Networking and Information Technology
Research and Development

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

Middleware and Grid Interagency Coordination (MAGIC) Meeting Minutes
June 1, 2022, 12-2 pm ET

Virtual

Participants

Ashok Srinivasan (NSF)	Katherine Austin Ph.D. (TTU)
David Martin (ANL)	Marc West
David Rager (NREL)	Marcy Collinson (Oracle)
Dhabaleswar K (DK) Panda (OSU)	Rich Carlson (DOE/SC)
Jay Park (NSF/OAC)	Seung-Jong Park (NSF)
Jeff Conklin (NCO)	Sharon Broude Geva (University of Michigan)
Juan Jenny Li (NSF/OAC)	Tevfik Kosar (NSF)
	Vic Baker (MATRIC)

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

Jeff Conklin introduced David Rager as the chief cloud technologist in the Computational Sciences Center at NREL.

Cloud Approach to Machine Learning at the National Renewable Energy Laboratory
David T. Rager, Chief Cloud Technologist, National Renewable Energy Laboratory

- David provided an overview of the presentation.
  - Developer / Systems and Data Architect for 25 years
  - Also a Cloud Solutions Architect for last 12 years
  - Specialization in cloud does overlap heavily with operations. It is a challenge to separate cloud development
  - support from dev-ops – API access to all functionality blurs the responsibilities/roles.
  - Approach to cloud:
    - A resource for developers and analysts at the lab to solve problems in an agile fashion
    - Encourage use of cloud provider managed application and analysis support services to avoid operational overhead, reduce complexity for maintenance and leverage cloud security
    - Leverage extensibility of cloud to "plug in" machine learning where cloud supports goals of a project
  - Multi-cloud tools are hard to find that make typical interactions seamless between environments. API and SDKs are highly opinionated and typically require user specialization to leverage effectively.
  - Once you embrace the SDK/APIs available cloud computing is empowering.

    - David stated that at NREL they operate the Stratus environment – which encompasses AWS, GCP and Azure to support research.
- David talked about how each of the major cloud architecture vendors, Amazon, ~~Azure~~Azure, and Google are aggressively growing with Amazon Web Services taking the larger part of the market share at this moment in time. He displayed a bar chart reflecting the Kaggle survey of cloud provider popularity over the last four years.
- David then displayed a diagram showing on premise services on the left and then each of the cloud services; Infrastructure as a service (IaaS), platform as a service (PaaS), and software as a service (SaaS) represented by the right three columns. He started that the goal is to move machine learning applications to the right side of the diagram.
    - Currently seeing a lot of machine learning activities using the left side or on-premise side of the diagram and using IaaS mostly versus PaaS or SaaS.
    - He sees great advantages of leveraging PaaS in the future and plug and play using SaaS.
- David displayed another bar chart illustrating ML product usage by occupation and stated that currently there is more interest at his facility for using Google's product but still most of the ML work is performed on premise.
- David displayed a slide illustrating Machine Learning focus on data science where it contrasts the areas important to the data scientist from greatest to least to the infrastructure that is needed least to greatest. David's group uses ML for areas like anomaly detection, security events and forecasting costs.
- David provided a slide describing the Stratus project his team is working on.
    - Big Data Analytics
        - data warehousing
        - data management tools
    - Containerized Applications
        - multiple scheduling systems
        - Docker containers at the edge
        - Docker serverless functions
    - Growth In
        - IoT support for field experiments
        - grid management studies
    - Ongoing Support for
        - data processing / ML workflows
        - public web applications / ML inference endpoints
        - Publishing of large open data sets
- David displayed a slide showing how Stratus is leveraged by illustrating through distributed workflows. On the left was the on-premise edge responsible for data collection, the middle represents Stratus in the cloud environment and on the right High Performance Computing environment (HPC) using the Eagle super-computer, where data modeling, ML modeling, training models takes place.
- David displayed a slide illustrating an example of a distributed workflow using cloud capabilities and ML extensibility.
    - Example: Use weather data to predict future energy costs.
        - Integrate AWS Forecast service OR A specialize model not provided by AWS using:
            - SageMaker Platform
            - Dockerized Model

        - These options allow us to use specialized hardware to optimize the job, which can easily change over time.
- David summarized his presentation as follows:
    - Cloud Computing generally supports our goal of commoditizing reliable machine learning at scale:
        - Highly available services
        - API controls for operational use of developed models
        - Scalable compute for large training jobs or heavy consumption of inference
        - Ability to manage and store very large datasets
        - Reliable / distributed event system and scheduler
        - On-demand model
        - FedRAMP security certifications
    - Challenges:
        - Learning curve for using managed services / cloud APIs and SDKs
        - Data scientists usually do not leverage cloud platforms, so models are deployed without the benefits cloud can provide when starting with their integrated platforms
- David stated their goal is to commoditize machine learning at scale.

Questions:

- Tevfik asked about the learning curve of ML in cloud services and if this learning curve would increase or decrease over time.
    - David stated we might continue to see it go up though as cloud providers continue to add more options for SaaS it may drop but at the same time the providers continue to add more and more functionality. Just looking at the classes for APIs there are hundreds of classes, so it just depends on the depth one would want to go to. David stated that what he's seen is that people gravitate towards a specific area and then they become a master in that area. So, in a nutshell it can be simpler using SaaS but also more complex because there is so much more functionality to offer.
- David Martin asked about charges, specifically if charges accrue when a job is submitted etc.
    - David stated they have the ability to forecast based on usage, but the versatile nature of the cloud could generate some unforeseen costs. Specifically, because of the nature of the environment the costs may not come through for more than 12 hours after the compute was executed. For the Google provider there does seem to be an ability to cap costs, where once you go over a certain threshold Google handles the cost differently. In AWS they have experienced unexpected cost spikes and he gave some examples of these experiences.
- Jeff Conklin asked about the preference or compare and contrast of the three largest cloud providers and the move of ML to the cloud.
    - David stated that Google cloud supports tensor flow and so there is some bias in that direction. Azure has a partnership with Data ~~Bricks~~Bricks, so some people support that approach,

<u>Cloud, Containers, and Discovery</u>

Vic Baker, Senior Systems Engineer, Mid-Atlantic Technology, Research, and Innovation Center (MATRIC)

- Multi Cloud – really means multi-environment
    - 'Cloud' is where you're not at
        - Cloud Service Providers
        - On-Prem Compute
        - Edge computing (sensors, cell phone, etc.)
    - Leverage APIs for service-to-service communications
    - Networking:
        - peered networks
        - commodity networks
        - dedicated / shared interconnects
        - consider egress fees (leverage compute where data lives)
    - Utilize cloud-native concepts
- Navigating on-prem to cloud (and local dev)
    - One of the biggest challenges: migrating from an "on-prem" development / deployment environment to cloud – HOW???
    - Utilize "cloud native" concepts
        - Containerization
        - Microservices
        - Scalability
    - Leverage enabling tools:
        - Deployment:
        - Kubernetes
        - Helm
        - Terraform
    - Development:
        - VS Code
        - Containerized environments
- Kubernetes
    - Supported by all major cloud providers
    - Configurable autoscaling based on load thresholds (CPU, RAM usage)
        - Services
        - Nodes
    - Managed service deployments via Helm
    - Seamless local, on-prem, cloud deployments via kubectl
    - Manage multi-environment clusters via GCP Anthos
- Everyone wants to jump to the top of the development pyramid
    - But… scientists still spend nearly 80% of their time acquiring, cleaning, and organizing data
    - The steps at the *bottom* must come first
    - They require leadership & support to ensure a solid data foundation for future R&D
    - "Invest 5% of research funds in ensuring data are reusable. Funders hold the stick: they should disburse no further funding without a data stewardship plan." – Nature, February 2020
    - Data are the energy for AI and analysis
- SmartSearch – Conquering the Data Avalanche
    - How do you currently search?
        - Type in a few keywords
        - Skim the top few results

        - Type in more keywords and try again
    - How do you find and connect to something relevant?
        - Open a file / web page
        - Read it (skim it)
        - Decide if it’s relevant
- SmartSearch automates data discovery by:
    - Analyzing ~~content~~content, you like
    - Finding new content via www, local, enterprise data stores
    - Telling you how relevant the discovered data is to what you like
- How does SmartSearch work?
    - Iterative cycle of Ingest -> Analyze -> Discover -> Catalog -> Recommend
    - Using AWS services – Parsing, Parallel Processing, Machine Learning, Natural Language Processing
    - AWS tools: Kubernetes, Spark, Docker
- Benefits of SmartSearch
    - Infinitely Scalable (Automated) Data Discovery
        - Analyze millions+ of files and generate comparison metrics
        - Generate topic models, categorization, recommendations
        - Desktop, cluster, cloud
    - Treat geospatial data like a document
        - Automatically extract text from geospatial data (shapefiles, geodatabases)
        - Compare textual vs geospatial data to identify relevancy
    - Search for meta tags within HTML body of discovered web sites
        - i.e., find map tags
    - Analyze archive files – even archives within archives (zips within zips, etc.)
        - Process every file – docs, spatial, etc.
- AI/ML in SmartSearch
    - SmartSearch built via cloud native design principles
    - Combines ‘cluster of clusters’, Spark, Kubernetes, and data lakes for massively scalable compute infrastructure
    - Natural Language Processing via SparkNLP
        - Distributed NLP processing via the Spark framework
        - Implemented via SparkML Pipelines
        - Provides thousands of pretrained models and pipelines (Glove, Bert, Onto, etc.)
        - Custom models can be implemented and trained within same distributed framework
    - Machine Learning via SparkML
        - Distributed ML processing via the Spark framework
        - SmartSearch Recommendation Engine
        - LDA Topic Modeling
        - Named Entity Recognition (NER)
        - Question Answering
        - Summarization
- SmartSearch In-Use
    - Global Open Oil & Gas Infrastructure Database
    - Carbon Storage data resources
    - Alloy property data
- Summary

  - SmartSearch automates the data discovery process by using a scalable compute environment coupled with NLP and ML
  - Will be integrated within EDX
  - SmartSearch supports ongoing research projects
  - SmartSearch used to evaluate cloud providers (GCP, AWS, Azure)

Questions:

- Tevfik had a question about searching content and talked briefly about the textual ~~approach, but~~approach but wanted to know if there was a way to search images and was metadata involved.
  - Vic stated they had the capability to search against textual metadata but at this time they did not have the capability to search images.

Designing Next-Generation Intelligent CyberInfrastructure: An Overview of the NSF-AI ICICLE Institute

Dhabaleswar K. (DK) Panda, University Distinguished Scholar of Computer Science and Engineering at the Ohio State University.

- High end computing has been evolving over the last three decades with multiple stages
- Stage 1 (1975 -): Scientific computing with Supercomputing/High performance computing (HPC)
- Stage 2 (2000 -): HPC and Big Data Analytics
  - Big Data changes the way people harness the power of data
  - Big Data and HPC started converging to meet large scale data processing challenges
  - Running High Performance Data Analysis (HPDA) workloads in the cloud has been gaining popularity
    - According to the latest OpenStack survey, 27% of cloud deployments are running HPDA workloads
  - Has evolved into Data Science
- Stage 3 (2010 +): HPC + AI (Machine Learning and Deep learning)
  - Machine Learning (ML) – the study of computer algorithms to improve automatically through experience and use of data
  - Deep Learning (DL) – a subset of ML that uses deep neural networks (DNNs)
    - Based on learning data representation
    - DNN examples: Convolutional Neural Networks, Recurrent Neural Networks, Hybrid Networks
  - AI-enabled Science, art, health, business
- Stage 4 (2015 -): Emergence of Computing Continuum
  - From HPC systems and data centers to Clouds, Edge, on field sensors, HPC
- DK displayed a slide depicting three slices of a pie representing HPC, Deep/machine learning, and Big Data/Science Data and the title "Increasing Usage of HPC, AI, and Data Science in multiple disciplines with distributed data and heterogeneous computing". It illustrates the convergence of HPC, Deep Learning and Data Science and the increasing need to run these applications in the cloud.
- The broad challenge is how to design the next generation intelligent cyber-infrastructure with plug-and-play capabilities to handle societal problems while taking advantage of heterogeneous high-performance computing and cloud resources.

- One solution is ICICLE – Intelligent Cyberinfrastructure with Computational Learning in the Environment http:// iscicle.ai – 20 ~~Million~~million USD for 5 years
- ICICLE Vision – A national infrastructure that enables AI at the flick of a switch, ICICLE will:
    - Democratize AI through integrated plug-and-play AI
    - Catalyze foundational AI/CI and transform application domains
    - Transparent and trustworthy infrastructure for AI-enabled future
    - Address societal problems and national priorities
    - Grow new generations of workforce and incubate sustainable and inclusive communities
- 14 organizations, 46 investigators, and many collaborators are participating in this project
- DK displayed a slide illustrating the ICICLE leadership team.
- DK displayed a slide entitled Objectives: Intelligent Cyberinfrastructure for Computing Continuum
    - On this slide is a layered diagram with the following:
        - Emerging Computing Continuum on the bottom: on field sensors (IoT), Edge and Near edge devices, Hybrid Cloud (on-premise/cloud), HPC systems and data centers
        - ICICLE as the middle layer
        - Inspired Science Domains as the top layer – Smart foodsheds, animal ecology, digital agriculture
- ICICLE DNA – Foundational Systems AI
    - Knowledge graphs – Multimodal, special-temporal
    - Model Commons – KG supported, precise profiling
    - Adaptive AI – context aware, interactive, continual learning
    - Federated learning – Heterogeneity, privacy preserving and robustness
    - Conversational AI – KG and model commons aware
- The Enabler@Edge – CI4AI
    - High Performance Training
    - High Performance Data Management
    - Edge Intelligence
    - AI-Adaptive Edge Wireless
    - Control and Coordination
- The Enabler@Scale-and-@Edge: AI4CI – This slide depicts architecture and how the parts of the system fit in with systems/middleware/application layers.
    - Systems – intelligent scheduling, intelligent wireless communication
    - Middleware – hyperparameter optimization, intelligent compute primitives
    - Application – Performance modeling/prediction
- The next two slides illustrate crop care examples
    - Nutrient and Pest management – demonstrate swarm of unmanned aerial systems to study crop stressors and insect infestations, etc. and produce crop maps to improve agriculture productivity
    - Water Management and Quality – Capture nutrient stress and moisture deficit in corn and soybean field and provide feedback control for automation of agriculture field machinery
- The net slide illustrates the ICICLE Software Architecture
- ICICLE as a Whole
- ICICLE Enables Global AI Leadership
    - Integrate into the National AI Ecosystem
    - Integrative and Interoperable
    - Leverages Existing Recognized Capabilities – Centers of Excellence, AI Institutes

    - Collaborative
    - Sustainable – Workforce development, broadening participation, ~~collaborative~~collaborative, and knowledge transfer
- Conclusion
    - AI solutions can help solve many societal problems
    - Increasing use of HPC, AI, Data Science with heterogeneous resources
    - Need for plug-and-play AI solutions which can democratize AI
    - The new ICICLE NSF-AI Institute aims to establish next generation cyberinfrastructure to provide comprehensive AI solutions to many societal problems

Questions:

- Tevfik asked are there any new challenges on this project that were not anticipated now that they are six months into it.
    - DK stated that the team is large and even though they have worked together on different projects in the past, there are still gaps in knowledge they are trying to overcome. There are a lot of meetings but with a project this size its hard to make good progress.

Roundtable

- The co-chairs talked about the NITRD 30th Anniversary event.

Next Meeting

July 6th (12 pm ET)