources), U Maryland (satellite land cover data repository)
- **Astronomy**
  - Integration of multiple digital sky surveys to support statistical analysis on astronomical objects, Caltech (continued on next slide…)

Astronomy (… continued)

- **Federate multi-frequency surveys**
- **2 million panels for the whole sky**
- **15 TB integrated catalog and image DB**
- **Computationally intensive analysis: "what is the morphology of galaxies"**
  - Digital Palomar Observatory Sky Survey
    - **optical**: 2 billion source catalog / 3 TB
  - 2-Micron All Sky Survey
    - infrared: 100 million source catalog / 10 TB
  - Sloan Digital Sky Survey
    - radio: NVSS (2 million sources) + FIRST (1 million)

Software Architectures

- **Computational Grid**
  - Distributed execution of applications across heterogeneous compute platforms linked by the Web
- **InterLib**
  - Uniform information discovery and data publication environment integrating access to heterogeneous data resources on the Web

Storage Environment

- **HPSS**
  - 2 IBM 3494 Magstar Tape Libraries which hold 2,000 tapes each, with 14 3590 tape drives
  - IBM Scalable POWERparallel System (SP2), 23 nodes
  - 90 HPSS servers (32 movers), 13 classes of service based on 23 storage classes, automatic CoS selection
  - Current Storage 50 TB, 65 TB with Cornell & Pittsburgh
- **STK**
  - 11 TB
  - IBM RS6000/590
- **IBM SSA disk subsystem, 1 TB**
- **260 GB Maximum Strategy disk**

SDSC HPSS Production System

- **4 Million files at present**
- **50 TB of data**
- **Movers average up to 4,000 to 5,000 requests**
  - Peaks at 10,000
- **20 concurrent users on average**
- **700 tape mounts / day on average**
- **Transfer rates: 8 MB / s over networks, 1 TB / day in & out of HPSS**
- **2 TB / month growth, 25 TB burst from PSC & CTC**

CTC HPSS System

- **13 TB from CTC on RS/6000 (2 x 2,500 tapes)**
  - 3 GB of HPSS metadata received over the vBNS
  - "bandwidth" of > 120 MB / s
  - 1200 3590 tapes & 1200 3480 tapes
- **2-3480 manual tape drives & access to 4-3590 tape drives in STK Library (retrofit)**
- **Migrating CTC user data into SDSC production system as a background job over within the year**
- **CTC users access data in Read Only mode**

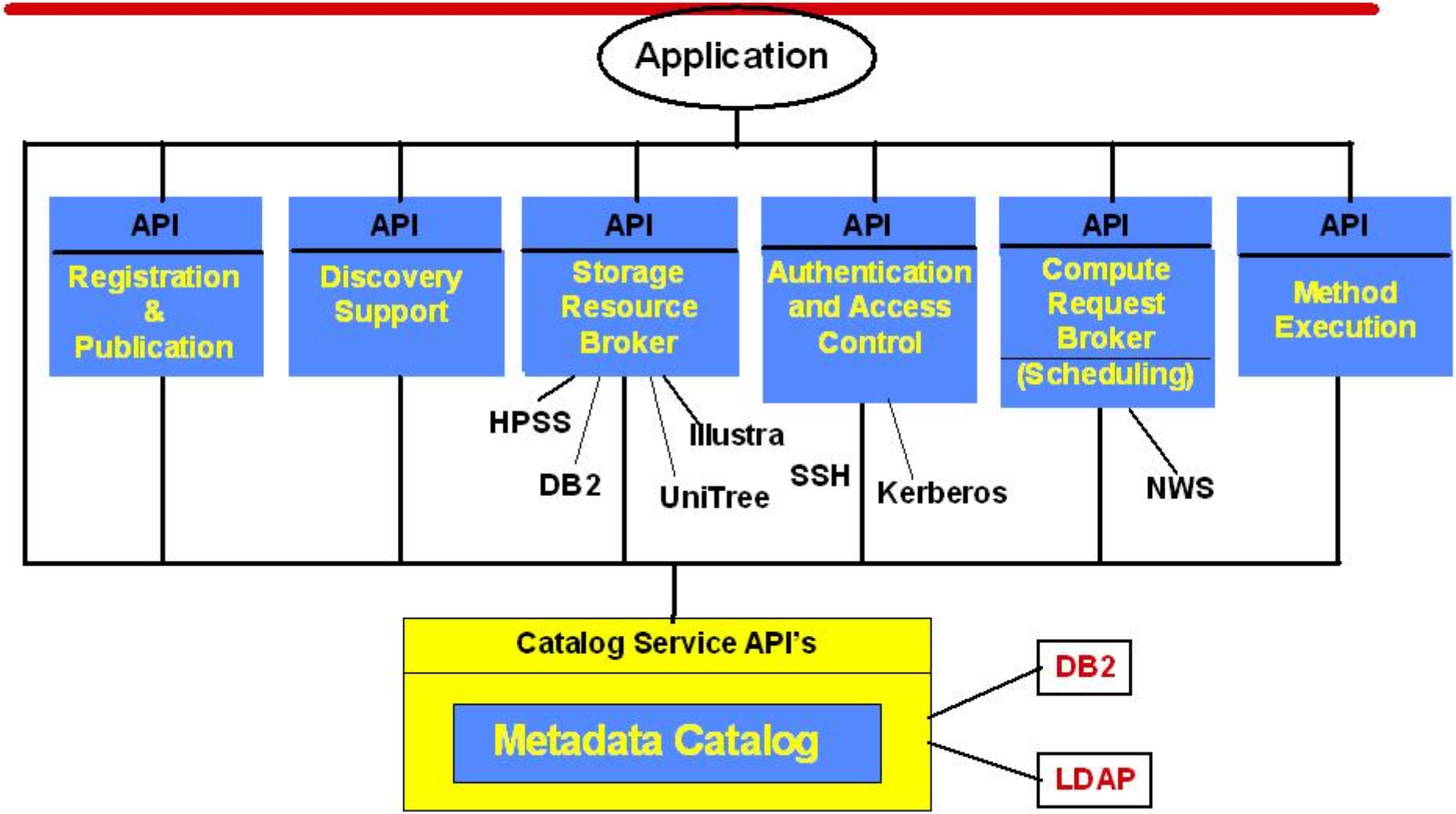
PSC Data Migration

- **All data over vBNS (OC-3 link at 155 Mb / s), 10 MB / s**
- **Transfers on a voluntary basis**
- **Transfers directly into HPSS tape**
- **Fully automated process**
- **About 12 TB to transfer total (5 more likely)**

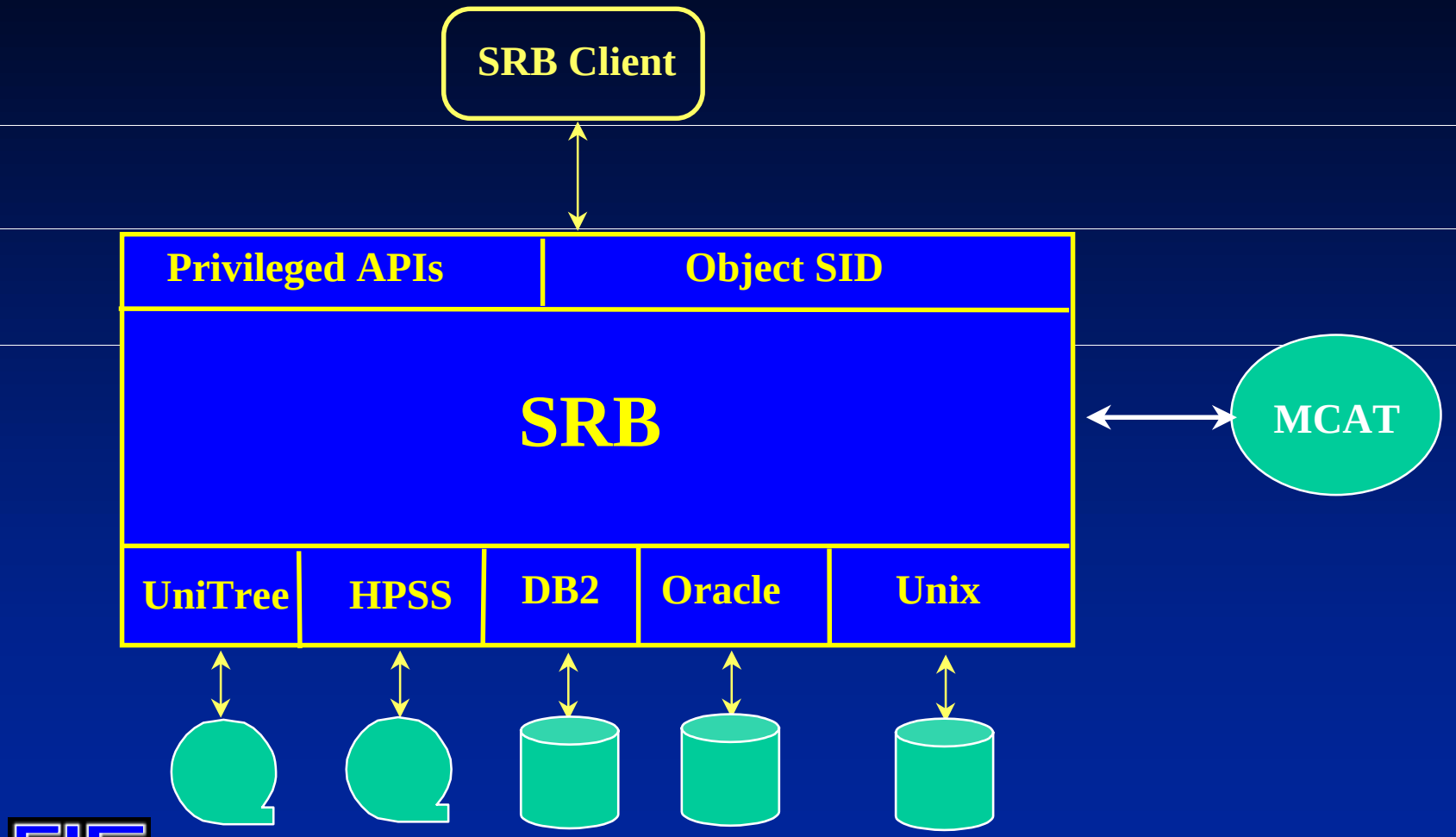
HPSS User Interfaces

- **FTP/PFTP**
  - Passwordless acess
  - Automatic COS selection
- **HSI (DCE and non-DCE)**
- **DB2 Universal Database**
- **Storage Resource Broker (SRB)**
  - http://www.npaci.edu/DICE

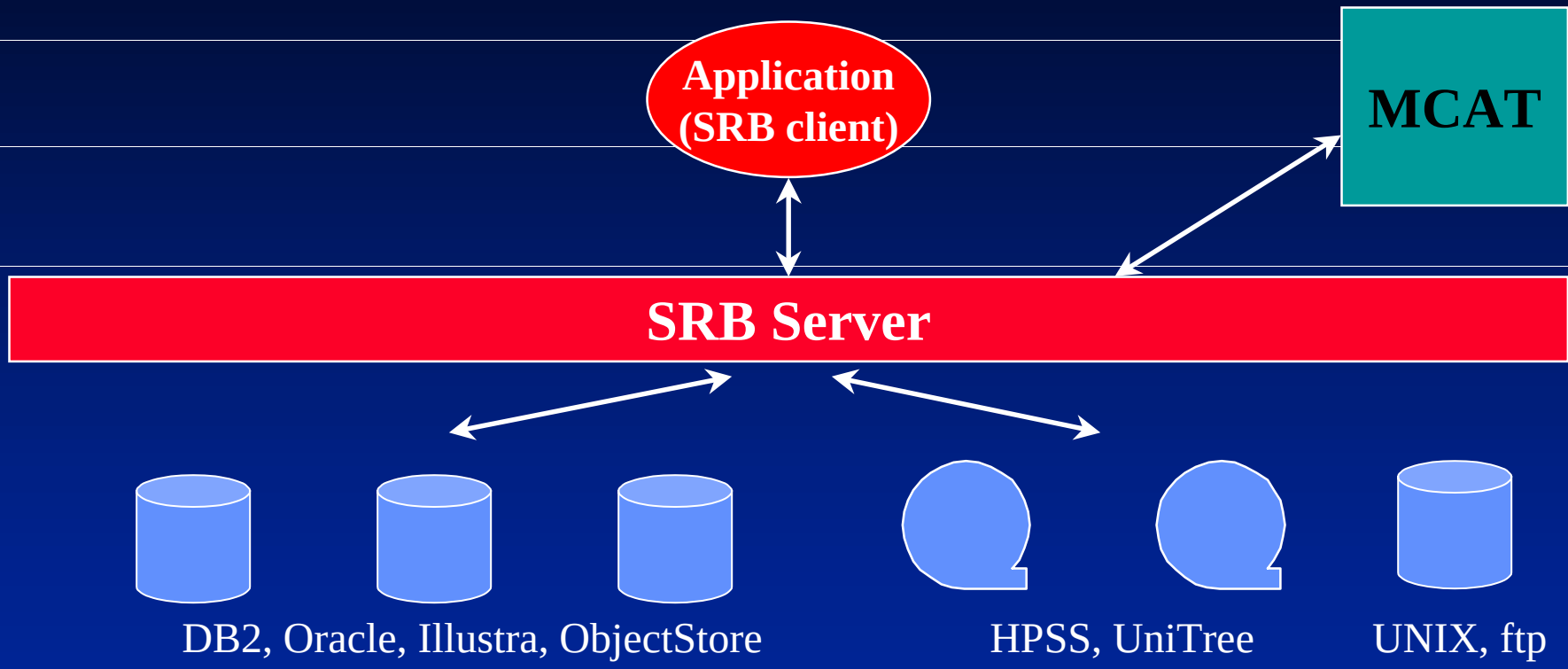
Information Based Computing Architecture



Software Architecture of the SRB



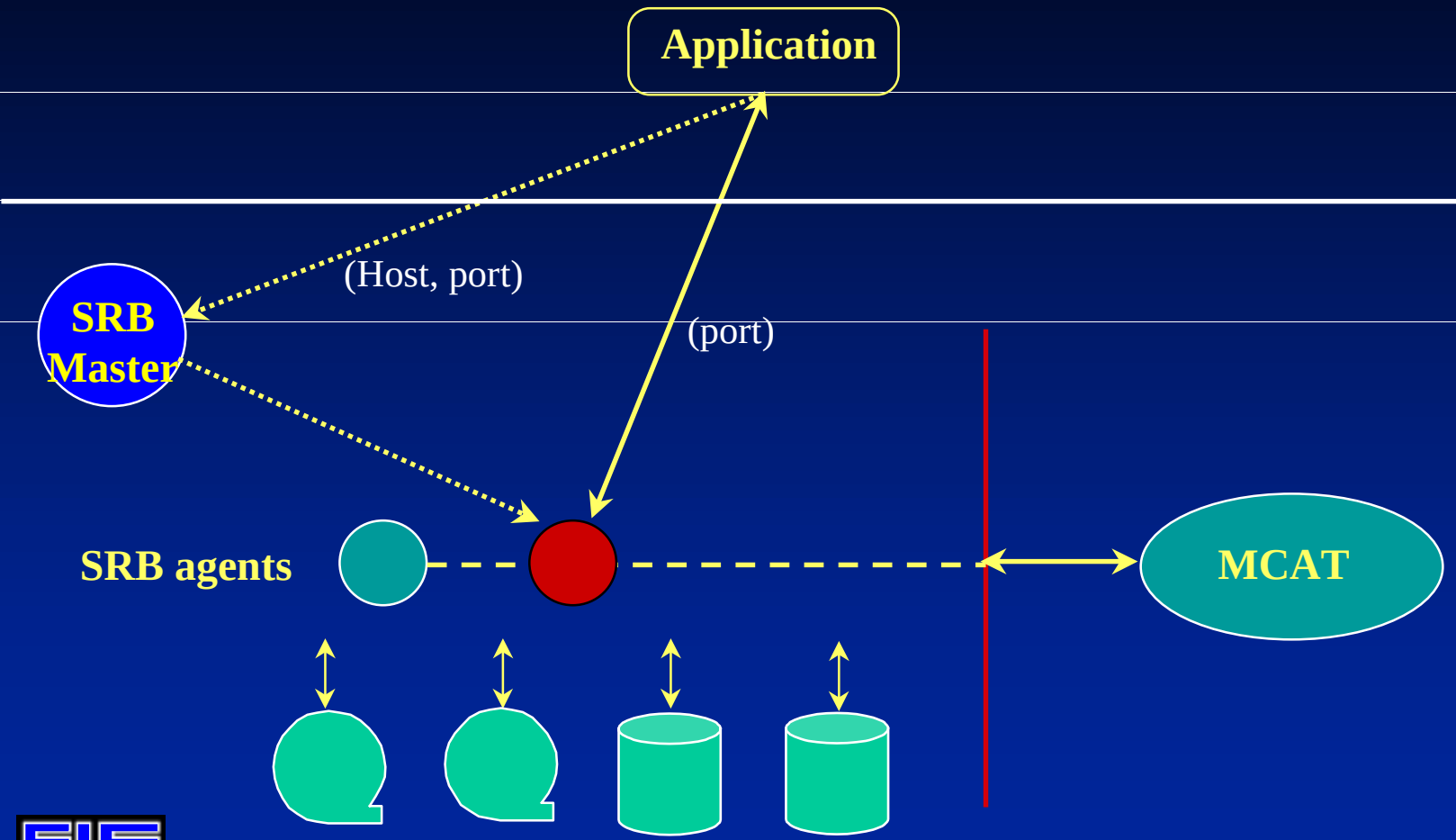
The Storage Resource Broker is Middleware



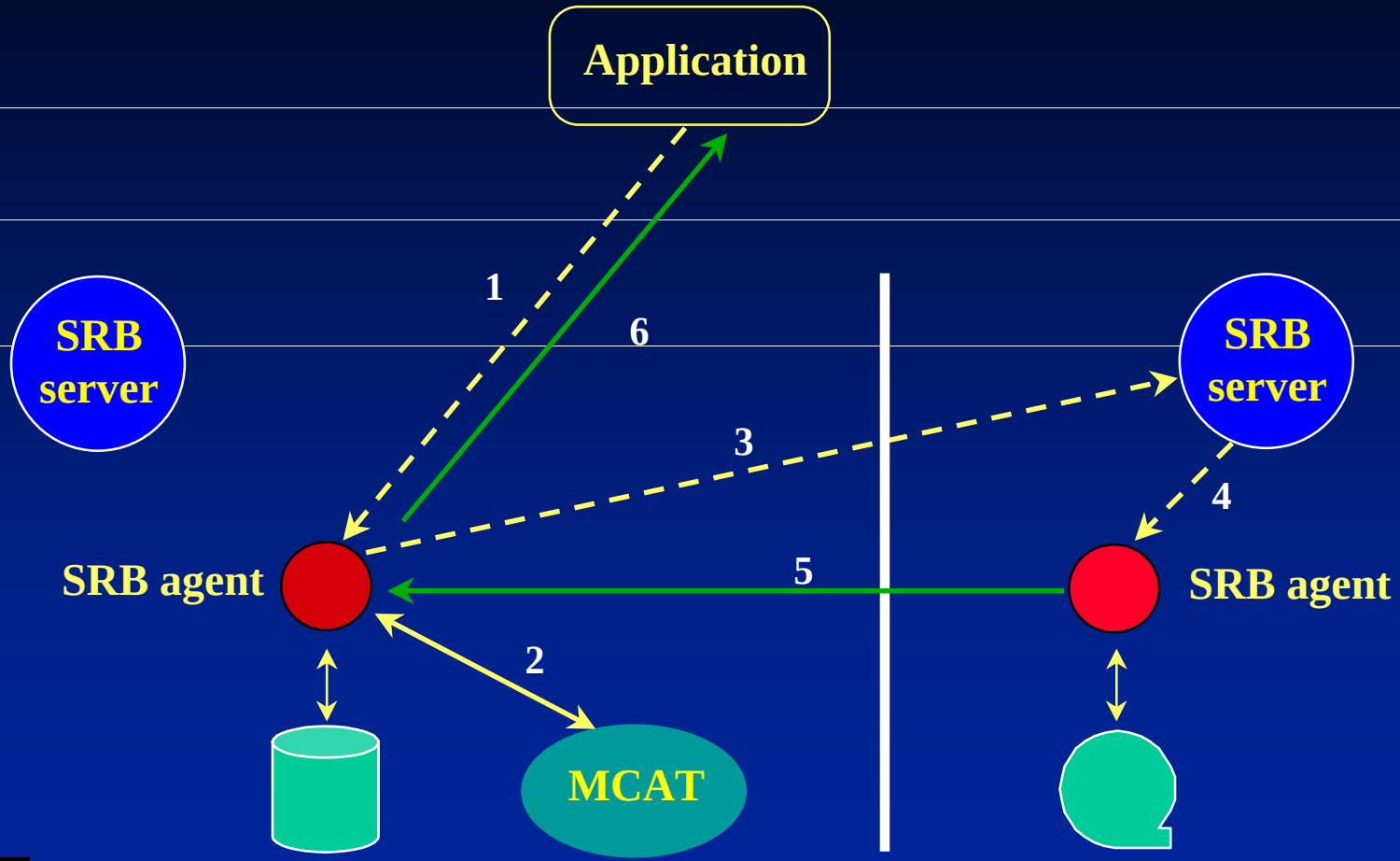
Distributed Storage Resources

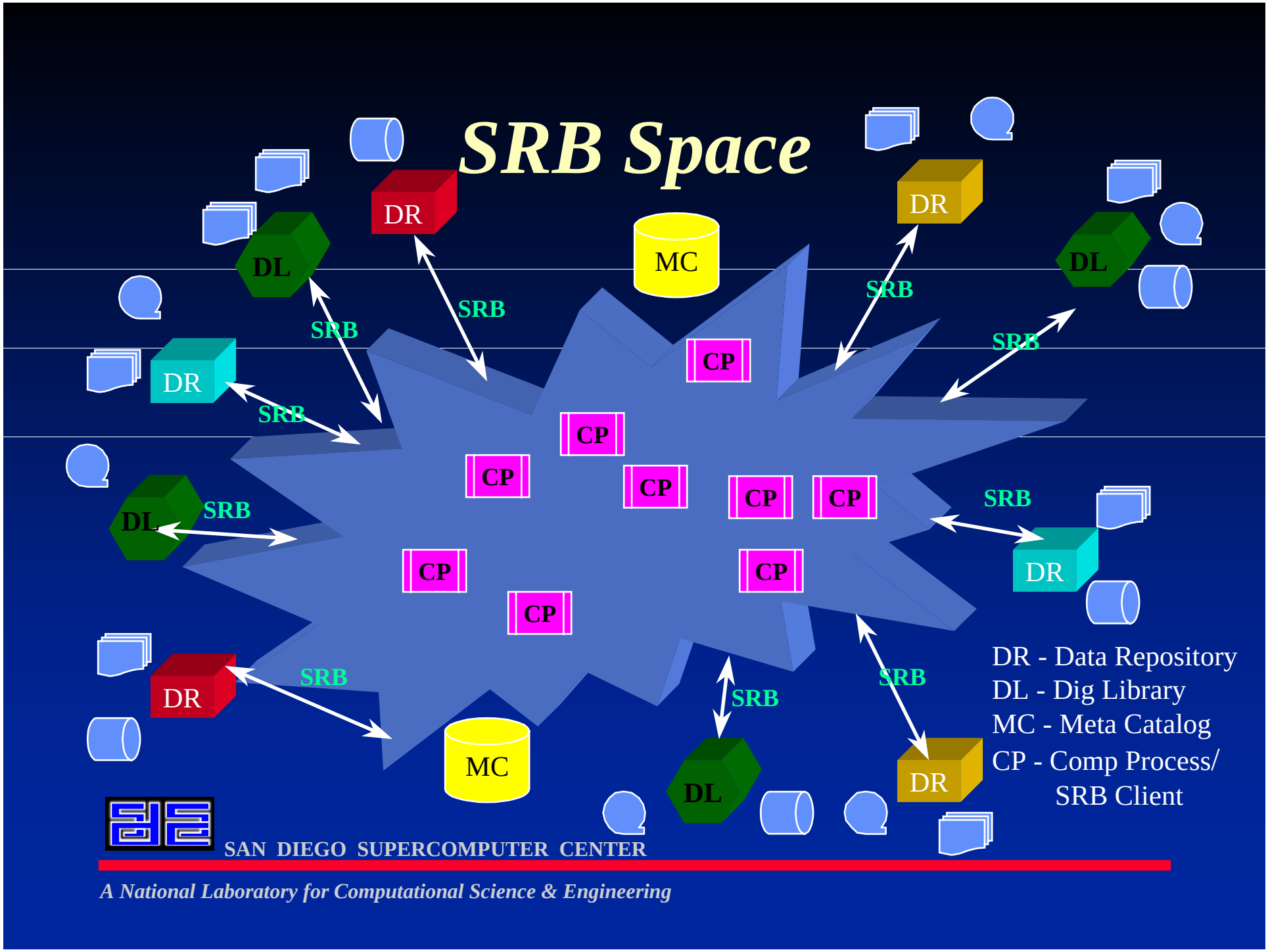
(database systems, archival storage systems, file systems, ftp)

The SRB Process Model



Federated SRB Operation




SRB Space
DR
DR
DR
DR
DR
MC
MC
DL
DL
DL
DL
SRB
SRB
SRB
SRB
SRB
SRB
SRB
SRB
SRB
SRB
CP
CP
CP
CP
CP
CP
CP
CP
CP
DR - Data Repository
DL - Dig Library
MC - Meta Catalog
CP - Comp Process/
SRB Client
SAN DIEGO SUPERCOMPUTER CENTER
A National Laboratory for Computational Science & Engineering

SRB V1.0 Features

- **Multi-platform (clients and servers)**
  - SunOS/Solaris, AIX, Cray C90, DEC OSF
- **API and command line interfaces**
- **“Low-level” and “high-level” APIs**
- **Storage systems supported**
  - *DB2, Illustra, Unitree, HPSS, UNIX*
- **Support for *federated* servers**
- **Released early September, 1997**

SRB V1.1 Features

- **In beta in DOCT. Released in January, 1998**
- **Ported to additional platforms - SGI, Cray T3E**
- **Incorporates the SDSC Encryption and Authentication (SEA) Library**
- **Ticket-based access control**
- **Graphical user interface - SRBTool**
- **Additional storage systems supported**
  - *Oracle, Objectstore, ftp, http*
- **Oracle-based MCAT**
- **Support for *proxy* operations, e.g. *move, copy, replicate***
- **Data replication using *Logical Storage Resource***

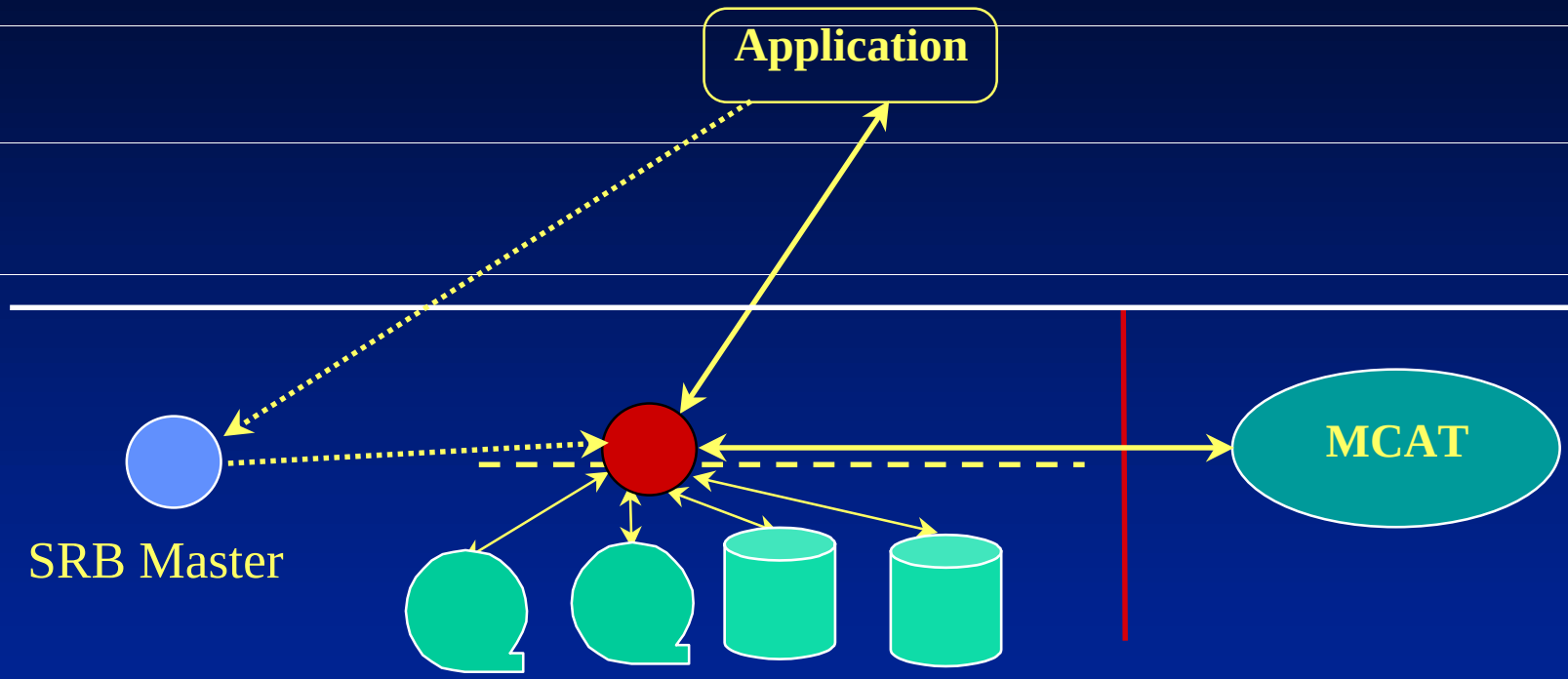
MCAT: Metadata Catalog

- **Stores metadata about**
  - Users, Data sets, Resources, Methods
- **Provides "collection" abstraction**
- **Stores detailed access control information**
- **Maintains audit trail information on data sets**
- **Implemented as a relational database with referential integrity constraints (currently uses DB2, ported to Oracle)**

SRB API

- **Programmatic API**
  - High-level API
  - Low-level API
  - SRB Manager API
- **Command Level Interface - Scommands**
- **Graphical User Interface - SRBTool**
- **Web Utilities**

SRB API Interface

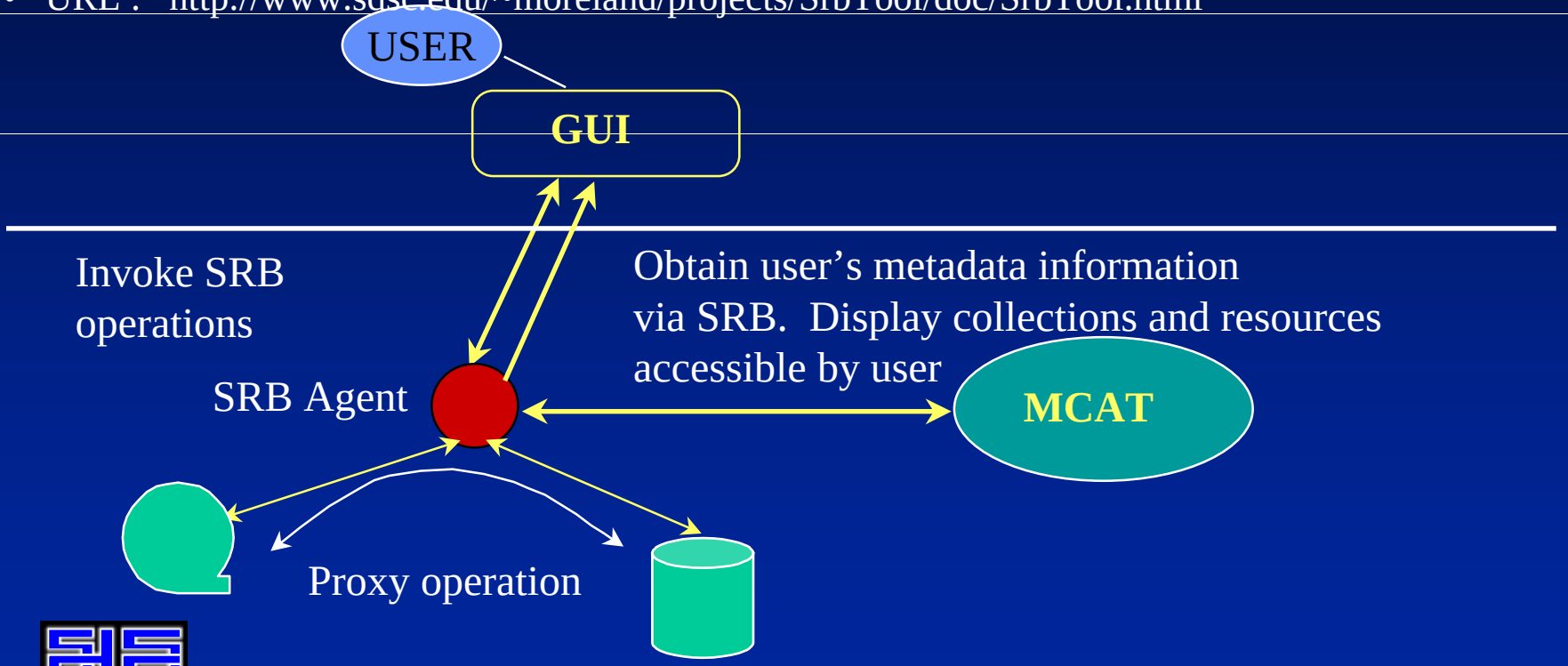


High & Low-level API

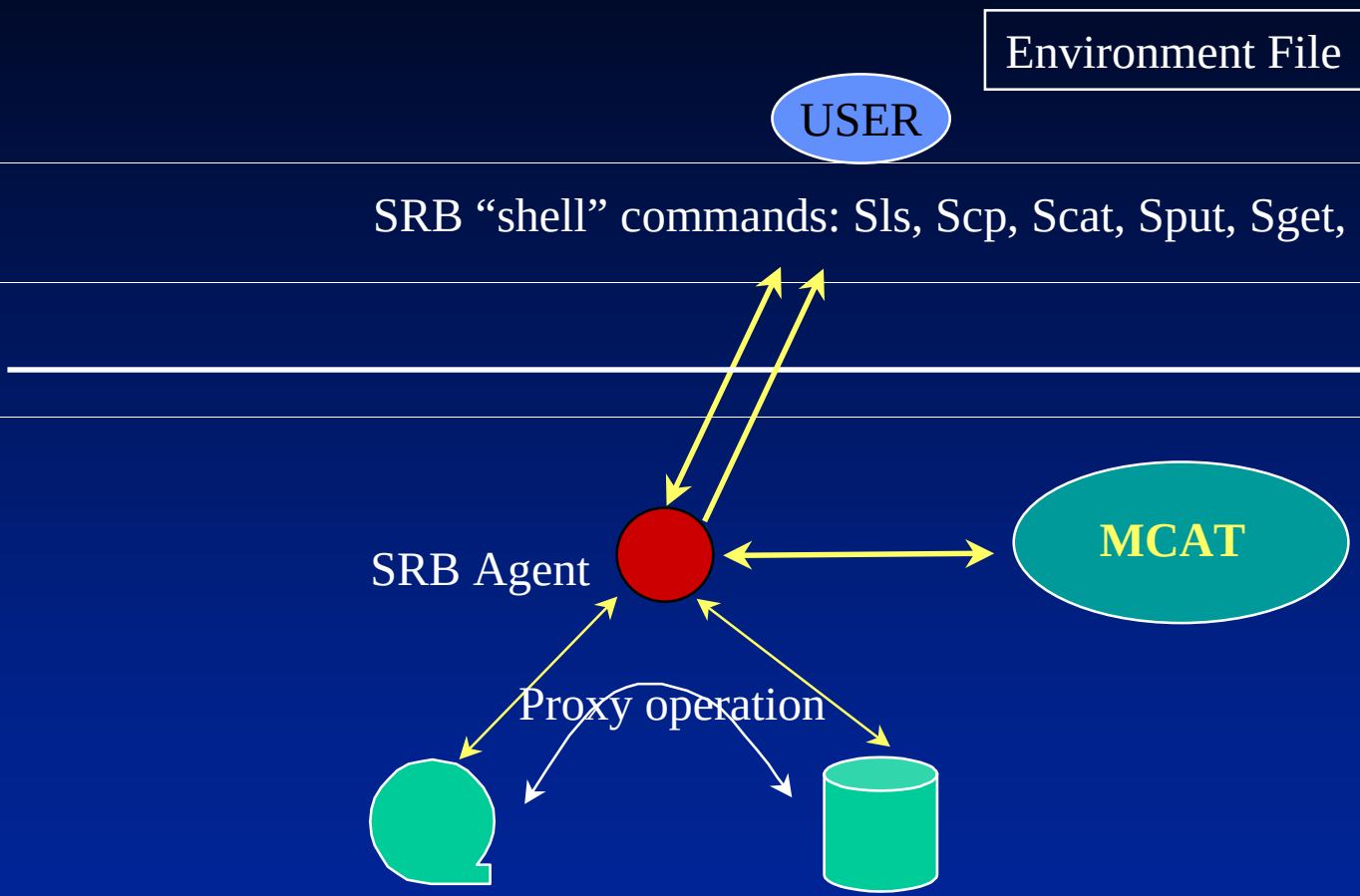
- **Low-level API**
  - talks to resource drivers
  - no registration of data sets in MCAT
  - no authentication through MCAT
  - User provides all information
- **High-level API**
  - Uses low-level API to access resources
  - Registers data management information in MCAT
  - Uses MCAT for authentication and meta information
  - Uses MCAT for resource and data discovery
  - Access/store data in remote SRB

SRB Graphical Interface

- The **SrbTool** is a prototype.
- It is a proof-of-concept demonstration for the **Uniform Visualization Architecture** .
- It incorporates designs from the **Interaction Infrastructure** and the **Data Arbitration System**.
- URL : http://www.sdsc.edu/~moreland/projects/SrbTool/doc/SrbTool.html



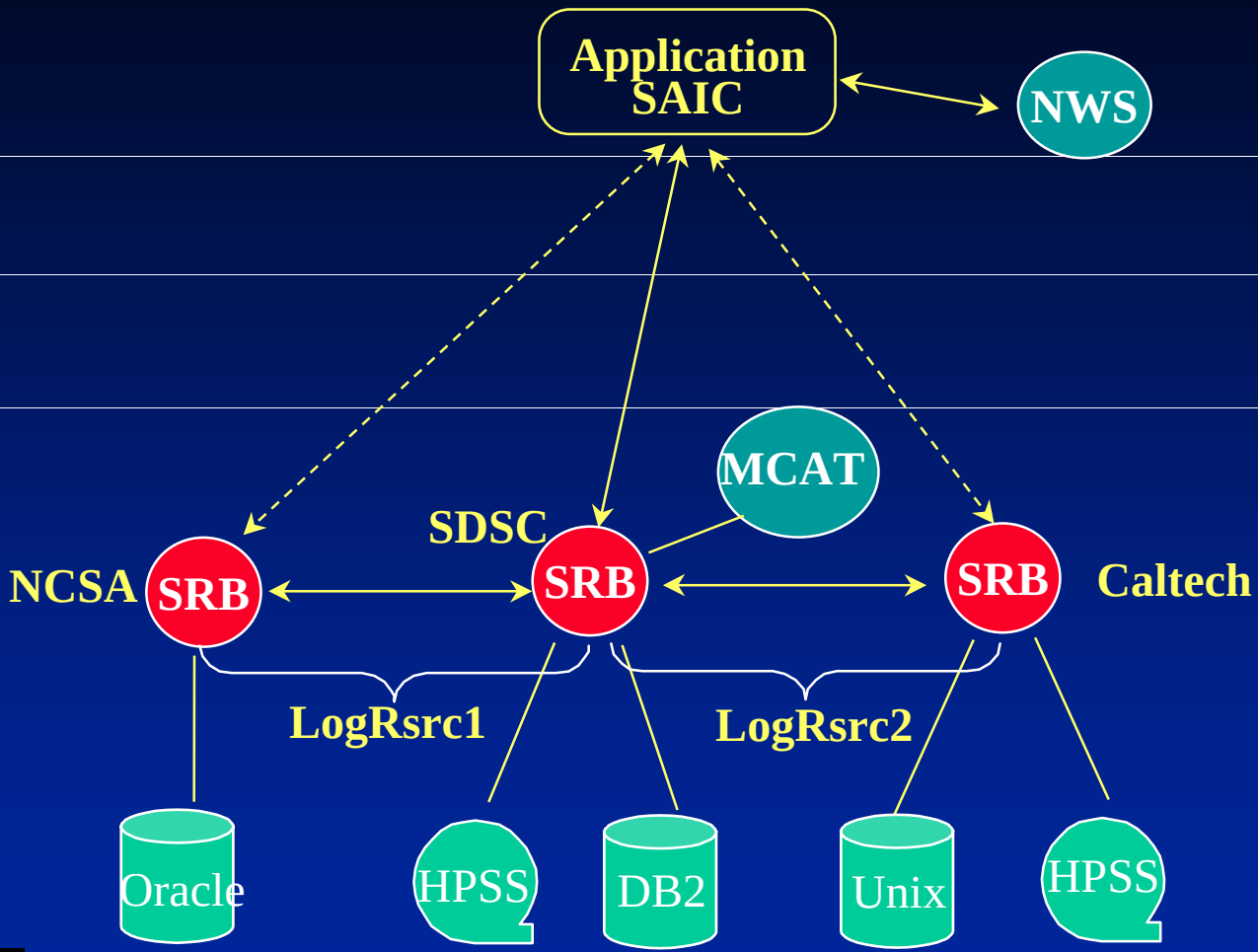
SRB Command Line Interface



SRB Data Replication Support

- **Replication via *Resource Set* definition**
- **Replication support integrated into *write* function**
- **srbObjReplicate API can be used for *post facto* replication**
- **Synchronous replication across all sites. Can choose any *k* out of *n***
- **Can choose specific replica on read operation**

Data Replication (DOCT)



SEA
(SDSC Encryption & Authentication)

- **Developed as part of DOCT**
- **Designed for Supercomputing/ MetaComputing Environment**
- **Based On RSA Public/Private Keys and RC5 Encryption Algorithm**
- **Integrated into SRB**
- **Being integrated into ‘pftp’ & ‘hsi’ - for Remote HPSS Access**

SEA Features

- **Secure User/Process Authentication Across Network (TCP Sockets)**
- **Optional Encryption As Independent Function**
- **Simple API**
- **Batch Support - Long-term Certificates**
- **Adjustable Key Lengths (Speed/Security Tradeoff)**
- **User-Adjustable Encryption Levels (Speed/Security Tradeoff)**
- **Multiple Initial User Registration Methods (Set By Administrator)**
  - Self-Introduction
  - Trusted Host
  - Password
- **Available for Cray T90, C90, T3E, SunOS, Solaris, IRIX, OSF1, AIX, CS6400, NextStep**
- **More Information: http://www.npaci.edu/DICE**

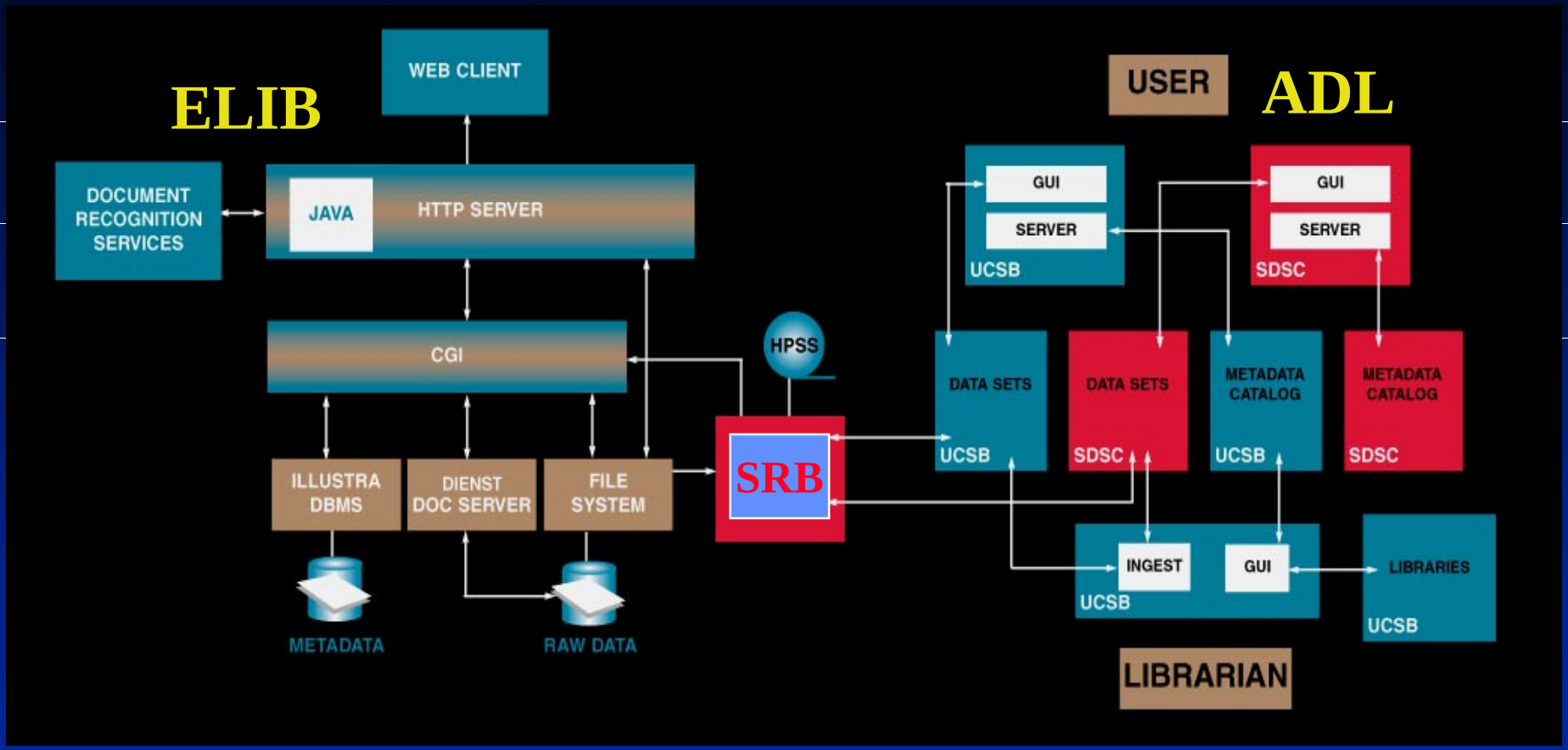
SRB Case Studies

- **Digital Libraries**
  - ELIB - Berkeley Digital Library (UCB)
  - ADL - Alexandria Digital Library (UCSB)
- **DOCT - Patent Workflow System**
- **Environmental Archives**
  - International Satellite Cloud Climatology Project Data
  - TIES Data Atlas - Chesapeake Bay Estuarine Studies

Digital Libraries

- **Access to images, documents & Tools**
- **Large Number of files -**
  - Images of various resolution
  - Documents of various types (valences)
- **Web-based access - form and spatial queries**
- **Domain Metadata - External DB**
- **Digital Objects replicated**
- **Uses SRB web interface and low-level API**

Digital Libraries and SRB



DOCT - Patent Workflow

- **Archiving Applications and Office-Actions**
- **Replicated Archiving**
- **Storage of Issued Patents in multiple forms**
  - SGML, DB Schema, HTML
- **Access of Patents from replicated storage**
- **Controlled Access for Applications and Office-Actions**
- **Uses SRB web utilities and high-Level API**
- **URL: http://www.sdsc.edu/DOCT**

TIES

- **Distributed Data Atlas for Cruise-Transects**
- **Data collected at Chesapeake Bay Estuarine Studies**
- **26 transects, 3 times/year, 6 variables**
- **Reference atlas of 2-D color plots**
- **Domain metadata stored in Oracle**
- **Each Object registered in MCAT**
- **3 partner sites - replication and staging in hierarchical storage (Unix and HPSS)**
- **Uses Scommands and srb web utilities**
- **URL: http://www.sdsc.edu/~marciano/DOCT/Atlas/doct.html**

TIES Data Atlas

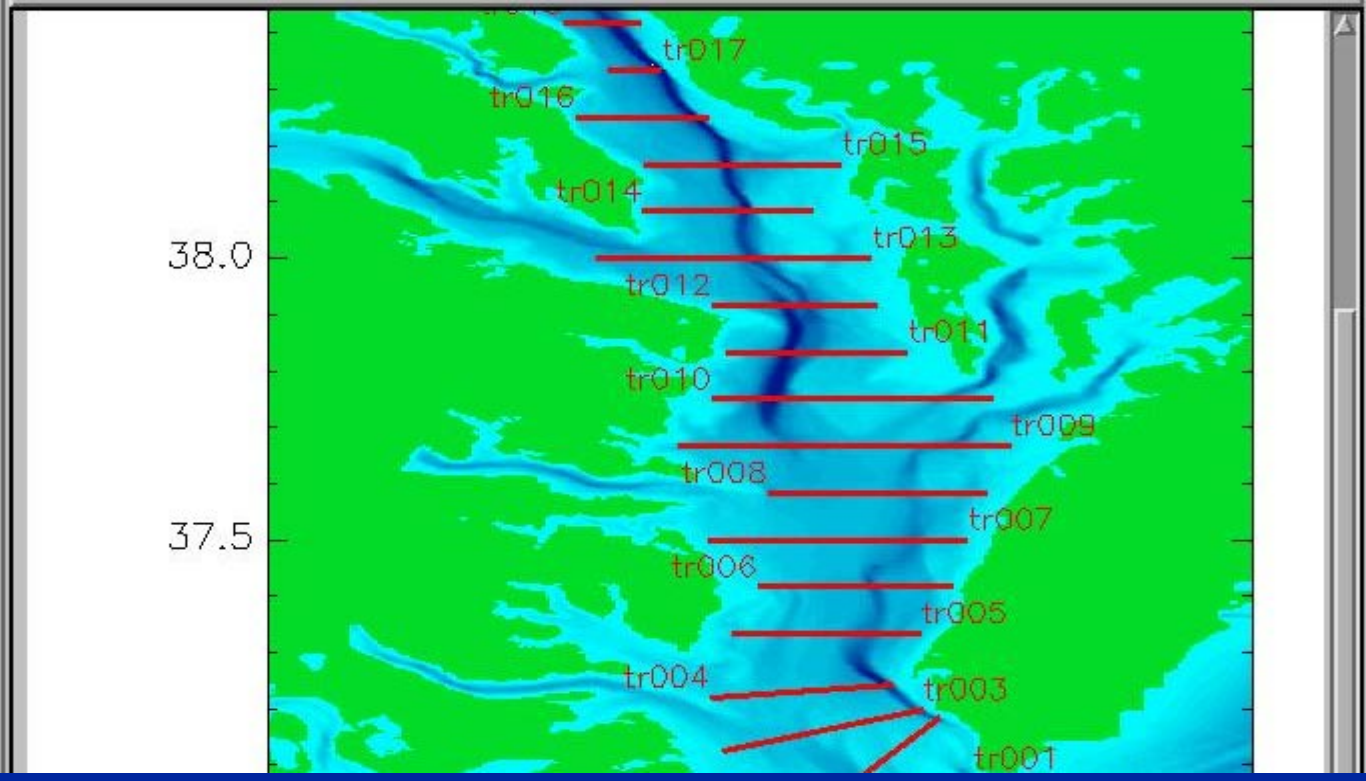
ODU: Cathy *Lascara*, Glen *Wheless* -- **SDSC:** Richard *Marciano*, Jon *Genetti*

Atlas Options

- Search Atlas
- MCAT Browse
- MCAT Copy
- Data Registration

Select Parameters of Interest

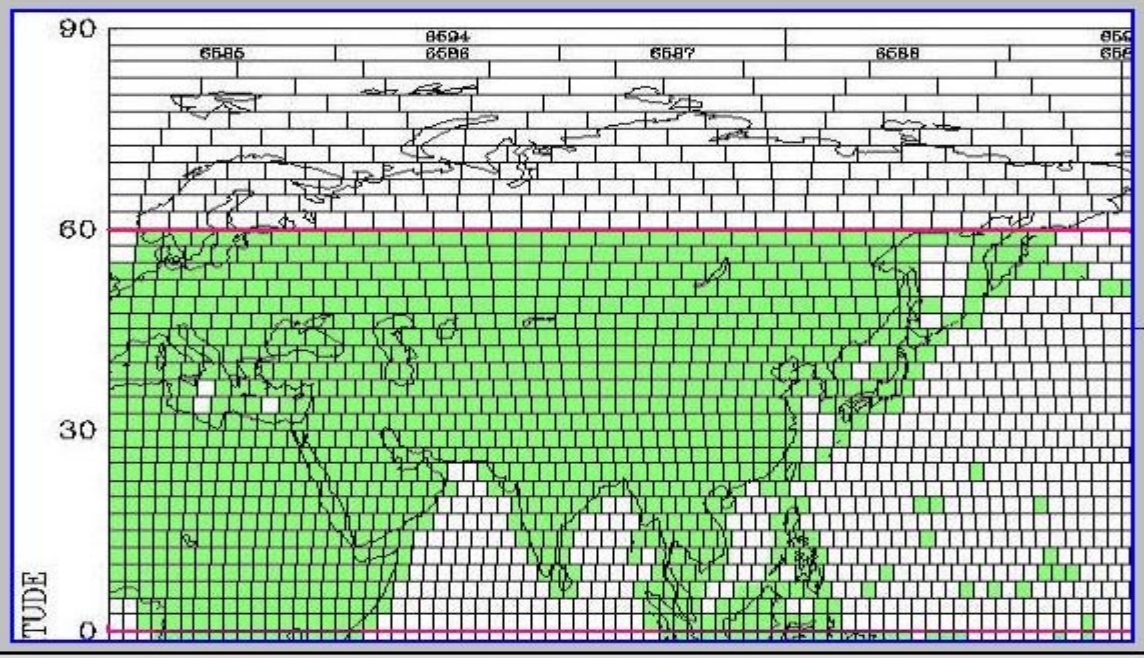
Year	Season	Variable	Transect	CHOICES
1997 1996	Spring Summer Fall	TEMP SAL SIGMAT CHLWSTAR	* tr001 tr002 tr003 Transect Map	Submit Choices Reset



Global Clouds Database

- **Storing cloud information throughout the world**
- **Tabular data - made online through SRB**
- **6,596 grid-cells over the globe**
- **200 variables per grid-cell**
- **Data collected every 3 hours over 4 years (89-92)**
- **Small metadata - stored in Flat file**
- **Each cell dataset (4 yrs data) is stored in SRB (HPSS)**
- **Uses Scommands & srb web utilities**
- **URL: http://www.sdsc.edu/~marciano/clouds/clouds.htm**

Click on desired LAND cell:



Cell INFO results:

Cell No.	Lat. Center	Lon. Center	% Land	DOWNLOAD
3619	96.25	81.82	25	Click Here NOW

3619

```
3619     92/01/01          0          3402     8
3619     92/01/01          3          2680    12
3619     92/01/01          6          2509    15
3619     92/01/01          9           908    42
3619     92/01/01         12          2452     8
3619     92/01/01         15          1791     8
3619     92/01/01         18          2068     4
3619     92/01/01         21          2121     8
3619     92/01/02          0          1433     4
```

Research Activities

- **Data Handling Infrastructure**
  - Parallel I/O Technology - ANL
  - Active Data Repository - UMd
- **Digital Library Infrastructure**
  - InterLib - Stanford, UCB, UCSB, CDL & others
- **Collaborations**
  - GDE Systems
  - NASA
  - ASCII
  - NLANR

Summary

- **Storing, Publishing, Sharing & Cooperating**
- **Distributed, Replicated, Heterogeneous Data Cache**
- **Unified access to Archival Storage, Database Storage, Disk Storage**
- **Information Discovery (application-level metadata)**
  - unifying meta-catalogs (future work)
- **Secure, encrypted controlled access & data movement**
  - integrate with other security systems (future work)
- **Scalability and performance modeling needed (see next slide…)**

Capacity Planning for MSS

- **Workload characterization**
  - trace analysis, get/put log files, resource monitoring, clustering
- **Baseline model creation**
  - analytical models
  - simulation models
  - others (prototype, numerical/statistical, hybrid)
- **Prediction model creation**
- **Examples:**
  - Analytical modeling:
    - Queuing Models of Tertiary Storage (Ted Johnson)
    - Analytical Performance of Hierarchical MSS (Daniel Menasce et al.)