Oct. 11, 2007
National Science Library
Beijing, China

The Chronopolis Demonstration Project:

A Grid-Based Digital Preservation Environment

Robert H. McDonald
Director of Strategic Data Alliances
San Diego Supercomputer Center
UC San Diego
mcdonald@sdsc.edu

CHRONOPOLIS
Consortium

Outline

- **SDSC at a Glance**
  - Data-Cyberinfrastructure
  - Leveraged Resources
  - Digital Preservation Initiatives
- **Chronopolis**
  - A Conceptual Preservation Framework
  - A Federated Partnership
  - Grid-Enabled Preservation
    - Grid Map
  - Collaborators
- **Chronopolis Demonstration Project**
  - Design Feasibility
  - Networking
  - Replication - Tape
- **Chronopolis Next Steps**
  - 2007-2008 NDIIPP
  - 2007-2008 Mass Transit

SDSC at a Glance

- **One of original 5 NSF supercomputer centers (1985)**
- **Supports High Performance Computing Systems and Data Intensive Computing**
- **Supports Data Applications for Science, Engineering, Social Sciences, Cultural Heritage Institutions**
  - 3 PB Disk Storage
  - 25 PB Tape Storage

Acronyms

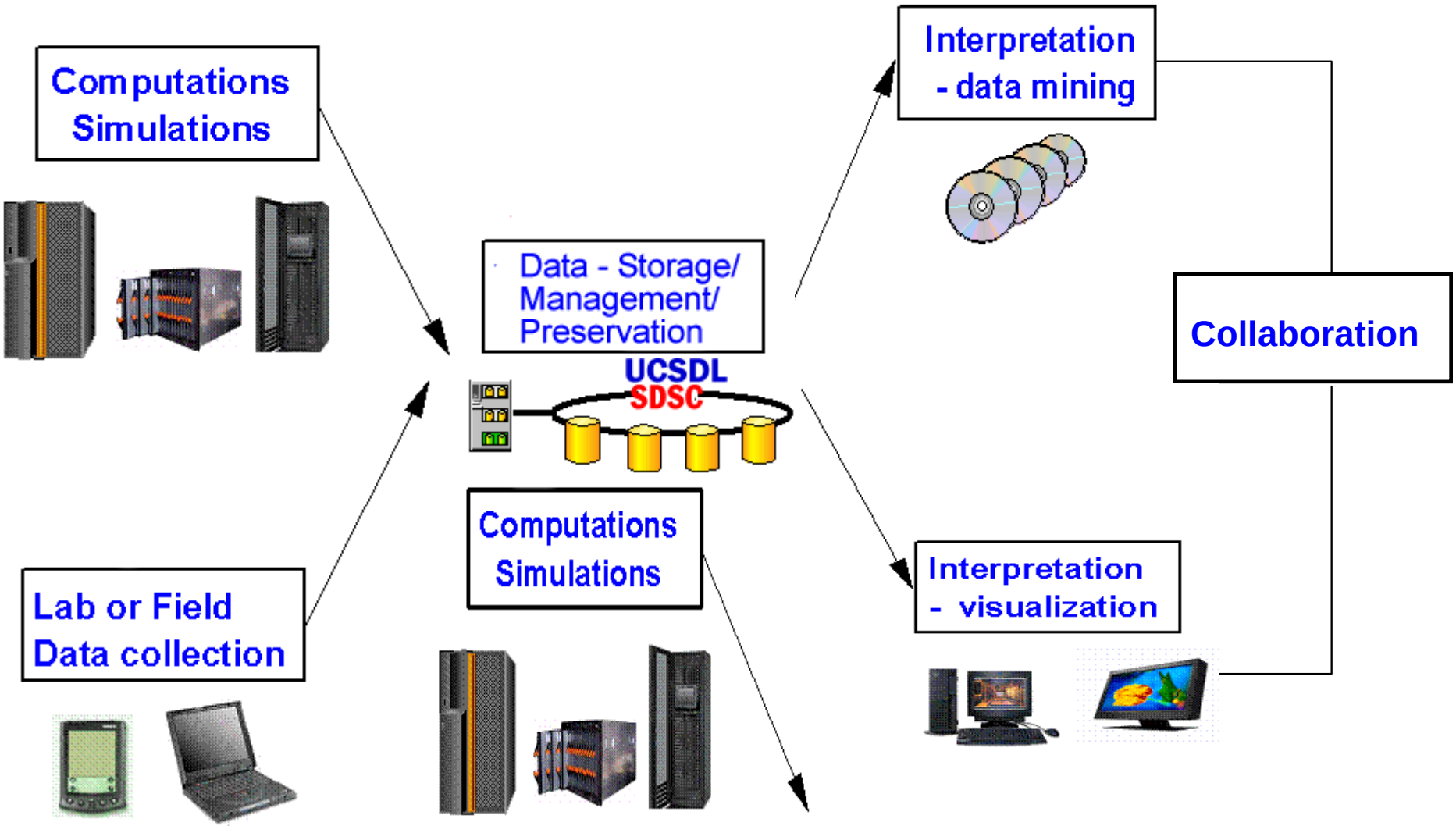
- **HPSS** – one of two archival storage systems at SDSC. Manages 6 StorageTek tape silos (25 PB).
- **SAM-QFS** – one of two archival storage systems at SDSC. Manages disk and tape storage resources at SDSC.
- **SRB** – Storage Resource Broker – middleware for managing data grid federations and heterogeneous storage hardware and software resources.
- **iRODS** – open source version of SRB with automated rules based storage management.
- **MCAT/iCAT** – metadata catalog for SRB and iRODS data management.
- **TeraGrid Network** – Dedicated 100G connectivity for NSF funded shared computing and storage resources on the TeraGrid.
- **I2 Abilene Network** – U.S. based dedicated educational network. Can have dedicated network bandwidth of 10G or 100G connectivity.

SDSC and Data Cyberinfrastructure

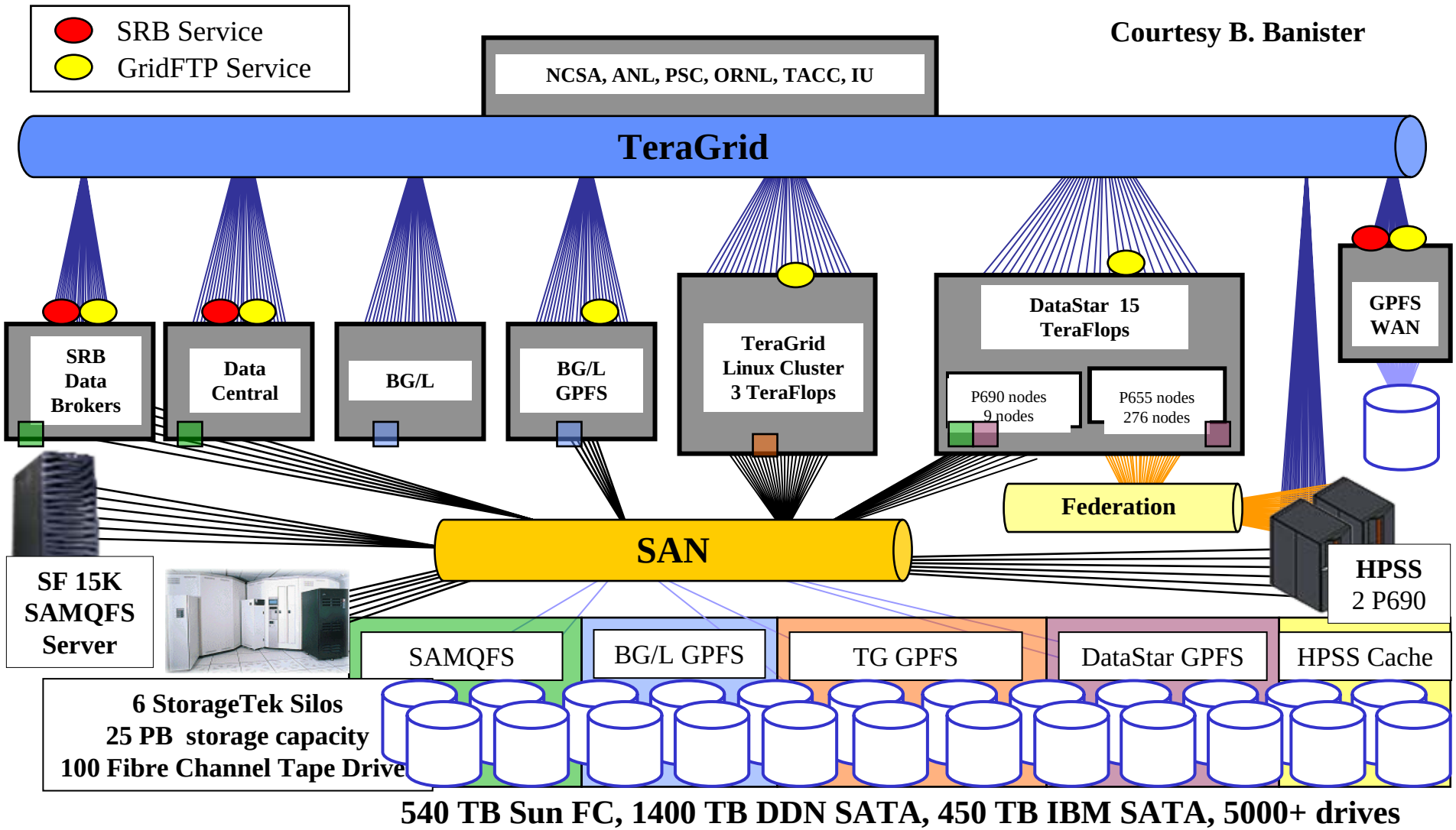
- *The mission of the San Diego Supercomputer Center (SDSC) is to empower communities in **data-oriented research, education, and practice** through the innovation and provision of Cyberinfrastructure*

Cyberinfrastructure = **resources** (computers, data storage, networks, scientific instruments, experts, etc.) **+ "glue"** (integrating software, systems, and organizations).

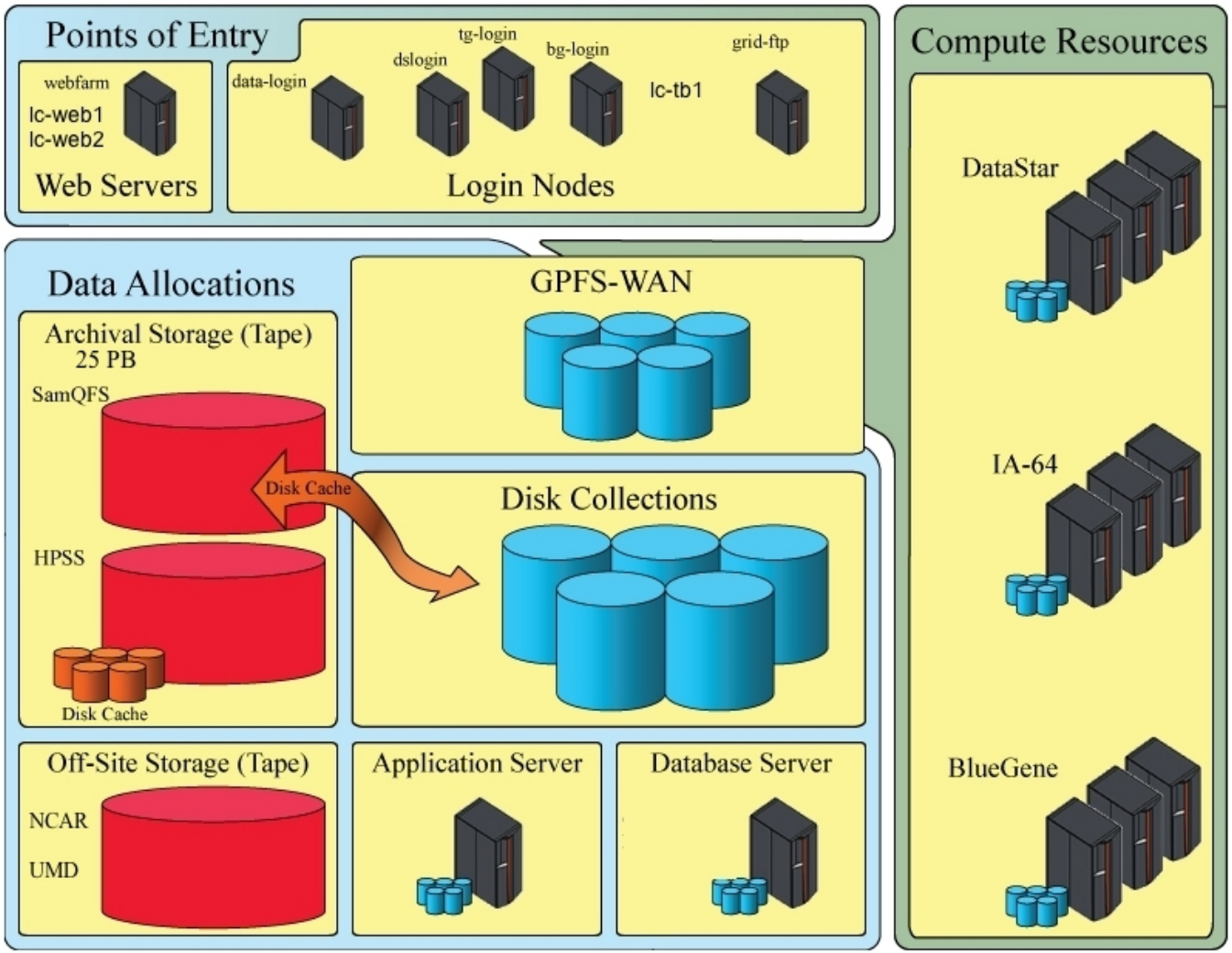
Virtually all modern research and education efforts are enabled by information and computational infrastructure


Computations
Simulations
Interpretation
- data mining
Data - Storage/
Management/
Preservation
UCSDL
SDSC
Collaboration
Computations
Simulations
Lab or Field
Data collection
Interpretation
- visualization

SDSC Centralized SAN and Storage Architecture



Data Structure at SDSC

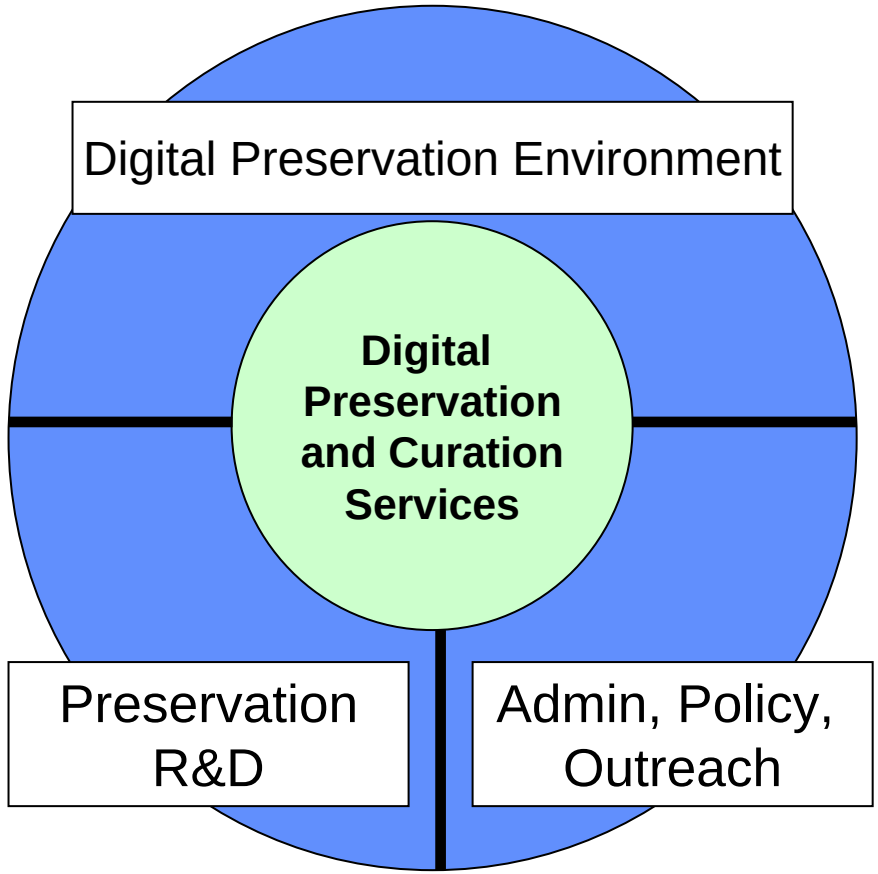


DPI at SDSC

- **Preservation Group within Production Systems Division at SDSC**
  - Charged with Developing and Supporting Digital Preservation Services for SDSC and projects supported by SDSC within the production systems division.
  - LC/NDIIPP
    - LC DATA Pilot Project
      - Content Transfer
      - Formalized Trust Relationships
    - NDIIPP Chronopolis Technical Architecture
      - Partnerships with:
      - CDL – California Digital Library
      - ICPSR – Interuniversity Consortium for Political and Social Science Research
  - California Digital Library (CDL)
    - Mass Transit Project

Chronopolis A Preservation Framework

- **Scaleable Production Digital Preservation Environment**
- **Digital Preservation Research and Development (R&D) Laboratory**
- **Management Framework for Preservation Administration, Policy, and Outreach**



A Federated 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

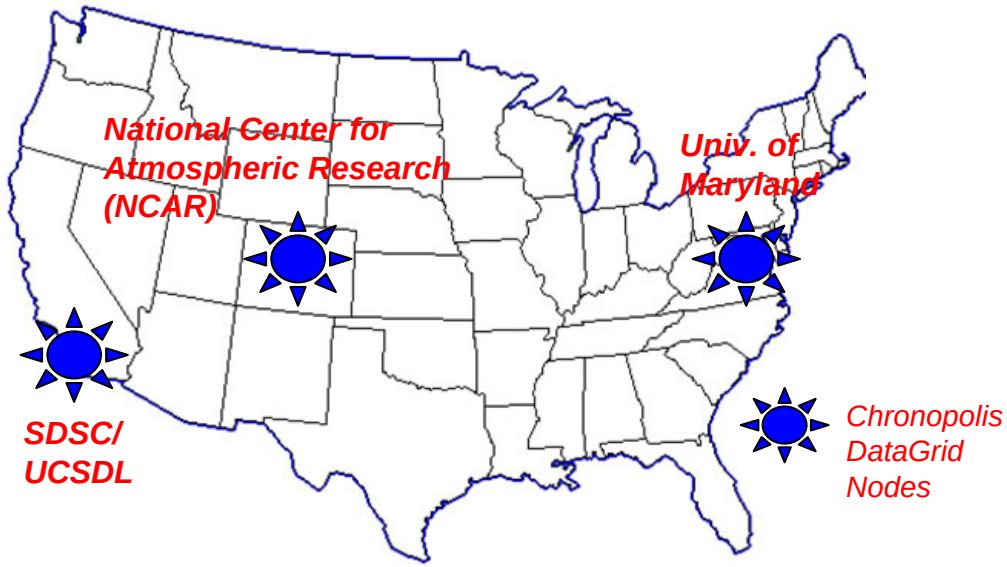
NCAR

Chronopolis Focus Areas

- Assessment of the needs of potential user communities and development of **appropriate service models**
- Development of **roles and responsibilities of providers, partners, users**
  - Development of Memoranda of Understanding (MOUs), Service Level Agreements (SLAs), etc. to formalize trust relationships and manage expectations
- Assessment and prototyping of **best practices** for bit preservation, authentication, metadata, etc.
- Development of appropriate **cost and risk models** for long-term preservation
- Development of appropriate **success metrics** to evaluate usefulness, reliability, and usability of infrastructure

Chronopolis Demonstration Data Grid

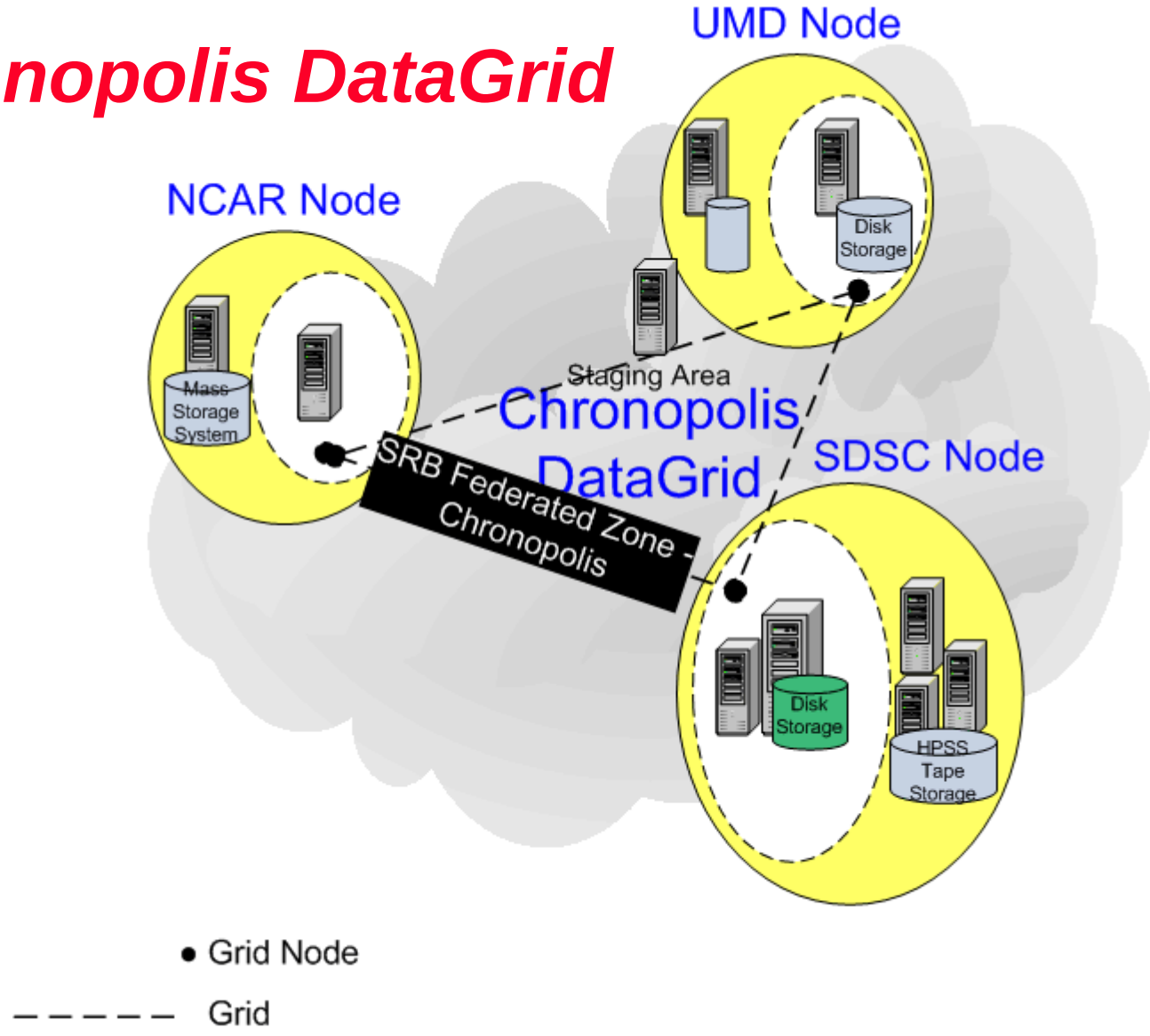
- The Chronopolis demonstration Data Grid is composed of 3 geographically distributed Chronopolis provider sites.
- Each provider takes on different roles with respect to a set of demonstration collections.



Demonstration collections include:

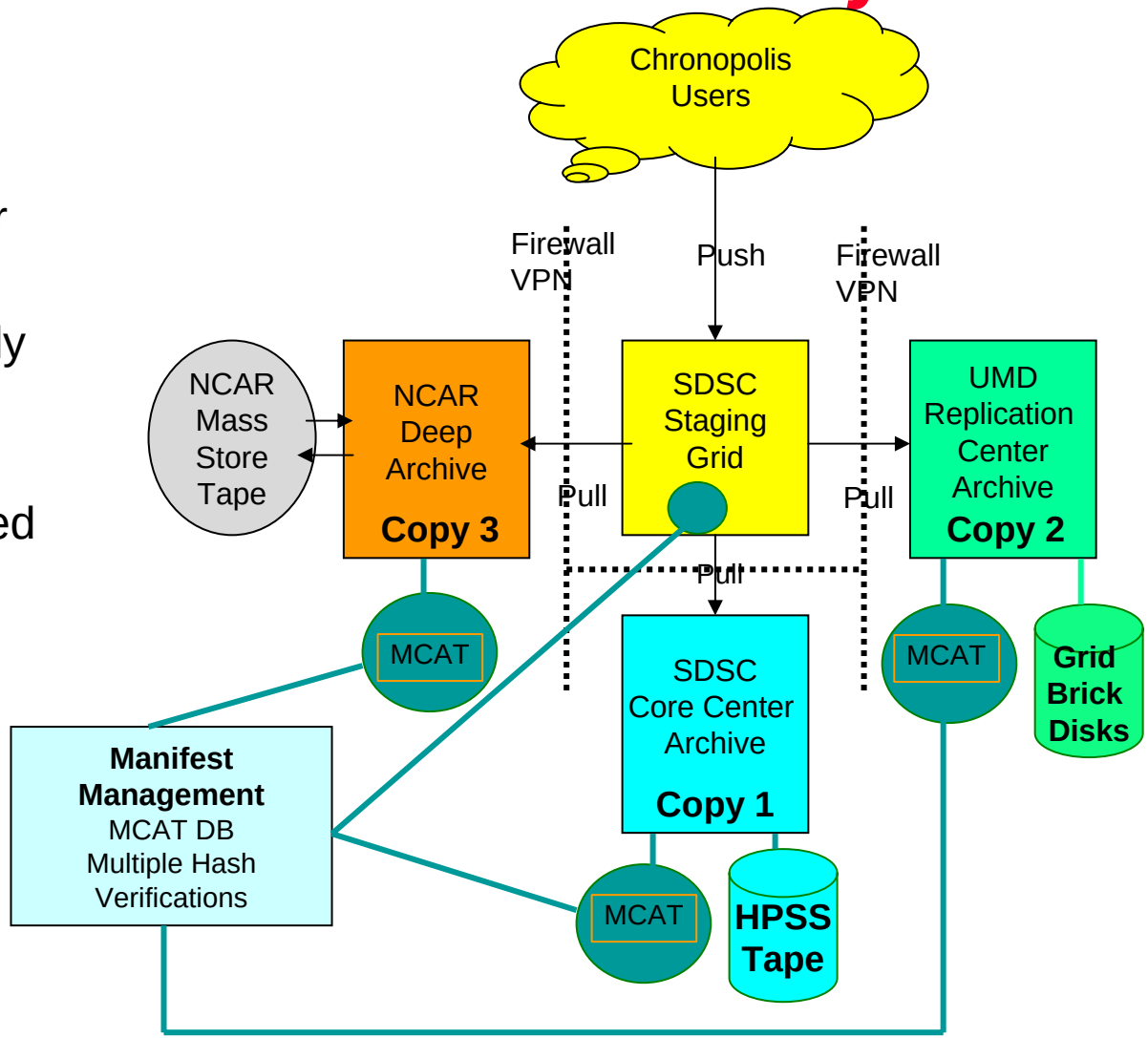
- National Virtual Observatory (NVO) [3 TB Digital Palomar Observatory Sky Survey]
- Library of Congress Image Collection [1 TB]
- Copy of Interuniversity Consortium for Political and Social Research (ICPSR) data [1 TB Web-accessible Data]
- NCAR Observational Data [3 TB of Observational and Re-Analysis Data]

Chronopolis DataGrid



Chronopolis Demonstration Project

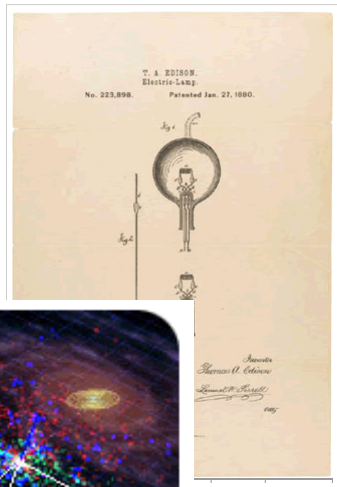
- Linked by main staging grid where data is verified for integrity, and quarantined for security purposes.
- Collections are independently “pulled” into each archive system.
- Manifest layer provides added security for database management and data integrity validation.
- **Benefits**
  - 3 independently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

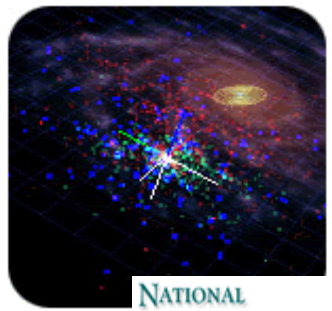


Chronopolis™ Demonstration Project Collections

- **National Virtual Observatory (NVO)**
  - Hyperatlas Images 3TB (partial collection)
- **Interuniversity Consortium for Political and Social Research (ICPSR)**
  - 2TB Web Accessible Data
- **Library of Congress PG Image Collection**
  - 1TB Prokudin-Gorskii Image Collection
- **NCAR Observational Data**
  - 3TB Observational Re-Analysis Data

Thomas Edison's Patent Application for the Light Bulb





Current Geographic Replications in Demonstration Project

RID PRESERVATION ENVIRONMENT

California	Colorado	Maryland

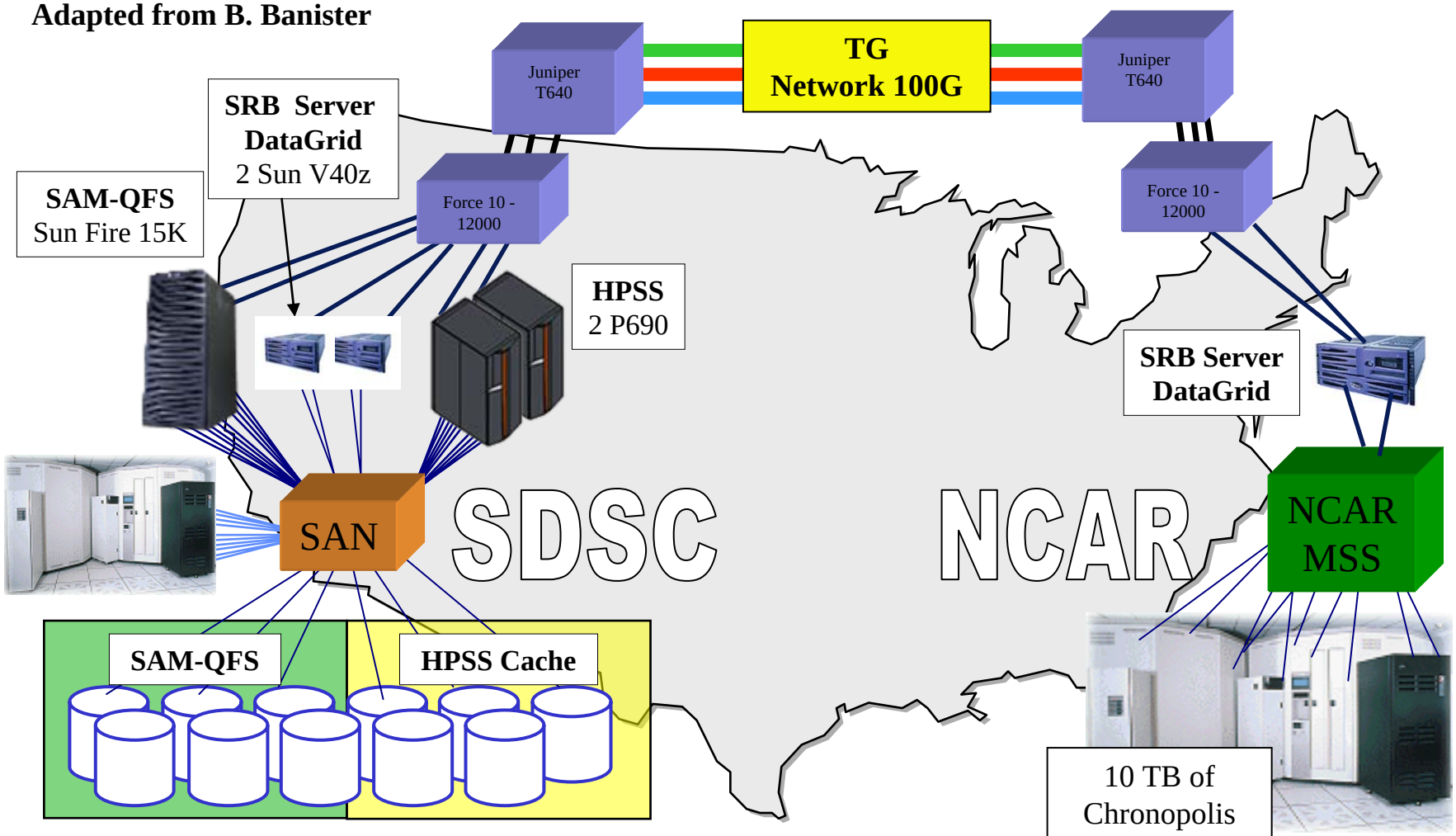
• **SDSC Node (10 TB)**
  • Library of Congress PG Image Collection
  • ICPSR Collection
  • NVO DPOSS Collection
  • NCAR Collection
	• **NCAR Node (10 TB)**
	  • Library of Congress PG Image Collection
	  • ICPSR Collection
	  • NVO DPOSS Collection
	  • NCAR Collection
		• **UMD Node (5.5 TB)**
		  • Library of Congress PG Image Collection
		  • ICPSR Collection
		  • NVO DPOSS Collection

Chronopolis Learning Curve

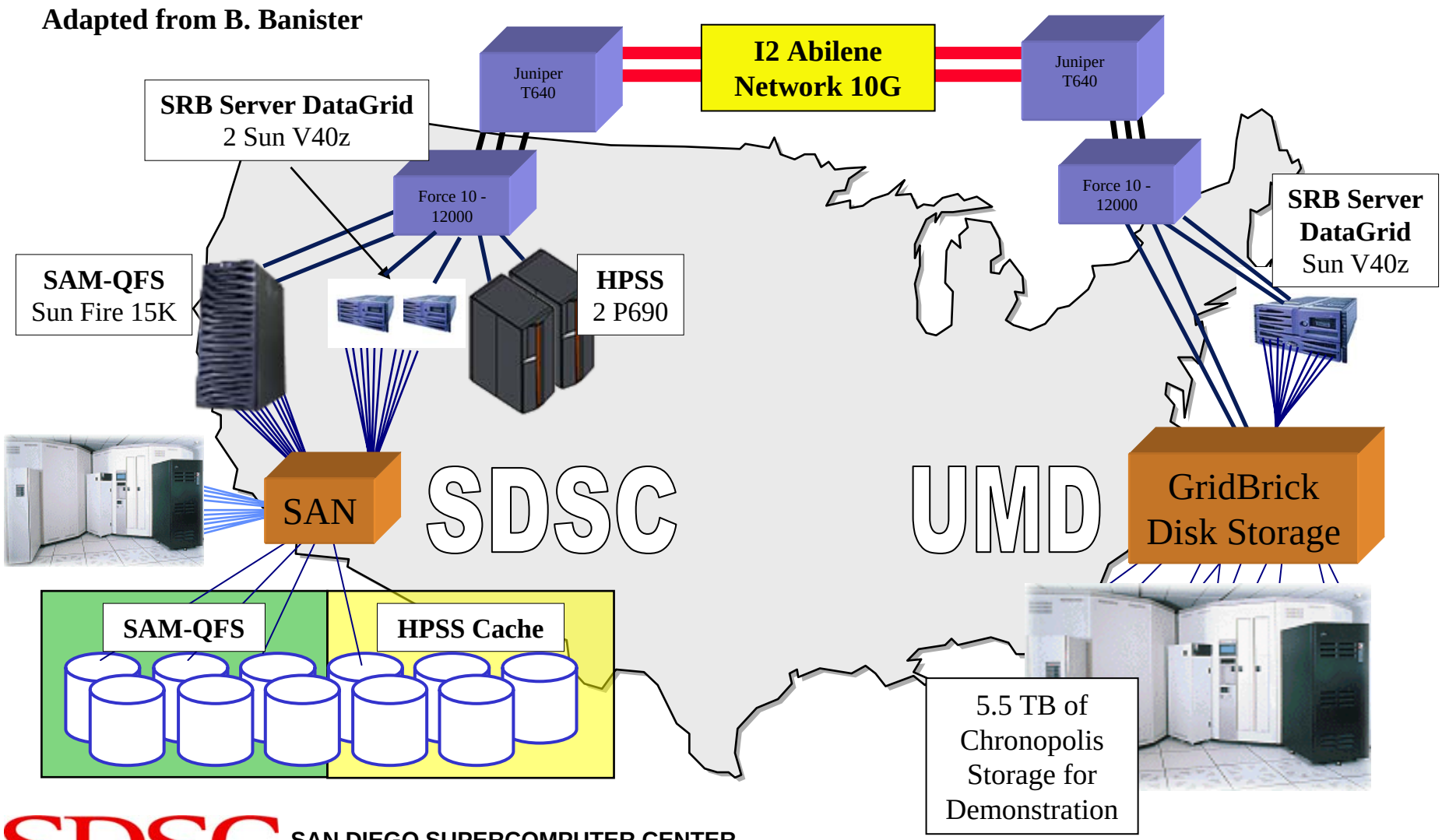
- **Studying Multiple Transmission Tools for Transferring Data**
- **Providing Ways to Improve High Speed Network Transmission**
  - Packaging for Content Transmission
  - Dissemination Information Package (DIP) that is fine tuned for high-speed transmission
- **Working on issues of data authenticity/provenance between federated archives**
- **Developing NARA/RLG TRAC Audit for Chronopolis Federation**
- **Creating a METS Profile for Preservation Metadata**
  - PREMIS compliant

Chronopolis Architecture SDSC to NCAR

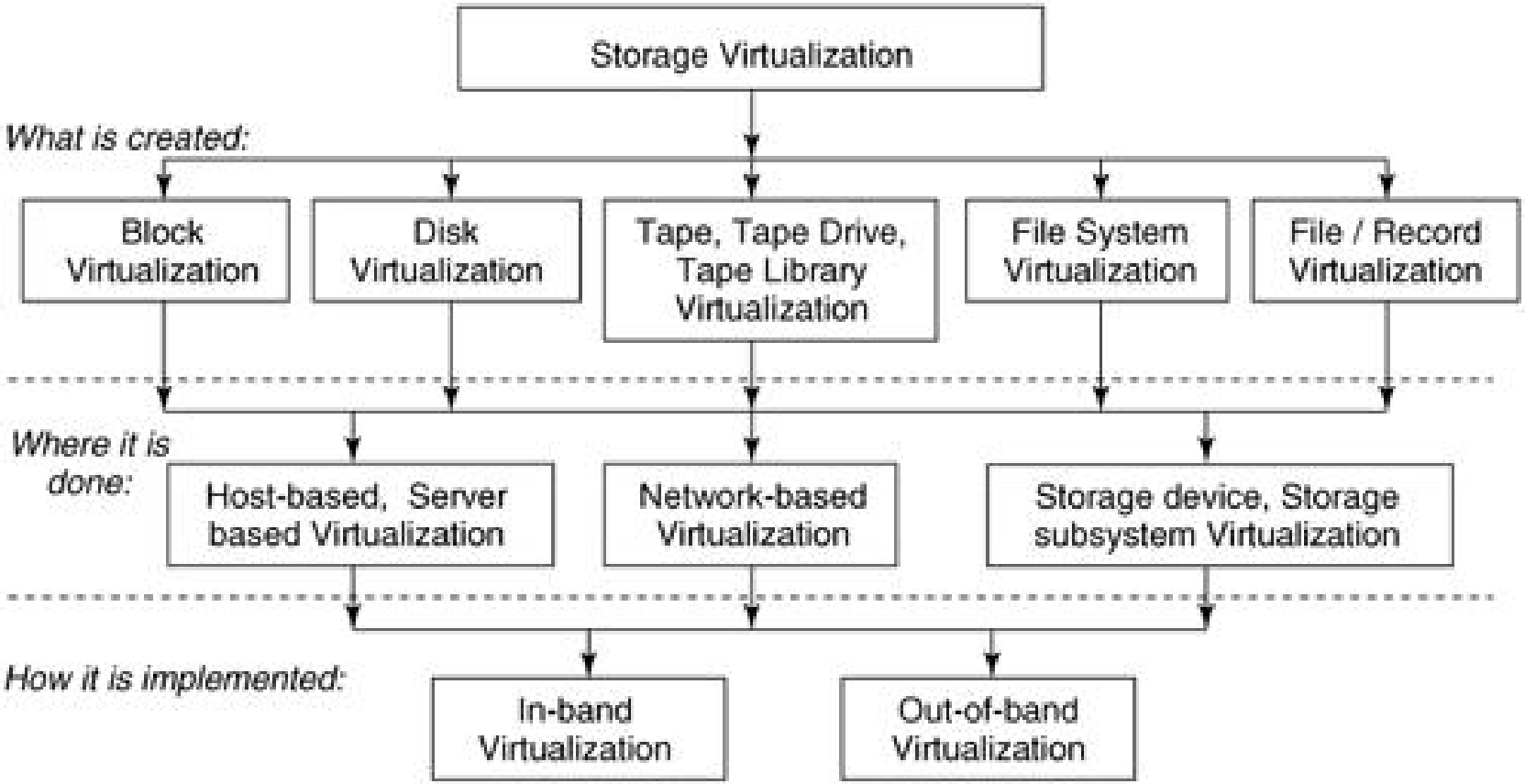
Adapted from B. Banister



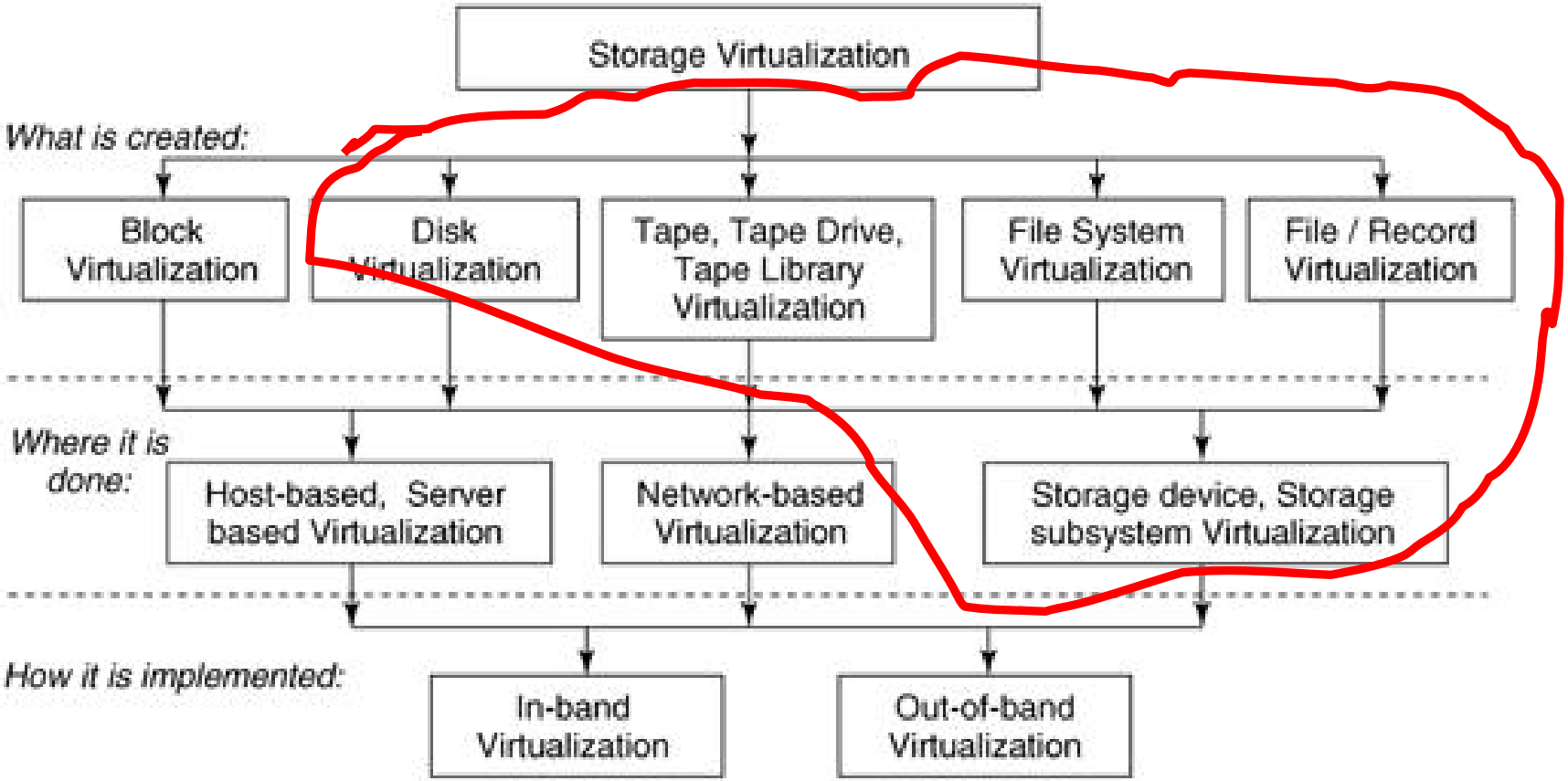
Chronopolis Architecture SDSC to UMD



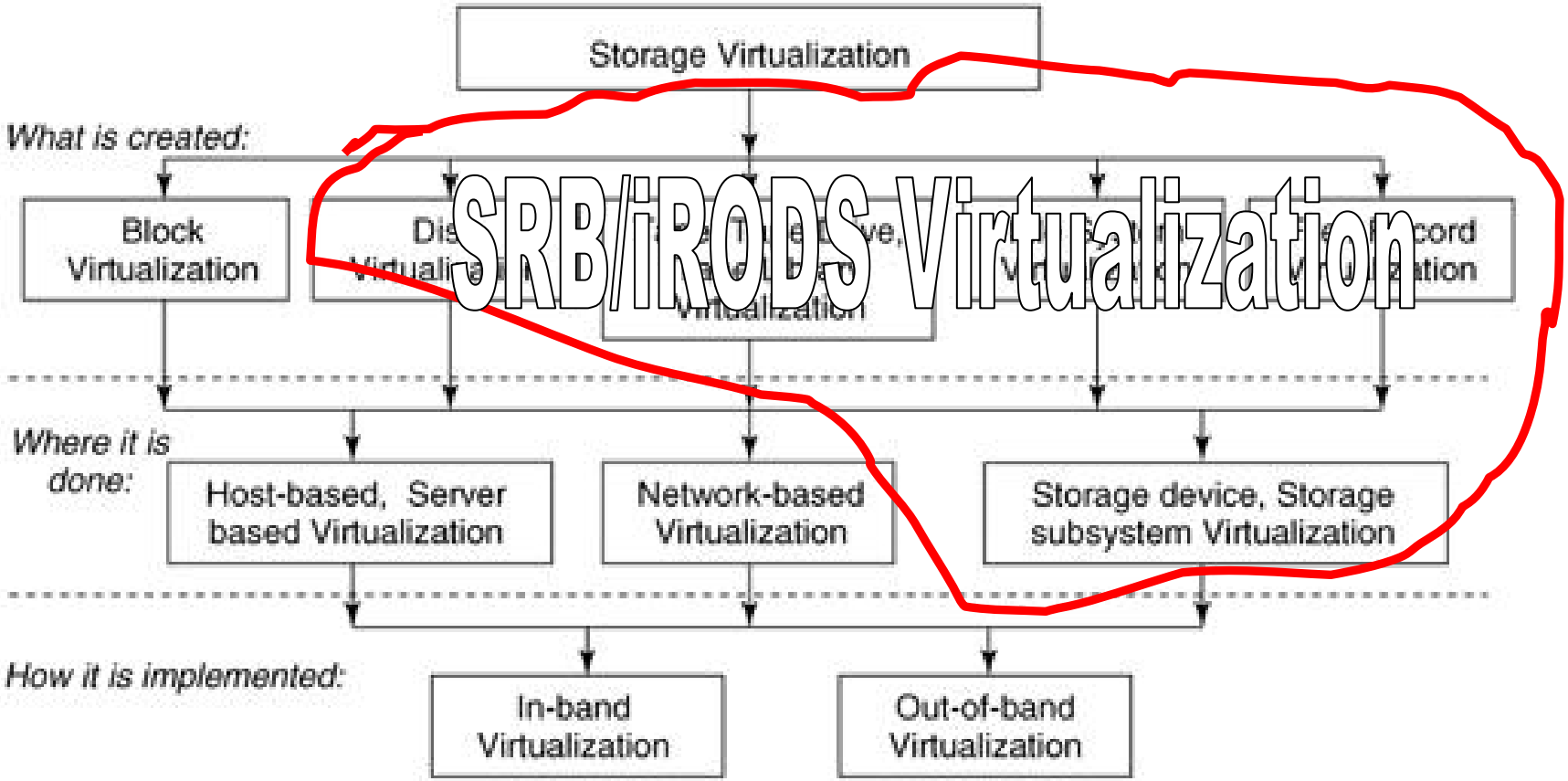
SNIA Storage Abstraction Layer



SNIA Storage Abstraction Layer



SNIA Storage Abstraction Layer



Chronopolis 2007-2008

- **Library of Congress NDIIPP Chronopolis Program**
  - Build Production Capable Chronopolis Grid (50 TB)
  - Further define transmission packaging for archival communities
- **California Digital Library (CDL) Mass Transit Program**
  - Enables UC System Libraries to transfer high-speed mass digitization collections across CENIC
  - Develop transmission packaging for CDL content
- **Migration Path for SRB/iRODS**
- **Interoperability with Community Based Archival Systems/Standards**
- **Interoperability with Network Virtualization Storage Layer**

Chronopolis Digital Preservation DataGrid

Administration for Policy and Outreach

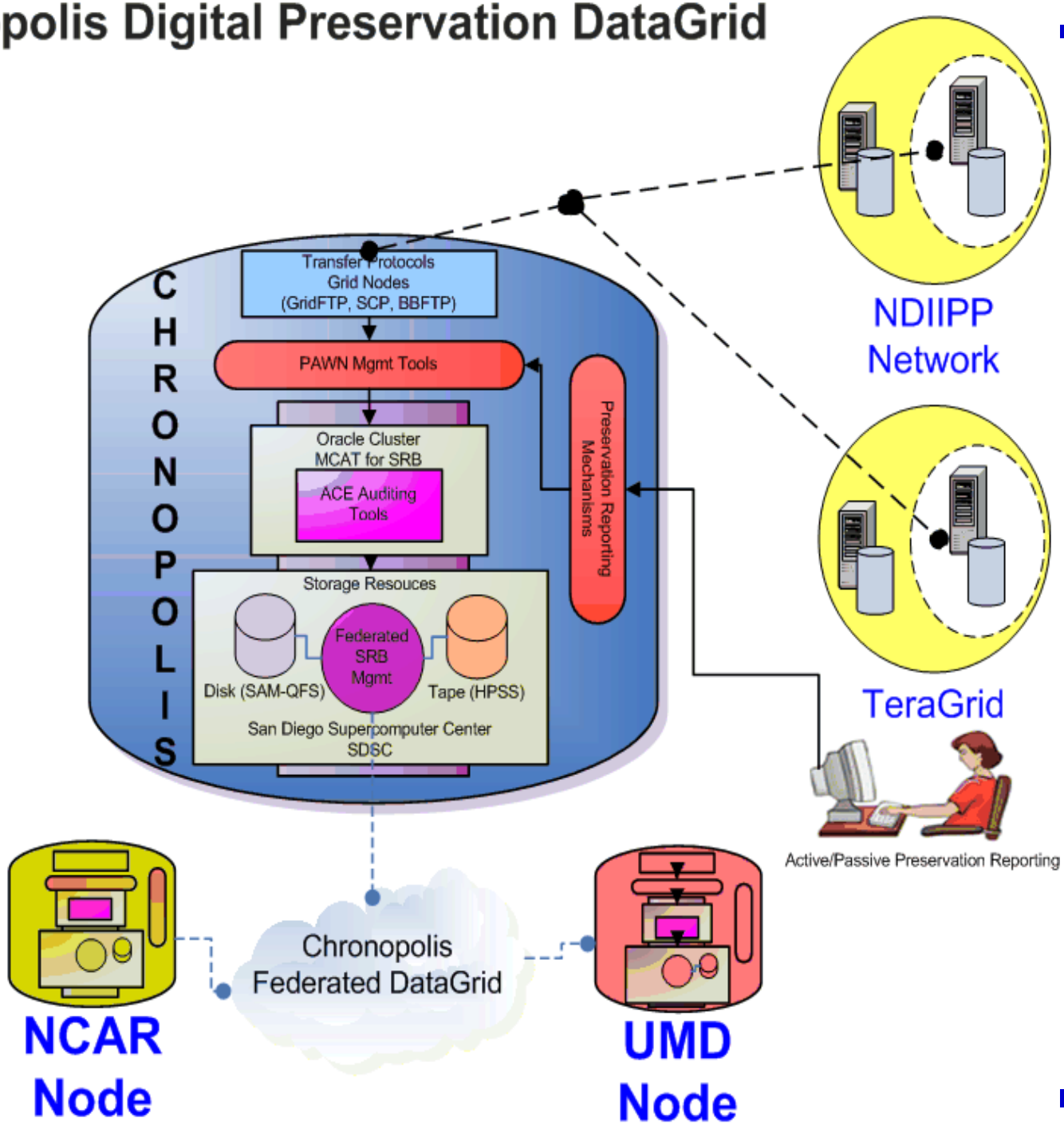
(Supports the overall partnerships and mgmt for preservation services and works as a liaison with Chroonpolis partners and other regional and national preservation programs)

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)



Chronopolis Collaborators

SDSC

Fran Berman
Richard Moore
Reagan Moore
Arcot Rajasekar
David Minor
Chris Jordan
Robert McDonald
Bryan Banister
Sheau Yen-Chen

NCAR

Don Middleton
Michael Burek

UCSD Libraries

Brian Schottlaender
Luc Declerck
Ardys Kozbial

UMD

Joseph JaJa
Mike Smorul
Mike McGann

Questions?