# *Leveraging High Performance Computing Infrastructure for Trusted Digital Preservation*

**12 December 2007**
**Digital Curation Conference**
**Washington D.C.**


## Richard Moore

**Director of Production Systems**
**San Diego Supercomputer Center**
**University of California, San Diego**
**rlm@sdsc.edu, http://www.sdsc.edu**

# *San Diego Supercomputer Center: A History of Data*

- **One of original supercomputer centers established by National Science Foundation (ca 1985)**
- **Supports high performance computing (HPC) systems with a focus on data-intensive computing**
- **Supports data applications for science, engineering, social sciences, cultural heritage institutions, etc.**

***SDSC is leveraging the infrastructure and expertise from its HPC experience for digital preservation***

# *SDSC is taking multiple paths to enable digital preservation*

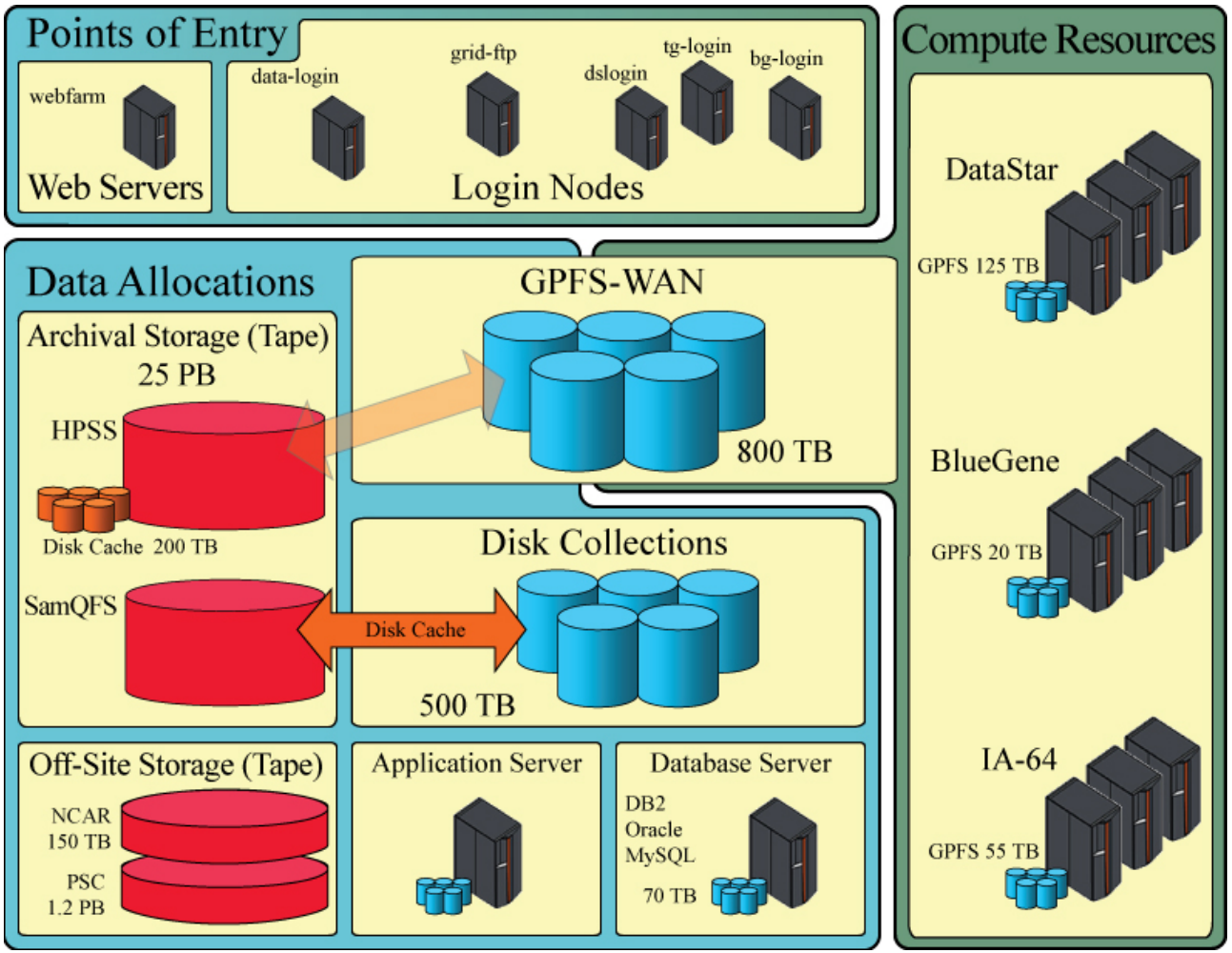
- **Production-level storage support**
  - Increased levels of reliability/verification
- **SRB & iRODS development (the *other R Moore*)**
- **Multiple preservation partnerships**
  - Demonstrating trust as digital repository
  - At local, regional, national and international levels
- **“Bit preservation” cost estimates/projections**
- **Developing sustainable business models**

# *Production Storage Services*

- **SDSC has extensive experience in the provision & operation of large-scale storage systems**
- **Size and stability allows SDSC to drive *economies of scale* and maintain deep base of staff experience**
- **SDSC has developed a stable infrastructure which allows for growth and innovation**
  - **Critical mass of staff & expertise**
  - **Use of data management tools**

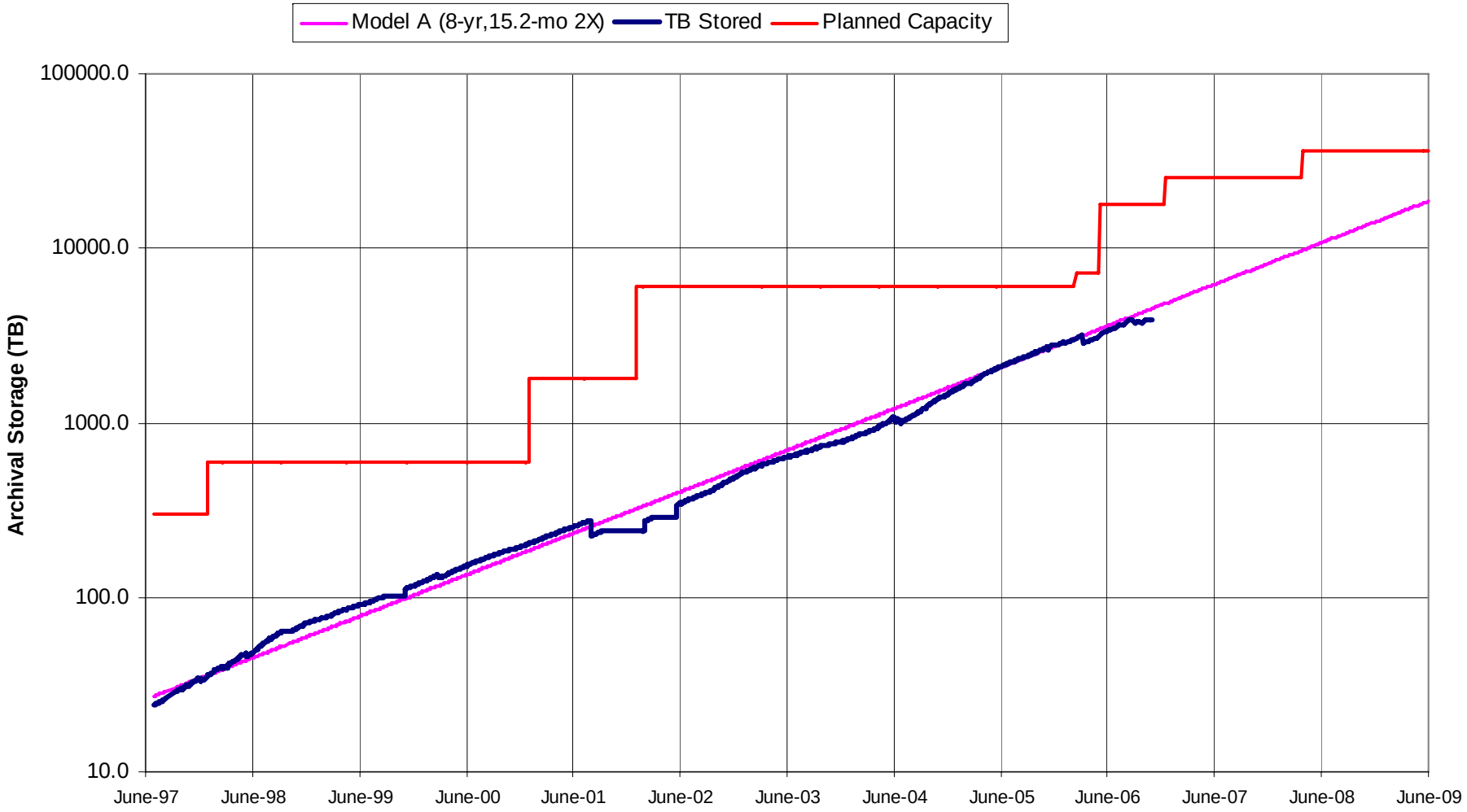
***The storage services are relatively generic – i.e. users can take advantage of the infrastructure without deep understanding of the implementation details***

# SDSC as High-Performance Data Center




- **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 Archival Capacity/Stored Data 1997 - 2007



# *Exemplary Digital Preservation Collaborations*

- **UC San Diego Libraries (local)**
  - Digital Asset Management System w/ SDSC Data Resources (SRB)
- **California Digital Library (regional)**
  - CDL Digital Repository
  - ‘Mass Transit” program, enabling Data Sharing among UC Libraries
- **Library of Congress (national)**
  - Pilot Data Center Project (2006-2007)
  - LC NDIIPP Partnership w/ ICPSR & CDL (2007-2008)
- **Southern California Earthquake Center (national)**
  - Managed library of HPC simulations & analysis
- **… Many others**

# *A Library of Congress-SDSC Pilot Project "Building Trust in a Third-Party Data Repository"*

"… demonstrate the feasibility and performance of current approaches for a production third-party digital Data Center to support the Library of Congress collections."

- One-year pilot project
- Ingest, store & serve 2 collections (~6 TB data total) w/ different usage models (e.g. static/dynamic, access patterns)
- Focus on "trust" issues - verification, change detection, audit trails
- Documentation and lessons learned

**Prokudin-Gorskii Photographs**
*http://www.loc.gov/exhibits/empire/*

**Internet Archive Web Crawls**
*http://lcweb2.loc.gov/cocoon/minerva/html/minerva-home.html*

# LC NDIIPP Chronopolis Program

## Chronopolis Digital Preservation DataGrid

**Administration for Policy and Outreach**

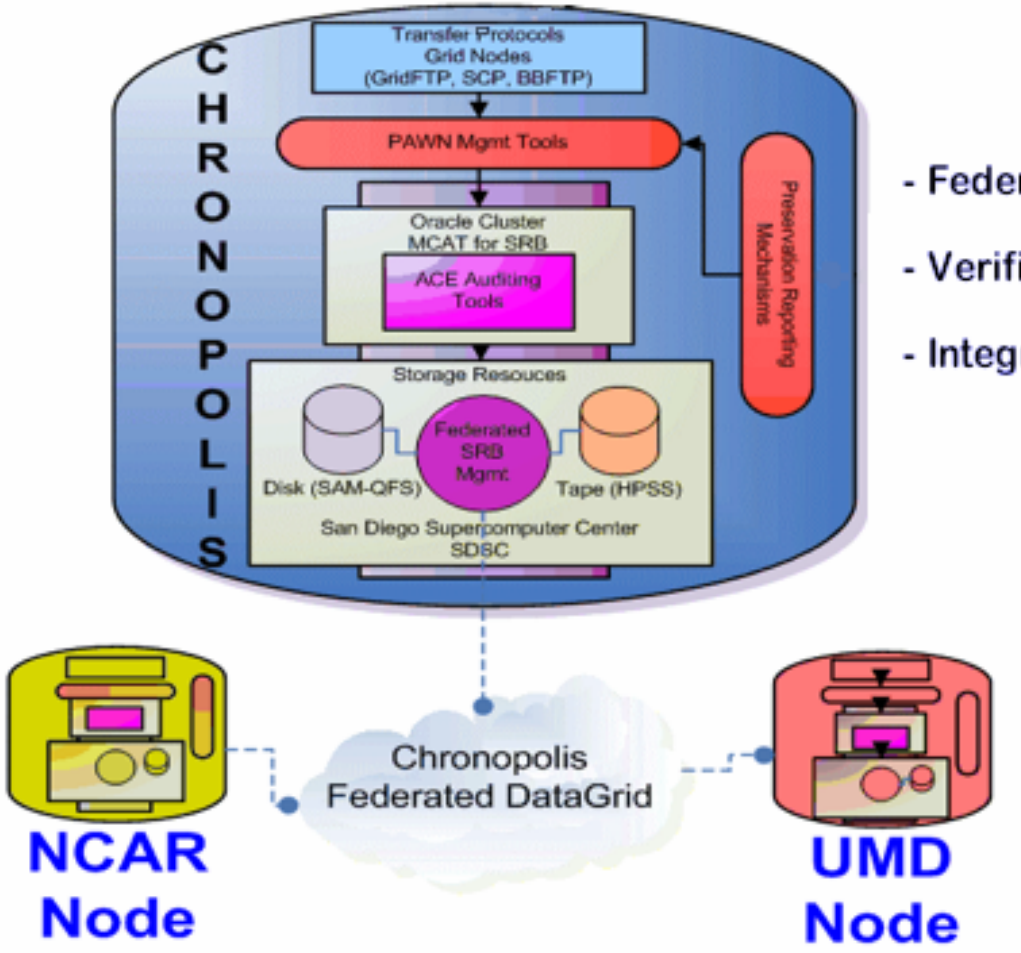
*(Supports the overall partnerships and mgmt for preservation services and works as a liaison with Chroonpolis partners and other regional and national preservation programs)*

**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

# *SDSC as Trusted Digital Repository*

- **Undertaking NARA/RLG TRAC audit 2007-2008**
- **Undertaking DRAMBORA audit 2007-2008**
- **Reagan Moore's group developing rules-based approach for NARA/RLG TRAC compliance in iRODS**
- **Developing best practices with NDIIPP partners for Federated Digital Preservation Management**
- **Establishing data reliability policies that fit both HPC Data Center and Trusted Repository Center needs**
- **Developing best practices for Data Collection packaging and transmission**

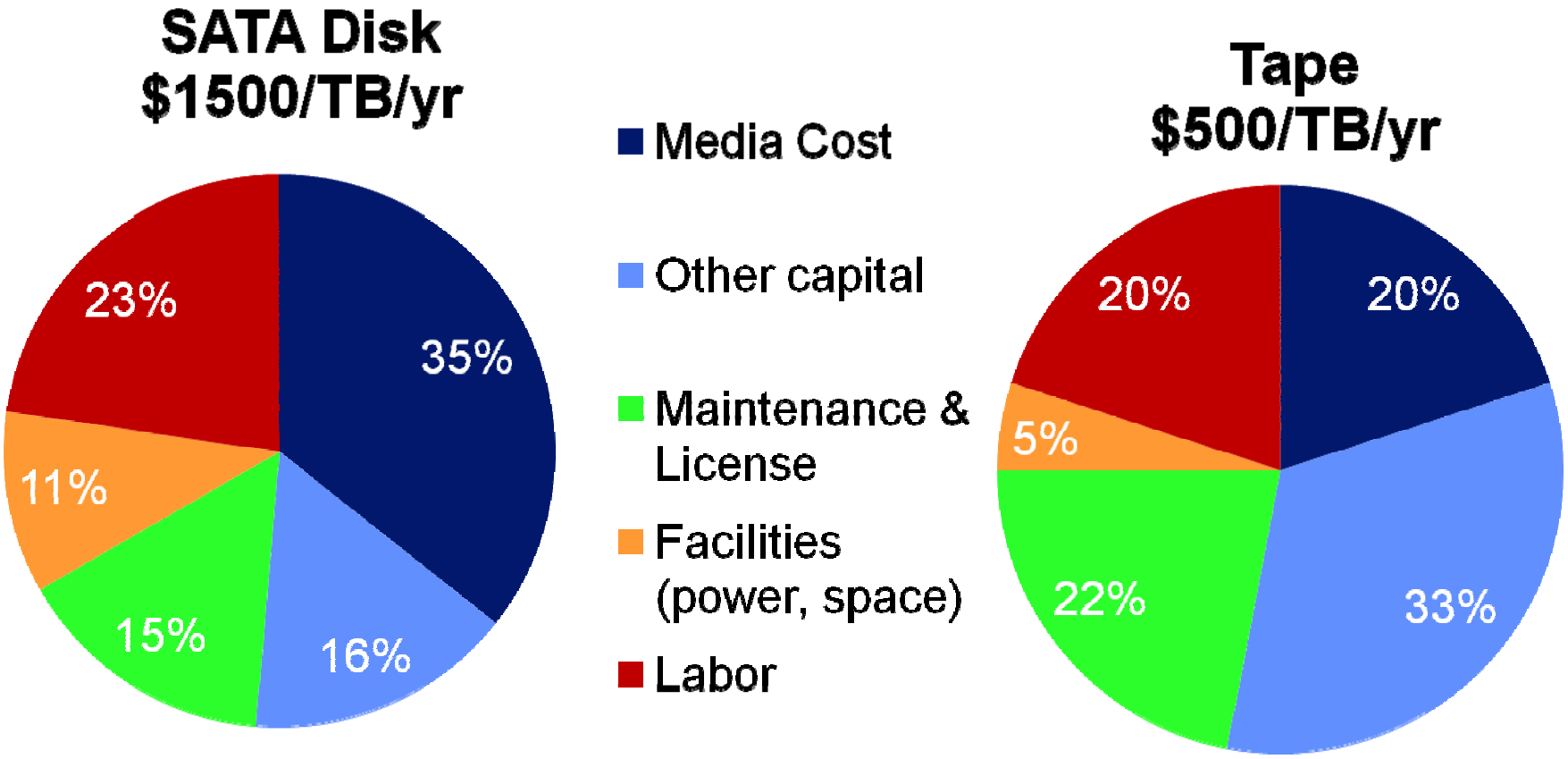
# A Three-Stage Model for A Digital Preservation Environment

**Ingest** → **Store** → **Use**

**'Bit Storage'**
- Capacity
  - Online (disk)
  - Archival (tape)
- Single-copy reliability
- Media/technology advances
- Data migration

- Replication
  - Geographically distributed
  - System diversity
- Verification & recovery
- Synchronization
  - 'Master' version
  - Propagating to replicas
- Audit trails
- Mitigation of termination risk

# Disk/Tape "Bit Storage" Cost Comparison: Relative Cost Elements



# *Future projections of "bit storage" costs*

- **If annual costs decline exponentially with a halving time ∆t, the cost to store data *in perpetuity* is finite (1.44 * ∆t * Current cost/yr)**
- **Expect that exponential declines in media costs and other IT equipment will continue for a while … current technologies as well as new technologies**
  - MAID targeting "disk archive": capital cost comparable to disk, but lower operations costs (utilities, floor space) and projections of extended lifetime
  - Disruptive technologies on horizon – e.g. holographic storage
- **Integrated cost ($/TB/yr) will decline, but how much?**
- **Critical issue is which cost elements scale with declining media costs and which do not?**
  - Most costs scale w/ media, but labor costs may not scale well
- ***Cost elements that do not scale well w/ media will dominate future costs, even at the 'bit storage' level, and undermine "finite cost"***
- ***And we expect that for total preservation 'storage' costs beyond bit storage - e.g. file management, curation, etc. - labor costs dominate!***

# *Thank You!*