This Presentation: http://libnet.ucsd.edu/nara/2005.04.15_DLF.ppt

200,000 Images A DSpace / SRB Use Case

Chris Frymann – cfrymann@ucsd.edu
University of California, San Diego Libraries

Digital Library Federation Meeting
San Diego, California
April 15, 2005

- Grant from:

  The National Archives and Records Administration (NARA)

- Collaboration with:

  San Diego Super Computer Center (SDSC)

  Massachusetts Institute of Technology (MIT)

Primary Goals

- Preservation
- Reusable (ETL) procedures

  Extraction Transformation and Loading
- Cross-collection discovery and access

The Collection

- 200,000 35mm slides

  associated MARC records in local ILS

- 200,000 TIFF files

  20 MB / file

  4 Terabytes

DSpace

- Needs no introduction

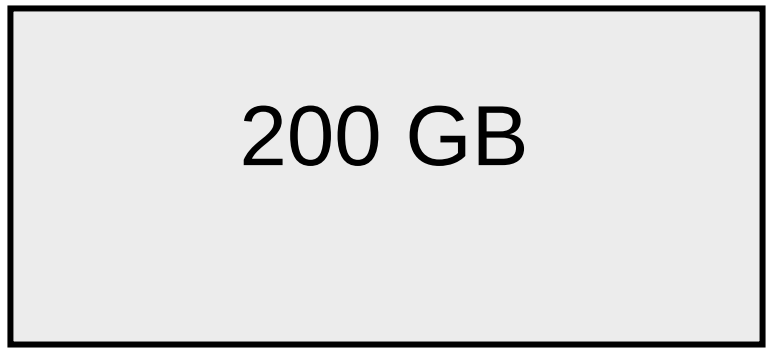
SRB

- Storage Resource Broker
- Developed at:
  San Diego Supercomputer Center

SRB

- Server software & programming interfaces (middleware)
- Enables applications that store and retrieve files
  to treat multiple and heterogeneous storage devices
  as a single logical resource
- Over the network this qualifies as “grid” technology

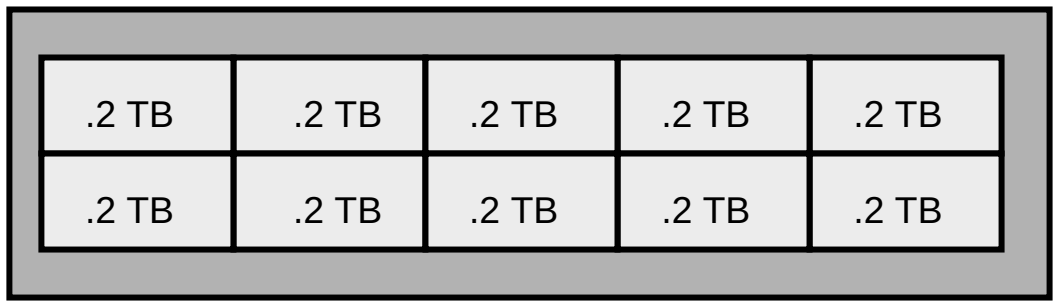
Basic Storage Resource



Inexpensive commodity disk drive

Storage Resource

10 drives
2 Terabytes/box
Grid Brick

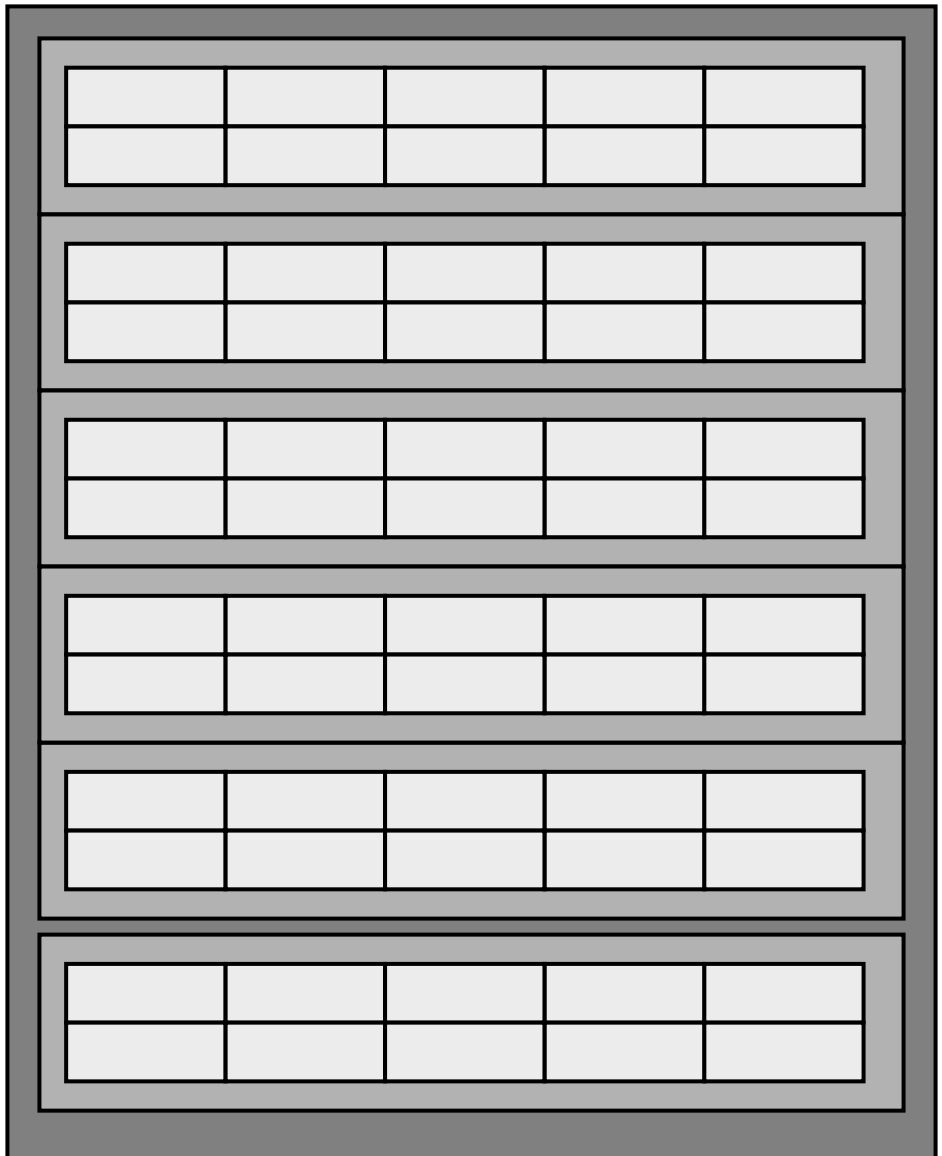


Rackmount Storage Server

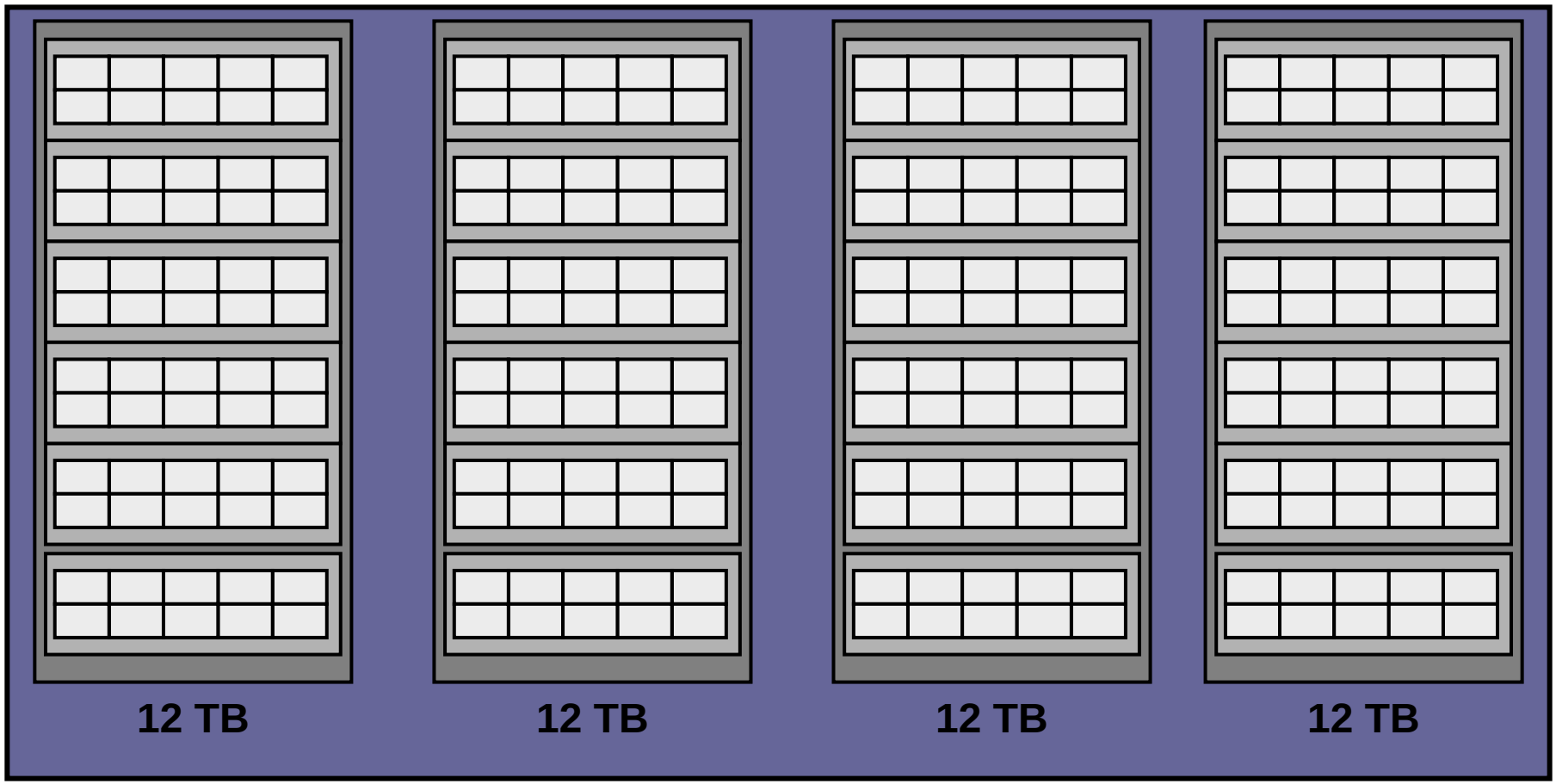
SRB lets us treat it as a single logical resource

Single Logical Resource – 12 TB

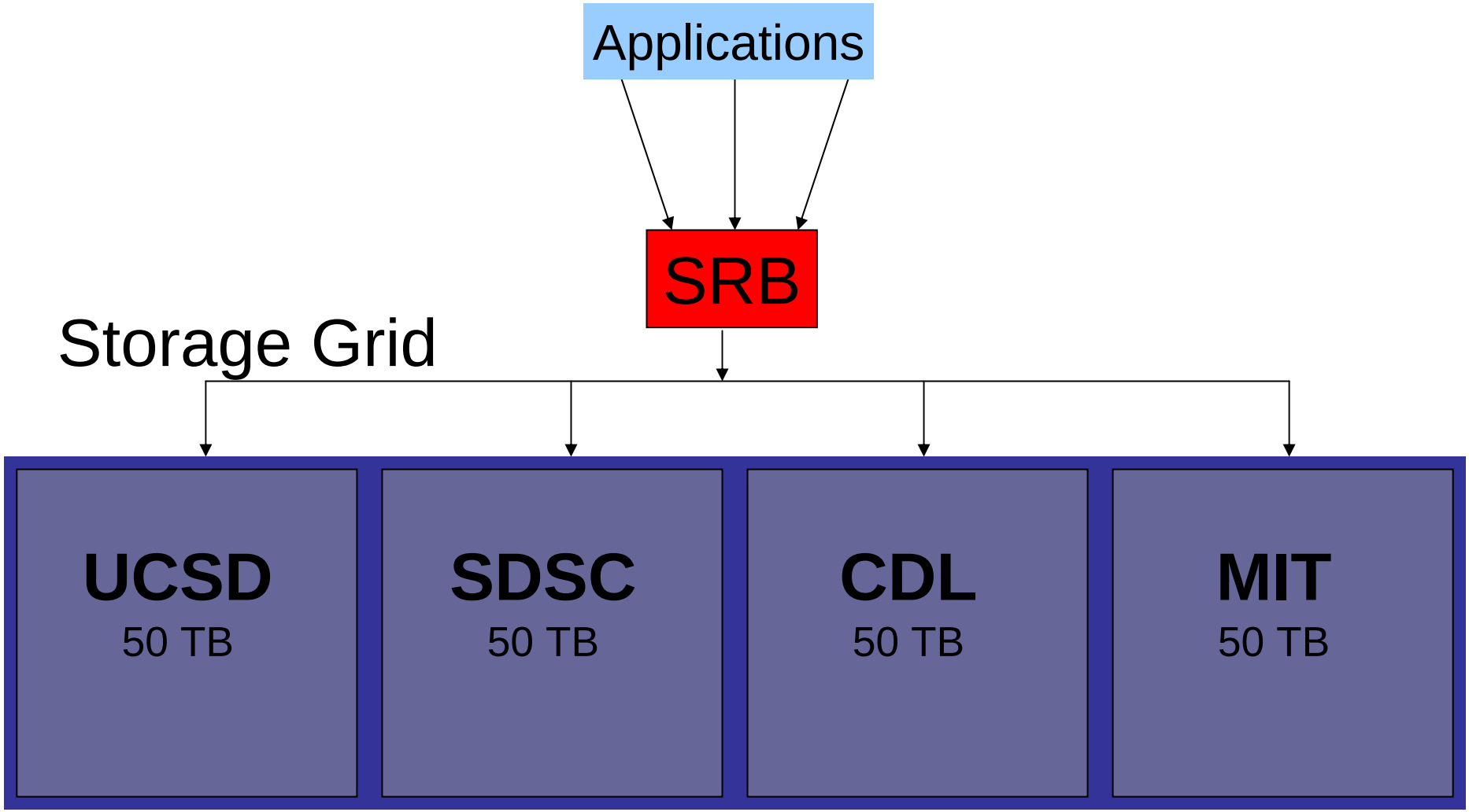
Rack of
Storage Servers
Grid Bricks



Single Logical Resource ~ 50 TB



Room of Racks



200 TB

Single Logical Resource

Approach

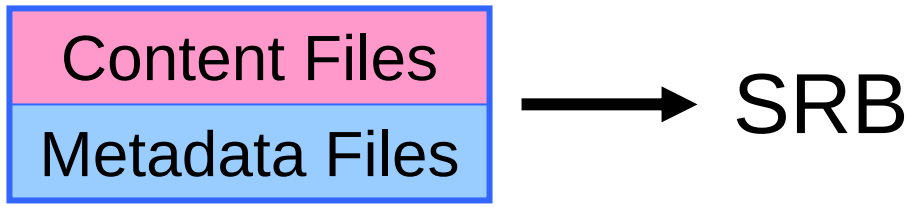
- Use SRB for
  - Economical storage
  - Grid-based replication
- Use DSpace for Digital asset discovery and access
- Modifiy Code to integrate DSpace and SRB
- Develop batch processes for ingesting into DSpace/SRB

Initial Focus on Preservation

Enabled us to think in terms of:

- “Dark Archive”
- Asset Store
- AIP

AIP



- The AIP requires us to address:
  - Metadata Encapsulation
  - File Naming

File Naming Requirements

- Generated Automatically
- Unique
- Semanticly opaque
- Bind content and metadata files
- **Consistent with CDL approach**
  - Archival **R**esource **K**ey - **ARK**

ARK Used for SRB File Naming

- Every digital object

  and all sub-components

  assigned names with common ARK-base

Details of ARK-based File Naming in SRB

- **Thanks to John Kunze for developing this approach**
- General form
  - ark:/NAAN/Name/NAAN-Name-ServiceComponent.Vnnn.Format
- Where
  - NAAN = Name Assignment Authority Number
    = 20775 for object named by UCSD
  - Name = ARK generated according to specified template
    e.g. [bb] [7 random digits] [checksum character]
  - ServiceComponent = string identifying a part or aspect of the object
    e.g. master, metadata-mets
  - Vnnn = version number; zero-padded positive integer of 3 or more digits
  - Format = mime-type format designator
- Example
  - ark:/20775/bb1234567k/20775-bb1234567k-master.v001.tif
  - ark:/20775/bb1234567k/20775-bb1234567k-metadata-mets.xml

ARKs Also Used in Implementing Actionable URLs

- Every digital object

and all sub-components

assigned URL with common ARK base

Details of ARK Assignment in Actionable URLs

- Prefix
  - http://libraries.ucsd.edu/
- Actionable reference to:
  - Object (item)
    - http://libraries.ucsd.edu/ark:/20775/bb1234567k
  - Component file (bit stream)
    - http://libraries.ucsd.edu/ark:/20775/bb1234567k/20775-bb1234567k-master.v001.tif

Integration of DSpace & SRB Introduces Multiple Layers of Name Indirection

- SRB
  - Physical
  - Logical
- DSpace
  - Physical name
  - Local handle
  - Global Handle

The AIP – Part II

- Metadata encapsulation

  and the obvious choice is …

METS

- Minimal mandatory metadata requirements (“low floor”)
- Support for almost unlimited complexity (“high ceiling”)
- Relational database independent
- File system oriented
- XML
- **Required for ingestion into:**
  CDL Digital Preservation Repository (DPR)

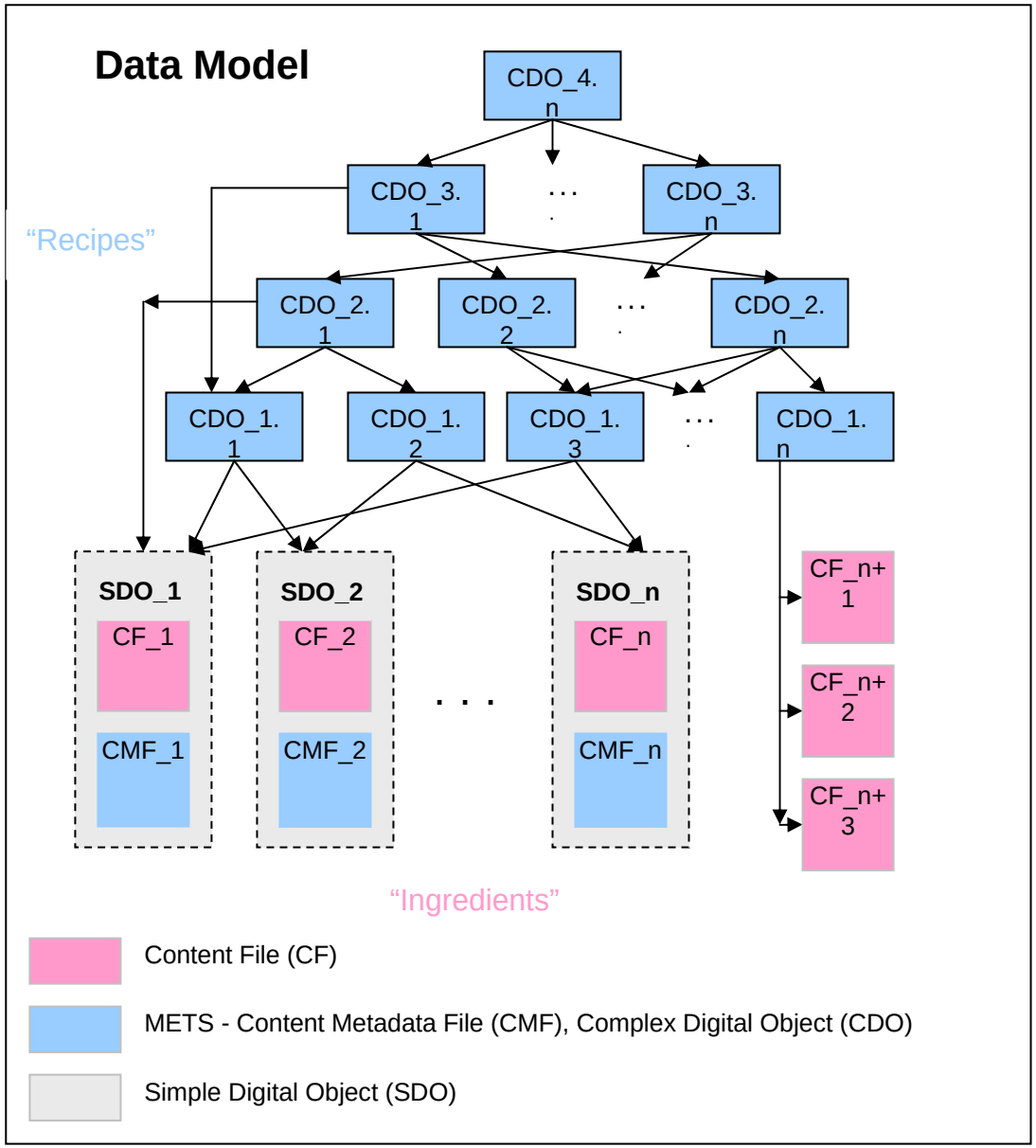
METS Profile

- Developed and refined over many months
- Used to submit objects to CDL DPR
- Ready for registration at LOC

<?xml version="1.0" encoding="UTF-8" ?> <!-- edited by Bradley D. Westbrook, Digital Library Program, University of California, San Diego. With the kind assistance of Rick Beaubien, Robert Dias, and Gabriela Montoya -->
- <METS_Profile xmlns:xsi="**http://www.w3.org/2001/XMLSchema-instance**" xsi:noNamespaceSchemaLocation="**http://www.loc.gov/standards/mets/profile_docs/mets.profile.v1-1.xsd**">
<URI LOCTYPE="**URL**">**http://???.ucsd.edu/mets/profiles/UCSD Single Still Image Profile**</URI>
<title>**UCSD Single Still Image Profile**</title>
<abstract>**UCSD digital objects composed of a single image use this METS profile. Multiple versions of the image may be included in a METS record conforming to this profile, but only one version is required. The profile does not prescribe a file format for the version(s), but it is suggested that the format of one file generally be of an archival quality, e.g., a tiff or high resolution jpeg.**</abstract>
<date>**2005-01-21T11:42:31**</date>
- <contact>
<name>**Digital Library Program Office**</name>
<address>**Geisel Library, UC, San Diego**</address>
<email>**DigitalLibraryProgram@ucsd.edu**</email>
</contact>
<related_profile RELATIONSHIP="**controlled vocabularies for USE attribute values and TYPE attribute values taken from**" URI="**http://www.loc.gov/standards/mets/profiles/00000004.xml**">**Model Imaged Object Profile**</related_profile>
- <extension_schema>
<name>**Metadata Object Description Schema (MODS)**</name>
<URI>**http://www.loc.gov/standards/mods/v3/mods-3-0.xsd**</URI>
<context>**mets/dmdSec/mdWrap/xmlData**</context>
<note>**Used for descriptive metadata representing the object.**</note>
</extension_schema>
- <extension_schema>
<name>**NISOIMG**</name>
<URI>**http://www.loc.gov/standards/mix/mix.xsd**</URI>
<context>**mets/amdSec/techMD/mdWrap/xmlData**</context>
<note>**Used for technical metadata about the characteristics, origin, and modification of the content file.**</note>
</extension_schema>
- <extension_schema>
<name>**METSRights**</name>
<URI>**http://cosimo.stanford.edu/sdr/metsrights.xsd**</URI>
<context>**mets/amdSec/rightsMD/mdWrap/xmlData**</context>
<note>**Used for recording intellectual property rights.**</note>
</extension_schema>
- <description_rules>
All applications of MODS in UCSD METS records adhere to the MODS User Guidelines published by the Library of Congress's Network Development and MARC Standards Office.
</description_rules>

Data Model

- Paired Content and Metadata Files
  with ARK-based names
- Metadata encoded in “standard” METS profiles
- Stand-alone METS files
  describing arbitrary levels of aggregation
  of lower level objects

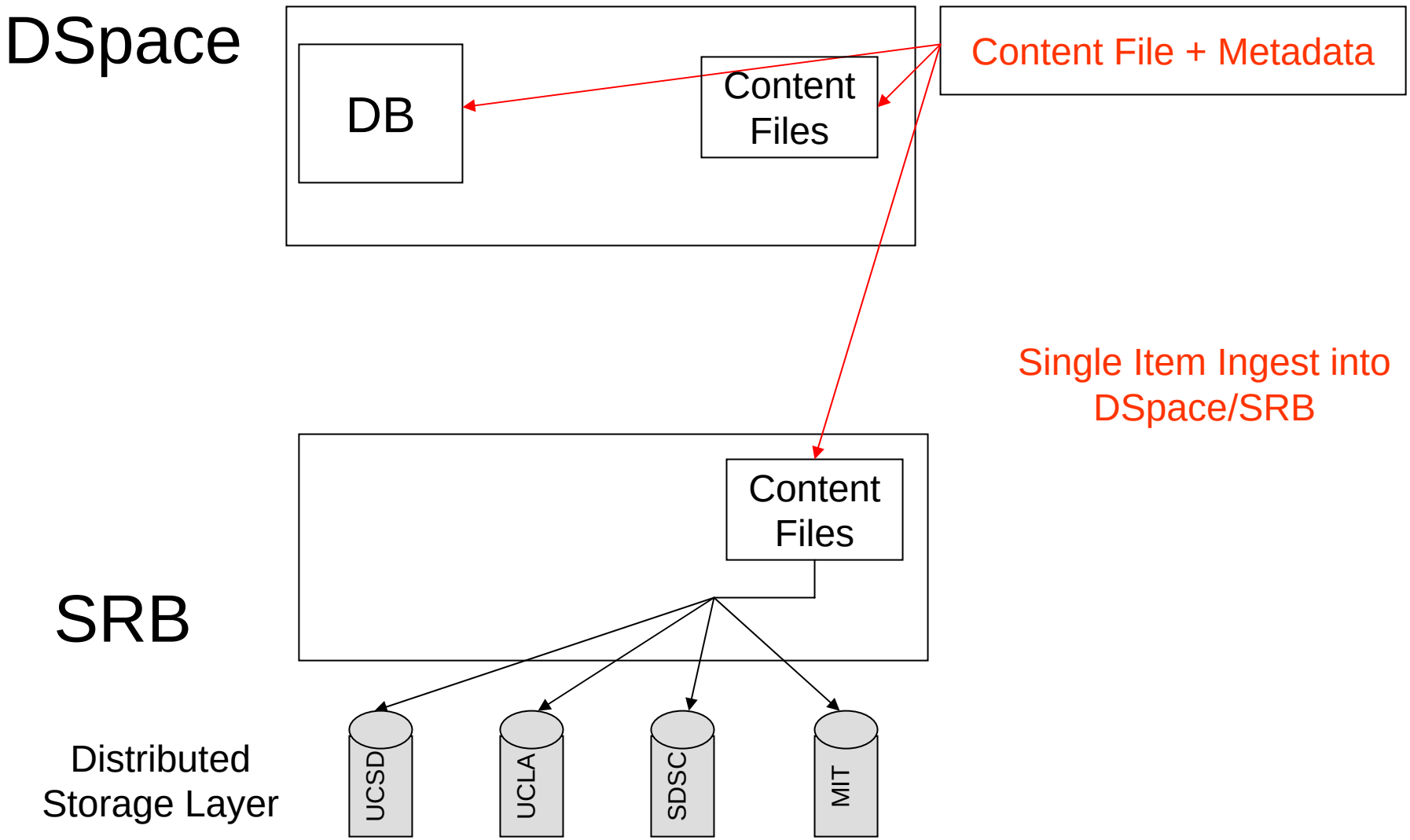

Data Model
CDO_4.n
CDO_3.1
CDO_3.n
"Recipes"
CDO_2.1
CDO_2.2
CDO_2.n
CDO_1.1
CDO_1.2
CDO_1.3
CDO_1.n
SDO_1
CF_1
CMF_1
SDO_2
CF_2
CMF_2
SDO_n
CF_n
CMF_n
CF_n+1
CF_n+2
CF_n+3
"Ingredients"
Content File (CF)
METS - Content Metadata File (CMF), Complex Digital Object (CDO)
Simple Digital Object (SDO)

DSpace/SRB Code Integration

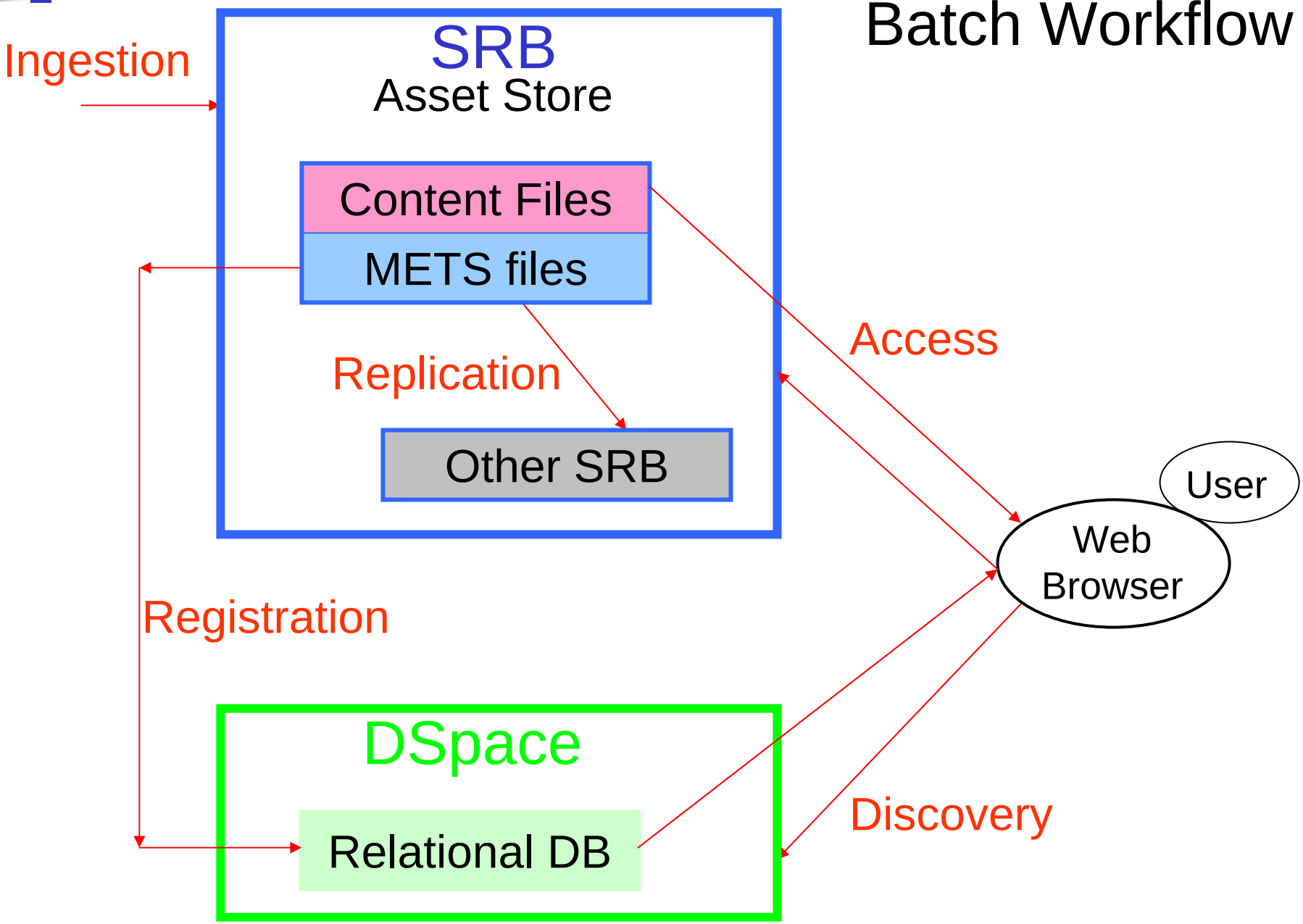
1. Replace DSpace file system calls
   with SRB access calls

2. Augment DSpace ItemImporter
   “register” SRB objects into DSpace

Single Item Workflow



Batch Workflow



DSpace 1.3 Code Patches

- March 17 - Submitted to Sourceforge
- April 8 - Accepted by DSpace committers

Extraction Transformation and Loading (ETL) Processes

- Load data into file staging area:
  - Extracted MARC record data from ILS
  - Vendor digitized TIFF files from 38 120 GB hard drives
- Create temporary staging database and insert all data needed to generate METS files:
  - MARC record data
  - Technical metadata from digitization vendor spreadsheets
  - Checksums
  - ARK names generated from NOID
- Use staging database to control repetitive transfer of objects to permanent Asset Store (SRB)
  - Transfer TIFF file to SRB and assign it an ARK-based name
  - Transfer METS file to SRB and assign it a paired ARK-based name
  - Update record status fields in staging database as steps are completed
- Use XSLT transformation to generate “DSpace Qualified Dublin Core” files from METS
- Register DS QDC files into DSpace
  - Use modified DSpace ItemImporter
    - Achieves results of Single item retrieval modifications to standard DSpace
- Use SRB-to-SRB copy to replicate at SDSC
- Ingest into CDL DPR
  - Common ARK-based naming
  - Possible SRB-to-SRB replication
- Continuous synchronization

Load Data into File Staging Area

- MARC records extracted from ILS
- 38 120 GB hard drives
  with vendor digitized TIFF files

Load Staging Database

- Includes everything needed to generate METS files:
  - MARC record data
  - Technical metadata from digitization vendor
  - Checksums
  - ARKs minted from John Kunze's NOID script

Transfer Data to Asset Store

- Staging database governs repetitive transfer of objects to permanent Asset Store (SRB)
  - Transfer TIFF file to SRB, assign ARK-based names
  - Transfer METS file to SRB, assign paired ARK-based name
  - Update record status fields in staging database
  - This transfer took nine days

Transfer Metadata to DSpace

- Use XSLT transform to generate "DSpace Qualified Dublin Core" files from METS
- Use ItemImporter to register SRB-based AIP

Last Step
Preservation Copies

- Do SRB-to-SRB replication at SDSC
- Do replication to CDL DPR
  - Java API
  - Possible SRB-to-SRB copy

Summary

- 200,000 digital objects preserved, discoverable and accessible
  - Asset Store with METS/ARK-based AIP
  - Repurposeable automated workflow processes
  - DSpace enabled discovery and retrieval
  - SRB enabled storage and grid integration

- Project website:

  http://libnet.ucsd.edu/nara

- This presentation:

  http://libnet.ucsd.edu/nara/2005.04.15_DLF.ppt