Overview of the Chronopolis Digital Preservation Framework

SAN DIEGO SUPERCOMPUTER CENTER

Robert H. McDonald

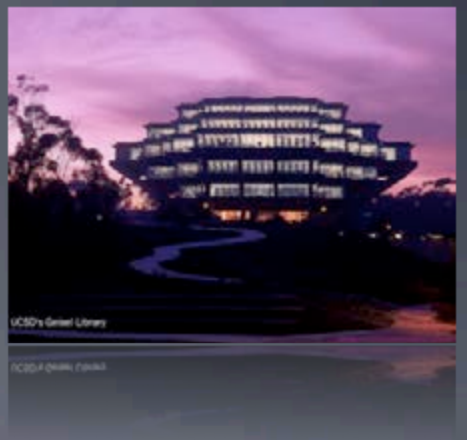
Director, Strategic Data Alliances

Digital Preservation Initiatives Group

San Diego Supercomputer Center

UC, San Diego Libraries

University of California, San Diego



UC SAN DIEGO LIBRARIES

UNIVERSITY OF MARYLAND INSTITUTE FOR ADVANCED COMPUTER SCIENCE

NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

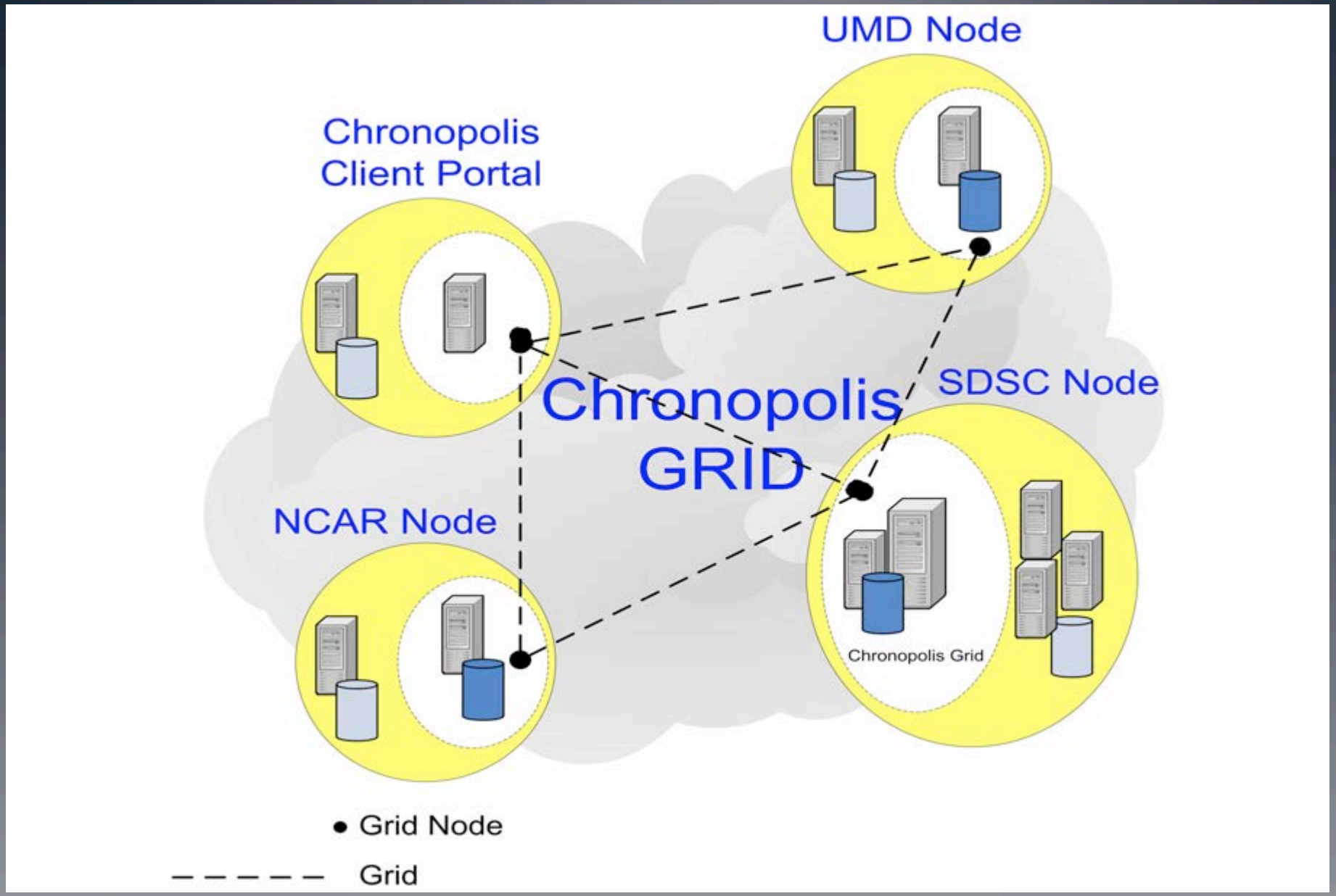
Outline

- Chronopolis a DataGrid Approach
- Partnerships and Collaborations
- Library of Congress and NDIIPP
- Current Status
- Future Initiatives

Grid Computing

- **Grid Computing** - is a component of distributed computing that allows researchers to harness the power of distributed computers, data storage systems and networks to create virtual supercomputers. Leveraging the power of many components to create one massive computational infrastructure.

- **Grid Examples**
  - **TeraGrid** – NSF sponsored grid computing framework for open scientific discovery combining leadership class resources at eleven partner sites to create an integrated, persistent computational resource.
  - **BIRN** – Biomedical Informatics Research Network - NIH sponsored grid.
  - **e-Science Grid** (UK) – Research Councils UK


UMD Node
Chronopolis
Client Portal
Chronopolis
GRID
SDSC Node
NCAR Node
Chronopolis Grid
Grid Node
Grid

Grid vs Cloud

- **Cloud Computing/Storage** – usually refers to grid resources that have been commoditized for commercial services and are sometimes referred to as utility or commodity computing. This can refer to compute power for business transactions or for mass-scale storage.

Popular Examples of DGs/Clouds

- **DataGrid Software**
  - Oracle Coherence
  - GigaSpaces XAP
  - IBM WebSphere ObjectGrid
  - Storage Resource Broker (SRB) (SDSC)
  - iRODS (Rule Oriented Data System) (SDSC) (Open-Source)
- **Compute/Storage Clouds**
  - Google Infrastructure
  - Amazon EC2/S3

Chronopolis: A Partnership

- **Chronopolis is being developed by a national consortium led by SDSC and the UCSD Libraries.**
- **Initial Chronopolis provider sites include:**
  - *SDSC and UCSD Libraries at UC San Diego*
  - *University of Maryland*
  - *National Center for Atmospheric Research (NCAR) in Boulder, CO*

SDSC
SAN DIEGO SUPERCOMPUTER CENTER

UMIACS

NCAR

Institutions and Roles - UCSD

SDSC

- Storage and networking services
- SRB support
- Transmission Packaging Modules

UCSD Libraries

- Metadata services (PREMIS)
- DIPs (Dissemination Information Packages)
- Other advanced data services as needed

Institutions and Roles - NCAR

National Center for Atmospheric Research

- Archives – Complete copy of all data
- Storage and network support
- Network testing

Institutions and Roles - UMIACS

University of Maryland – Institute for Advanced Computer Studies

- Archives - Complete copy of all data
- Advanced data services
  - PAWN: **P**roducer – **A**rchive **W**orkflow **N**etwork in Support of Digital Preservation
  - ACE: **A**uditing **C**ontrol **E**nvironment to Ensure the Long Term Integrity of Digital Archives
- Other advanced data services as needed

SDSC Chronopolis Program

Chronopolis Digital Preservation DataGrid

Administration for Policy and Outreach

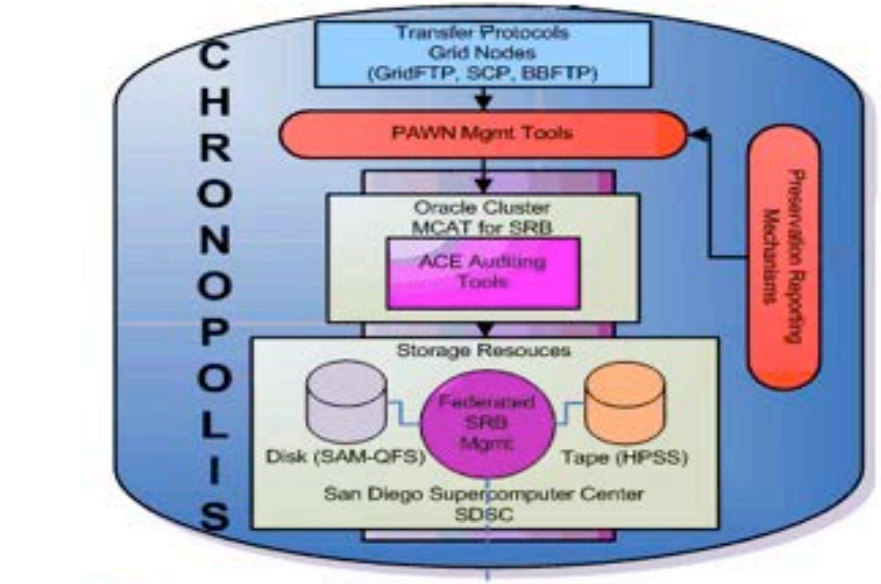
(Supports the overall partnerships and mgmt for preservation services and works as a liaison with Chroonpolis partners and other regional and national preservation programs)

Research and Development

(Research and development for rules-based preservation mgmt and technology forecasting for continual technology migration and mgmt)

Production Digital Preservation

(Long-term preservation with geographic replications and preservation services)



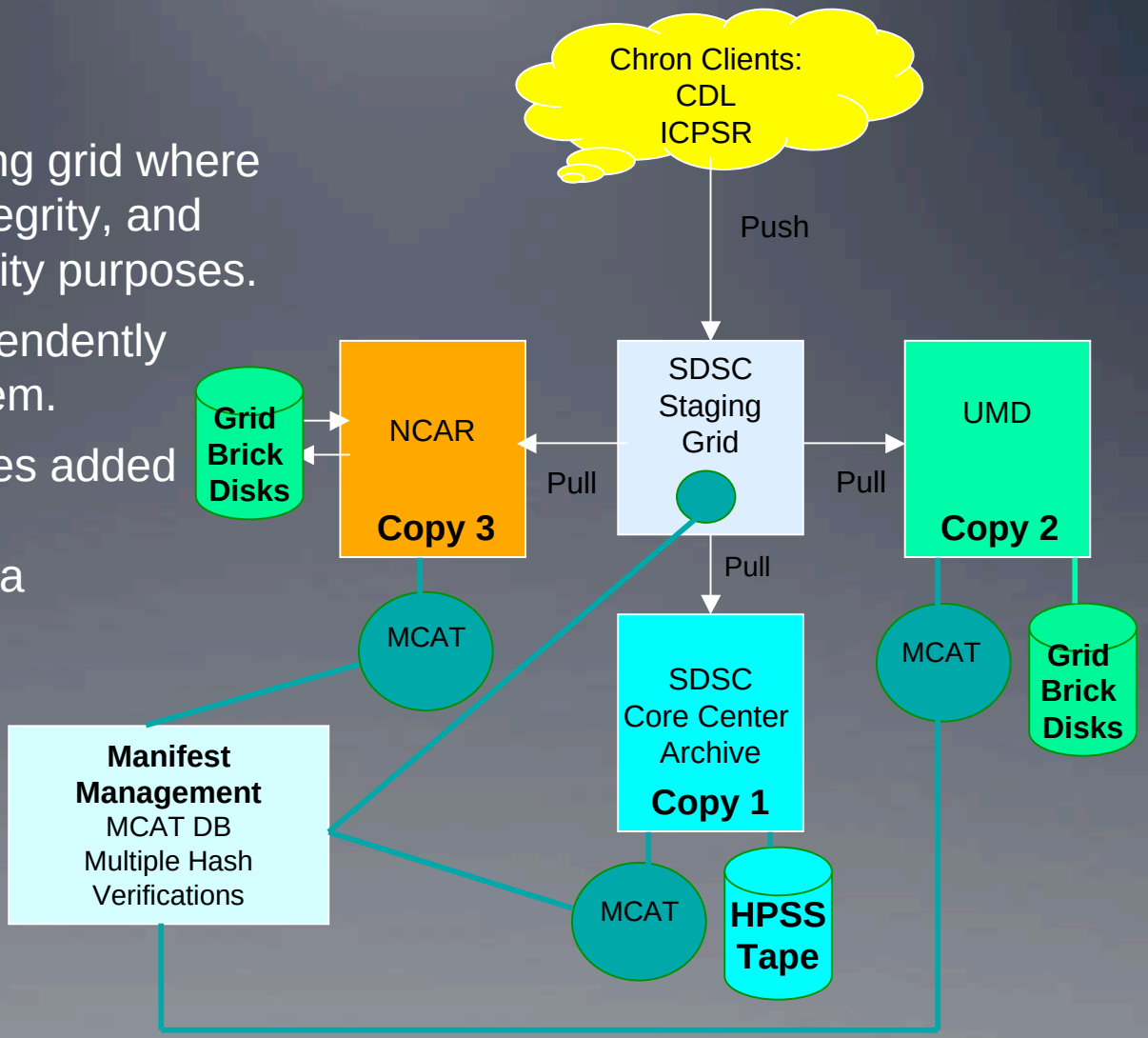
- Federated Replication
- Verification Tools
- Integrated Management

Chronopolis Vocabulary

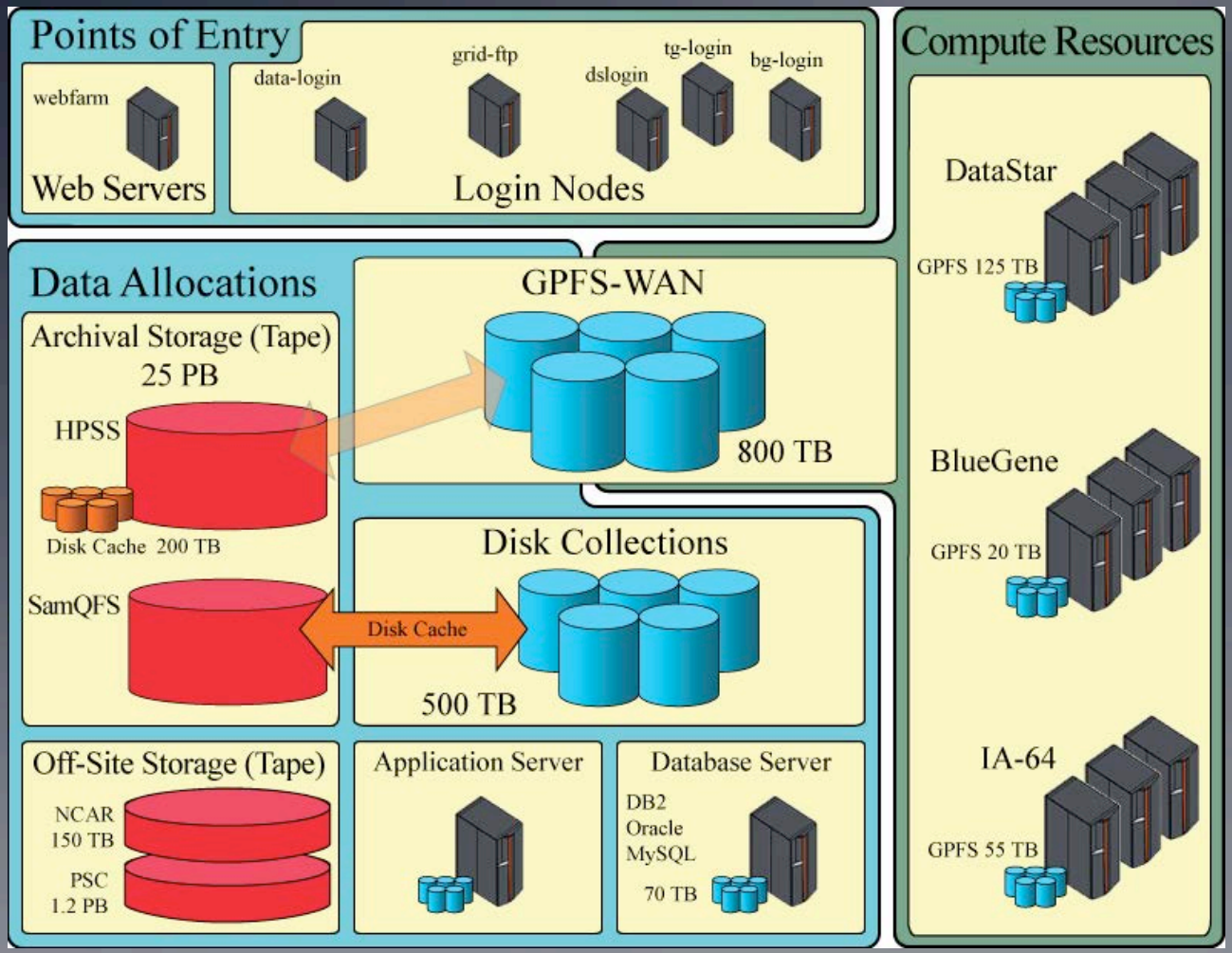
- **Partners – UCSD Libraries, National Center for Atmospheric Research, University of Maryland Institute for Advanced Computer Studies all provide grid enabled storage nodes for Chronopolis services.**
- **Clients – ICPSR, CDL– contribute content to the Chronopolis preservation network.**
- **SRB – Storage Resource Broker – datagrid software.**
- **iRODS – integrated Rule Oriented Data System – datagrid software.**
- **ACE – Audit Control Cnvironment – part of the ADAPT project at UMD.**
- **PAWN – Producer Archive Workflow Network – part of the ADAPT project at UMD.**
- **INCA – user level grid monitoring - executes periodic, automated, user-level testing of Grid software and services – grid middleware.**
- **Bagit – Transfer specification developed by CDL and the Library of Congress.**
- **GridFTP – parallel transfer technology - moves large collections within a grid wide-area network.**

Chronopolis: Inside

- Linked by main staging grid where data is verified for integrity, and quarantined for security purposes.
- Collections are independently pulled into each system.
- Manifest layer provides added security for database management and data integrity validation.
- Benefits
  - 3 independently managed copies of the collection
  - High availability
  - High reliability



SDSC Leveraged Infrastructure



Adapted from Richard Moore (SDSC)

- Serves Both HPC & Digital Preservation
- Archive
  - 25 PB capacity
  - Both HPSS & SAM-QFS
- Online disk
  - ~3PB total
  - HPC parallel file systems
  - Collections
  - Databases
- Access Tools

SDSC DPI Group

- Digital Preservation Initiatives Group
  - Charged with Developing and Supporting Digital Preservation Services within the Production Systems Division of SDSC.
  - http://dpi.sdsc.edu
  - Cross-Organizational Group
    - SDSC Personnel/UCSD Libraries Personnel
      - Libraries
      - Archives
      - Technology
      - Information Science

Chronopolis Demonstration Project

- Demonstration Project 2006-2007
  - Demonstration Collections Ingested within Chronopolis
    - National Virtual Observatory (NVO)
      - 3 TB Hyperatlas Images (partial collection)
    - Library of Congress PG Image Collection
      - 600 GB Prokudin-Gorskii Image Collection
    - Interuniversity Consortium for Political and Social Research (ICPSR)
      - 2TB Web Accessible Data
    - NCAR Observational Data
      - 3TB Observational Re-Analysis Data

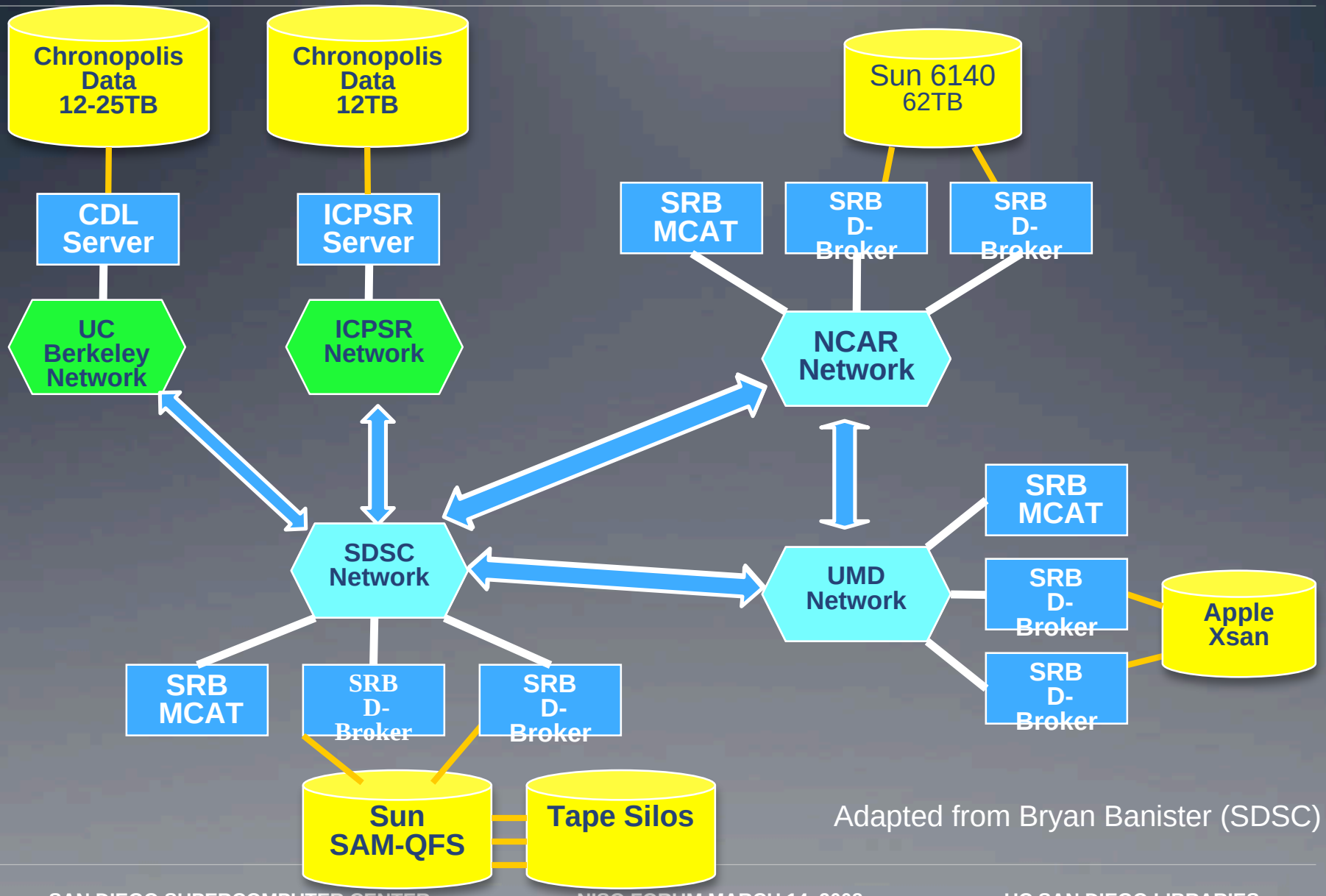
Partnerships and Collaborations

- **LC Pilot Project – Building Trust in a 3rd Party Repository**
  - Using test image collections/web crawls ingest content to SDSC repository
  - Allow access for content audit
  - Track usage of content over time
  - Deliver content back to LC at end of project
- **Library of Congress NDIIPP Chronopolis Program**
  - Build Production Capable Chronopolis Grid (50 TB x 3)
  - Further define transmission packaging for archival communities
  - Investigate best network transfer models for I2 and TeraGrid networks
- **California Digital Library (CDL) Mass Transit Program**
  - Enable UC System Libraries to transfer high-speed mass digitization collections across CENIC/I2
  - Develop transmission packaging for CDL content
- **UCSD Libraries' Digital Asset Management System**
  - RDF System with data managed in SRB at SDSC

NDIIPP Chronopolis Project

- Creating a 3-node federated data grid at SDSC, NCAR and UMD – up to 50 TB data from CDL and ICPSR
- Installing and testing a suite of monitoring tools using ACE, PAWN, INCA
- Creating Appropriate Transmission Information Packages
- Generating PREMIS definitions for data
- Writing Best Practices documents for clients and partners

Chronopolis Grid Framework



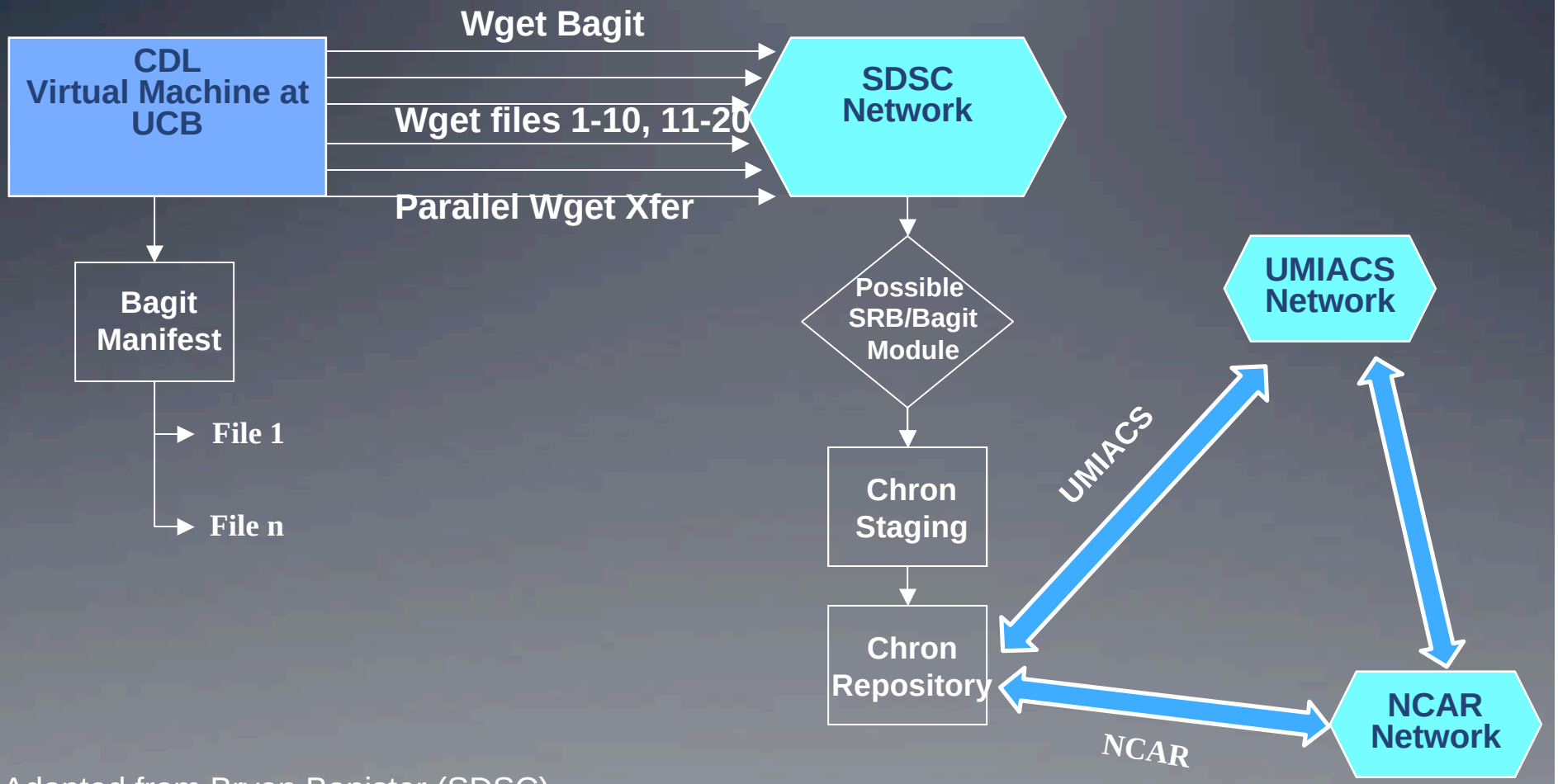
Adapted from Bryan Banister (SDSC)

NDIIPP Chronopolis Clients-CDL

California Digital Library

- A part of UCOP, supports the University of California libraries
- Providing up to 25TB of data: Web-At-Risk project
  - Five years of political and governmental websites
  - ARC files created from web crawls
  - Using Bagit Transfer Structure
  - Trisha Cruse-John Kunze-Stephen Abrams

Diagram of CDL Data Transfer



Adapted from Bryan Banister (SDSC)

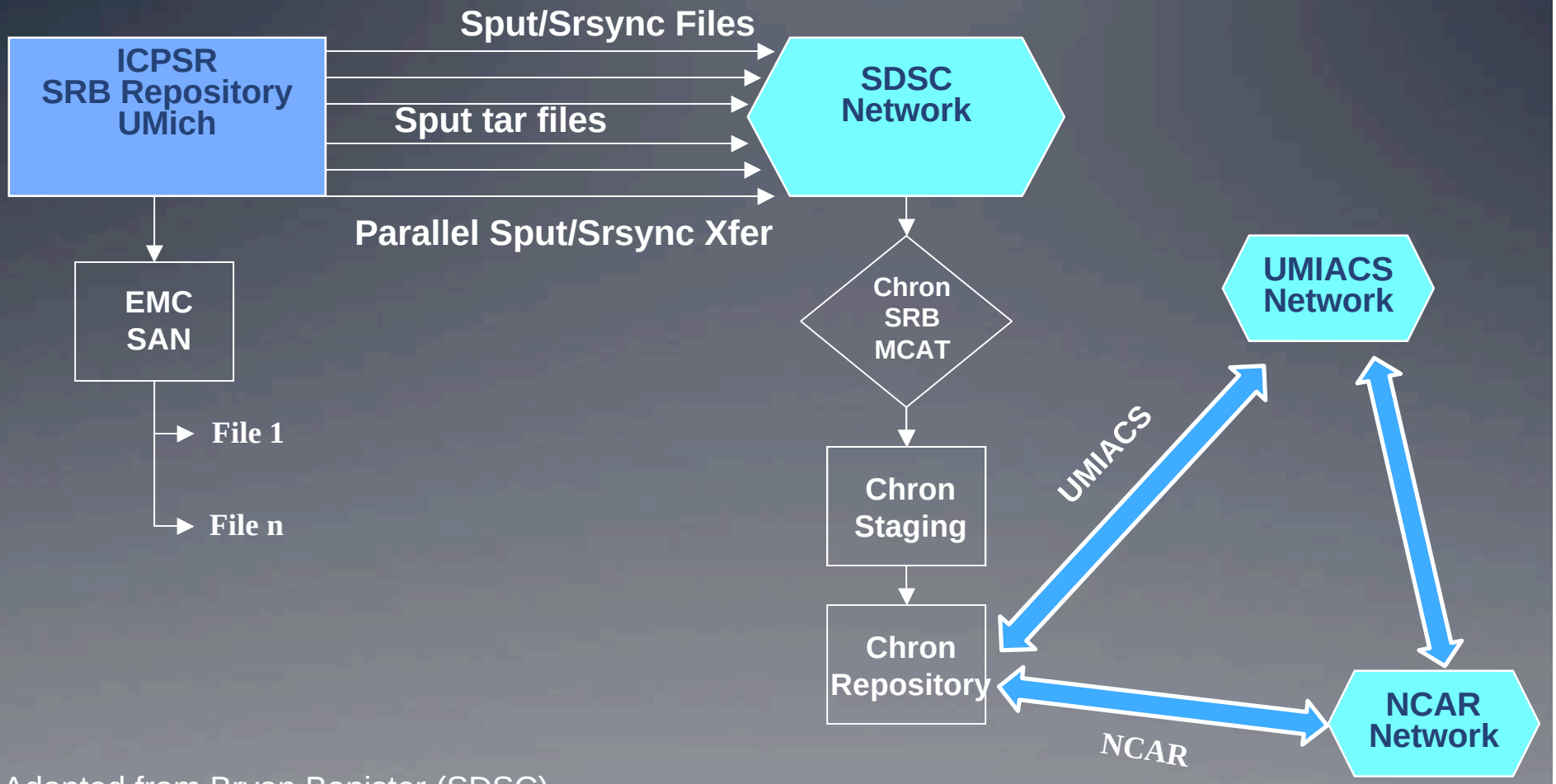
ICPSR NDIIPP Chronopolis Clients-ICPSR

Inter-University Consortium for Political and Social Research, University of Michigan

- Providing @12TB of data: Wide variety of types
- Already working with SDSC using SRB
- Bryan Beecher-Nancy McGovern

ICPSR

Diagram of ICSPR Transfer



Adapted from Bryan Banister (SDSC)

Ongoing and Future Initiatives

- Migration of Chronopolis Grid from SRB to iRODS
- Develop Interoperability with Community Based Archival Systems/Standards
  - DataNet
  - NDIIPP
  - State Archive E-Records
- TRAC compliance for SDSC Production Preservation Services/Chronopolis Consortium

Looking for Partnerships

- Repositories interested in moving large digital collections among heterogeneous repository systems.
- Fedora, DSpace or E-Prints sites interested in managed datagrid storage.
- Institutions interested in personnel swaps to conduct TRAC audit assessment compliance.
- Community Needs for Mass-Scale Data Transmission and Storage.

Chronopolis Credits

- SDSC
  - Fran Berman
  - Richard Moore
  - David Minor
  - Chris Jordan
  - Jim D'Aoust
  - Robert McDonald
  - Don Sutton
  - Brian Banister
  - Phong Dinh
  - Jay Dombrowski
  - Emilio Valente
- UCSD Libraries
  - Brian Schottlaender
  - Luc Declerck
  - Ardys Kozbial
  - Brad Westbrook
  - Arwen Hutt
- NCAR
  - Don Middleton
  - Michael Burek
  - Linda McGinley
- UMIACS
  - Joseph JaJa
  - Mike Smorul
  - Mike McGann
- Library of Congress
  - Martha Anderson
  - Lisa Hoppis
- CACI
  - Mike Ivey

http://chronopolis.sdsc.edu

SDSC
SAN DIEGO SUPERCOMPUTER CENTER

Thank You!