Networking and Information Technology
Research and Development

MAGIC Meeting Minutes
September 4, 2013

Attendees

Shane Canon	NERSC
Rich Carlson	DOE/SC
Gail-Joon	
Dan Gunter	LBL
Shantenu Jha	Rutgers U.
Chris Jordan	TACC
Dan Katz	NSF
Miron Livny	U. Wisc
David Martin	Northwestern U.
Grant Miller	NCO
Sarp Oral	ORNL
Monish Parashar	Rutgers
Iaone Raicu	IIT
Lavanya Ramakrishnan	LBL
Alan Sill	TTU
Rich Wagner	SDSC
Von Welch	Indiana U.

Action Items

1. Please contact Lavanya Ramakrishnan with points of contact for university data storage managers, distributed data centers, and Federal agency data storage managers (NASA, NOAA, USGS,…).
2. Grant Miller should approach Jim Bottom of Clemson U. to discuss the results of his workshop on Identity Management.

Proceedings

This MAGIC Meeting was chaired by Rich Carlson of DOE/SC and Dan Katz of the NSF. Lavanya Ramakrishnan organized an introductory session on Storage and Data Management at National Labs. Other discussants included Sarp Oral, Shane Canon, Rich Wagner, and Chris Jordan.

Discussion of Storage and Data Management at National Labs

Current systems for storing, managing, and analyzing data are being stressed by current applications. Supercomputing Centers are reevaluating their storage centers and workloads to identify what is currently available and to identify gaps in capabilities. Specific topics for discussion include:

- File systems
- Support for scientific databases
- Active storage, SSD

- Scientific I/O middleware for data management movement
- Support for analysis tools, like Hadoop

Discussion identified that other groups should be involved in this discussion including University data storage facility managers where the university stores data of their own researchers. Other entities that should be involved in the discussion include distributed data centers and other agencies (in addition to DOE) that have their own data storage resources (e.g., NASA, NOAA, USGS).

AI: Please contact Lavanya Ramakrishnan with points of contact for university data storage managers, distributed data centers, and Federal agency data storage managers (NASA, NOAA, USGS,…).

Costs of storage is a major issue. Policies are just as important as costs. For example, XSEDE has lots of resources but policies impeded use of those resources. Access mechanisms are also an issue: Do I have the right interface to access and use specific data sets?

There are commercial resources such as Amazon and figshare outside the academic realm that can contribute to resources and considerations of costs and policies.

Most applications have solved their immediate storage and computational needs but long-term archiving is needed. The total life-cycle of the data and application need to be considered and, generally, long-term archiving has not been considered adequately. TACC has 3-4 separate resources that apply to different aspects of the data/application lifecycle including an archive system, local cluster file system, global cluster system, and data publication storage. SDSC is an XSEDE site with a large parallel file system and cloud storage based on Open Stack available for both campus and remote users. NERSC has, primarily, global resources. It does not have iRODS for identifying resources. Users develop their own metadata resources. ORNL has middleware, ADIOS, used to support several scientific codes. A scratch file system provides archival storage.

Middleware needs to be considered. Do sites provide Hadoop or other analysis tools? The Genomic Institute uses the Hadoop Distributed File System (HDFS) in production mode now.

Digital Object Identifiers (DOIs) have to be part of the process for long-term publishing and documentation of applications and results. We need the means to track provenance and to provide reproducibility.

Access control is needed to share data or when data is released or data is kept in a private space. This also requires Identity Management. Proprietary data will require more stringent controls. We need general mechanisms for sharing data and collaborations as collaborations with industry become more common. Typically now collaborations and access are controlled by one-on-one discussions and agreements. This does not scale for the future. We need to work with users to identify mechanisms for access and identity management that will work in the future. How do we provide backups and second copies of data sets?

We need a discussion forum to discuss the development of global policies and automated engines for the future. iRODS is part of the longer-term automatic system. We also need standard forms for metadata for moving data. Dublin Core is a start. We also need discussion of business models and funding models.

We need to work with users to ask them, how much of your funding are you willing to devote to long-term archiving of data versus running your current application with the data?

Jim Bottom, CIO of Clemson U., held a workshop on Identity Management and how identity management is supported on campuses. The workshop identified the need for a discussion framework to address Identity Management. We need to address machine-to-machine identity management and what objects have identity in our ecosystem.

We should ask for presentations form GENI, Data1 and Leico.
We need to distinguish between data management and storage management; sharing data between two groups is very different from archiving and moving data across sites. We need to resolve storage issues if we want to solve data issues. Costs are a critical parameter in discussions. We need a methodology to discuss storage and cost issues.

AI: Grant Miller should approach Jim Bottom of Clemson U. to discuss the results of his workshop on Identity Management.

MAGIC tasking for FY15

MAGIC Tasks for LSN

MAGIC Tasks for FY13 included:

- MAGIC Priorities: 2012-2013
    - Identity management
        - Operational modes: Identify best practices for federated ID management
        - Operational methods: Identify operational issues for faults and failures
        - Wide scale deployment
        - National and International Standards
    - Cloud/Grid Computing Challenges
        - Computing and storage interactions
        - Programming and infrastructure control models
        - Operational methods: Testbeds, benchmarks, standards; identity management, privacy and security
        - Faults and resiliency

Based on the MAGIC participant discussions, MAGIC tasks for FY14 should include:

- Identity Management
  --Provide a forum for discussion of common policies for identity management, current practices, access control, and authorization
- Data Storage and Computational Environments; Provide a forum for discussion of:
    - Policies for interoperation and sharing data
    - Best practices
    - Costs
    - Tradeoffs of data management versus long-term archival needs
    - Metadata needs
    - Distributed environments, collaboration, and interoperability with commercial resources

Upcoming Meetings

September 16-19 OGF 39, Madrid, Spain

September 19-20	Workshop on Scaling Terabit Networks, Washington, DC
October 23-25	Escience meeting, Beijing, China
November 11-15	InCommon Identity Week, Silicon Valley

OSG and XSEDE are offering a summer school to provide understanding of the principles, concepts and applications. A link to this meeting is provided on the XSEDE Web page.

Next MAGIC Meetings

- October 2, 2:00-4:00, NSF, Room II-415
- November 19, 1:30-3:30 at SC13, Room 507