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

Middleware and Grid Interagency Coordination (MAGIC) Meeting Minutes
August 1, 2018, 12-2 pm
NCO, 490 L'Enfant Plaza, Ste. 8001
Washington, D.C. 20024

Participants (*In-Person Participants)

James Burke	Valiant	Thomas Morton*	DoD/OSD
Richard Carlson	DOE/SC	Valerio Pascucci	Utah
Bev Corwin		Rajiv Ramnath	NSF
Kaushik De	UTA	Don Riley	UMD
Dan Gunter	LBNL	Sonia Sachs	DOE/SC
Gregor von Laszewski	IU	Matyas Selmeci	UW-Madison
Joyce Lee*	NCO	Alan Sill	TTU
Brian Lin	UW-Madison	Derek Simmel	PSC
Grant Miller	Retired	Wei Yang	SLAC/Stanford

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

Speaker Series: DevOps and the Scientific Community

- *DevOps in ATLAS* - Kaushik De, Professor of Physics, University of Texas at Arlington
- *What is DevOps* - James Burke, Senior Security Architect/Engineer, Valiant Solutions
- *Education and Research Computing in DevOps*, Gregor Von Laszewski, Assistant Director of the Digital Science Center, Adjunct Associate Professor of the Intelligent Systems Engineering Department Indiana University

Speaker Presentations
DevOps in ATLAS - **Kaushik De**
ATLAS experiment at CERN's Large Hadron Collider (LHC): Largest scientific instrument ever built

ATLAS Computing Challenge
ATLAS's intense computing needs: 1 PB data streaming out/second.
ATLAS Software stack: Learned the importance of DevOps through its 6M lines of code (Slide 3)

DevOps as Evolution (slide 4)
Atlas is 25 years old: Emerged as natural working model that followed the natural evolution of the experiment.

- Moved from initially focusing on design and development to realizing, in the past 10 years of experimental data collection, that software and computing had to fully embrace the DevOps model to succeed.

    - Science-driven: need high efficiency in collecting, processing data. Need to analyze data and publish data quickly.
    - Exploring and collecting data that had never been collected. Needed to constantly change with software changes. Software and computing team became a large distributed, multi-national, collaborative critical to meeting these challenges.
- DevOps is a natural solution to large data science: Need DevOps at every level (e.g., 5M lines of activation software: continuous testing monitoring and analysis is a complete DevOps cycle).

<u>DevOps practices and principles used in ATLAS experiments: 3 examples (slide 6 – 8):</u>
Our distributed computing environment was 1M lines of code developed organically with developers, operations team and users. How did we do it?

1) PanDa (Production and Distributed Analysis system) is part of an ecosystem of 1M lines of python code (lends itself to DevOps). Developed and used our code from the beginning. First called "Iterative development cycle", then "Agile", now "DevOps".
    a. Scale: Team of developers, operations and support work together with users (thousands) who also write code work together. They are part of ADC (Atlas Distributed Computing), the most important organization in Atlas which practices DevOps ideas daily. Made possible by ensuring that DevOps is practiced throughout.
    b. PanDa System image (Slide 8)
2) "Ops" working with "Dev": The National Energy Research Scientific Computing Center exemplifies how DevOps can be agile and work together to achieve physics goals. (Slide 9)
    a. Started with 500 nodes to 1500 nodes (150 cores each); limit due to shared file system. Jumped to 1500/2000 nodes by going to containers for input data and input code. Then by iterating with Dev and Ops in an integrated manner, increased to 3k nodes by putting containers and loopback file system for output and quick development cycle. Able to ramp up to .5M cores within hours to finish a task.
3) Looking to the future: DevOps at service layer at the edge.
    a. New Slate project to automate deployment and operation of distributed services (Rob Gardner, UChicago). Working towards service layer at the edge of HPC or campus computing resources to equip Science DMZ with service orchestrated platform to allow services at the edge and enable more efficient use of these resources. (Slide 10)

<u>Lessons Learned</u> (Slide 11)
Need to work together (not silos or separate teams), and continuously, to ensure DevOps' success.

<u>LHC Challenges Ahead</u> (slide12)
Need real DevOps to meet growing resource needs (much higher than expected with our natural cycles of Moore's law). In 7-8 years, our resources will be insufficient for our experimental data. Challenge: to practice and use DevOps in more ways than previously thought possible.

<u>Conclusion</u>
DevOps is important for large scale data science. Culture is important from the beginning. An integrated DevOps that works together in the ADC environment has enabled our success in this challenging environment.

What is DevOps **– James Burke** (Working on 2020 Census at Census Bureau)
Development, IT operations, and Information security in a tightly-integrated way to successfully execute the 2020 Census.

DevSecOps
Focus on DevOps security – how to overlay security in the development process? Ensure security is included.

DevOps as "CLAMS"

- **Culture**: Security is about breaking down barriers between teams to share same cultural mindset.
- **Lean**: Minimize tools, meetings, large scale development. Traditional development in commercial world takes long time – break down the development of large codes, large applications, systems into "tools, meetings, sprints". Overlay security in each phase (compliance).
- **Automation** of codes and services to streamline process: Continual integration/ Continual development for infrastructure deployment and development (CI/CD).
- **Measurement**: to track progress.
- **Sharing** progress and where we are from a development and also security perspective.

Process: Pre-commit, commitment (deployment) and continuous delivery (tools). Use tools to check our status to locate risks and minimize them.

Common DevOps & DevSecOp Strategies (Cultural)
IT ops, development and security. Bi-weekly staff meetings. 24/7 accountability: Everyone is accountable for certain components and systems.

Treat Infrastructure like Code
Automated, repeatable operations with predictable outcomes. Build security into every stage of DevOps.

How to Get There
First, be lean. Do not have a lot of practices/procedures that slow you down. This is a high-level presentation; the audience is invited to ask about specific tools.

Education and Research Computing Supported by DevOps - Gregor von Laszewski, Fugang Wang, Allan Streib, Geoffrey Fox

Ideas: Observations (Slide 2)

- DevOps supports a variety of users
- Single data center or large-scale data center is insufficient. Need edge computing and educational centers that contribute to these large data centers.
- Need templated, state-of-art images assisted by Containers
    - Added client tools (Cloudmesh client) to enhance the DevOps experience

Intelligent System Engineering Department: Indiana University's new department features a cross cutting engineering discipline emphasis on intelligent systems (Slide 3)

Observations about Users:

- When deploying systems, knowledge is greater in the administrative community than the research community, which has greater integration needs.
- Delegated usage: the Administrator and User determine what is ultimately put on the machine. Uses templates to guide this activity. (Slide 4)

Systems support research (Slide 5)
There are a variety of machines to address various needs (storage, GPU, Containers, etc). A staff member employs DevOps to provision machines based on specific applications.

Research Compute Resources (Slide 6)

- Integrated portal in DevOps: not only administers deployments, but also has user deployments due to their increased privileges, which allows us to focus more on research and system development.

Summary of System Admin Usage (Slide 7)

- Initial provisioning of new cluster nodes

Types of resources (Slide 8)
DevOps provides a service not provided by other resources Need all 3 to support nation's research goals.

RAIN: Templated Images (see links, Slide 9-12)
RAIN is a provisioning system using DevOps.

- Need templated images integrated through DevOps framework that is updated continuously to maintain security.
- Enabled the re-provisioning of cluster(s) to MPI, OpenStack or Hadoop Mode, which was integrated into the concept of virtual clusters.
- Continued this concept in the NSF SDSC Comet project. Users completely control what they can install. (Slide 12)

Observation: How to leverage these tools (Ansible, Chef, Puppet, CFEngine), including virtualized software environments (e.g. python) and the deployment in a virtual cluster (Slide 15)

Why use Ansible? (Slide 16)

- Evolving focus: Developed Cloudmesh client to switch easily between VM providers.
- Also using DevOps to deploy containers. No need for second platform to develop client tools.

DevOps in support of NIST Service Abstractions (Slides 19-24)

- NIST is developing big data reference architecture
- Architecture will be integrated into DevOps; implement prototype supported by DevOps
- Architecture: enable testing to ensure functionalities work. Start with Cloudmesh compute node (running Hadoop on top) with DevOps hidden behind to make it happen. Users can start Hadoop by themselves on a number of machines; can integrate Ansible, etc.

NIST Usage Scenarios & Links (Slides 25-27)
Single deployment example: employs entire Hadoop fingerprint stack.

Summary (Slide 28)

- DevOps user support allow the configuration and provisioning of sophisticated environments.
- Reimplementing Cloudmesh based on NIST to automatically derive codes to specifications; (e.g., can get templates to implement python, etc.).
- Total integration: we are bringing something new to the table. Need to bring DevOps to users and researchers, which is why we need interfaces such as Cloudmesh.

Q&A

How large of community is participating in Cloudmesh?

- No commercial sponsors. The Indiana University community is doing the development; it is in the early stages of forming a community.

What can the science community or government/ government-assisted IT communities do to foster the development of re-useable tools (accessible (open source or otherwise rapidly re-deployable)?

- Intrinsic weaknesses of DevOps: We end up developing tools that become stock and trade within community of developers and operations people who are working together, but never make it out of that community. Would be useful to pay more attention to getting some of these tools and practices out into other communities. In HEP, almost fully open source and open code – publicly available, but not as useable as would like. We are paying more attention to it now and the open source community is the right way to go. In Atlas, DevOps is more of a culture than tools.

Federal side:

- Code.gov is the government's effort to open source code it is developed; contains a registry of ongoing projects. Standing up code.mil (index of source code that will be making open source and re-useable). Will make efforts more known. GSA and OMB are trying to be proactive in promoting that site for government -related activities.

Secure DevOps or DevSecOps: consensus term to describe it? Folks will discuss offline.

- Different scopes: Secure: ensure that DevOps as a whole is being done securely vs. setting up operations in support of security services (e.g., what is the DevOps for deploying a specific security instrumentation).

CY19 Tasking

CY18 Tasking- Containerization and DevOps series, possibly workshop.

Discussion

- Topics: Cross-agency activities, data gathering in distributed computing environments, provenance issue and verification, edge services, academic community (weaving into MAGIC activities), identity management.
- Put forward a couple of large tasks (several month series) and single sessions (identity management).
- Option for publication of presentations, possibly a special issue of a publication. Gregor Von Laszewski will try reach out.
- Data gathering
    - May include provenance issues.
    - Should deal with data gathering prior to addressing ED services.
    - Impact of streaming data
    - No problem with data gathering, but funding is not available for analyzing data (e.g., space probes. Not good at producing tools to making data accessible, curated and turned into knowledge.
    - Conclusion: Looking at "Data life cycle" which is consistent with ongoing projects; curation and sustaining data over time is increasingly becoming the greatest challenge.
        - NSF pushing convergent research involving data. May add more complexity to data gathering and analysis.

- May be addressed by data life cycle. Universities have been building repositories and archives part of funding mandates to include these components in proposals. The data life cycle is not removed from the design of the knowledge discovery process; data acquisition processing and analysis, data product development reflects the original intent of research. View it as part of process intended to produce knowledge or another agency goal?
- Starting point: Ask scientists to identify the newly discovered science that is driving the analysis?

Summary

Data Life Cycle series:

- 4-5 month series on specific aspect of life cycle (gathering, triaging, analyzing, archiving, and reusing data) and their connections and various aspects of the life cycle. Include merging data life cycle issues and the changing nature and resulting impact of technologies.

Remaining sessions: single item topics such as identity management

Academic community: Roundtable or other activities to keep abreast of curriculum changes and, by extension, workforce development.

Discussion of formal workshop on the Data life cycle

- Need to sharpen the focus and purpose of the workshop (e.g., knowledge discovery and agency missions). Set the agenda and desired outcome.
- What makes data a useful resource and how to support activities that use data productively?

At the September 5 meeting, we will refine the topic and discuss series and get a better idea of whether a workshop would be beneficial. Send an email to Joyce Lee (NCO) if you would like to focus on a specific area.

Roundtable

Coalition of Advanced Scientific Computing (CASC): Held a Bird of Feather on return on investment on academic computing; at Practice & Experience in Advanced Research Computing (PEARC) conference. Funding models should be included in CASC report

August 6-7, National Research Platform, Bozeman, MT

November 11 – 16, Supercomputing18, Dallas, TX

November 12, SC18 Data center automation workshop will cover data center automation technologies, from server control to provisioning systems.

November 14, MAGIC meeting 1:30-3:30pm CT, Room D175

Next meeting: September 5 (12 noon EDT), National Coordination Office.