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

Middleware and Grid Interagency Coordination (MAGIC) Meeting Minutes
May 3, 2023

Virtual Meeting

Participants

Cong Wang (New Jersey Institute of Technology)	Mallory Hinks (NCO)
David Martin (ANL)	Marcy Collinson (Oracle)
Debbie Bard (LBNL)	Martin Swany (Indiana University)
Hal Finkel (DOE SC)	Miron Livny (Wisconsin)
Ian Karlin (NVIDIA)	Rafael Ferreira da Silva (ORNL)
Jay Park (NSF)	Ravi Madduri (ANL)
Jonathan Skone (LBNL)	Rich Carlson (Retired DOE)
Kevin Thompson (NSF)	Sheikh Ghafoor (Tennessee Tech University)

Introductions: This meeting was chaired by Jay Park (NSF) and Hal Finkel (DOE SC)

An Overview of AI Workflows

Rafael Ferreira da Silva (ORNL)

Rafael discussed the traditional scientific workflows and their capabilities of analyzing petabyte-scale datasets composed of millions of individual tasks. He also talked about the modern scientific workflows, which include AI, and he acknowledged several colleagues who provided slides for the presentation.

Rafael discussed the use of AI in empowering HPC workflows and shared an interesting slide from Shantanu Jha from Rutgers and Brookhaven that showed how combining machine learning with molecular dynamics executions can help in reaching beyond regular HPC workflows in a shorter time scale. Rafael emphasized that AI not only helps in finding new solutions, but can also accelerate workflows, leading to a disruptive impact in the ecosystem.

Rafael shared three examples of modern scientific workflows, including:

- An autonomous multi-scale workflow that uses AI to create surrogate models and physical models to analyze large-scale datasets - from Timo Bremer at Lawrence Livermore National Lab
- A workflow that runs across different facilities and involves cloud computing, HPC, and edge computing, and a distributed workflow that is managed by AI and predicts future resource allocation - from Rosa Badia at the Barcelona Supercomputing Center

- A workflow that involves privacy preserving federated learning using a distributed set of workflows and AI. Users can access a web app service to train and compute models at different computing sites, with coordination across different HPC sites – from Ravi Madduri at Argonne National Laboratory

He also summarized the changes in modern scientific workflows from traditional static DAGs to dynamic and evolving workflows that involve input from humans and scientists.

Rafael discussed the challenges with AI workflows including the lack of support for heterogeneity in computer resources and difficulty with fine-grained data management within workflow systems. He also discussed the challenges faced in AI HPC workflows, including the lack of support for heterogeneity in computer resources and fine-grained data management. He also mentioned the need to develop AI benchmark workflows. He introduced the survey of AI HPC workflow applications, meter, and performance, which aims to create common execution motives of AI coupled HPC workflows. The survey categorizes the workflows into multistage and multistage backlining, including examples of multi-scale and multi-physics, inverse design, concurrent duality, distributed models and dynamic data, and adaptive execution for training. Rafael emphasized the growing complexity of workflows mixing different motifs and involving different infrastructures. He also mentioned ML Commons' benchmarks for machine learning.

Rafael explained the need to develop benchmarks and tools for assessing deep neural network Surrogate models and suggested a common SDK for testing different workflow systems. He also discussed the creation of a workflow community initiative and the characteristics of AI workflows, including the lack of fine-grained data management, large volumes of data and metadata, and pseudo-random access to datasets. He emphasized the challenges that arise when human input and decision-making are involved. He also discussed the classification of AI workflows and its impact on the community.

He discussed the challenges in developing AI workflow benchmarks and the need for a paper defining the requirements and use cases of AI workflows. He also emphasized the lack of existing benchmarks for workflows and the need for common terminology to categorize AI enabled workflows. The discussion highlighted the complexity of mixing different motifs and infrastructures, as well as the challenges of human input and decision-making.

Rafael explained the concept of AI enhanced workflows, which involve using AI in the simulation code to help with problem solving. These workflows mix AI in multiple aspects, including helping with the workflow system, the simulation, and generating the final model. This can lead to a complex integration of AI throughout the workflow.

Rafael discussed the different categories of AI-enabled workflows, including inner loops, outer loops, and coupled loops. He acknowledged that there is disagreement regarding these categories. He also emphasized the challenges of integrating different technologies and the constant development of new AI tools.

Rafael discussed a project targeting integration of workflow and application services, including data management frameworks, visualization frameworks, and AI tools. The project involves engaging with different stakeholder communities, computing centers, and facilities to meet requirements.

Questions:

- The Q&A session revolved around the need for integration between workflow systems and AI training frameworks, specifically for large language model training in PyTorch. The group discussed the challenges of creating a new programming language for workflow systems versus using existing tools like MLflow and the importance of better communication between developers and ML tool users.

Managing a highly heterogeneous workload at NERSC: How we provision resources for batch and urgent workflows

Debbie Bard, NERSC

Debbie's presentation focused on resource provisioning and resilience at NERSC. She discussed strategies for balancing batch workload and urgent workflows at NERSC, as well as the importance of resilience in HPC facility operations. With 9,000 users from various science areas and a wide range of applications, NERSC's user base and workflows reflect the priorities of DOE program managers. She provided an overview of NERSC's computing infrastructure, including their supercomputers, storage, and side cluster. She also discussed the deployment of appropriate technology to serve the needs of scientists with complex workflows. NERSC is currently running nearly 3,000 compute jobs of varying sizes.

Debbie discussed the range of scales and workloads that run on NERSC's systems and how the SLURM workload scheduler is used to manage them. She focused on the challenges of supporting real-time, urgent compute demands from experimental and observational facilities while also maintaining high utilization and not disrupting regularly scheduled batch nodes. NERSC cannot afford to have idle capacity, so maximizing system utilization is key.

Debbie discussed the importance of maximizing system utilization while also provisioning urgent requests. She focused on the real-time queue, reservations, and preemptible jobs as strategies for addressing this challenge. The real-time queue is reserved for scientists with running experiments that have genuine urgency, with a small number of nodes set aside for this workload. Jobs in the real-time queue enter the queue with very high priority. She also discussed the use of the real-time queue with high priority for urgent requests, but noted the importance of not letting them disrupt larger jobs that are already draining the system. NERSC also has a reservation system for urgent computing, but works closely with science teams to schedule access to resources.

Debbie talked about NERSC's use of reservations for science teams with predictable compute needs, but noted that during downtimes in reservations, nodes sit idle and are wasted. To

address this issue, NERSC has implemented the use of preemptible jobs that fill in the gaps in reservations when nodes are not being used by science teams. This has been working well and is advertised to users. However, charging for preemptible jobs is still being figured out.

She also mentioned the challenge of accurately charging for compute time used during reservations and non-reservation jobs, and their efforts to integrate user space checkpoint restart tools for easier preemptible job management. She emphasized the need for a large preemptible workload to handle urgent compute requests, such as in the event of a supernova, and how NERSC works closely with SLURM to balance various workload needs.

Debbie discussed the challenges of resilience for experiment teams and the need for a truly resilient workflow. NERSC has made improvements to increase robustness and resilience, such as increasing generator capacity and deploying non-disruptive software patches. However, a truly resilient workflow needs to span multiple computing centers, which is a challenge. Bard also discussed the growing complexity of workflows, which now involve simulation, AI, and data analysis, and the need to design the next system around accelerating end-to-end workflows.

Questions:

The Q&A session focused on various challenges related to accommodating and managing AI workloads in a supercomputer environment. Debbie Bard, Ian Karlin, and Jay Park discussed the need for common interfaces for basic operations, the potential increase in AI workloads, and the challenges related to hardware allocation and infrastructure. The group also discussed the potential integration of Kubernetes infrastructure into the HPC system, the balancing of resource allocation between workloads and Kubernetes, and the cultural challenges related to dealing with failures and evolving application design. The session highlighted the need for agencies, vendors, and the community to work together to support these efforts.

- Debbie Bard and Ian Karlin discussed the challenges of making workflows portable across different machines and sites. They discussed the importance of common interfaces for basic operations and gathering community consensus around what interfaces are appropriate. This effort and supporting it will require agencies and vendors to be very supportive, and community work will be required to get that going.
- Jay Park and Debbie Bard discussed how the existing SLURM queue accommodates small AI workloads and treats them like any other compute job. They mentioned that there are currently no specific requirements for AI workloads on their GPU supercomputer, but they anticipate challenges when combining AI with data analysis in a running experiment.
- Debbie Bard and Jay Park discussed the potential increase of AI workloads and the challenges it presents in terms of hardware allocation and infrastructure. Debbie also commented on the challenge of adapting cloud computing approaches for scientific applications and the potential integration of Kubernetes infrastructure into the HPC system.
- Debbie Bard, Hal Finkel, and Miron Livny discussed the potential integration of Kubernetes infrastructure into the HPC system and the challenges of balancing resource

allocation between current workloads and Kubernetes-style workloads. They acknowledged the difficulty of predicting when larger partitions of the machine will be needed for certain jobs and the manual time-consuming process of checkpointing and restarting jobs. Miron Livny commented on the cultural challenge of dealing with failures in the HPC system and the need for developers to evolve their thinking about application design.

Roundtable

- Oracle is sponsoring two working groups with the Research Data Alliance. They just finalized the key statement for the first working group. Marcy gave a heads up that they will be launching the second working group in a couple of months. https://www.rd-alliance.org/group/rda-ofr-mapping-digital-research-data-infrastructure-landscape-wg/case-statement/rda-ofr

Next Meeting June 7, 2023