Integrating Clouds and Cyberinfrastructure for CDS&E: Research Challenges

Manish Parashar, Rutgers University (parashar@rutgers.edu) and Geoffrey Fox, Indiana University, (gcf@indiana.edu)

With contributions form:

Kate Keahey, Argonne National Laboratory, keahey@mcs.anl.gov

Alan Sill, Texas Tech University, alan.sill@ttu.edu

Scott Brim, Internet2, swb@internet2.edu

Michael Nelson, mnelson@pobox.com

**I. Introduction**

Aggressive cloud computing technology development has resulted in multiple classes of cloud services that provide attractive solutions for many different types of business applications. It is expected that cloud services will join more traditional research cyberinfrastructure (CI) components, such as high-performance computing system, clusters and Grids in supporting scientific exploration and discovery.

Cloud computing services can support Computational and Data-Enabled Science and Engineering (CDS&E)2,1 applications in multiples ways. It can provide a platform for applications, for example when local infrastructure is not available. It can also supplement existing platforms to provide additional capacity or complementary capabilities to meet heterogeneous or dynamic needs. For example, Clouds can service as accelerators, or provide resilience to scientific workflows by moving the execution of the workflow on alternative or fewer resources when a failure occurs. The simplicity of the Cloud abstraction can alleviate some of the problems scientific applications face in the current HPC environment. For example, many-task computing applications (e.g., ensemble applications) can benefit from the ease of use, an abstraction of elastic and/or readily accessible resources and the ability to easily scale up, down or out. Finally, Cloud computing can not only help scientists address today’s problems more effectively, but also allow them to explore new ways of formulating their application using the abstraction of on-demand access to elastic resources.

However, before CDS&E can fully realize the potential benefits of a hybrid cyberinfrastructure that integrates cloud services, several research issues remain. In 2011/2012, the NITRD MAGIC[[1]](#footnote-1) subcommittee (chaired by Gabrielle Allen and Daniel Katz, NSF and Rich Carlson, DOE/SC) explored best practices in cloud, distributed, and Grid computing and specifically explored these research issues. These discussions are summarized in this report.

The discussions in this report are also based on two reports. The first is by Gannon and Fox[[2]](#footnote-2), in which they reviewed and classified applications suitable for clouds. The second is by Parashar et al.[[3]](#footnote-3), in which they explored how a hybrid HPC/Grid + cloud cyberinfrastructure can be effectively used to support real-world science and engineering applications, presented illustrative scenarios, and discussed limitations and research challenges.

**II. Findings**

The NIRTD MAGIC subcommittee discussions in April and May 2012 focused on focused on the *Challenges and Opportunities of Integrating Clouds with Cyberinfrastructure* and were led by Manish Parashar (minutes of these meeting can be found at https://connect.nitrd.gov/nitrdgroups/index.php?title=MAGIC\_Meetings\_2012). During these calls the committee discussed experiences from projects such as FutureGrid, Open Science Data Cloud, Megellan, XSEDE and GENI, as well as core research challenges. The research challenges were grouped into 6 categories: (1) Algorithms and Application Formulations for Clouds, (2) Programming Models, Abstractions and Systems, (3) Middleware Stacks and Services, Management Services, (4) Data Management in the Cloud, (5) Security Policies and Mechanisms, and (6) Deployment/Transition to Practice. These categories and associated research challenges are presented below. Note that the discussions below are from the CDS&E perspective and are complementary to more general cloud research challenges.

**(1) Algorithms and Application Formulations for Clouds**

Clouds provide abstractions, usage modes and execution behaviors that are different form more traditional CI platforms, and it is important to understand these differences from an application perspective and thier impact on how algorithms are developed and applications are formulated. Key research challenges in this category include exploring application formulations that can effectively utilize clouds and addressing, at the algorithmic level, implications of cloud characteristics such as elasticity, on-demand provisioning, virtualization and multi-tenancy. Furthermore, specific science/engineering applications classes lead to specific challenge, such as, for example, privacy in case of bio- and medical-informatics applications and bursty data streams in case of sensor-actuator applications.

**(2) Programming Models, Abstractions and Systems**

A key research challenge is developing appropriate programming models and systems that can enable CDS&E applications to take advantage clouds as well as cloud + CI. These include developing programming abstractions and tools to support the federation of clouds and CI and elastic access to cloud services, extending existing cloud programming models and platforms, e.g., MapReduce, to support scientific computing, as well as exporting entire applications, applications patterns and kernels, optimized libraries, and/or specialized middleware as a service. Research addressing tools for application debugging, validation management and performance engineering is critical.

**(3) Middleware Stacks and Services, Management Services**

Middleware stacks and service are essential for supporting CDS&E application formulations and hybrid usage modes targeted to cloud + CI environments, including support for dynamic cloud bursting and infrastructure federation. A key research issue is the provisioning, scheduling, management and optimization (multiple objectives including performance, energy, cost, reliability, etc.). Another key issue is interoperability with and integration of cloud storage models and solutions. Developing and deploying cloud software stacks providing platforms for testing and research will be important.

**(4) Data Management in the Cloud**

Data management research challenges are broadly due to different types of cloud storage solutions and the nature of cloud connectivity. Specific research challenges include support for selecting between the large variety of storage options with varying service levels, networks architectures to support data transport needs and their interaction with cloud storage offerings, and the co-location of compute and data.

**(5) Security Policies and Mechanisms**

Clouds tend to emphasize the need for quality security mechanisms due to the sharing of storage and computing. In addition to crosscutting cloud security challenges, specific issues to CDS&E and cloud + CI integration include the interoperability with broader CI security mechanisms and policies, such as for example, single sign-on, federated identify management (e.g., inCommons, cilogin, SCIM, etc.), and security policies and mechanisms for specific applications, e.g., differential privacy and data anonymization requirements for bio/medical informatics applications.

**(6) Deployment/Transition to Practice**

Research challenges associated with the deployment and transition to practice of cloud technologies included the definition of community standards, the development of community testbeds and benchmarks, documentation of experiences and best practices, and the developing of curricula, and training modules.

**III. Actions and Timelines**

The MAGIC committee discussions highlighted both short term and longer-term activities to address the research challenges outlined above. Key **short term** activities included: (1) development of pilot projects to identify use cases and best practices, as well as to determine requirements; (2) definition of standards and develop services to enable interoperability between CI (i.e., XSEDE, OSG, FutureGrid) and cloud services (e.g., EC2, Azure, etc.) as well as their integration; (3) creation of community research and experimentation testbeds as well as establishment of community benchmarks and common metrics; (4) development of processes for the translation of research innovation into software frameworks tools that can be used by applications; and (5) creation of community forums for exchange of ideas and artifacts. **Longer-term** activities include setting up programs to address the fundamental research challenges discussed above and incorporating resulting research innovations into sustainable software system that can be deployed and used by the community.

1. https://connect.nitrd.gov/nitrdgroups/index.php?title=Middleware\_And\_Grid\_Interagency\_Coordination\_(MAGIC) [↑](#footnote-ref-1)
2. G. Fox, D. Gannon, “Cloud Programming Paradigms for Technical Computing Applications,” Technical Report, <http://grids.ucs.indiana.edu/ptliupages/publications/Cloud%20Programming%20Paradigms.pdf>, 2012, [↑](#footnote-ref-2)
3. M. Parashar, M. AbdelBaky and I. Rodero, “Cloud Paradigms and Practices for CDS&E,” Technical Report, 2012, <http://cometcloud.org>, 2012. [↑](#footnote-ref-3)