MAGIC Meeting Minutes
September 6, 2017

Attendees

Rich Carlson	DOE/SC
Wo Chang	NIST
Dan Gunter	LBL
Gilberto Pastorello	LBL
Shantenu Jha	Rutgers
Joyce Lee	NCO
Brian Lyles	ORNL
Lavanya Ramakrishnan	LBL
Rajiv Ramnath	NSF
Don Riley	UMD
Alan Sill	TTU
Craig Tull	LBNL

Proceedings
This MAGIC meeting was chaired by Rich Carlson (DOE/SC) and Rajiv Ramnath (NSF).

Administrative
- August 2017 Meeting minutes approved

Wo Chang: NIST Big Data Standards Activities (Slide presentation available)
NIST Big data public working group (started 4 years ago) and ISO/IEC working group

NIST Big data public working group: Identify reference architecture for Big Data (needs to be vendor-neutral, technology and infrastructure agnostic so scientists and others could use this architecture to advance their data discovery, etc.

- Identify high level key components, define preliminary interfaces between components and use interfaces to build Big Data applications
- 5 Subgroups and produced 7 documents last year (getting to version 2 draft)
  - Version 2: September 21 deadline for public comment
- Goal for architecture to have commonalities with BD architecture and have unified architecture to support BD reference architecture (info from IBM, ORACLE, etc)
- 9 documents available on URL (https://bigdatawg.nist.gov/V2_output_docs.php (Nov 2017)). Version 3 will be final product.
- Challenges:
  - Many available computing tools; challenge is to support different kinds of computing stack in more agnostic way
  - How to support integrations in 3 different areas (computing stack/analytic stack and data source).
- Focus: supporting data scientists
  - Step 1: Identify reference architecture-Version 1 (See slide diagram).

    - Step 2: Look at interfaces- Version 1 (51 unique use cases and requirements)
    - Step 3: How to manage ecosystem – reuse analytic tools
- Use case implementation (Version 2)
    - Use DevOps environment, implement using container (Docker); Microservices for communication. Map use cases in scenario e.g., Healthcare fraud detection
    - Trying to marry HPC + BD stack using architecture and microservices as an implementation strategy
    - Seeking to share strategies with implementers/collaborators
- NIST is getting inputs from BD working groups and ISO/IEC Working Group (e.g., within NIST and ISO subcommittees and working groups, IEEE, etc)

ISO/IEC working group → WG9-BD Working Group (26 countries)
Would like closer collaboration with subcommittees and working groups; multimedia (MPEG, JPEG)
2 projects:
1) Big data definition project – potentially become international standard by early 2018 and
2) Reference architecture project – 5 parts

Discussion (See Reference Architecture slide)
Standard roadmap; want to utilize existing standards and identify gaps in standards. Found 16 gaps in the Big Data reference architecture (tables of available standards listed in part 5 of Part 5 of ISO/IEC 20547). Continuing to refine the table and include other available standards (e.g., API framework).

NIST Big Data-Public Working Group
Subgroup co-chairs- Those interested in participating can contact Wo or the co-chairs.

- Definitions and Taxonomies- Dr. Nancy Grady, Principal Data Scientist and Technical Fellow, SAIC
- Use Cases & Requirements - Prof. Geoffrey Fox, University of Indiana
- Security & Privacy -Dr. Arnab Roy, Member of Research Staff, Fujitsu Laboratories of America
- Reference Architecture- David Boyd, InCadence Corp.
- Reference Architecture Interface – Gregor von Laszewski, University of Indiana
- Standards Roadmap – Russell Reinsch, Center for Government Interoperability
- Adoption & Modernization – Russell Reinsch, Center for Government Interoperability

Craig E. Tull - SPOT Suite tool (Presentation available)

- scalable autonomous data management and scientific workflows for X-Ray Light Sources run by BES Facilities (slide presentation available)
- serves wide range of science and technology that take advantage of Lightsources

Trends and changing landscape:
First timers due to expansion of Advanced Light Source (ALS) user base; affordability issue
Exploding data volumes, automated systems; new mathematical techniques being applied, new hardware architectures, new paradigms, machine learning)

SPOT suite- integration of ALS, NERSC facility and ESnet networking into "proto-super facility" allows for querying database, browsing data files, visualizing information in real-time or later.
Successful prototype; now pseudo production system –ran over 388,000 data sets through system.

Automation – automated system enabling real-time review of data analysis which informs the next sample. Data automatically archived on tape system and made available through disc and data access capabilities. Remote experiments/system automation – UK scientists used spot suite on smart phone

Data set processing for microtomography beamline. 2 problems:

1) Submitting jobs on distributed workflow on batch queue system has long wait time
2) Scientists want fast, not necessarily good quality feedback. So, added faster path.

X-Swap program: Extreme scale scientific workflow analysis and prediction

- Dropped queue wait time by introducing RabbitMQ into system, which allowed a number of tasks to be repeatedly fed into a single node without incurring batch queue penalty.
- Faster feedback though SPOT tomographic processing, but tradeoff on quality
- Statistical modeling of data transfer information to recognize effects of configuration changes
- Missing data – shows anomalies, but need more information to diagnose

DEDUCE project – Distributed dynamic data analytics infrastructure for collaborative environments;

- Goal: understand whether can detect scientifically meaningful differences in data and if can data can be taken from multiple beamlines and facilities.
- Able to detect significant temporal and spatial changes of image sequences (e.g., crystallization), potentially decreasing number and size of images needing to be processed.

Organic Photovoltaic (OPV) – collaboration between CAMERA, CRD, ESN, ORNL, ALS and NERSC. Not as efficient as other PVS, but can be printed on flexible material. Still needs work for production capability

Xi-cam project – Currently under discussion.

- Can do full analysis on single data set and capture workflow that can be theoretically executed thousands of times on different data bases by SPOT on HPC resources.

Summary

- Importance of combining HPC network, advanced math algorithms and superfacility infrastructure to advance the science of DOE. Beamlines are benefiting from it.
- More to be done. Composible API-based ecosystem allows contributions from many sources.
- Using NERSC primarily, but have run on Oakridge and other resources. Plan to expand to using cloud resources and custom resources (e.g., other GPU clusters)
- Room for advances in mathematical techniques, ML, etc.
- Significant: if we can reach community standards and common facilities API

Discussion

Global light sources- Trying to move some work flow to a French workflow system which has longer provenance. Light sources in China and Japan as well. Discussed importance of common tools and work flows and ways to encourage broadly used community developments

Potential MAGIC Tasking-

Group members finalized the list of proposed tasks and workshop to be presented at LSN's Annual Planning meeting.

Slide 1: MAGIC Tasks

- **Task 1**: Explore containerization, virtualization technologies that allow the use of resources with a prescribed environment; i.e., an application environment that moves with the job.
    - Subtask: Explore usability for HPC applications.

- **Task 2**: Examine a broad range of challenges and current status of containerization and virtualization issues in the context of creating a federated, distributed cloud computing environment for science.

- **Task 3:** Explore data repositories and data oriented work groups worldwide. How do we constitute, compose and implement resources at the nation's cyberinfrastructure facilities to mesh well with emerging capabilities (e.g., data sharing, data transfer, data management).
    - Subtask: Coordinate with Research Data Alliance and other relevant resources, as needed.

- **Slide 2: Workshop:** Develop and provide recommendations on how to provide a flexible, accessible cloud environment across applications, user groups, and cyberinfrastructure to support big science.

Discussion:

- Task 2: NIST and IEEE could collaborate on deeper topics on federated cloud computing. Discussed inviting NIST to present plans on associated clouds. NIST also has an event on cloud computing as well as big data.
- Task 3: Big data management (BDGMM) formed 4 months ago. Explore synergies with NIST. Hackathon – Alan and Wo will discuss.
- Workshop: December 5-8 – Cloud computing conference (uccc-conference.org). Rich and/or Rajiv could potentially lead a panel discussion

MAGIC Roundtable

TTU: Allan Sill

UCC 17, the 10th IEEE/ACM International Conference on Utility and Cloud Computing, co-located with the 4th International Conference on Big Data Computing, Applications and Technologies, will be held in Austin, TX in December. Rich and Rajiv will explore leading a panel.

NSF: Rajiv Ramnath

Survey results/whitepapers from the September 6-7 NSF Large Facilities Cyberinfrastructure Workshop have been posted.

Meetings of Interest

Dec 5-8, 2017 UCC 2017 Austin, TX

Next MAGIC Meeting

October 4, 2017, 12:00-2:00 p.m. EDT (Remote)