Chronopolis Overview

Joseph JaJa
UMIACS

Robert H. McDonald
Indiana University

David Minor
San Diego Supercomputer Center

Ardys Kozbial
UCSD Libraries

Don Sutton
San Diego Supercomputer Center

SAA Research Forum
August 26, 2008

Chronopolis: A Partnership

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

SDSC
SAN DIEGO SUPERCOMPUTER CENTER

UMIACS

NCAR

NDIIPP Chronopolis Project

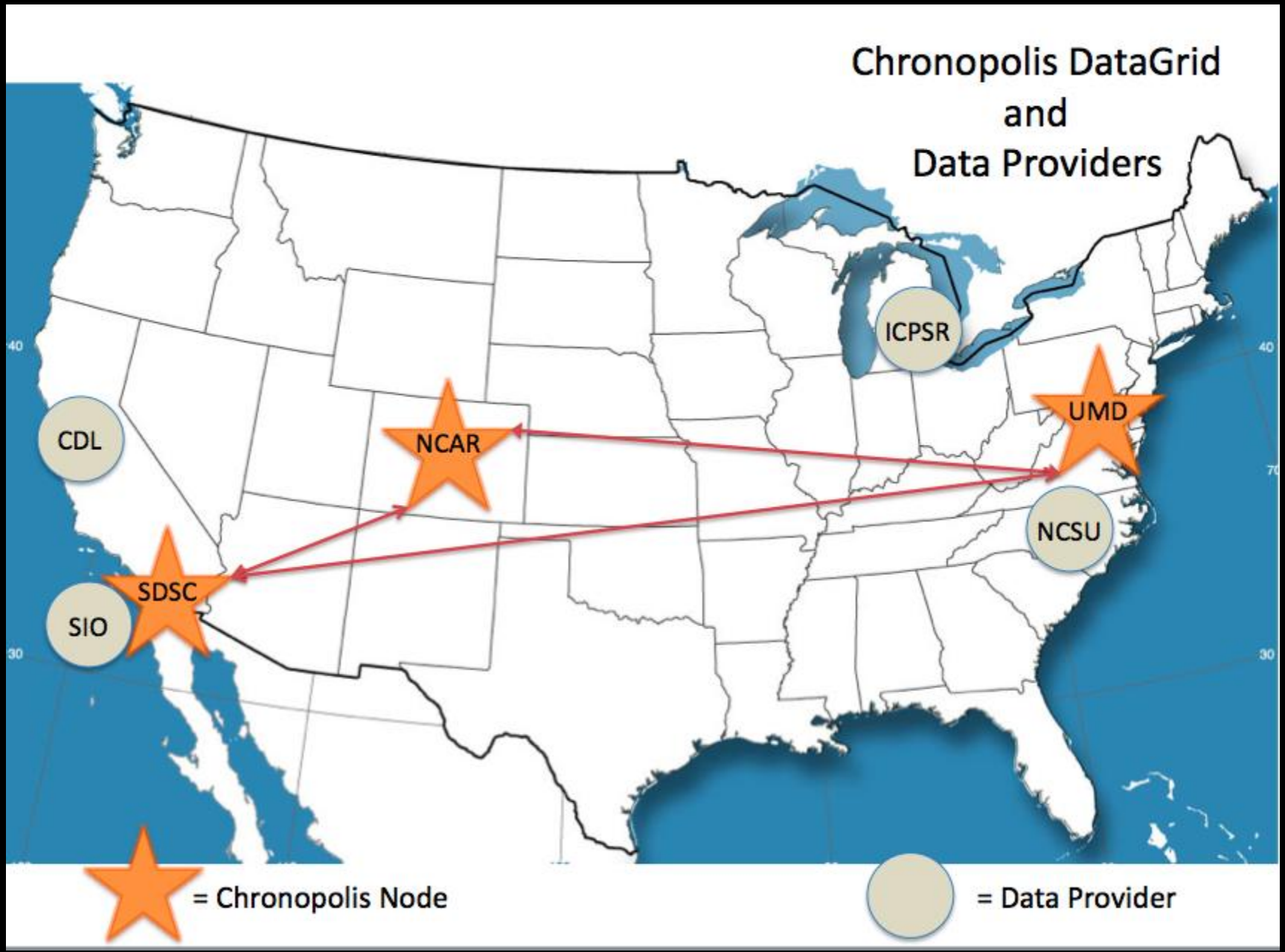
- Creating a 3-node federated data grid at SDSC, NCAR and UMIACS with up to 50 TB of data from the California Digital Library (CDL), the Inter-university Consortium for Political and Social Research (ICPSR), Scripps Institution of Oceanography (SIO), and North Carolina State University (NCSU)

- Installing and testing monitoring tools using ACE and the Replication Monitor

- Creating appropriate transmission information packages

- Generating PREMIS definitions for metadata

- Writing best practices documents for clients and partners


Chronopolis DataGrid
and
Data Providers
ICPSR
UMD
CDL
NCAR
NCSU
SDSC
SIO
= Chronopolis Node
= Data Provider

Institutions and Roles at the Nodes

SDSC, UMIACS, NCAR

- Storage and network support
- Transmission packaging modules
- Complete copy of all data
- Network testing
- SRB support (SDSC, UMIACS)
- Advanced data services (UMIACS)
  - ACE: **A**uditing **C**ontrol **E**nvironment to ensure the long-term integrity of digital archives

UCSD Libraries

- Metadata expertise (PREMIS)
- DIPs (Dissemination Information Packages)

Institutions and Roles: Data Providers

California Digital Library

- 6 TB of data
- Web-at-Risk Project
- Crawls of political and government Web sites
- ARC files, uniform size
- BagIt protocol for data transfer

ICPSR

- 10-12 TB of data
- 40 years of social science research
- Millions of files
- Currently using SRB

Institutions and Roles: Data Providers

NCSU

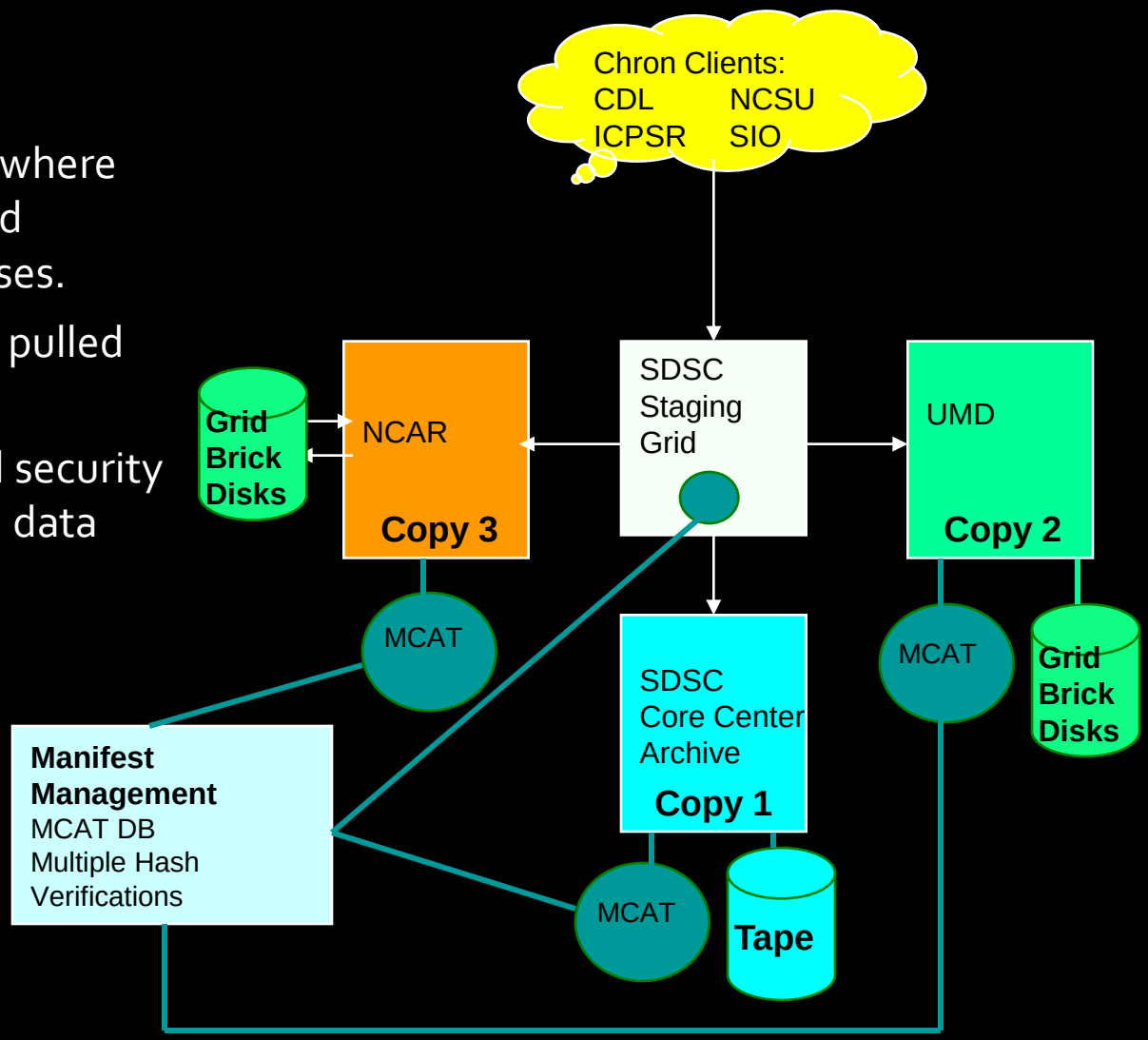
- 5 TB of data
- State and local geospatial data
- BagIt protocol for data transfer

SIO

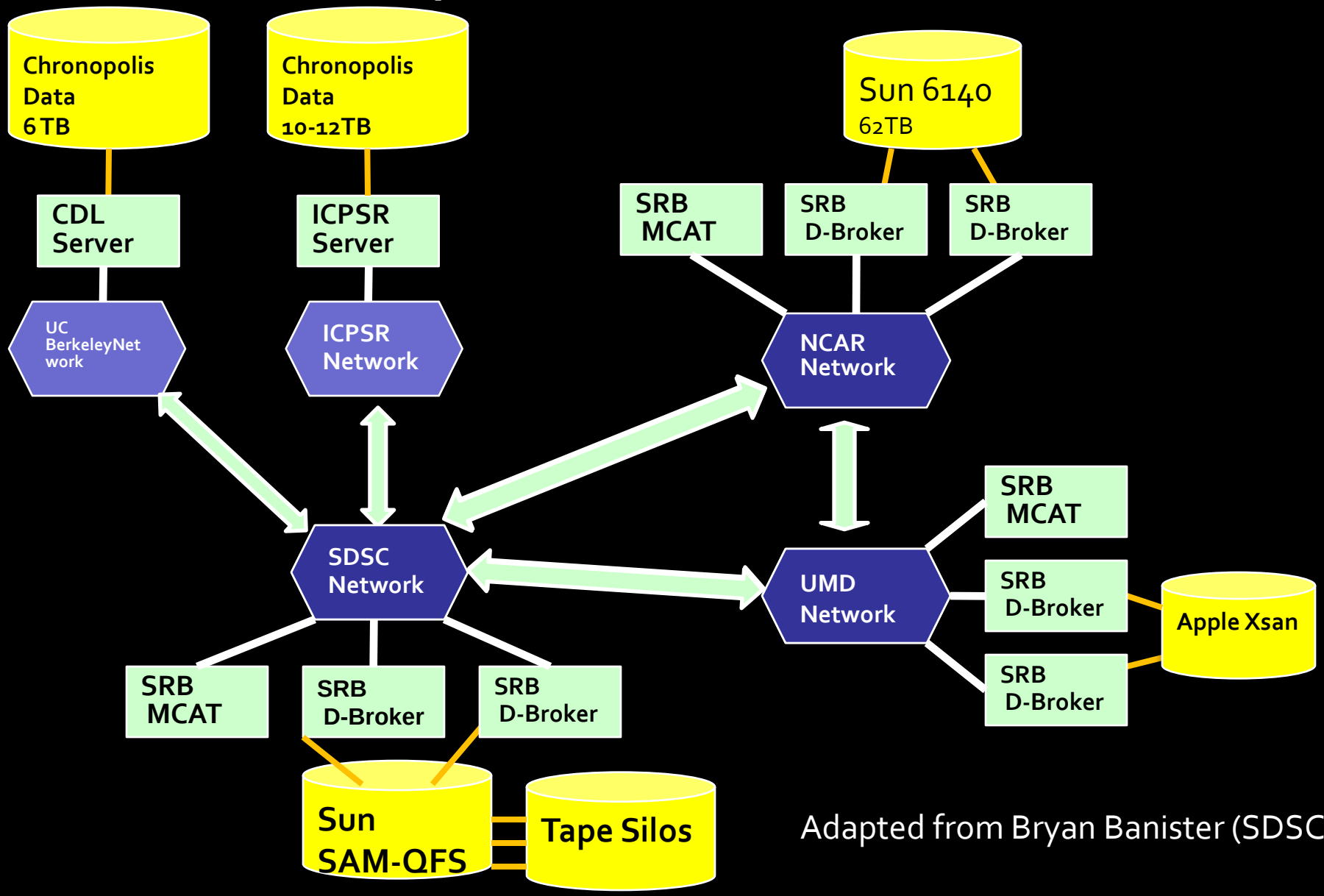
- 2 TB of data
- 50 years of data from SIO research cruises
- Currently using SRB

Chronopolis: Inside

- Linked by a 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



Chronopolis Grid Framework



Adapted from Bryan Banister (SDSC)

Current Status

Data Provider Content Ingested to SDSC

- CDL – 6 TB
- ICPSR – 12 TB
- SIO – 2 TB

Replicated Content

- SDSC →UMIACS – **15 TB (Copy 2)**
- SDSC →NCAR (forthcoming)

Transmission Speed for Ingest

- ICSPR – Approx 1 TB per day
- CDL – BagIt tests using the LC python scripts (15 processes)
  - City Bag – 46.22 Mb/sec – 500 GB per day
  - State Bag – 42.88 Mb/sec – 500 GB per day

Chronopolis Credits

SDSC
- Fran Berman
- Richard Moore
- David Minor
- Chris Jordan
- Jim D'Aoust
- Robert McDonald
- Don Sutton
- Bryan Banister
- Phong Dinh
- Jay Dombrowski
- Emilio Valente

UCSD Libraries
- Brian Schottlaender
- Luc Declerck
- Ardys Kozbial
- Brad Westbrook
- Arwen Hutt

NCAR
- Don Middleton
- Michael Burek
- Lynda McGinley

UMIACS
- Joseph JaJa
- Mike Smorul
- Mike McGann

Library of Congress
- Martha Anderson
- Lisa Hoppis

CACI
- Mike Ivey

http://chronopolis.sdsc.edu