

Replication Policies for Federated Digital Repositories



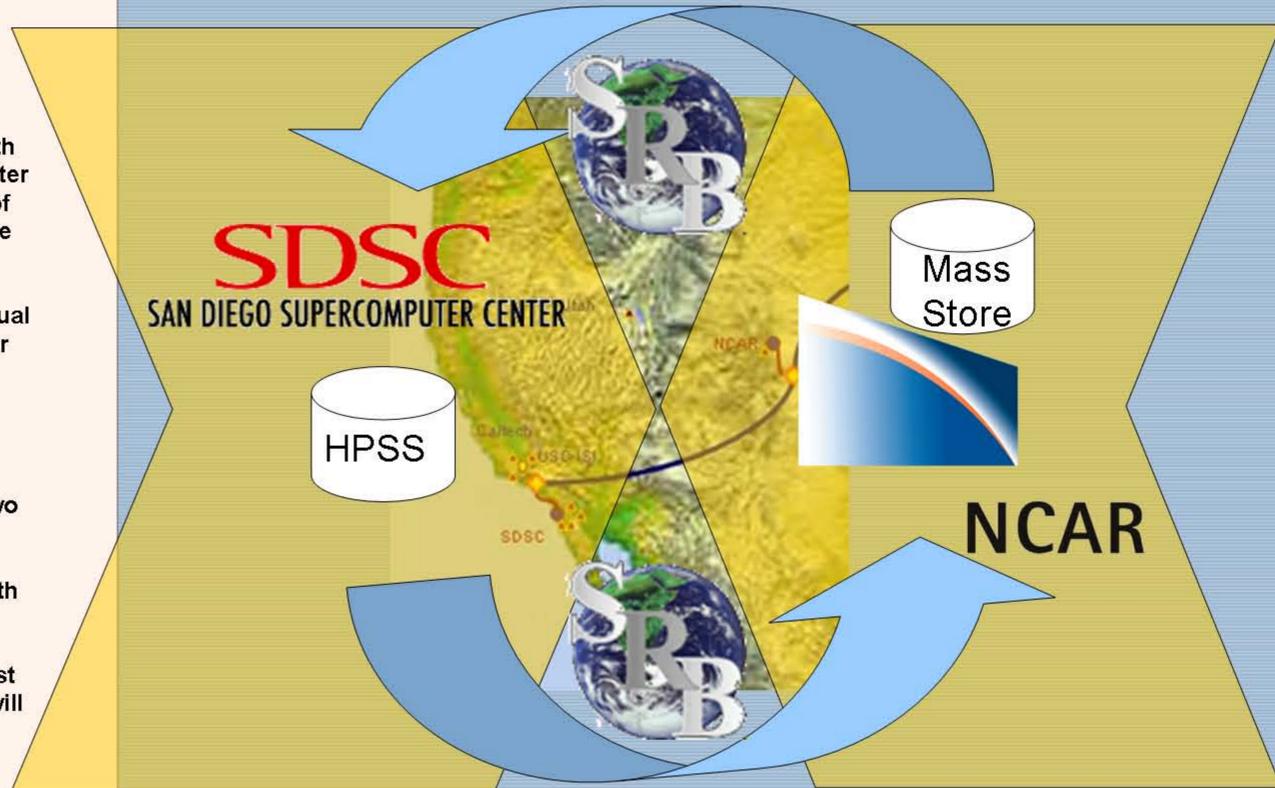
Robert H. McDonald and Christopher Jordan
San Diego Supercomputer Center (SDSC)
U.C. San Diego, La Jolla, CA - USA

Prepared for Science Archives in the 21st Century Workshop – April 2007, University of Maryland.

Introduction

The San Diego Supercomputer Center (SDSC) along with the U.C. San Diego Libraries (UCSDL), the National Center for Atmospheric Research (NCAR), and the University of Maryland Institute for Advanced Computer Science have formed a collaborative partnership called Chronopolis. The underlying goal of the Chronopolis digital preservation environment will be to curate this intellectual capital at a national-scale level in order to preserve it for the next generation of scientists and scholars.

As a first step process toward developing the Chronopolis working prototype SDSC and NCAR have signed a memorandum of understanding (MOU) to facilitate cross-replication of collections between the two sites. In its initial stages this redundancy serves as merely storage swap. However, each institution is developing its own policies for collection replication with the goal of creating policies that can interact independently as well as cross-institutionally for the replication of critical data. This iterative process is a first step-towards a model cross-institutional strategy that will eventually extend to all partners working in the Chronopolis preservation environment.



Materials and Methods

The storage swap methodology described in this presentation is based on Storage Resource Broker (SRB) zone federation and utilizes both the Mass Storage system at NCAR and the HPSS storage system at SDSC. The data is replicated over the TeraGrid network connection between NCAR and SDSC.

NCAR Mass Storage System (MSS) - The Mass Storage System is a central, large-scale data archive that stores data used and generated by climate models and other programs executed on NCAR's compute servers. The NCAR MSS holds more than 3 petabytes (PB) of stored data and has a net growth rate of 60-70 terabytes (TB) per month.

SDSC HPSS – HPSS is IBM developed software that manages data across disk and robotic tape libraries. SDSC HPSS is used to provide highly flexible and scalable hierarchical storage using the nearly 25 petabyte (PB) capacity of tape storage available at SDSC.

STORAGE GROWTH

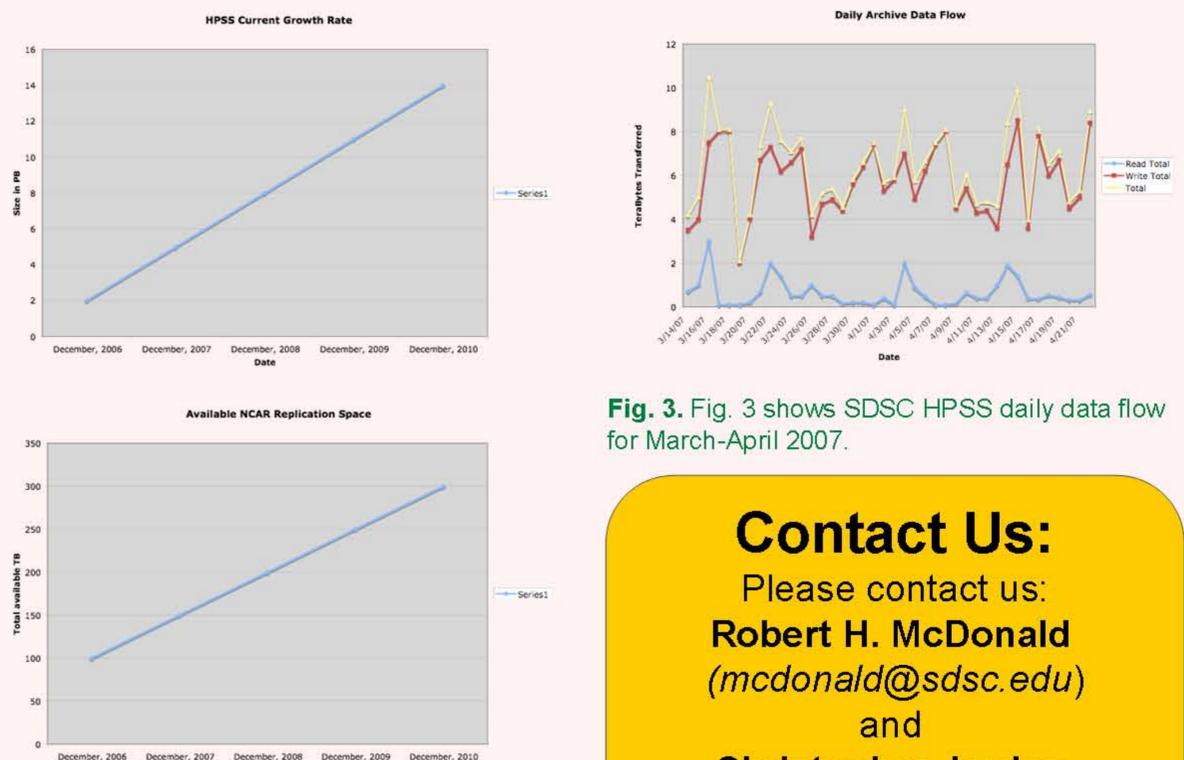


Fig. 3. Fig. 3 shows SDSC HPSS daily data flow for March-April 2007.

Figs. 1 and 2. Fig. 1 shows SDSC HPSS current growth rate. Fig. 2 shows expected growth for SDSC NCAR Replication Space.

Contact Us:

Please contact us:
Robert H. McDonald
(mcdonald@sdsc.edu)
and
Christopher Jordan
(ctjordan@sdsc.edu)

Conclusions

Data growth in the science and engineering domains will continue to outpace the capability of any one institution for replication infrastructure. Models for shared data-cyberinfrastructure are needed. SRB offers one solution for shared infrastructure replication. Future work is needed on automating repository replication policies and workflows within the storage and preservation environment.

Acknowledgments

This project is a synopsis of work done by a variety of personnel at SDSC and NCAR. The authors would like to thank all participants for their input especially the SDSC Data Reliability Committee.