Timing is Right for SDSC Cloud

New Storage System Supports NSF Data Policy

October 5, 2011

Jan Zverina

Successfully managing, preserving, and sharing large amounts of digitally-based data has become more of an economic challenge than a technical one, as researchers must meet a new National Science Foundation (NSF) policy requiring them to submit a data management plan as part of their funding requests, said Michael Norman, director the San Diego Supercomputer Center (SDSC) at the University of California, San Diego.

"Data management has become an even more challenging discipline than high-performance computing," Norman said during remarks delivered at the 50th anniversary meeting this week of the Association of Independent Research Institutes (AIRI) in La Jolla, California. "The question used to be 'what's the essential technology?' but is now 'what's the sustainable cost model?'"

The revised NSF policy, which went into effect early this year, asks researchers to submit a two-page data management plan on how they will archive and share their data. According to the policy, "investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants."

Norman said this revised policy was one of the key drivers that shaped SDSC's planning for a new Web-based, 100 percent disk data storage system called the SDSC Cloud, which was announced late last month. Believed to be the largest academic-based cloud storage system in the U.S. to-date, the SDSC Cloud is primarily designed for researchers, students, and other academics requiring stable, secure, and cost-effective storage and sharing of digital information, including extremely large data sets. While SDSC's primary motivation to create its own data cloud was to provide an affordable resource for UC San Diego researchers to preserve and share their data, the resource is being made available to all academic researchers. "Whatever we want to call it - the data deluge, the data tsunami, or the data explosion - the fact is that we are now in the era of data-intensive computing and SDSC has been working to solve a major challenge for a whole collection of scientific disciplines: the cost of data storage and sharing," he said. Standard "on-demand" storage costs for UC researchers on the SDSC Cloud start at only $3.25 a month per 100GB (gigabytes) of storage. A "condo" option, which allows users to make cost-effective long term investment in hardware that becomes part of the SDSC Cloud, is also available. Full details can be found at https://cloud.sdsc.edu/hp/index.php.

'Bit Cemetery' Historically, data management has been a project-related cost for major research facilities, which traditionally have been funded to cover the cost of preservation and access, said Norman. "The NSF's Office of Cyberinfrastructure (OCI), historically focused on high-performance computing, while data management was secondary and consisted of archiving in a tape-based silo," he said. "We now call that a 'bit cemetery' because use and retrieval rates were so low," Norman told AIRI attendees. "In fact, many researchers were writing their data once and reading it never. That's not data management - that's data burial, and not what active researchers need."

Over the last few years, however, the research infrastructure for data-enabled science has been widely discussed at the NSF, leading to the new data management and sharing policy. The document that is charting the course is called Cyberinfrastructure Framework for the 21st Century Science and Engineering (CIF21). "This document works across all the NSF Directorates and finally makes data-enabled science a first-class citizen," said Norman. "And during the last year and a half, the NSF has been moving from vision to policy to action."

Still, Norman warned that researchers will likely never be able to afford to save all their data, and should focus on saving and sharing only what is intellectually valuable, while creating a sustainable business model. He referenced a KRDS report (Keeping Research Data Safe) that said the cost of long-term data stewardship is as much as six times the cost of bit preservation. "So it's not the costs of storing the bits - it's the cost of hosting the hardware, all of the administration costs, and the costs of migrating the data.

In late 2007, the Blue Ribbon Task Force on Sustainable Digital Preservation and Access was commissioned by the NSF and The Andrew W. Mellon Foundation to study the economic sustainability challenge of digital preservation and access. The Task Force, which worked in partnership with the Library of Congress, the Joint Information Systems Committee of the United Kingdom, the Council on Library and Information Resources, and the National Archives and Records Administration, published both an Interim and Final report, which can be found at http://brtf.sdsc.edu/.

"SDSC, like other data resource centers, has a long-term obligation to steward that data, and maintenance costs are needed to keep that data persistent," said Norman. "It's like real estate. You can either rent out your rooms or sell your condos, but if you're not recovering costs as a landlord, you go out of business."

Media Contacts: Jan Zverina, SDSC Communications, 858 534-5111 or jzverina@sdsc.edu Warren R. Froelich, SDSC Communications, 858 822-3622 or froelich@sdsc.edu