Rights Management for Shared Collections

Reagan W. Moore
moore@sdsc.edu
http://www.sdsc.edu/srb

Abstract

- **National and international collaborations create shared collections that span multiple administrative domains.**
- **Shared collection are built using:**
  - Data virtualization, the management of collection properties independently of the storage system.
  - Trust virtualization, the ability to assert authenticity and authorization independently of the underlying storage resources.
- **What rights management are implied in trust virtualization?**
- **How are data integrity challenges met?**
- **How do you implement a “Deep Archive”?**

State of the Art Technology

- **Grid - workflow virtualization**
  - Manage execution of jobs (processes) independently of compute servers
- **Data grid - data virtualization**
  - Manage properties of a shared collection independently of storage systems
- **Semantic grid - information virtualization**
  - Reason across inferred attributes derived from multiple collections.

Shared Collections

- **Purpose of SRB data grid is to enable the creation of a shared collection between two or more institutions**
  - Register digital entity into the shared collection
  - Assign owner, access controls
  - Assign descriptive, provenance metadata
  - Manage audit trails, versions, replicas, backups
  - Manage location mapping, containers, checksums
  - Manage verification, synchronization
  - Manage federation with other shared collections
  - Manage interactions with storage systems
  - Manage interactions with APIs

Trust Virtualization

- **Collection owns the data that is registered into the data grid**
  - Data grid is a set of servers, installed at each storage repository
  - Servers are installed under a SRB ID created for the shared collection
  - All accesses to the data stored under the SRB ID are through the data grid
  - Authentication and authorization are controlled by the data grid, independently of the remote storage system

Rights Management

- **Shared collection approach**
  - Manage authentication of all users who will have privileges beyond that of public “read”
  - Map users to groups
  - Provide access controls on users and groups
- **Virtual Organization Management approach (VOM)**
  - Authenticate users, and provide certificate asserting membership in a group
  - Manage access controls on groups
  - Provide rules for access based on membership in a group

Authentication / Authorization

- Collection owned data
  - At each remote storage system, an account ID is created under which the data grid stores files
- User authenticates to the data grid
- Data grid checks access controls
- Data grid server authenticates to a remote data grid server
- Remote data grid server authenticates to the remote storage repository
- SRB servers return the data to the client

Authentication Mechanisms

- **Grid Security Infrastructure**
  - PKI certificates
- **Challenge-response mechanism**
  - No passwords sent over network
- **Ticket**
  - Valid for specified time period or number of accesses
- **Generic Security Service API**
  - Authentication of server to remote storage

Trust Implementation

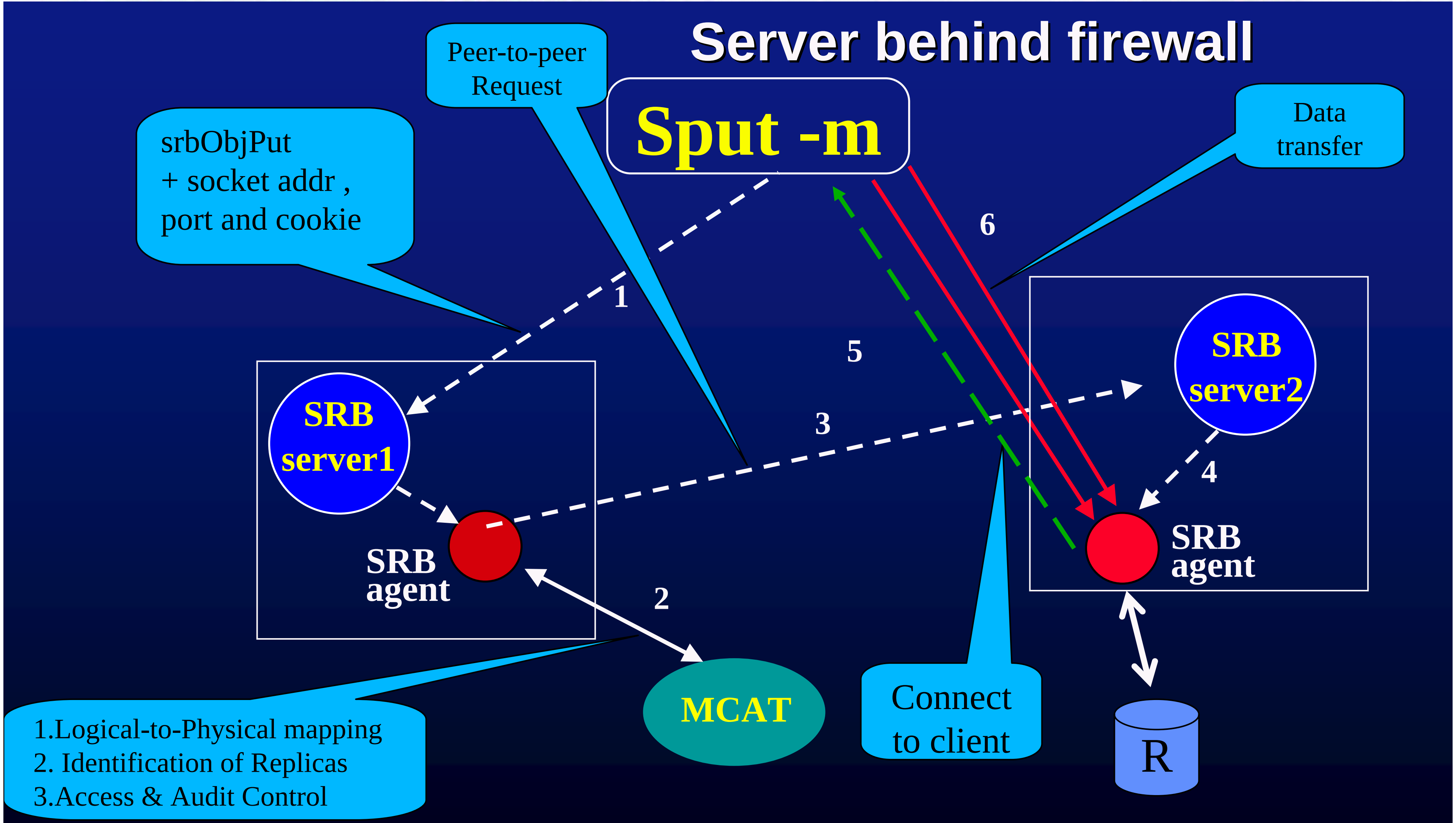
- **For authentication to work across multiple administrative domains**
  - Require collection-managed names for users
- **For authorization to work across multiple administrative domains**
  - Require collection-managed names for files
- **Result is that access controls remain invariant. They do not change as data is moved to different storage systems under shared collection control**

Network Devices

- **Access to data must handle security requirements of network devices**
  - Firewalls - typically requires access be initiated from within a firewall
  - Network routers - location of data may change based on load leveling
  - Private virtual network - location of data not known explicitly known
- **Handled by creating network transport protocols to meet requirements of each network device**

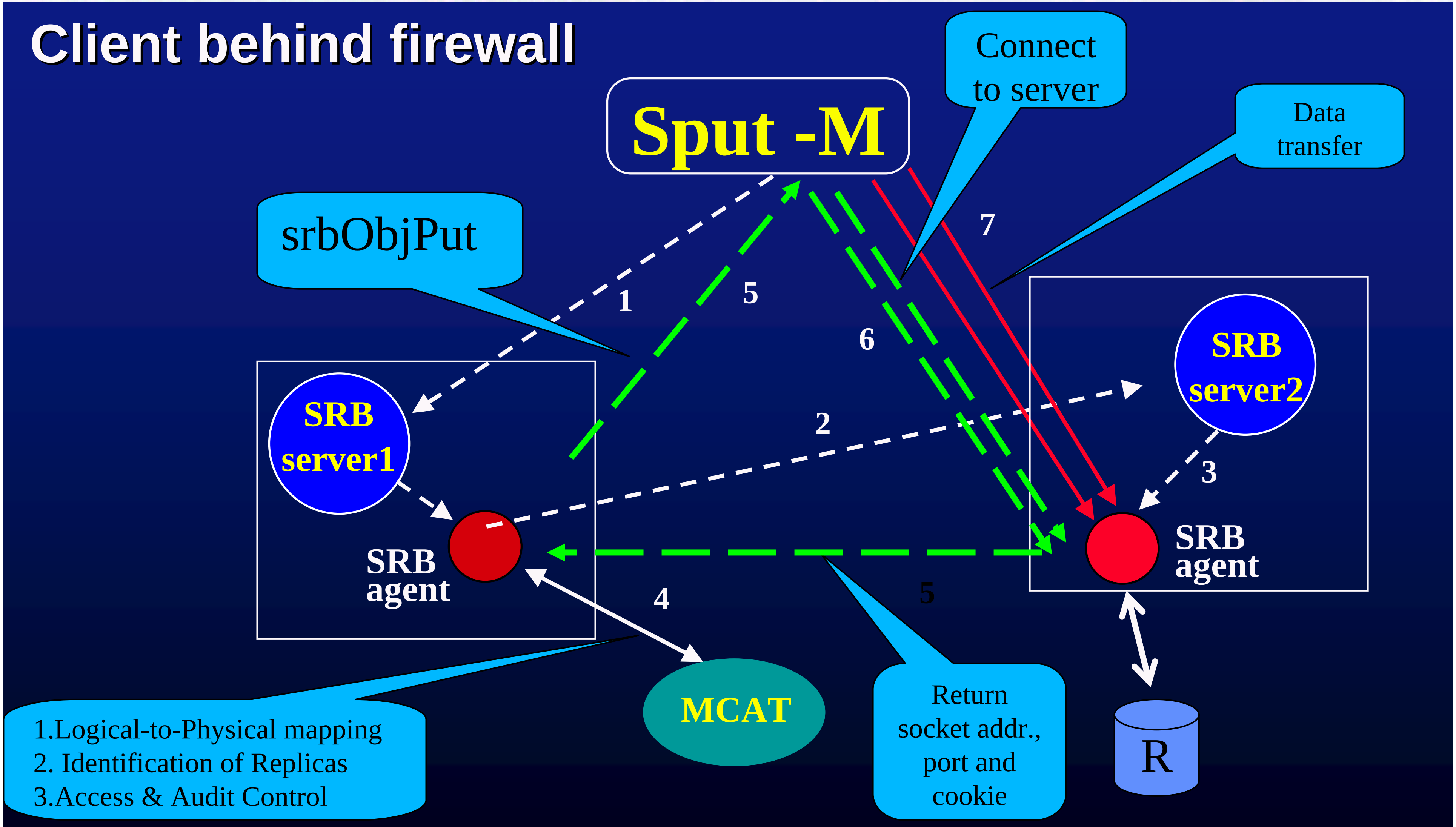
Parallel mode Data Transfer – Server Initiated

Server behind firewall



Parallel mode Data Transfer – Client Initiated

Client behind firewall



1.Logical-to-Physical mapping
2. Identification of Replicas
3.Access & Audit Control

UCSD

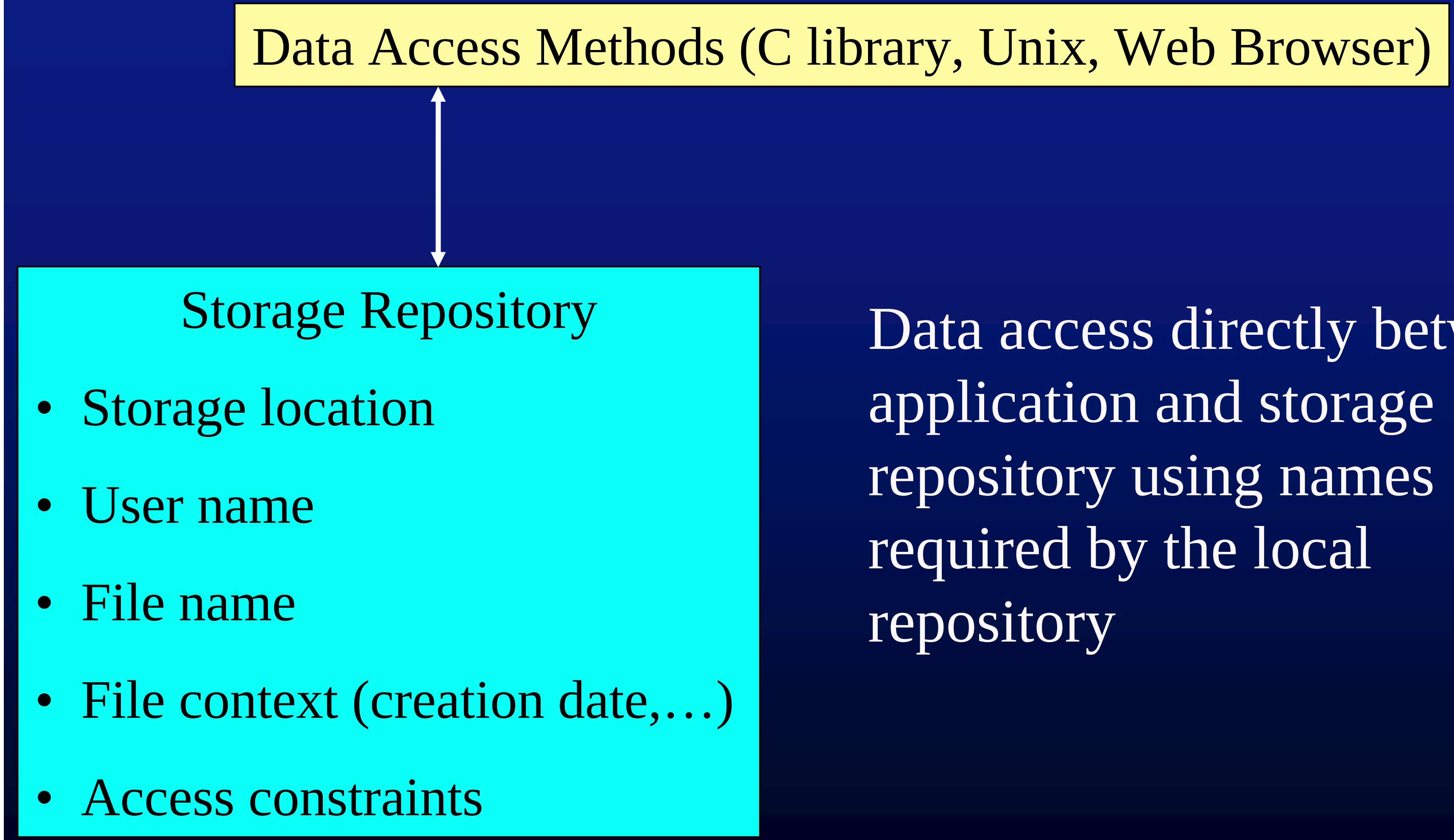
SDSC

HIPAA Patient Confidentiality

- Access controls on data
- Access controls on metadata
- Access controls on storage systems
- Audit trails
- End-to-end encryption (manage keys)
- Localization of data to specific storage

- Access controls do not change when data is moved to another storage system under data grid control

Logical Name Spaces



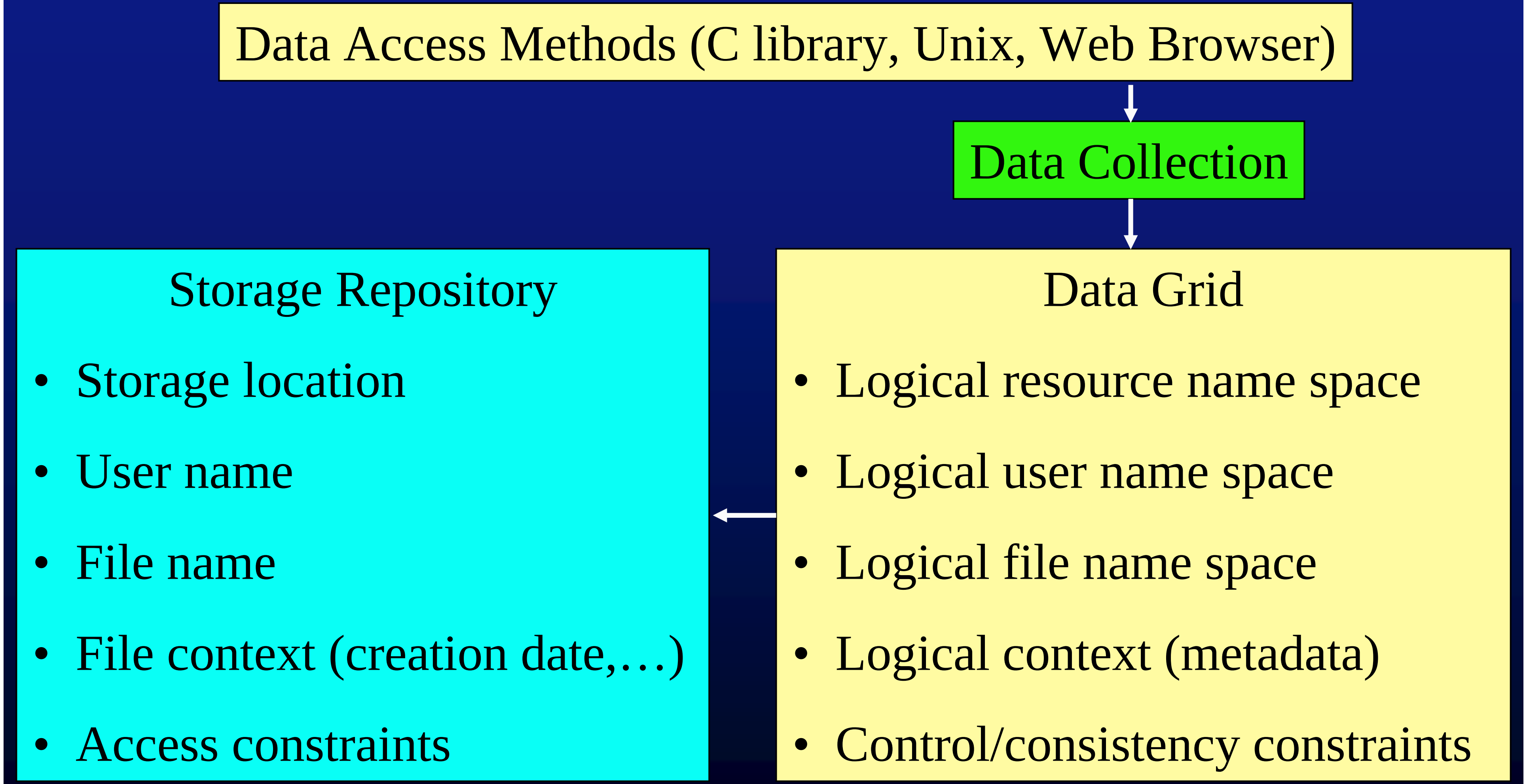
Data Access Methods (C library, Unix, Web Browser)

Storage Repository

- Storage location
- User name
- File name
- File context (creation date,…)
- Access constraints

Data access directly between application and storage repository using names required by the local repository

Logical Name Spaces



Data Access Methods (C library, Unix, Web Browser)

Data Collection

Storage Repository

- Storage location
- User name
- File name
- File context (creation date,…)
- Access constraints

Data Grid

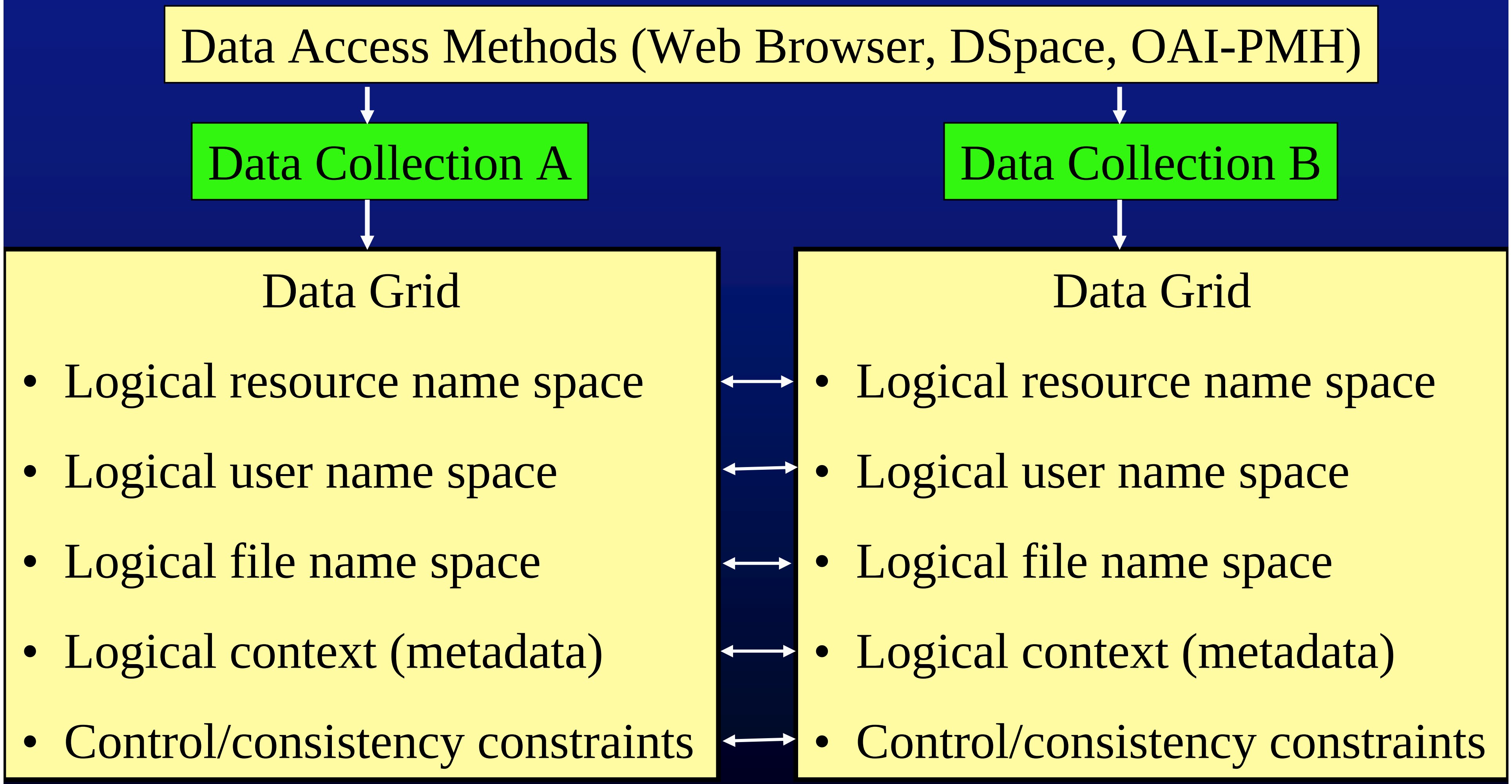
- Logical resource name space
- Logical user name space
- Logical file name space
- Logical context (metadata)
- Control/consistency constraints

Data is organized as a shared collection

Logical Resource Names

- **Logical resource name represents multiple physical resources**
- **Writing to a logical resource name can result in:**
  - Replication - write completes when each physical resource has a copy
  - Load leveling - write completes when the next physical resource in the list has a copy
  - Fault tolerance - write completes when “k” of “n” resources have a copy
  - Single copy - write is done to first disk at same IP address, then disk anywhere, then tape

Federation Between Data Grids



Data Access Methods (Web Browser, DSpace, OAI-PMH)

Data Collection A

Data Collection B

Data Grid

- Logical resource name space
- Logical user name space
- Logical file name space
- Logical context (metadata)
- Control/consistency constraints

Data Grid

- Logical resource name space
- Logical user name space
- Logical file name space
- Logical context (metadata)
- Control/consistency constraints

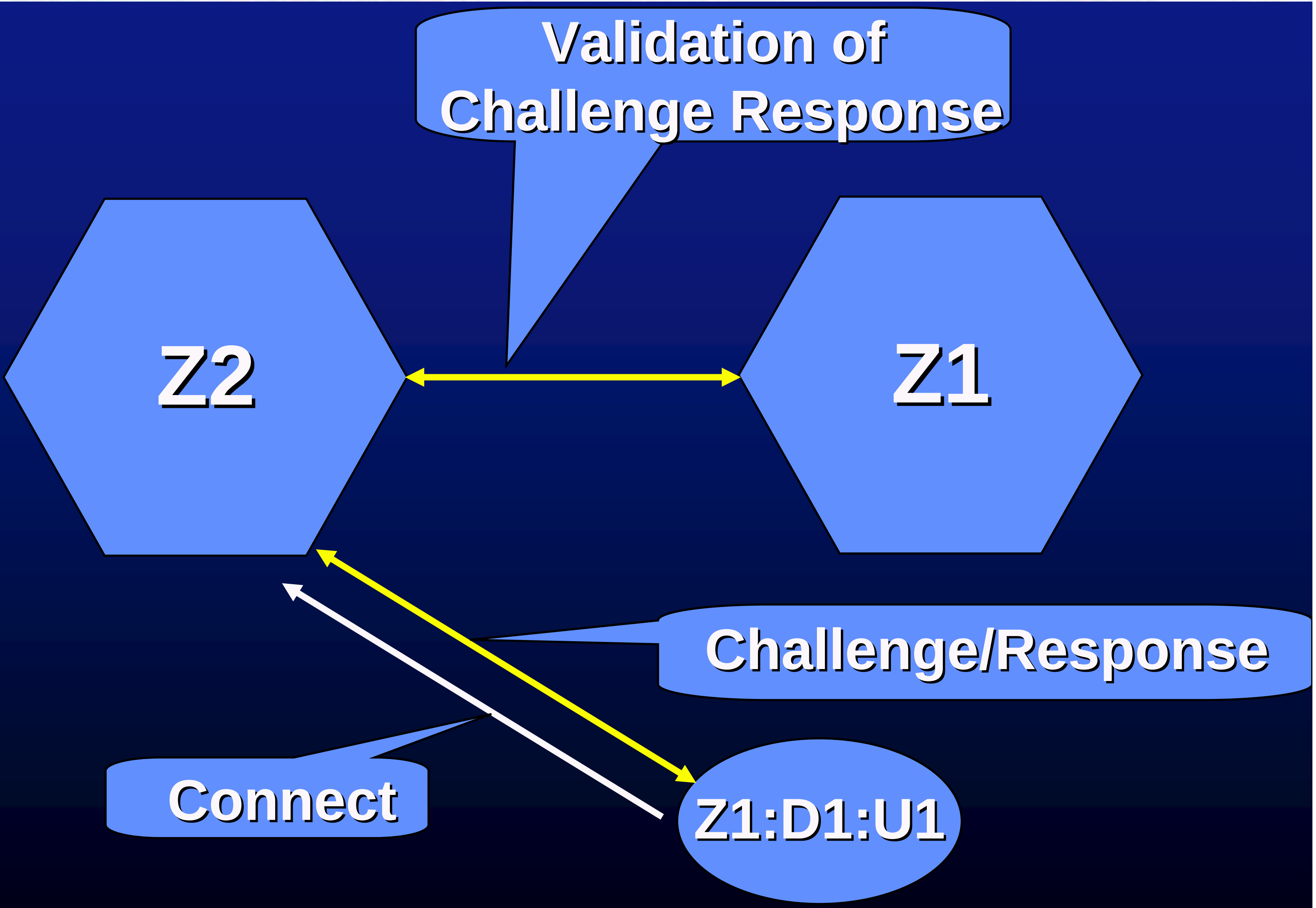
Access controls and consistency constraints
on cross registration of digital entities

Authentication across Zones

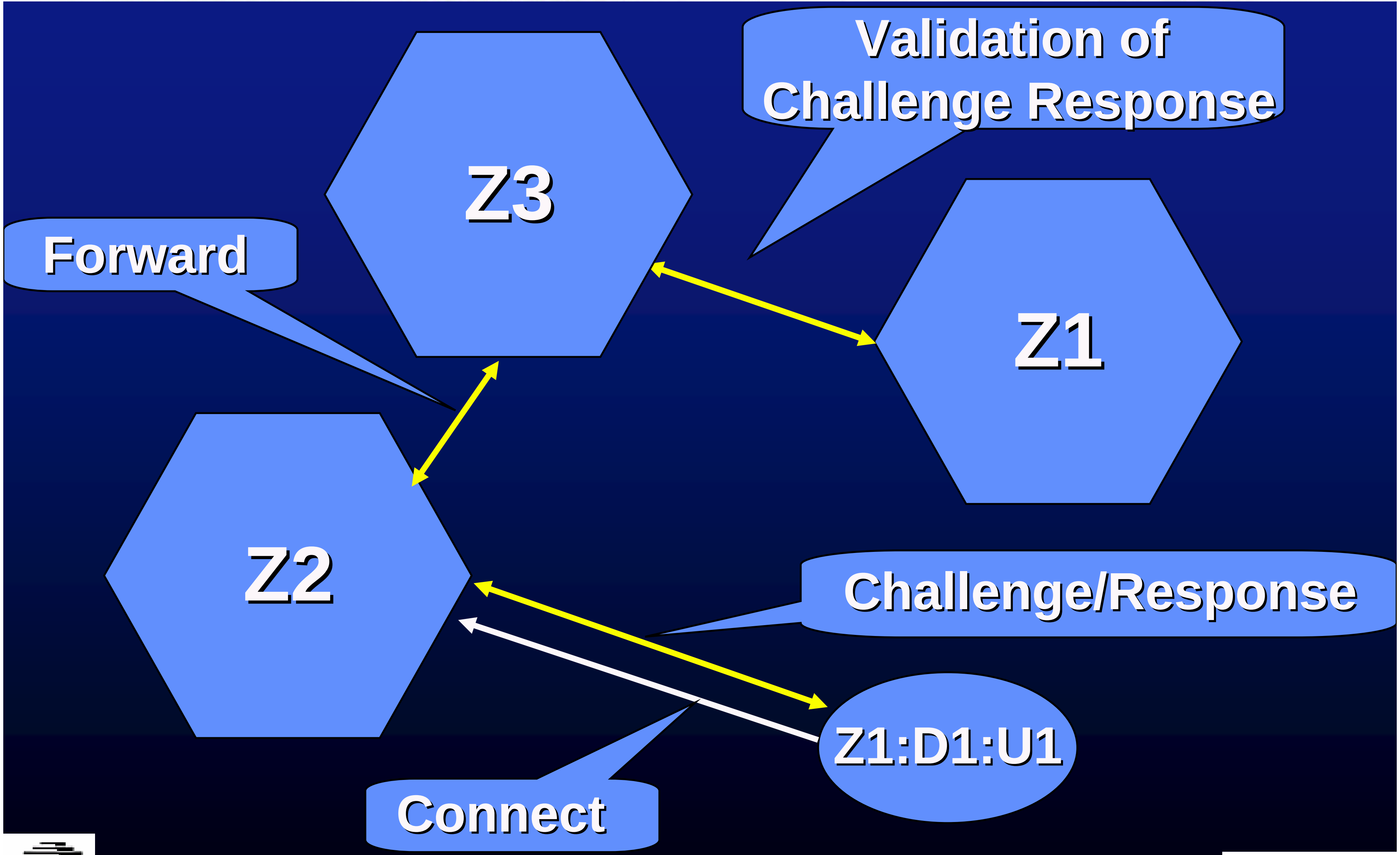
- **Follow Shibboleth model**
- **A user is assigned a “home” zone**
  - User identity is

    Home-zone:Domain:User-ID
- **All authentications are done by the “home” zone**

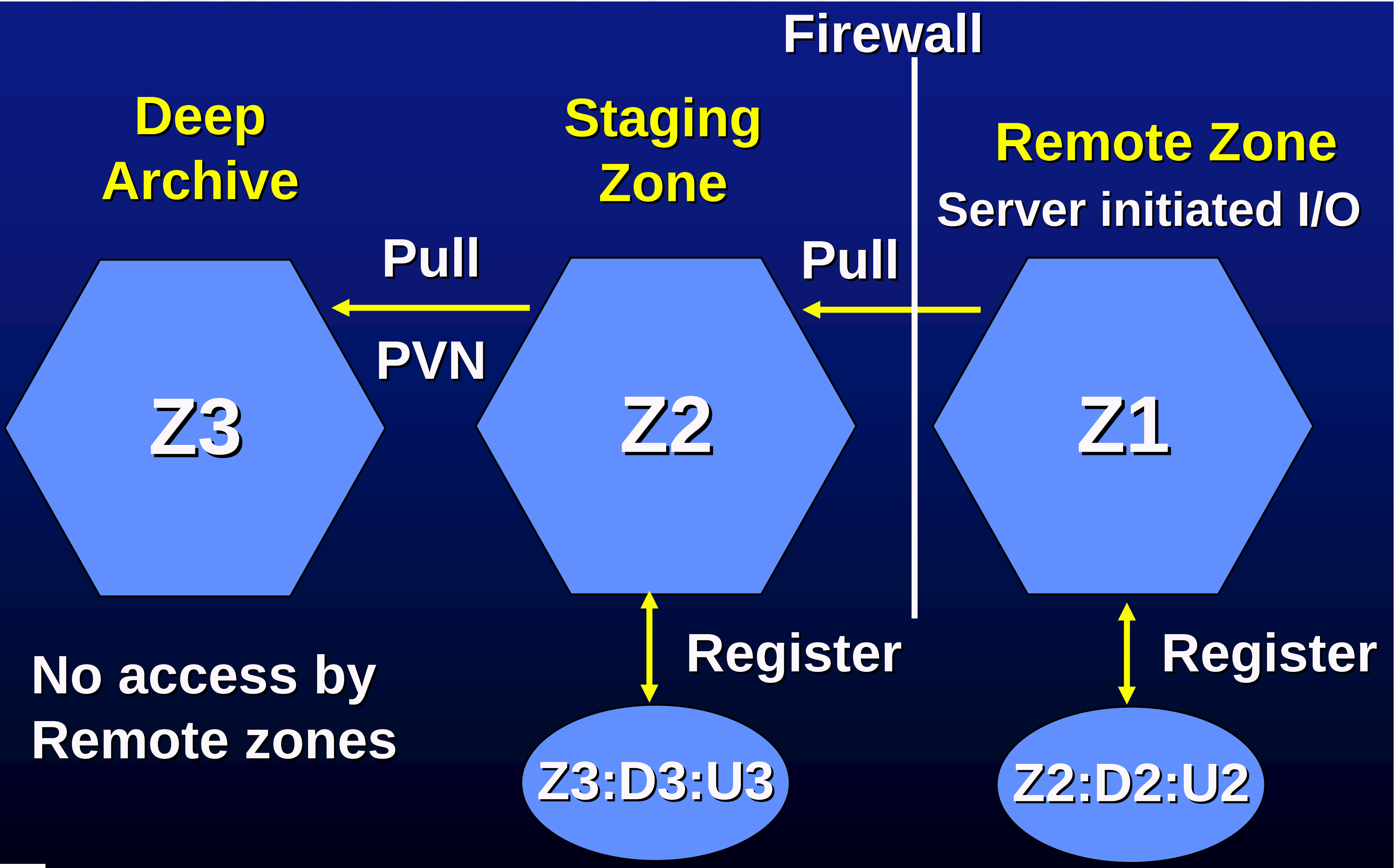
User Authentication



User Authentication



Deep Archive

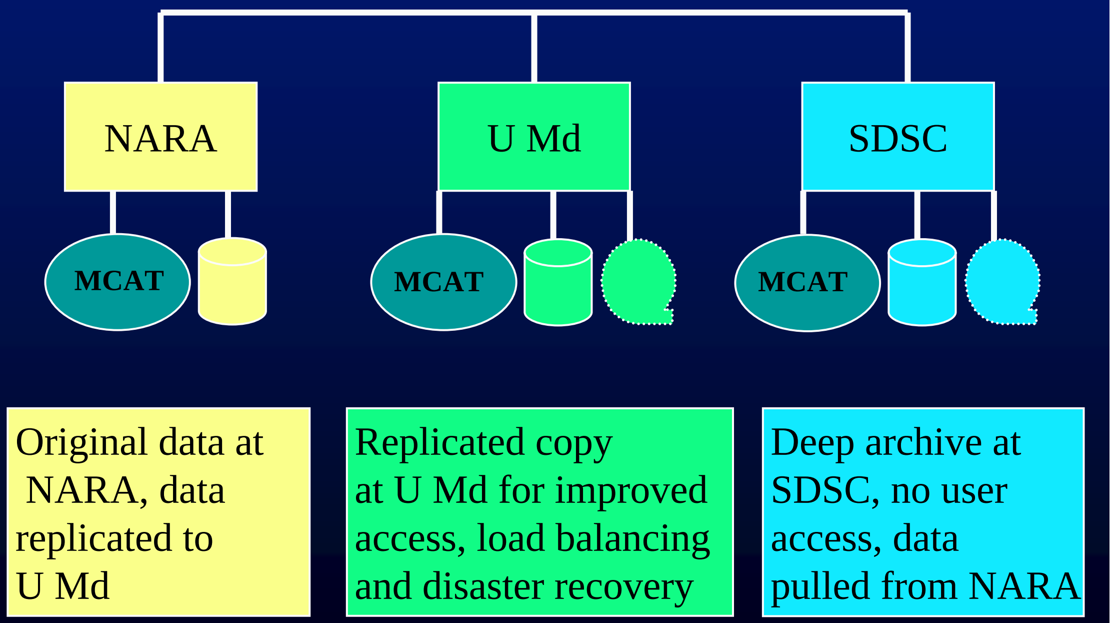


NARA Persistent Archive

Demonstrate preservation environment

- Authenticity
- Integrity
- Management of technology evolution
- Mitigation of risk of data loss
  - Replication of data
  - Federation of catalogs
- Management of preservation metadata
- Scalability
  - Types of data collections
  - Size of data collections

Federation of Three Independent Data Grids



Original data at NARA, data replicated to U Md

Replicated copy at U Md for improved access, load balancing and disaster recovery

Deep archive at SDSC, no user access, data pulled from NARA

For more Information on Storage Resource Broker Data Grid

Reagan W. Moore

moore@sdsc.edu

http://www.sdsc.edu/srb/