**MAGIC Meeting Minutes**
January 7, 2015

**Attendees**

| | |
|---|---|
| Bob Bonneau | DOD |
| Rich Carlson | DOE |
| Larry Davis | DOD/HPCMP |
| Martin Grenet | NCSE |
| Shantenu Jha | Rutgers Un. |
| Dan Katz | NSF |
| Ji Lee | NCO |
| Miron Livny | Un of Wisconsin/OSG |
| Mark Luker | NCO |
| David Martin | Northwestern/Argonne |
| Don Middleton | |
| Grant Miller | NCO |
| Don Preuss | NIH/NCBI |
| Don Riley | Un of Maryland |
| Ray Scott | PSC |
| Alan Sill | TTU |
| Justin Webb | Un of Tennessee |

**Action Items**

**Proceedings**

This MAGIC Meeting was chaired by Rich Carlson of DOE and Dan Katz of the NSF... The meeting heard presentations from the Federal agencies on their research and infrastructure programs for distributed computing. The MAGIC objective for 2015 is to take a detailed look at distributed computing to identify the current status and where we need to go over the near future.

**Distributed Computing at NSF: Dan Katz**

NSF supports both infrastructure and research programs for distributed computing. Research programs in distributed computing include several CISE core programs including Computer Systems Research (CSR), Networking Technology and Systems (NeTS), and Exploiting Parallelism and Scalability (XPS). About 17% of new FY14 CISE projects support research into cloud computing including:

- Computer systems
- Computer networks
- Security and privacy
- Data management
- Applications and software engineering

Additional relevant programs at NSF include BIGDATA and Computational and Data-Enabled Science and Engineering (CDS&E).

NSF also supports research partnerships with SRC (9 jointly funded projects in security) and Intel (CPS security, 2 larger-scale projects in the future)

NSF also supports infrastructure projects including XSEDE, OSG, Campus Cyberinfrastructure – Data, Network and Innovation (CC-DNI), CI platforms (iPlant, nanoHUB, OOI,…) and Major Research Instrumentation (MRI).l XSEDE provides access to advanced digital services to support open research. It supports a large number of production class projects each year. NSF's Advanced Computational Initiative supports large-scale compute platforms at 8 facilities throughout the US. OSG (NSF in coordination with DOE) provides computing facilities and services integrating distributed reliable and shared resources. It is connected to campus grids through BOSCO and OSG Connect. OSG provides up to 2 million CPU hours each day. Chamelion supports a large-scale reconfigurable experimental environment for cloud research. It is large-scale, responsive, reconfigurable providing a one-stop shop for experimental needs. CloudLab is also an NSF FutureCloud built on EmuLab and GENI that explores emerging and extreme cloud architectures. Funding platforms include NSF FutureCloud, CRI, and some MRI. Domain science research also provides support.

NSF programs are forward looking emphasizing transition to practice and federation and make sure resources are available supporting distributed computing and distributed infrastructure to support research. NSF is also preparing to support the infrastructure supporting the Internet of Things with cloud and HPC servers for research users.
For the complete briefing, please see the January 2015 MAGIC meeting site at: https://www.nitrd.gov/nitrdgroups/index.php?title=MAGIC_Meetings_2015

**DOE Distributed and Collaborative Computing R&D Activities: Rich Carlson**

DOE Supports Grid research and standardization:

- Globus and Grid pilot projects
- Partnerships with HEP, FES, NP, Climate
- Distributed computing for DOE scientists

The DOE Magellan project explores how the DOE science community can best exploit cloud-based technologies.

ASCR provides leadership-class supercomputing and networks. The facilities division supports open science research and provides staff knowledge to port codes. The ASCR research division funds Science discovery research activities and provides collaborative communities with unique instruments and compute and storage resources. They are currently developing a usable exascale computer.

- Big Panda explores using US ATLAS Workload Management Software on the Titan Supercomputer at Oak Ridge
- Data driven computing requirements explores the future of distributed computing
- dV/dt helps scientists use non-indigenous compute, data, and applications.

Science applications are evolving from simple jobs to large complex workflows. The ASCR CS/NGNS workshop developed concepts to deal with massive data generated by simulations and instruments. It identified a need for workflow scientists.

Analytical modeling for extreme-scale computing environments supported 4 large projects starting in FY14 including:

- IPPD: Integrated end-to-end performance prediction and diagnosis
- Panorama: Predictive modeling and diagnostic monitoring
- Ramses: Models for science at extreme scales
- X-Swap: Extreme-scale scientific workflow analysis and prediction

Potential future directions include:

- Near-time processing of simulation, experimental and observational data
- Interactive supercomputing with multiple simultaneous users
- Distributed storage infrastructures

For the complete briefing, please see the January 2015 MAGIC meeting site at: https://www.nitrd.gov/nitrdgroups/index.php?title=MAGIC_Meetings_2015

**DoD High Performance Computing Modernization Program: A Distributed Computing Program: Larry Davis**

The DoD High Performance Modernization Program accelerates the development and use of advanced computational environments to advance defense technologies. Their priorities are to maintain leadership, expand the base, increase customer productivity, exploit next-generation technologies, and develop the workforce. The Service/Agency Approval Authorities allocate controlled CPU hours on HPCMP systems, implement requirements, surveys, resource allocations, resource monitoring and resource reallocation. They provide guidance to users on which HPC assets are appropriate for their projects.

HPCMP provides:

- Development of next-generation scientific and engineering software for advanced computing
- Advanced computing capabilities via 5 centers and an integrated network
- Fostering development of a computational workforce
- Integrating these capabilities to solve complex DoD problems

HPCMP currently provides:

- About 12 large HPC systems from a variety of vendors
- Connectivity to these systems
- Expert help in accessing and using the systems
- Access to commercial applications software packages
- Access to create design modeling tools

In FY14 HPCMP supported many hundreds of DoD projects

The five HPCMP centers include:

- ARL
- AFRL
- ERDC
- Navy
- MHPCC

These centers provide technical services for customer assistance, system administration, data analysis and operators and operations support. Projects and services include:

- User dashboard
- Software configuration management
- HPC portal: Web supercomputing
- Advanced reservation service across all production systems
- Dedicated support partition to support large computational work needing a dedicated system partition
- Storage management: ~47000 TB in FY14

HPCMP also provides Dedicated HPC Project Investments for modest-sized HPC systems for projects that cannot otherwise be supported.

DREN is the Defense Research and Engineering Network that is focused on S&T and T&E with high bandwidth (up to 10Gbps), low latency. It connects the DoD supercomputer centers and users. It is a component of the global Information Grid and a sister network to NIPRnet/SIPRNet. HPCMP gathers requirements for services to assure users receive their needed support and upgrades are provided in a timely fashion. The HPCMP allocation allocates 100% of all resources at DoD, about 70% to service agencies and 35% to challenge plus service-agency high priority projects.

Tactical cloudlets are being developed to support tactical missions using mobile ad hoc networking.

For the complete briefing, please see the January 2015 MAGIC meeting site at: https://www.nitrd.gov/nitrdgroups/index.php?title=MAGIC_Meetings_2015

**NIH Distributed Computing Programs: Don Preuss**

NIH consists of 27 institutes, each with its own mission. NIH overall has a budget of $30 Billion supporting about 50,000 projects each year. NIH is estimated to spend over $1 Billion on data per year. In 2014 the NIH Director appointed Phil Bourne the first Associate Director for Data Science. He is leading an NIH-wide initiative to address the exponential growth of biomedical research datasets. The mission is to provide an ecosystem to enable biomedical research to be conducted digitally. The NIH mandate provides a team of 8 people to implement intramural participation and funding through Big Data to Knowledge (BD2K). The program is intended to provide:

- Provide dropbox-like storage
- Develop quality metrics
- Expedite taking the computing to the data resources
- Provide a place to collaborate and discover

A cloud broker moderates among the cloud providers.

A cancer genomics cloud has 3 pilot projects:

- Broad Institute
- Seven Bridges Genomics
- Institute for Systems Biology

The clouds are interoperable and users may be able to work between the clouds.

The Entrez Programming Utilities (eUtils) are a set of seven server-side programs that provide a stable interface into the Entrez query and database system at the National Center for Biotechnology Information. The eUtils use a fixed URL syntax that translates a standard set of input parameters into the values necessary for various NCBI software components to search for and retrieve the requested data. eUtils has about 80,000 users per day over a simple API and has supported over 10 million requests.

BLAST (Basic Local Alignment Search Tool) is an algorithm for comparing primary biological sequence information, such as the amino-acid sequences of different proteins or the nucleotides of DNA sequences. A BLAST search enables a researcher to compare a query sequence with a library or database of sequences, and identify library sequences that resemble the query sequence above a certain threshold.

An SRA toolkit provides remote I/O for data at the NCBI. It transparently locates and caches data needed for cloud computation (remote I/O, slicing, encryption at rest…)

NCBI users download about 3 TB of data per day through 2 x 100G high reliability networking

Discussion among the participants asked:

- How do we scale up to large numbers if we are successful in providing cloud environments?
- How do we scale up?
- How do we add limits, if necessary?
- Do we need alternative mechanisms?
- Can we identify methods for large users to implement sets of resources in the cloud where the user can pay the costs of using those resources?
- How do we decide the total dollar amount we are willing to put into cloud computing resources?
- How do we schedule resources? (Based on policy)

**Upcoming Meetings:**

**Next MAGIC Meetings:**
February 4, 2015, NSF
March 4, 2015, NSF