The government seeks individual input; attendees/participants may provide individual advice only.

Middleware and Grid Interagency Coordination (MAGIC) Meeting Minutes[1]
April 7, 2020, 12-2 pm ET

Virtual

Participants

Alisa Manning (MGH)	Keith Beattie (LBL)
Alison Derbenwick Miller (Oracle)	Kevin Thompson (NSF)
Arjun Shankar (ORNL)	Mallory Hinks (NCO)
Derek Weitzel (University of Nebraska, Lincoln)	Matyas Selmeci (UW-Madison)
Eric Lancon (BNL)	Miron Livny (OSG)
Evan Wooldridge (Morgridge Institute for Research)	Nicholas Goldsmith (NSF)
Frank Wuerthwein (UCSD)	Peter Nugent (LBL)
H Birali Runesha (U of Chicago)	Ravi Madduri (ANL)
Jack Wells (NVIDIA)	Richard Carlson (DOE)
Jason Lopez (Xodiax)	Stefan Robila (Montclair State U)
Jessica Li (Illinois)	Tevfik Kosar (NSF)
Kate Evans (ONL)	Tim Cartwright (UW – Madison)
Kate Keahey (ANL)	Todd Shechter (UW – Madison)
Kathy Austin (TTU)	

Introductions: This meeting was chaired by Richard Carlson (DOE/SC) and Tevfik Kosar (NSF)

Cloud Speaker Series

The OSG Fabric of Services and Cloud Resources

Miron Livny, Director of the Software Assurance Marketplace and the Technical Director of the Open Science Grid (OSG).

Main Question

- How do cloud resources fit into the distributed high throughput computing (dHTC) model of the fabric of services provided by the Open Science Grid (OSG)?
    - Naturally! It's all about offering access points and services to deploy execution points

Background on OSG

- Consortium of stakeholders governed by a council
- Consortium itself does not own or operate any resources

[1] Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Networking and Information Technology Research and Development Program.

- Provides a fabric of dHTC services

Fabric of Services

- Community Building
- Research Computing Facilitation
- Operation

Open Science Pool (OSPool)

- Access Points (APoints) are open to any US researcher and a distributed HTCondor pool
- OSG Compute Federation sites contribute resources to the OSPool by running execution points (XPoints)
- Showed map of compute and storage capabilities that OSG can bring to the OSPool and other pools

Democratizing Access

- We view Access Points as holding the key to addressing the challenge of democratizing access to national research computing resources
    - Can be deployed and operated by a single PI laboratory or by campuses or science collaborations

OpenStack Environment

- Jetstream cloud is an OpenStack, NSF-funded academic cloud operated by Indiana University and TACC
- Started in December and have already contributed 750,000 core hours.

Other Pools

- Organizations like science collaborations and campuses leverage OSG services to deploy and operate private distributed HTCondor pools

Bring your Own Resources (BYOR)

- Members of PATh working on enhancing BYOR services for commercial cloud resources
- Commercial clouds offer a new level of elasticity, BYOR support must be “at scale”

Q&A

- Birali Runesha – Can you comment on the example of HTAC that runs on commercial cloud? Can you also comment on the support of running java? Any differences in running on the commercial cloud and running on the typical resources at a university
    - Miron – It may be good for the group to hear from Frank and his crew on their experience because they ran it several times and they looked not only into the cost of the compute, but also the cost of the network and the cost of the storage.
    - Frank Wuerthwein (UCSD) – We are giving a talk at GDC 21 on benchmarking applications that we want to *audio distorted*. People who are interested may want to check that out.
    - Frank – In general, costs are so strongly dependent on what you’re doing that saying something that is generic is almost impossible. When I’m doing something quick, for a limited amount of time, I’m better off doing it in the cloud. Because buying something takes a long time and is a pain. There are clear advantages of cloud and it’s more complicated than just saying one is more expensive than the other.

- Frank – I know of only one calculation that included everything (power consumption, space, operations effort, etc.) It was done by Burt Holtzman from Fermilab. It's a few years old by now.
- Kate Keahey – If you are interested in pricing, we compared pricing for premium resources, commercial resources. They published a paper called overcast and it's about structuring experiments across Chameleon and commercial clouds. Texas Advanced Computing Center also did that comparison.

Multi-Messenger Astronomy and the Merger of HPC and Cloud Computing

Peter Nugent, Senior Scientist, Division Deputy for Scientific Engagement & Dept. Head for Computational Science, Lawrence Berkeley National Laboratory

Multi-Messenger Astronomy

- Observations that take place in two different, not just wavelength regimes, but two different sets of particles.
- Today it almost always focuses on what's happening with gravitational waves

Q1: Nucleosynthesis

- Nucleosynthesis (neutron star mergers alongside heavy element production)
- Light curve of this event is very red
- Able to get a nice spectrum and the spectrum almost always just exists in the infrared – showed that we had a bunch of heavy elements (Au, Ag, U)
- Showed graphic of element origins
  - Merging neutron stars
  - Dying low mass stars
  - Exploding Massive stars
  - Exploding white dwarfs
  - Big Bang
  - Cosmic ray fission
- DECam is the best instrument in the southern hemisphere to follow-up GWs
  - Speed is the key - Other people searching your data in real time - Need to be first
- AWS Benefits
  - No queues – start processing immediately
  - Instances are "yours'
  - Everything is containerized
  - Sophisticated batch system
  - We got a grant from them!
  - No limits on resource usage
- AWS negatives – nothing is ever free
  - Charges: $0.005 iops-month – 300,000 files = $1500
  - $0.08 Gb-month – 9600 GB = 768
  - Trigger each month
- Work around – process everything at AWS and delete data immediately. Save 100x100 pixel cut-outs around best candidates – reprocess later at NERSC

Opportunities for NIH cloud interoperability approaches to improve outcomes of pediatric diseases
Alisa Manning, Researcher at Massachusetts General Hospital (MGH) and the Broad Institute of MIT and Harvard.

Cloud Analysis and Researcher Journey

- 2017-2018: First researcher to perform a GWAS using Fire Cloud
    - Interested in using cloud-based resources because the data was really large
- 2018-2019: Collaborative Development of Cloud-based Workflows
- 2020: Collaborative Analysis in NHLBI's BioData Catalyst

Problem: User with a research question and analysis plan

- Find data
    - In the past – manual download of data in the dbGap
    - Current – Web portal in NHLBI's BioData Catalyst data
- Set up place to do analysis
    - In the past – Local and institutional compute
    - Current – Analysis workspace in NHLBI's BioData Catalyst
- Authorization to use data
    - Automatic checks in place to verify that you still have permission to access the data set

BioData Catalyst

- Data use agreements limit the ability to analyze data at various institutions
    - Moving data is difficult because files are very large
    - Expensive to extract data if it's sitting in the cloud
    - Want to perform analysis with collaborators at different institutions
- NHLBI initiated the effort for BioData Catalyst 3 years ago
- Ecosystem of platforms

NIH Cloud Platform Interoperability Effort (NCPI)

- Goal: To establish and implement guidelines and technical standards for empower end user analyses across participating platforms and facilitate the realization of a trans-NIH federated data ecosystem
- Motivation: researchers want to access data across ICs/stacks
- Early 2020: Interoperability between NIH stack was quite sparse
- Standards-based interoperability features in the works
    - Global Alliance for Genomics & Health
        - GA4GH Passports
        - Data Repository Service v1
- Goals for 2021:
    - Finish connections

Important Data Governance Lessons

- Bring together the Pediatric Cardiac Genetics Consortium Study data for the first time in the cloud for researchers
- Challenge to enabling this cross-platform data access is maintaining each program's data governance
- Interim solution for data access for analysis in GMKF data resource or BioData Catalyst workspace

Q&A

- Rich Carlson: Are you able to bring your own analysis tools into the environment or are you required to use the analysis tools that are being provided?
    - In the cases of BioData Catalyst I can bring in my own research tools. Comes with pluses and minuses. When I write my own analysis workflows they may not be the most efficient. So within BioData Catalyst, although people can bring their own analysis tools, there's a desire to have these tools registered and vetted to reach various levels of quality control
- Alison Derbenwick Miller: Do you have any sense of how widely this ultimately will be used when you get it built?
    - It's a hard-won battle to get researcher to switch their analysis paradigm because there are costs involved in performing cloud-based research.
- Miron Livny: At the high-level model, how does this relate to what the high energy physics experiment is doing? Because what you call huge data doesn't impress everyone.
    - Ravi Madduri: One of the key capabilities that NIH has done because of the presence of something called ERA Commons is that system made it possible for implementing and enforcing an authorization authentication model that watches for researchers identity and validates the identity and implements a federated identity model that spans institutions and using BioData Catalyst and Stripes program, transfer to cloud-based authentication models. I think the identity management authorization management is still the most important challenge.
    - Ravi: In high energy physics, one of these experiments cannot be done individually. To some extent, the incentives and the system is created in a way that people come together and do this and be able to share data. In life sciences, teams are typically 5 to 10 researchers across two organizations validating a hypothesis.
    - Miron: We are moving away from identity management to capability. And maybe that's another lesson that can be explored. Moving more into token base authorization. POSSIBLE FUTURE TOPIC?

Roundtable

- Jack Wells (NVIDIA) – NVIDIA GPU Technology Conference is next week (M-F). Keynote is on Monday 11:30AM ET, rebroadcast at 9PM ET

Next Meeting

May 5 (12 p.m. ET)