UC San Diego
UC San Diego News Center

September 22, 2011 | By Jan Zverina

SDSC Announces Scalable, High-Performance Data Storage Cloud

Web-based System Offers High Durability, Security, and Speed for Diverse User Base

The San Diego Supercomputer Center (SDSC) at the University of California, San Diego, today announced the launch of what is believed to be the largest academic-based cloud storage system in the U.S., specifically designed for researchers, students, academics, and industry users who require stable, secure, and cost-effective storage and sharing of digital information, including extremely large data sets.

“We believe that the SDSC Cloud may well revolutionize how data is preserved and shared among researchers, especially massive datasets that are becoming more prevalent in this new era of data-intensive research and computing,” said Michael Norman, director of SDSC. “The SDSC Cloud goes a long way toward meeting federal data sharing requirements, since every data object has a unique URL and could be accessed over the Web.”

SDSC’s new Web-based system is 100% disk-based and interconnected by high-speed 10 gigabit Ethernet switching technology, providing extremely fast read and write performance. With an initial raw capacity of 5.5 petabytes – one petabyte equals one quadrillion bytes of storage capacity, or the equivalent about 250 billion pages of text – the SDSC Cloud has sustained read rates of 8 to 10 gigabytes (GB) per second that will continually improve as more nodes and storage are added. That’s akin to reading all the contents of a 250GB laptop drive in less than 30 seconds.

Moreover, the SDSC Cloud is scalable by orders of magnitude to hundreds of petabytes, with aggregate performance and capacity both scaling almost linearly with growth. Full details about the new SDSC Cloud can be found at http://cloud.sdsc.edu.

Conceived in planning for UC San Diego’s campus Research Cyberinfrastructure (RCI) project, the initiative quickly grew in scope and partners as many saw the technology as functionally revolutionary and cost effective for their needs. At launch, users and research partners include, among others, UC

San Diego's Libraries, School of Medicine, Rady School of Management, Jacobs School of Engineering, and SDSC researchers, as well as federally-funded research projects from the National Science Foundation, National Institutes for Health, and Centers for Medicare and Medicaid Services.

"The SDSC Cloud marks a paradigm shift in how we think about long-term storage," said Richard Moore, SDSC's deputy director. "We are shifting from the 'write once and read never' model of archival data, to one that says 'if you think your data is important, then it should be readily accessible and shared with the broader community.'"

"UC San Diego is one of the most data-centric universities in the country, so our goal was to develop a centralized, scalable data storage system designed to meet performance, functionality, and capacity needs of our researchers and partners across the country, and to evolve and scale with the needs of the scientific community," said Dallas Thornton, SDSC's division director of cyberinfrastructure services. "Developing this resource in-house atop the OpenStack platform allows for highly-capable and flexible, yet extremely cost-effective solutions for our researchers."

OpenStack is a scalable, open-sourced cloud operating system jointly launched in July 2010 by NASA and Rackspace Hosting, which today powers some of the largest public and private cloud computing services using this scalable and proven software.

Durability and Security

Data stored in SDSC's new cloud is instantly written to multiple independent storage servers, and stored data is validated for consistency on a round-the-clock basis. "This leads to very high levels of data durability, availability, and performance, all of which are of paramount importance to researchers and research organizations," said Ron Joyce, SDSC's associate director of IT infrastructure and a key architect of the system.

The SDSC Cloud leverages the infrastructure designed for a high-performance parallel file system by using two Arista Networks 7508 switches, providing 768 total 10 gigabit (Gb) Ethernet ports for more than 10Tbit/s of non-blocking, IP-based connectivity. The switches are configured using multi-chassis link aggregation (MLAG) for both performance and failover.

"This network configuration allows us to unshackle extreme-scale/extreme-performance storage from individual clusters and instead make data available at unprecedented speeds across our university campus and beyond," said Philip Papadopoulos, SDSC's division director of UC systems. "In addition to incredibly fast data transmission speeds, our goal was to build a high-performance storage system right from the start that was completely scalable to meet the evolving needs and requirements of the campus, as well those within industry and government."

The environment also provides high-bandwidth wide-area network connectivity to users and partners thanks to multiple 10Gb connections to CENIC (Corporation for Education Network Initiatives in California), ESNet (Energy Sciences Network), and XSEDE (Extreme Science and Engineering Discovery Environment). This allows huge amounts of data, such as sky surveys or mapping of the human genome, to be rapidly transported simultaneously to/from the SDSC Cloud.

In addition to large storage capacity and high-speed transmissions, the SDSC Cloud provides:

- **Cost advantages:** Standard "on-demand" storage costs start at only $3.25 a month per 100GB of storage, and there are no I/O networking charges. 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. Users will soon have the option to have additional copies of their data stored offsite at UC Berkeley, one of SDSC's partners in the project.
- **Anywhere, anytime accessibility and wide compatibility:** Every data file is given a persistent URL, making the system ideal for data sharing such as library or institutional collections. Access permissions can be set by the data owner, allowing a full spectrum of options from private to open access. The HTTP-based SDSC Cloud supports the RackSpace Swift and Amazon S3 APIs and is accessible from any web browser, clients for Windows, OSX, UNIX, and mobile devices. Users can also write applications that directly interact with the SDSC Cloud.

- **Enhanced security:** Users set their own access/privacy levels. Users know and can coordinate precisely where their data is stored in the cloud, including replicated copies. In addition, a HIPAA and FISMA compliant storage option launches on October 1st in partnership with the Integrating Data for Analysis, Anonymization and SHaring (iDASH) program at UC San Diego, a National Center for Biomedical Computing (NCBC) project funded in 2010 under the NIH Roadmap for Bioinformatics and Computational Biology.

Working in Tandem with Other SDSC Storage Systems

The SDSC Cloud is configured to work in tandem with other innovative storage technologies at the supercomputer center. One is the *Data Oasis* system, a Lustre-based parallel file system designed primarily for high-performance, low-latency scratch and medium-term project storage, ideal for researchers conducting data-intensive operations on SDSC's *Triton*, *Trestles*, and *Dash* high-performance computing (HPC) systems.

SDSC's *Data Oasis* is currently capable of speeds of 50GB/s, meaning that researchers can today retrieve a terabyte of data – or one trillion bytes – in less than 20 seconds. By early 2012, *Data Oasis* will be expanded to serve SDSC's *Gordon*, the first supercomputer within the HPC community focused on integrating large amounts of flash-based SSD (solid state drive) memory. As *Gordon* enters production in January 2012, SDSC will double the speed of *Data Oasis* to 100GB/s, making it one of the fastest parallel file systems in the academic research community. While *Data Oasis* is used for in-

process HPC storage, the SDSC Cloud is designed to accommodate any storage needs either prior to or afterward, delivering durable, secure storage that can be shared within SDSC or across the country with ease.

MEDIA CONTACT

Jan Zverina, SDSC Communications, 858 534-5111 or jzverina@sdsc.edu

UC San Diego's Studio Ten 300 offers radio and television connections for media interviews with our faculty, which can be coordinated via studio@ucsd.edu. To connect with a UC San Diego faculty expert on relevant issues and trending news stories, visit https://ucsdnews.ucsd.edu/media-resources/faculty-experts.