



MAGIC Meeting Minutes

July 1, 2015

Attendees

| | |
|--------------------|---------------------------------|
| Jim Basney | NCSA |
| Richard Carlson | Richard.carlson@science.doe.gov |
| Gary Crane | SURA |
| Shantenu Jha | Rutgers Un. |
| Dan Katz | NSF |
| Jim Kirby | SDP |
| Bertran Ludaescher | OSG |
| Grant Miller | NCO |
| Reagan Moore | RENCI |
| Mike Nelson | |
| Rajiv Ram | NSF |
| Ruth Pordes | FNAL/OSG |
| Frank Wurthwein | UCSD |

Action Items

Proceedings

The meeting was chaired by Rich Carlson, DOE and Dan Katz, NSF.

RENCI Coordination Environment: Reagan Moore

RENCI maintains the DataNet Federation Consortium, a collaboration among the Un. of North Carolina/Chapel Hill, UCSD, Arizona State Un., Un. of Arizona, Un of Virginia, and Drexel Un. to support a wide range of data intensive science programs including:

- Odium Institute
- Institute for the environment
- RENCi
- Data Intensive Cyber Environments Center (DICE)
- Science Observation Network
- Temporal Dynamics of Learning Center
- iPlant Collaborative
- HydroShare
- Semantic ontologies

Three mechanisms are sufficient for federating existing data management systems:

- Shared Name spaces (Federated data grids)
- Direct interaction using API of the remote system (encapsulate APIs in micro-services)
- Indirect interaction (communicate through a message bus)

For federation of data and services you can move data to the remote service or (increasingly common) move the service to the local data

A theory of data science should support:

FOR OFFICIAL GOVERNMENT USE ONLY

c/o National Coordination Office for Networking and Information Technology Research and Development

Suite II-405 · 4201 Wilson Boulevard · Arlington, Virginia 22230

Phone: (703) 292-4873 · Fax: (703) 292-9097 · Email: nco@nitrd.gov · Web site: www.nitrd.gov

- Characterization of the data management system through changes to state information by operations
- Identification of assertions
- Prediction of the probability of success in maintaining the assertions
- Prediction of the sustainable workload
- Identification of the assertions maintained
- Prediction of the probability of success and sustainable workload of a federation

Required infrastructure components include a policy-based system, persistent state information, and event tracking. Policy components for the iRODS system include properties, policies, procedures and persistent state information for digital objects and attributes. iRODS provides the needed components to support operations, state information updates and policies to control micro-services. High performance tracking of events is accomplished using C++ and sending event messages to an external index. iRODS interactions with new technologies are encapsulated in plug-ins (API, authentication systems, databases, micro-services, networks, storage systems, and Zonereport). The pluggable rule engine provides the plug-in services.

Policy sets provide:

- Event auditing
- External indexing
- Protected data management
- Preservation
- Digital library
- Data sharing

The iRODS Consortium membership provides a sustainable infrastructure composed, largely of pharmaceutical companies and storage vendors.

The DataNet Federation Consortium pursues federation across cyberinfrastructure projects and federal agencies. Data grids include NASA, NOAA, NSF, EPA,NIH, NIEHS, NARA,... Service federation is provided through the Discovery Environment, e.g. iPlant. It tracks provenance of workflows within iRODS. Currently iRODS virtualizes data collections and workflows. It can also virtualize data flows.

Participant Roundtable

Shantenu Jha described his work on Expeditions in applied distributed computing in Research in Advanced Distributed Cyberinfrastructure and Applications Laboratory (RADICAL). TeraGrid, followed by XSEDE exploit distributed task-level parallelism to speed up simulations of detailed environments. Shantenu applies non-equilibrium physics to enable parallelism and reduce simulation time. He is supporting the AIMES project for ASCR to develop integrative middleware for extreme-scale computing. He is investigating improving distributed execution of simulations on heterogeneous Distributed Computation Infrastructure (DCI). He has identified that you need a deep understanding of the capabilities of the computer infrastructure as well as the needs of the simulation. He is investigating: "If I give you a workload, what can I achieve with the existing infrastructure". .

Rajiv Ram works with Dan Katz at the NSF in the area of software sustainability. He is seeking, through NSF programs to provide game-changing advances in software development, sustainability, and reliability.

Jim Kirby is the Co-chair of the Software Development and Productivity (SDP) NITRD Subgroup. The SDP is focused on improving software reliability, development time and cost (fast, cheap, good). We

can generally focus on any 2 of those characteristics at the expense of the third. How do we provide improvements in all 3 of these areas. Software defects cost the US economy billions of dollars each year. How do we eliminate defects without exorbitant costs and time. If we produced ultra-reliable software (at a great cost) we would not want to change the software because it might decrease the reliability. SDP is focused on fostering fast secure, reliable software at a reasonable cost.

Gary Crane: SURA

SURAGrid is being sunsetted now. They are now partners in the XSEDE program providing outreach to new and underserved customers. They are creating programs to address campus needs. They are providing a pilot for VIVO that will create a federated search across data bases to support research. They identify researchers, tools, data, and faculty in research areas. They are also using SURA to bridge communities to provide large user communities within their institutional purview. They issue reports and coordinate plans and procedures for security.

Gary Ludaescher: NCSA

Gary Ludaescher works at NCSA on data integration and workflow issues. He is working to provide provenance and querying capabilities. They provide annotate scripts, e.g., Python and elevate the script to make it look like a workflow while still using scripts.

Frank Wurthwein: UCSD, OSG Director

OSG provides infrastructure, tools, and software to provide an open facility for advanced software infrastructure and an open software stack. They disseminate knowledge across researchers, IT professionals and software developers. They accomplish this through the OSG which currently expedites about 120 million transfers per month and about 800,000 jobs per day. OSG supports communities and also individual PIs OSG recently added a login platform that enables users, in less than an hour, to achieve access to OSG resources and to initiate their projects.

Meetings:

August 13-14: BER virtual laboratory meeting, Washington DC: How to design a virtual laboratory

September 28-29, WSSPE meeting, Boulder, CO: Software sustainability

October 15-16: Software management workshop, Washington DC

Next MAGIC Meeting

August 5, 2015, NSF, Room TBD.