Cyberinfrastructure for Managing Research Data at the University of Colorado Boulder

Thomas Hauser, thomas.hauser@colorado.edu
Director of Research Computing
University of Colorado Boulder

Cyberinfrastructure

- Cyberinfrastucture consists of the computer hardware, software, networks, expertise and integration that supports and encourages research, collaboration and discovery.
- An effective sociotechnical infrastructure should “fit the needs, activities, and contexts of the people who use it, as well as those of the people who create it, operate it, and contribute to its content ”

Research Data Challenges

- Research data can be 'big' in different ways [1]
  - Volume that challenges our computing, storage and network infrastructure
  - Lasting significance, e.g. clinical trial, environmental data
  - Descriptive challenges, e.g. experimental setup
- Research Computing's (RC) approach to research data challenges
  - Partner with faculty and researchers
  - Networking (RC-DMZ) for large data transfers
  - PetaLibrary storage infrastructure
  - Large scale compute
  - Research Data Services (RDS)
    - Collaboration between RC and the CU-Boulder Libraries
    - Under development

[1] C. Lynch, "Big data: How do your data grow?," *Nature*, vol. 455, no. 7209, pp. 28–29, Sep. 2008.

Data Intensive Projects at CU-Boulder

- National Snow and Ice Data Center
- High Energy Physics (HEP) group runs an Open Science Grid (OSG)
- BioFrontiers institute that collaborates with the Anschutz Medical Center in Denver
- Researchers in the Department of Computer Science analyze twitter data
- Intermountain Neuroimaging Consortium (CU-Boulder + Mind Research Network)
- Museum of Natural History

Priorities for Data Management

- Top 5 issues from EDUCAUSE ACTI members:
  1. Data storage services and technologies (local and in the cloud).
  2. Shared services issues (e.g., funding, organization, administration, governance, support, policies and processes).
  3. Data management training for faculty and students.
  4. Current data management models and how to migrate to the future.
  5. Metadata management and development of best practices.
- What's missing in the above list?

Scientific Data Movement

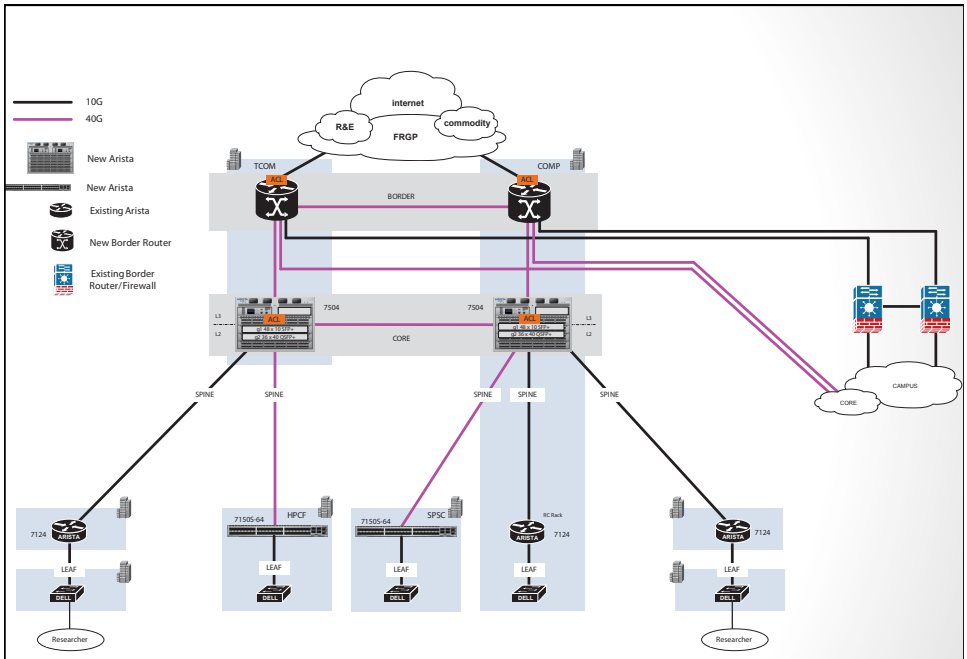
- The Task: Large Data Transfer
  - "End to End" – Disk, Network Card, OS, Application Protocols, LAN, WAN
  - Topologically and Physically complex (e.g. multi-domain)
- The Concerns
  - Machine/OS/Protocol Tuning (e.g. TCP as the typical choice in this space)
- High Energy Physics is prime example that a collaboration and "end to end" tuning and performance debugging is necessary

Science DMZ

- Concept involves some important players:
  - Architectural Split
    - Enterprise vs. Science use
    - Migration of "big data" off of the LAN (mutually benefits the "regular users" too)
  - Security and Networking
    - Paradigm shift – learn to trust things vs untrust of everything
    - Router filters are faster than firewalls
  - "Using the Right Tools"
    - Monitoring of the network – perfSONAR
    - Dynamic allocation of bandwidth – DYNES, OSCARS, OpenFlow
    - Proper data movement applications
      - SCP = Bad
      - **GridFTP = Good**
    - Well tuned servers (hardware, software, and protocol stack)

CU-Boulder's Science DMZ: Current and future

- RC-DMZ: Collaboration with OIT Networking group and RC
- Current RC-DMZ
  - Campus wide dedicated 10 gig
  - Dedicated uplinks
  - Single path single point of failures
- This summer (funded through NSF CC NIE)
  - Upgrade border routers to be 100G and OpenFlow capability
  - Add redundant paths and more paths between key sites using Arista switches and MLAG
  - Dedicated perfSonar and BRO nodes



GridFTP and Globus Online

- Gridftp resources (Globus Online)
  - Four Dell PowerEdge R710s as GridFTP servers
  - Dedicated 10Gb ethernet per node
- External access via science DMZ
  - colorado#gridftp
- Internal access via dedicated private vlans
  - colorado#jila, colorado#nsidc
  - --data-interface <vlan>

Projects Enabled by RC-DMZ

- DYNES
- High speed access to central research data storage
  - Globus Online
- Enabling different groups to have data driven high speed workflows applications
  - HEP – openscience grid node
  - NSIDIC: sharing of data
  - JILA: Transfer to and from XSEDE resources

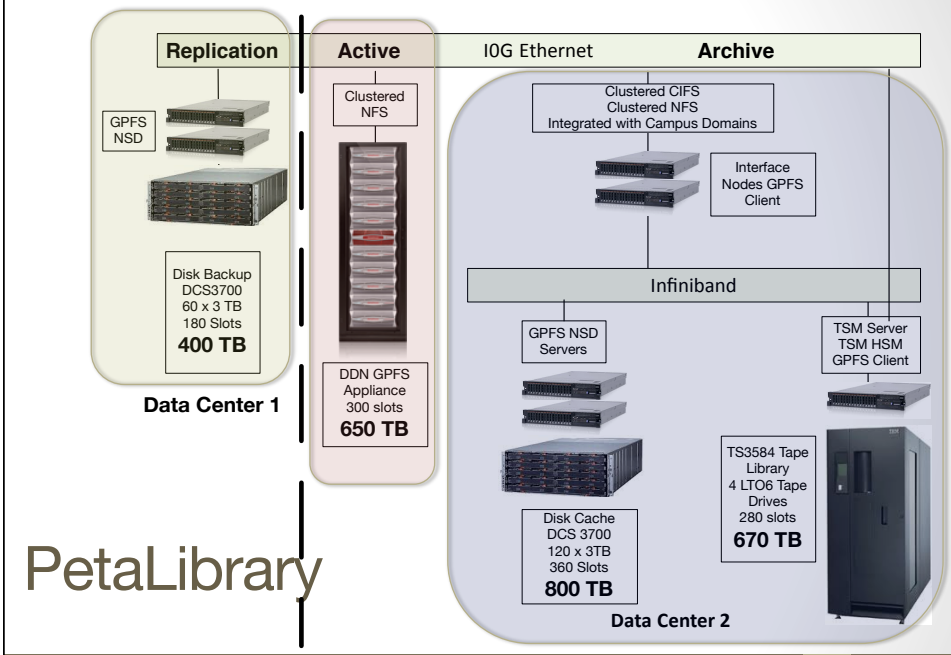
Performance Numbers

- Over a period of about 9 month

Data transferred from colorado#gridftp	122.5 TB
Data transferred to colorado#gridftp	21.6 TB
Peak transfer rate between distinct endpoints	2.9 Gb/s
Peak transfer rate to/from Janus (disk)	5.9 Gb/s
Peak transfer rate to/from Janus (memory)	9.5 Gb/s

Research Data Storage: NSF Funded Petalibrary

- NSF Major Research Instrumentation grant
- Provide a condominium model for large storage needs
  - Data collections from faculty and students
  - Deposition and discovery of data
- Provide expertise and services around this storage
  - Data management & Curation
- Partners include: *Libraries, LASP, Museum, BioFrontiers, NSIDC, APS, JILA, CIRES, High Energy Physics, Institute for Cognitive Science, Aerospace Engineering, Mechanical Engineering, INSTAAR, and Renewable & Sustainable Energy Institute*



PetaLibrary Cost for Researcher

Service offered	Cost/TB/yr
Active	$100
Active plus replication	$200
Active plus one copy on tape	$120
Archive (HSM)	$60
Archive plus one copy on tape	$80

- No HIPPA, FERPA, ITAR data
- Infrastructure guaranteed for 5 years
- Reporting requirement from PI about research activities supported

Data Management

- Completed data management task force
  - Report: http://hdl.handle.net/10971/1398
- Working on implementing recommendation
  - PetaLibrary
  - Data management services
  - Governance
  - Policies

DMTF recommendation: Highlight and Encourage Research Data Management

- **Alignment of faculty review systems** with the NSB guiding principles. Deans, Chairs, faculty governance, and tenure and promotion evaluation committees should consider the intellectual value of data creation, sharing, and stewardship in their evaluative work (e.g., faculty who demonstrate the positive impact of their shared data or open publications are rewarded at review cycles);
- **Faculty to consider various forms of Open Access publishing** and archiving of publications, and to consider policies like those of Harvard and other Tier One universities given the close connection between Open Access publishing and Open Data sharing noted by the NSB in point 2 above, and;
- **Faculty to adopt policies for sharing data in the most open way possible,** given the norms of each discipline and the rights and responsibilities of individual researchers. The DMTF recognizes that the results of some research cannot be shared openly, but encourages faculty to do so whenever possible.

Create a Research Data Services (RDS) Organization

- Advisory Committee
  - Campus policy
  - Broad representation of faculty
- Executive committee
  - Support the RDS operations
  - AVCR, Director of Research Computing and Senior personnel from the University Libraries
- Operations
  - Virtual organization
  - Research Computing, Libraries and others with relevant experience, e.g. NSIDC

Data Management Services

- http://data.colorado.edu
- Data management plan consulting
  - DMPTool with CU-Boulder credentials
  - Consulting with Research Computing and the CU-Boulder Libraries
- Data archiving support
  - Find appropriate archives
  - Long term goal: use PetaLibrary storage to serve a data repository

Develop research data governance and procedures

- Clearly define “researchers” and “research data”
- Clarify ownership of intellectual property rights for research data
- Provide legal (e.g., HIPAA, ITAR) and ethical (e.g., privacy) guidance
- Recommend security measures
- Determine appropriate periods of retention
- Work with broad faculty representation to define what types of data (e.g., raw data, publishable data, metadata) should be shared
- Promote widespread access to research data while accounting for disciplinary and community norms, and ethical and legal considerations
- Determine the roles and responsibilities of researchers and the institution

Future Goals

- Establish a sustainable sociotechnical infrastructure necessary for full-lifecycle data management
- Fully integrated social and technical infrastructure will be needed to support
  - Ingest
  - Full archiving
  - Access
  - Curation
  - Citation support

Questions?