MSSTC Tutorial
May 3, 2010
Lake Tahoe

Integration of Cloud Computing and Cloud Storage

Patrick Dreher
Director Advanced Computing Infrastructure and Systems
Adjunct Professor Dept. of Computer Science, NCSU

Policy-based Data Management

Reagan Moore
University of North Carolina at Chapel Hill

Scenarios for interaction of cloud computing and massive storage.

Data intensive computing requires the ability to parse and analyze massive data collections. In the traditional grid computing approach, the data are read from the archive and streamed to a compute server for processing. In this tutorial, we will explore an alternate approach that relies upon the integration of cloud computing and cloud storage. Specifically, we will examine on-demand computing applied at remote storage locations to generate the data subsets of interest, and the analysis of the data subsets within a cloud computing environment. Demonstration of some of the efforts underway will be included in the tutorial for hands-on experience for the audience. Both data subsetting and data processing will be demonstrated. The specific technologies include:

- Policy-based data management, and the use of server-side workflows to apply data subsetting commands on petabyte data collections. (Moore)
- Cloud storage as a data caching environment for on-demand computing
- Cloud computing, with examples of creation of complex data products from data cached in cloud storage (Dreher)
- Cloud computing, with examples of data processing pipelines that execute in the cloud and that deposit results into the petabyte data collection

When collections grow to the multi-petabyte scale, it becomes very hard to move the data to another site. This means the creation of derived data products to satisfy user requests needs to be done at the archive through use of on-demand computing resources (cloud computing). We will need mechanisms to do the data sub-setting at the storage resource, and processing on the data subset though cloud computing.

Data processing pipelines that have varying data rates can use cloud computing to meet on-demand processing. Data will need to be cached near the cloud computing resource. The results of the processing will need to be archived in the multi-petabyte storage environment. Data will need to be retrieved from the archive to compare current measurements with previous measurements.

Thus a plausible scenario is to use cloud computing to do the processing needed to display and manipulate data stored in the massive archive, and to do the processing needed on ingestion of data into the massive archive.
Multiple communities want to promote collaborations while controlling the properties of the shared data. Policy-based data management systems are being developed to enforce the policies at the remote storage location.
Related challenges are integration of data from multiple communities (ontologies, information registration resources, format registration sources), manipulation of data in the distributed environment, automation of administrative functions, and validation of the properties of the shared collection (what it represents).

Each community has a driving purpose behind the formation of an extended data collection that contains digital documents from multiple institutions. The purpose could be to promote collaborative research, or to integrate digital holdings to form a more comprehensive collection, or to mitigate risk of data loss by replicating to another site, or to assemble an authoritative collection or to build a reference collection. For each collection, policies are needed to ensure that the driving purpose has been met and that the collection has the desired properties. These policies control the ingestion of material into the collection, and the distribution, disposition, retention, access, and allowed data manipulation.

Policy-based data management provides the management virtualization that is needed to ensure your institution's policies are being enforced on your data within the distributed environment.

Many communities are now developing policy-based management of their collections. Some examples will be discussed and results of their efforts reviewed during the tutorial:

- NASA National Center for Computational Sciences
- JPL Planetary Data System
- Teragrid
- NOAA National Climatic Data Center
- NARA Transcontinental Persistent Archive Prototype
- NSF Oceanography Observatories Initiative
- NSF Science of Learning Centers
- Australian Research Collaboration Service
- SHAMAN Sustaining Heritage Access through Multivalent Archiving
- Carolina Digital Repository
- NHPRC Distributed Custodial Archiving Preservation Environment
- RENCI data grid
- Texas Digital Library

"Sustainable Economics for a Digital Planet: Ensuring Long-term Access to

Digital Information and a National Conversation on Digital Preservation"

Elizabeth A. Cohen, Vice Chair for Education
of the Science and Technology Council
of the Academy of Motion Picture Arts and Sciences™

Ann Kerr
Vice Chair International Symposia
IEEE MSSTC
AKConsulting

In the Information Age, digital information has revolutionized almost every aspect of our lives, from the way we access or store our favorite music and family photographs, to how our society conducts commerce, research and education. Underlying the potential of the Information Age and its paradigm-shifting access to digital information is the assumption that key information will be there when we want it, where we want it, and for the foreseeable future.
Realizing the potential of the Information Age spawns a series of daunting challenges for the future, how will we ensure the long-term preservation and access to our digital information, growing exponentially with each passing day? How will we successfully migrate data as technology moves from one preservation medium to the next? Who should determine which digital data should be saved, and what criteria will be used to make those decisions?

Perhaps even more challenging is the issue of economic sustainability. What is the cost to preserve valuable data and who will pay for it? Broadly speaking, economic sustainable digital preservation will require new models for channeling resources to preservation activities; efficient organization that will make these efforts affordable; and recognition by key decision-makers for the need to preserve, with appropriate incentives to spur action.

To address these issues, the Blue Ribbon Task Force on Sustainable Digital Preservation and Access was created in late 2007. In its work over two years, the BRTF-SDPA explored the sustainability challenge with the goal of delivering specific recommendations that are economically viable and of use to a broad audience, from individuals to institutions and corporations to cultural heritage centers. The BRTF-SDPA is funded by the National Science Foundation and the Andrew W. Mellon Foundation, 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.

The Final Report from the Blue Ribbon Task Force on Sustainable Digital Preservation and Access, called "Sustainable Economics for a Digital Planet: Ensuring Long-term Access to Digital Information", is the result of a two-year effort focusing on the critical economic challenges of preserving an ever-increasing amount of information in a world gone digital. The full report is available online http://brtf.sdsc.edu/. Released in February 2010 it includes a wide range of recommendations for decision makers and stakeholders to consider as they seek economically sustainable preservation practices for digital information.

The Task Force held a one day Symposium, Called **"A National Conversation on the Economic Sustainability of Digital Information"** on April 1, 2010 in Washington, D.C. It convened a diverse group of speakers from the academic, private, and public sectors to discuss one of the most pressing issues of the Information Age: identifying practical solutions to the economic challenges of preserving today's deluge of digital data. A spectrum of national leaders from the Executive Office of the President, the Academy of Motion Picture Arts and Sciences, the Smithsonian Museum, Nature Magazine, Google, and other organizations for whom digital information is fundamental for success spoke at the meeting.

Findings and Recommendations from the Report and the April Symposium will be reviewed as a basis for further discussion with the Tutorial Attendees. Active discussion on the limitations of current storage technologies and need for developing more robust and reliable systems for to insure preservation of critical digital content will be held.

New initiatives in Cloud Computing – Library of Congress NDIIP Program

Leslie Johnston, Manager of Technical Architecture Initiatives
National Digital Information Infrastructure & Preservation Program
Office of Strategic Initiatives Library of Congress
lesliej@loc.gov

The Library of Congress National Digital Information Infrastructure and Preservation Program (NDIIPP) is working to is to develop a national strategy to collect, preserve and make available significant digital content. NDIIPP is based on an understanding that digital stewardship on a national scale depends on public and private communities working together. The program focuses on three areas: Capturing, preserving, and making available significant digital content; building and strengthening a network of partners; and developing a technical infrastructure of tools and services. The Library has built a preservation network of over 130 partners from across the nation to tackle the challenge, and is working with them on a variety of initiatives. An overview of the program to date will be presented, showcasing a variety of stewardship models. New initiatives will also be discussed, emphasizing those in support of cloud computing.