Chronopolis: Present and Future Storage Environments

David Minor

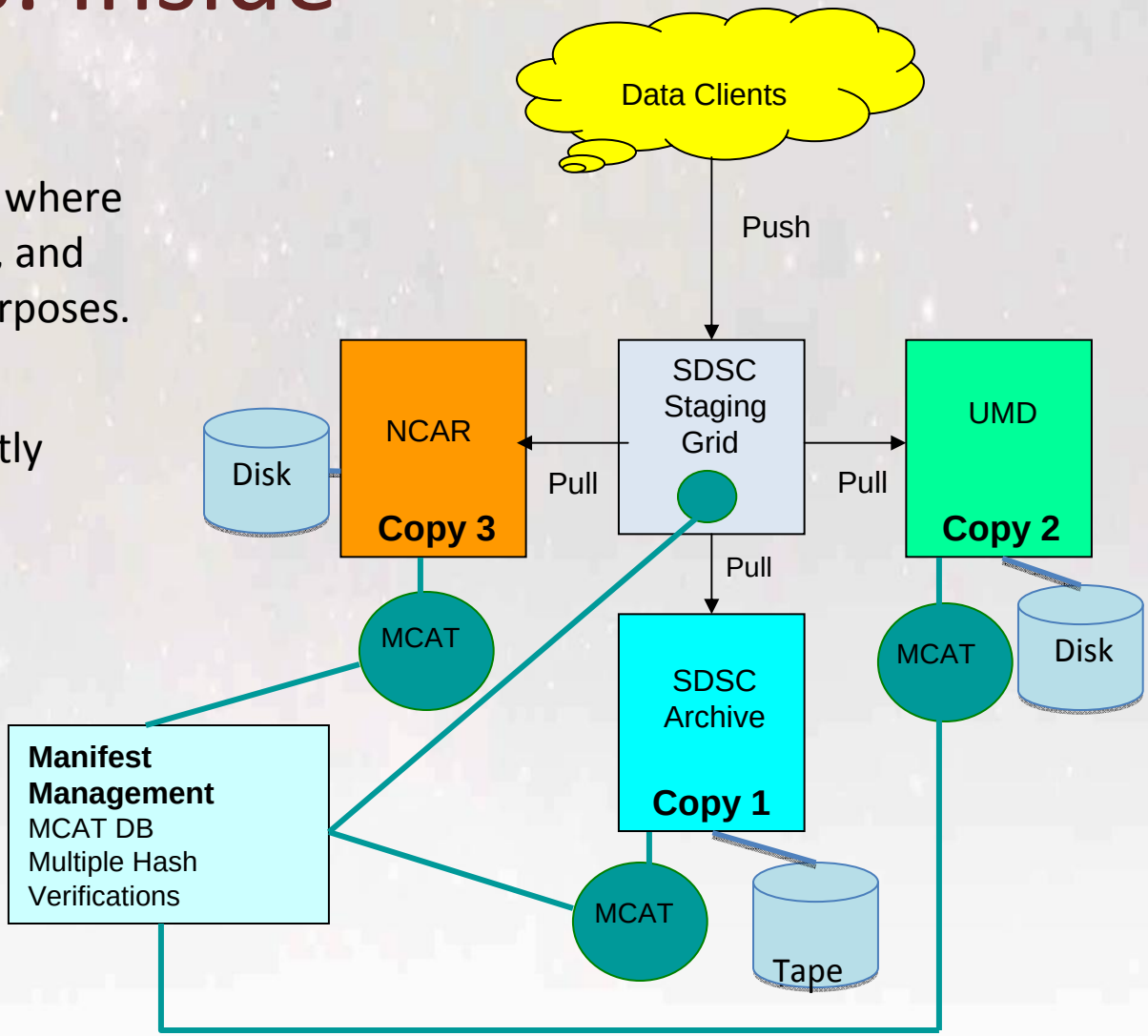
Chronopolis Project Manager

Chronopolis: Basic Facts

- Three node federated data grid at UCSD/SDSC, NCAR and UMIACS
- Capacity for up to 50 TB of data per node (150 TB total)
- Use Storage Resource Broker (SRB) for data management
- Use BagIt file packaging format and SRB protocol to transfer data
- Use Auditing Control Environment (ACE) for integrity checking
- Use SRB Replication Monitor for automated replication
- Analyzing metadata created by the various parts of the system

Chronopolis: Inside

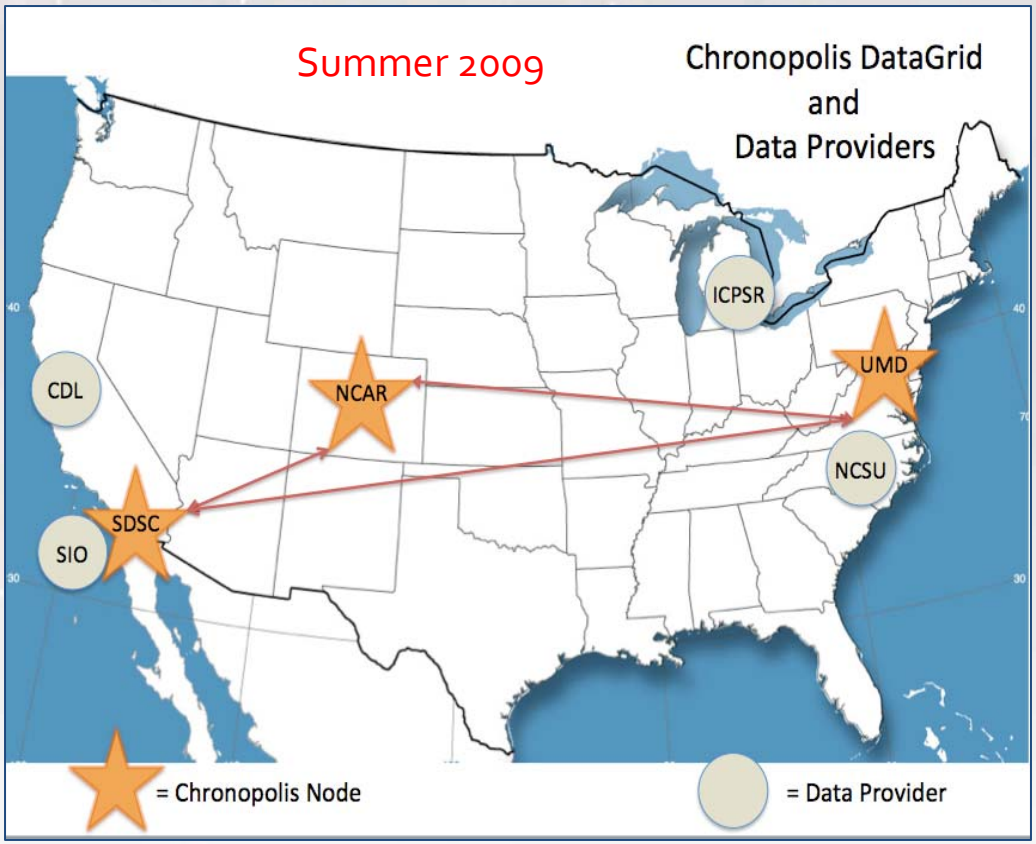
- Linked by main staging grid where data is verified for integrity, and quarantined for security purposes.
- Collections are independently pulled into each system.
- Benefits:
  - 3 independently managed copies of the collection
  - High availability
  - High reliability



Current Chronopolis collections

Data Providers:

- **Inter-university Consortium of Political and Social Research** – preservation copy of all collections including 40 years of social science data and Census
- **California Digital Library** – political and government web crawls, Web-at-risk collection
- **SIO Explorer** – data from 50 years of research voyages
- **NCSU Libraries** -- State and local geospatial data



http://chronopolis.sdsc.edu

Chronopolis next generations

- New data providers and storage nodes
- Actively pursue migration of tools
- ***More interactive planning and physical ties with other initiatives***

Chronopolis and MetaArchive

Technical Processes and Issues:

- Working with SRB and LOCKSS
- Using BagIt file packaging format as a key tool
- Active system (MetaArchive) going into an archival system (Chronopolis)

Chronopolis and MetaArchive

Philosophical Issues:

- Are we only backing up data, or also systems and infrastructure?
- Legal issues with data owners
- How is data validated – as it came in from original data owners or within the system?

Storage initiatives

- Transition to more disk-based archives
- New technologies – e.g. new HPC machine (Dash) uses SSDs
- We're already seeing customers with petabyte-sized data preservation needs

Data movement and processing

- Current data transfer rates are “good enough” for now, but ….
- Current data amounts allow for robust checksum processes and monitoring, but ….

Our current models may need to be adjusted

Broadly speaking

Focus on interoperability

- Almost all our recent proposals have been for network building
- Most of our initiatives now are multi-institution, widely geographically dispersed
- Focus will be on technologies which allow for large scale, widely distributed solutions