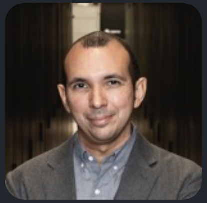
An Overview of AI Workflows

for HPC Systems



Rafael Ferreira da Silva, Ph.D.
Senior Research Scientist – Oak Ridge National Laboratory
Executive Director – Workflows Community Initiative

https://rafaelsilva.com

Outline

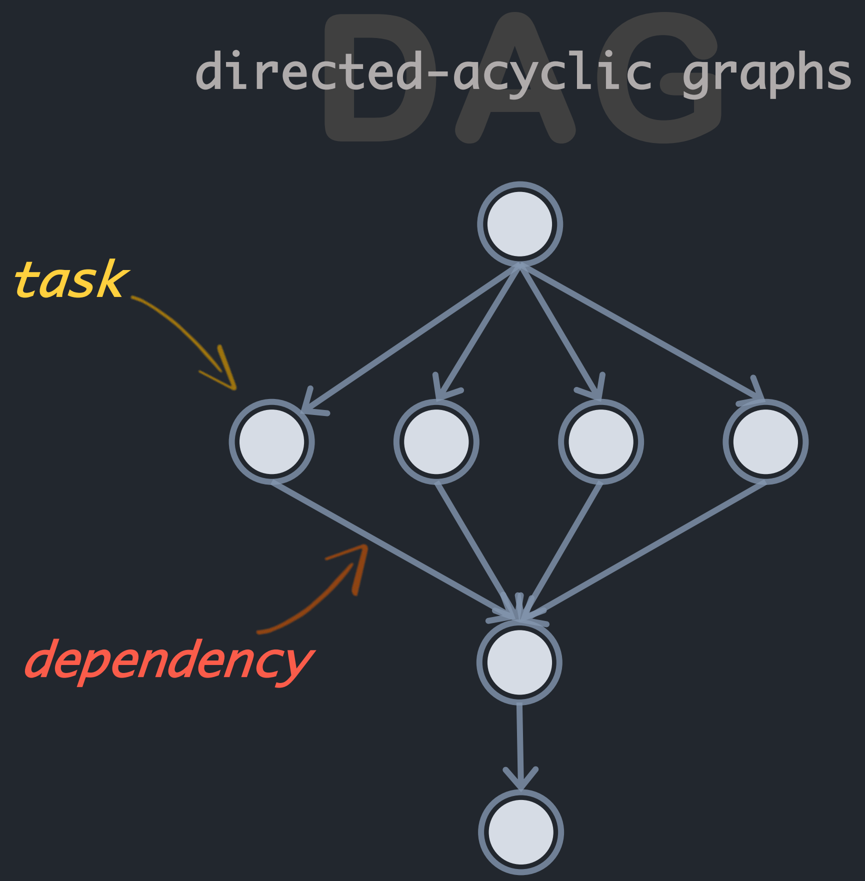
Traditional Scientific Workflows

Modern Scientific Workflows

AI and **HPC** Workflows

Workflows **Community**: sustaining workflows development and research

Traditional Scientific Workflows



Tasks often represents a **program** (or script) written in any programming language (**closed box****)

Dependencies are typically based on the **data flow**. It can also be expressed as **conditions, exceptions, user triggered action,** etc.

State-of-the-art of Traditional Scientific Workflows

Workflows are becoming more **complex** and require more **sophisticated** workflow management capabilities

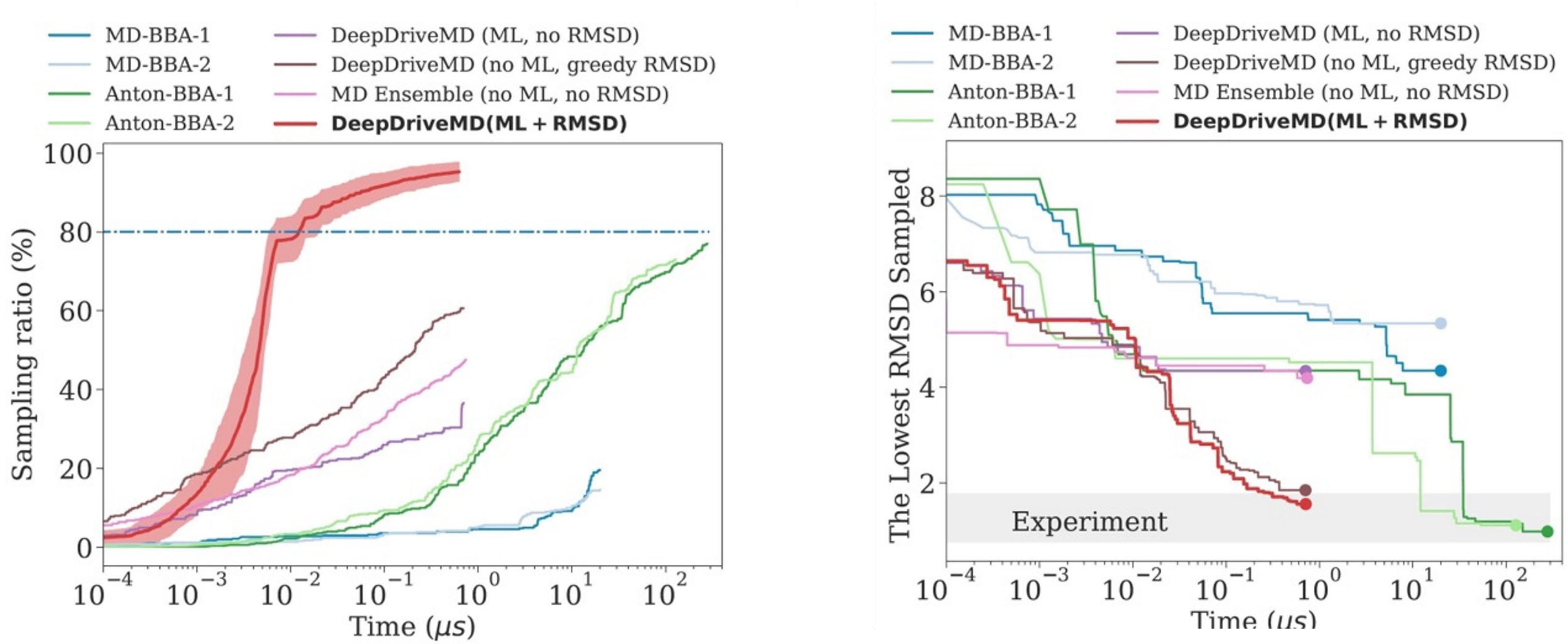
Workflows can now analyze **petabyte-scale datasets**, be composed of **millions of individual tasks** that execute for milliseconds up to several hours on distributed heterogenous platforms

Catering to these workflow features and demands requires **WMS research** and **development at several levels**, from algorithms and systems to user interfaces

AI and HPC Workflows

Empowering HPC workflows with AI

Revolutionizing the way we approach HPC workflows, **AI's disruptive impact** has enabled large-scale simulations

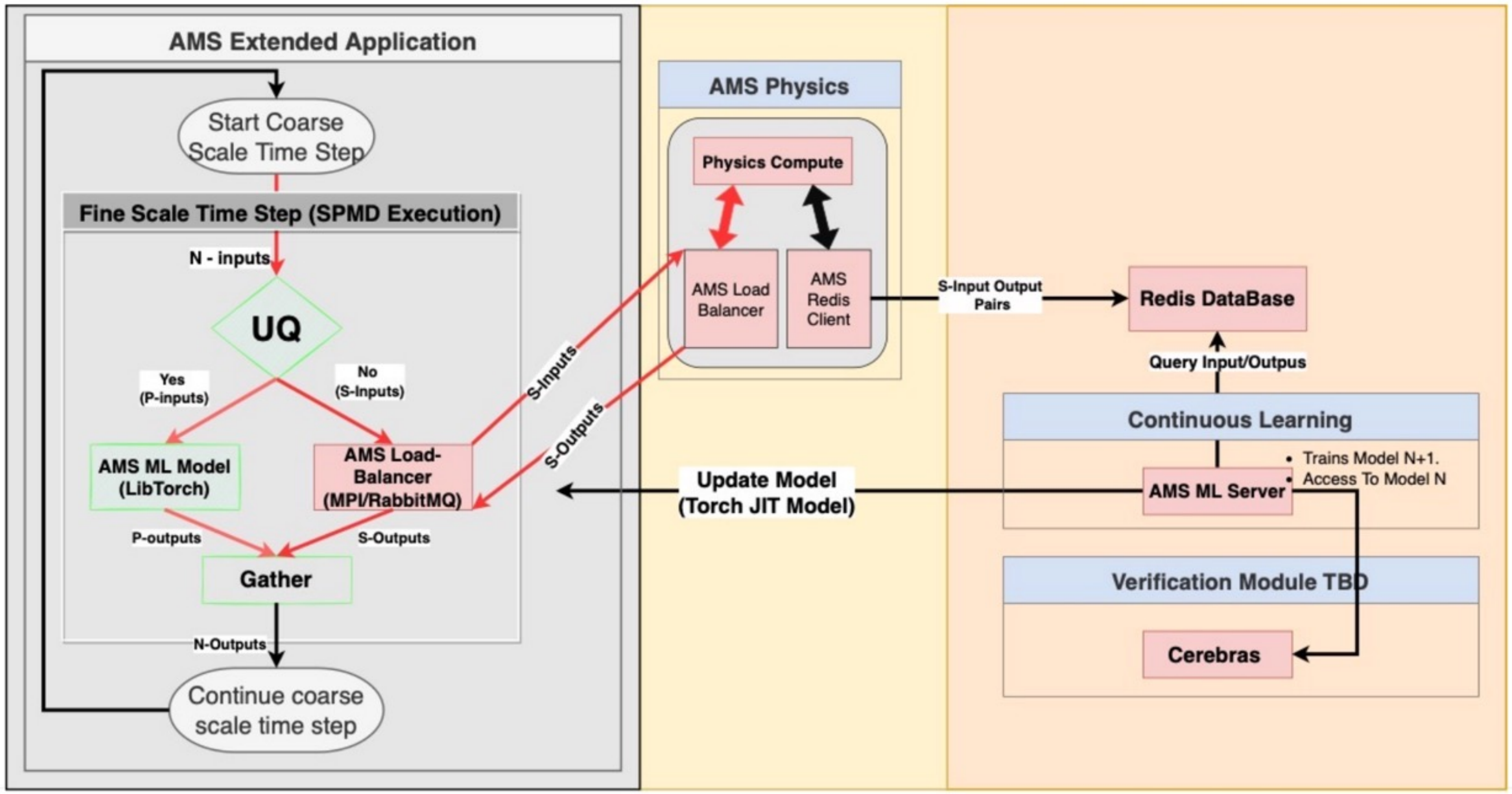


Achieving 100X faster simulations of complex biological phenomena by coupling ML to HPC ensembles. Alexander Brace, Hyungro Lee, Heng Ma, Anda Trifan, Matteo Turilli, Igor Yakushin, Todd Munson, Ian Foster, Shantenu Jha, Arvind Ramanathan: https://arxiv.org/pdf/2104.04797.pdf

Courtesy of Shantenu Jha (Rutgers/BNL)

Modern Scientific Workflows

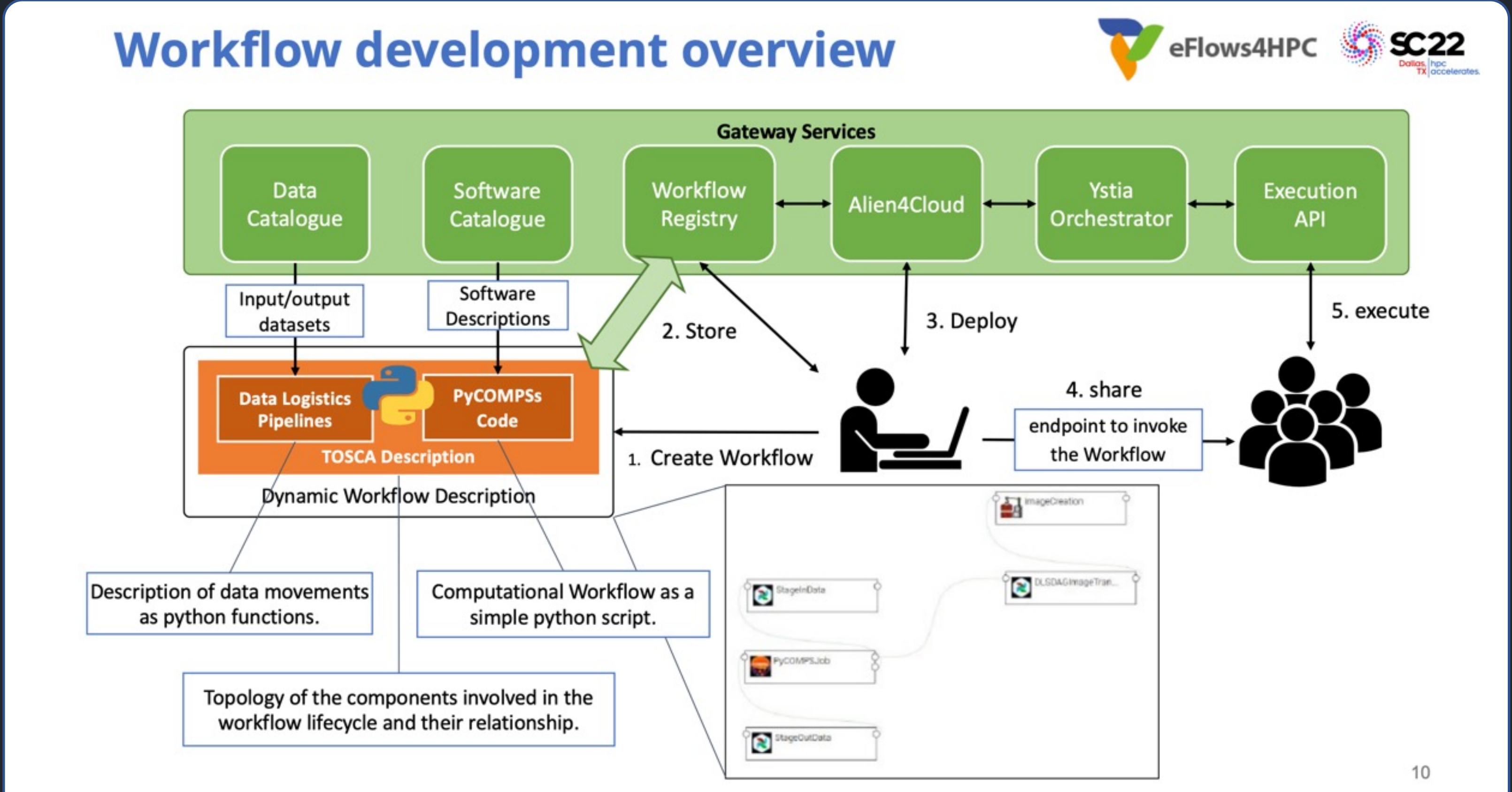
Autonomous Multiscale Aims to Break this Pattern by Directly Integration **Simulation, Data Collection**, and **Training**



Courtesy of Timo Bremer (LLNL)

Modern Scientific Workflows

A dynamic workflow that spans **multiple facilities** and **data management** services managed by AI

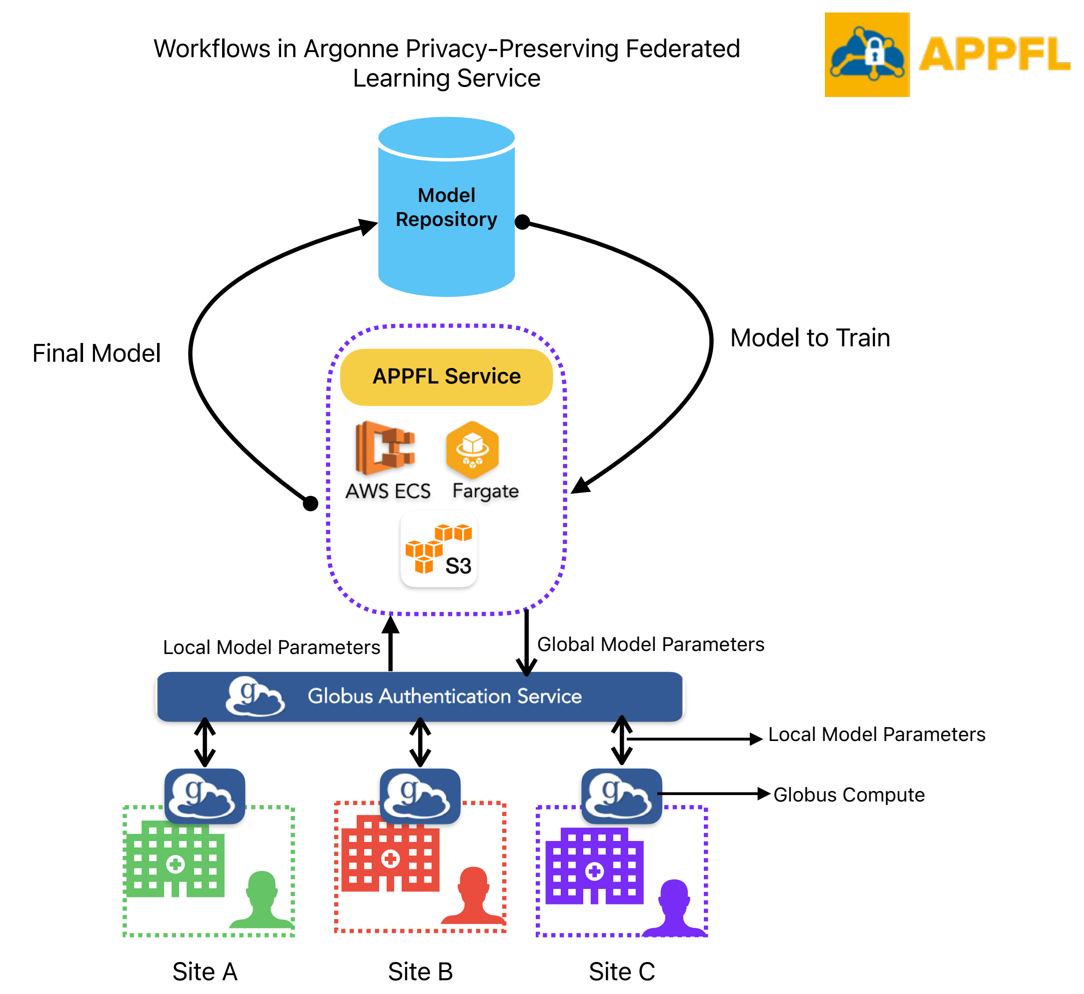


Courtesy of Rosa M. Badia (BSC)

Modern Scientific Workflows

Privacy-preserving Federated Learning via Web App (UI) and model training on HPC

Comprehensive report for each federation learning experiment including:
- training log
- hyperparameters validation results
- training metrics
- tensorboard visualization



Courtesy of Ravi Madduri (ANL)

The workflows community response to these emerging workflows

The 2021 Community Roadmap

arXiv.org > cs > arXiv:2110.02168

Search...

Help | Advance

Computer Science > Distributed, Parallel, and Cluster Computing

[Submitted on 5 Oct 2021 (v1), last revised 8 Oct 2021 (this version, v2)]

A Community Roadmap for Scientific Workflows Research and Development

Rafael Ferreira da Silva, Henri Casanova, Kyle Chard, Ilkay Altintas, Rosa M Badia, Bartosz Balis, Tainã Coleman, Frederik Coppens, Frank Di Natale, Bjoern Enders, Thomas Fahringer, Rosa Filgueira, Grigori Fursin, Daniel Garijo, Carole Goble, Dorran Howell, Shantenu Jha, Daniel S. Katz, Daniel Laney, Ulf Leser, Maciej Malawski, Kshitij Mehta, Loïc Pottier, Jonathan Ozik, J. Luc Peterson, Lavanya Ramakrishnan, Stian Soiland-Reyes, Douglas Thain, Matthew Wolf

The landscape of workflow systems for scientific applications is notoriously convoluted with hundreds of seemingly equivalent workflow systems, many isolated research claims, and a steep learning curve. To address some of these challenges and lay the groundwork for transforming workflows research and development, the WorkflowsRI and ExaWorks projects partnered to bring the international workflows community together. This paper reports on discussions and findings from two virtual "Workflows Community Summits" (January and April, 2021). The overarching goals of these workshops were to develop a view of the state of the art, identify crucial research challenges in the workflows community, articulate a vision for potential community efforts, and discuss technical approaches for realizing this vision. To this end, participants identified six broad themes: FAIR computational workflows; AI workflows; exascale challenges; APIs, interoperability, reuse, and standards; training and education; and building a workflows community. We summarize discussions and recommendations for each of these themes.

Comments: arXiv admin note: substantial text overlap with arXiv:2103.09181

Subjects: **Distributed, Parallel, and Cluster Computing (cs.DC)**

Cite as: arXiv:2110.02168 **[cs.DC]**

(or arXiv:2110.02168v2 **[cs.DC]** for this version)

We summarize findings of discussions within the community by presenting a consolidated view of the **state of the art, challenges**, and potential efforts, which we eventually synthesize into a **community roadmap**

https://arxiv.org/abs/2110.02168

Challenges

https://arxiv.org/abs/2110.02168
A Community Roadmap for Scientific Workflows Research and Development

Theme	Challenges
FAIR Computational Workflows	• Define FAIR principles for computational workflows that consider the complex lifecycle from specification to execution and data products • Define metrics to measure the FAIRness of a workflow. • Engage the community to define principles, policies, and best practices
AI Workflows	• Lack of support for heterogeneity of compute resources and fine-grained data management features, versioning, and data provenance capabilities • Lack of capabilities for enabling workflow steering and dynamic workflows • Integration of ML frameworks into the current HPC landscape
Exascale Challenges and Beyond	• Resource allocation policies and schedulers are not designed for workflow-aware abstractions, thus users tend to use an ill-fitted job abstraction • Unfavorable design of resource descriptions and mechanisms for workflow users/systems, and lack of fault-tolerance and fault-recovery solutions
APIs, Reuse, Interoperability, and Standards	• Workflow systems differ by design, thus interoperability at some layers is likely to be more impactful than others • Workflow standards are typically developed by a subset of the community • Quantifying the value of common representations of workflows is not trivial
Training and Education	• Many workflow systems have high barrier to entry and lack training material • Homegrown workflow solutions and constraints can prevent users from reproducing their functionality on workflow tools developed by others • Unawareness of the workflow technological and conceptual landscape
Building a Workflows Community	• Define what is meant by a "workflows community" • Remedy the inability to link developers and users to bridge translational gaps • Pathways for participation in a network of researchers, developers, and users

Proposed Community Activities

Theme	Community Activities
FAIR Computational Workflows	• Review prior and current efforts for FAIR data and software with respect to workflows, and outline rules for FAIR workflows • Define recommendations for FAIR workflow developers and systems • Automate FAIRness in workflows by recording necessary provenance data
AI Workflows	• Develop comprehensive use cases for sample problems with representative workflow structures and data types • Define a process for characterizing the challenges for enabling AI workflows • Develop AI workflows as a way to benchmark HPC systems
Exascale Challenges and Beyond	• Develop documentation in the form of workflow templates/recipes/miniapps for execution on high-end HPC systems • Specify benchmark workflows for exascale execution • Include workflow requirements as part of the machine procurement process
APIs, Reuse, Interoperability, and Standards	• Identify differences and commonalities between different systems • Identify and characterize domain-specific efforts, identify workflow patterns, and develop case-studies of business process workflows and serverless workflow systems
Training and Education	• Identify basic sample workflow patterns, develop a community workflow knowledge-base, and look at current research on technology adoption • Include workflow terminology and concepts in university curricula and software carpentry efforts
Building a Workflows Community	• Establish a common knowledge-base for workflow technology • Establish a *Workflow Guild*: an organization focused on interaction and good relationships and self-support between workflow developers and their systems

Community approaches for tackling *(some)* of these challenges

Common Execution Motifs of AI-coupled-HPC Workflows

Multistage (and typically multiscale & multiphysics) pipelining

Molecule selection, virtual screening

ML models steering ensembles of simulations

Advanced sampling, swarm methods

Inverse design (from observations, determine causal factors)

Molecule or material design given property(s); e.g., Structure → Sequence

Concurrent duality

Concurrent HPC simulation and AI/ML based Digital Twins

Distributed models and dynamic data

Distributed AI/ML based reduction/analysis coupled to HPC simulations
Diverse models on edge-to-exascale infrastructure

Adaptive execution for training

HPO, NAS, LLM (e.g., Adaptive LLM Training for Molecular Design)

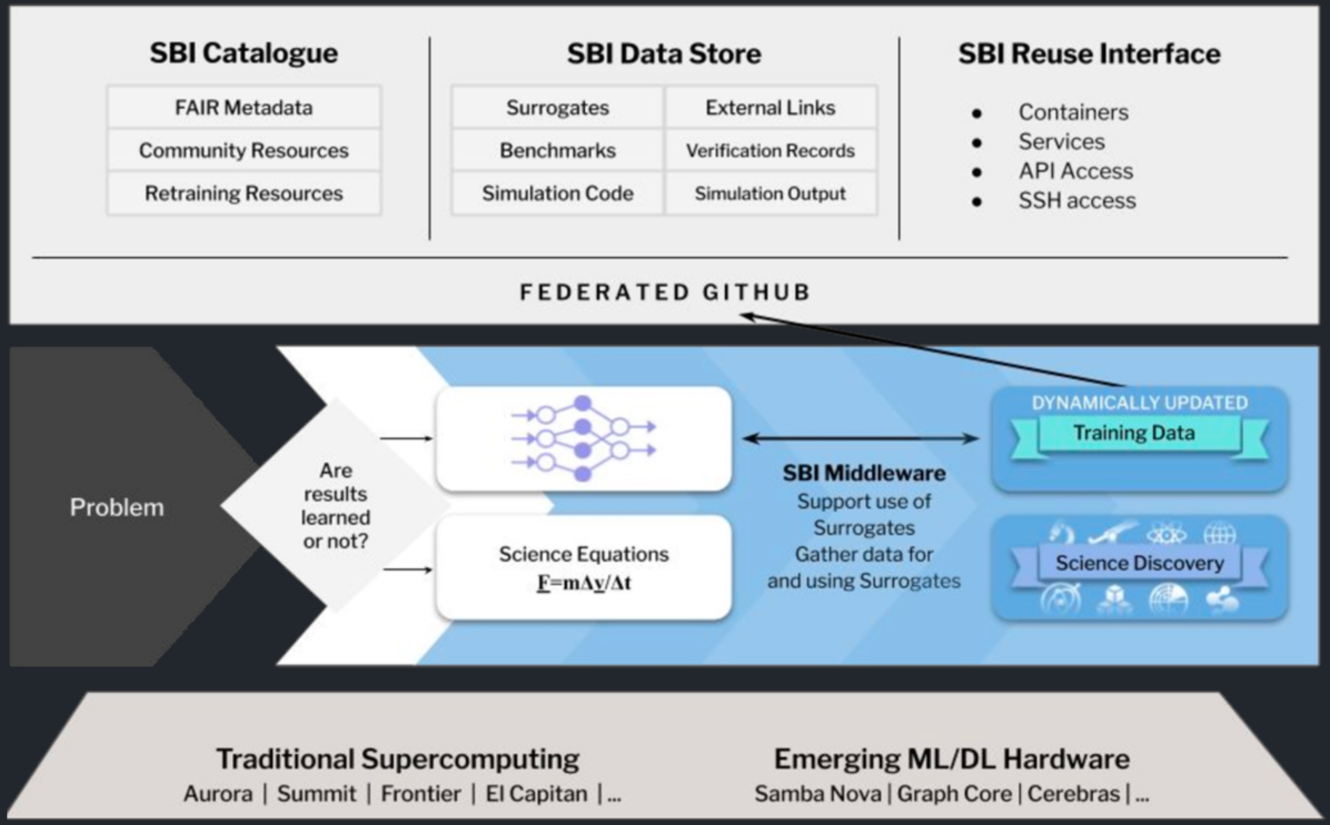
Gainaru and Sha et al. *A Survey of AI-HPC Workflow Applications, Middleware and Performance*, 2023 (under development)

Benchmarks

DOE ASCR's Surrogate Benchmark Initiative

Benchmarks and tools for assessing deep neural network *surrogate* models

Generalize to other AI-coupled-HPC motifs



Courtesy of Shantenu Jha (Rutgers/BNL)

Enabling Workflows at Exascale

Provide a production-grade **Software Development Kit (SDK)** for exascale workflows

The **SDK democratizes access** to hardened, scalable, and interoperable workflow management technologies and components

Approach

- Community policies for software quality (based on E4S)
- Open community-based design and implementation process
- Ensure scalability of components on **Exascale Systems**
- Standard packaging and testing
- Work toward shared capabilities in the SDK

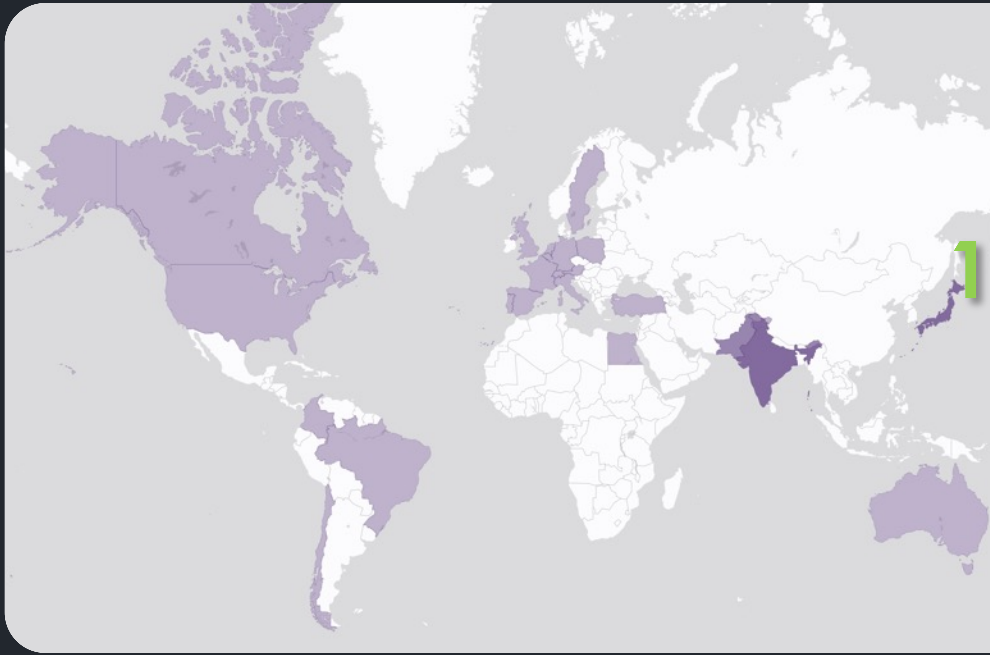
Community building and towards sustainability

Workflows Community Initiative

https://workflows.community

The **Workflows Community Initiative (WCI)** aims to unite the workflows community, including **users**, **developers**, **researchers**, and **facilities**, by offering shared resources and capabilities

This fosters the discovery of relevant ***software products***, ***events***, ***technical reports***, and other related efforts while also promoting collaboration to address significant challenges in the field



29 workflow systems

170 community members

6 working groups

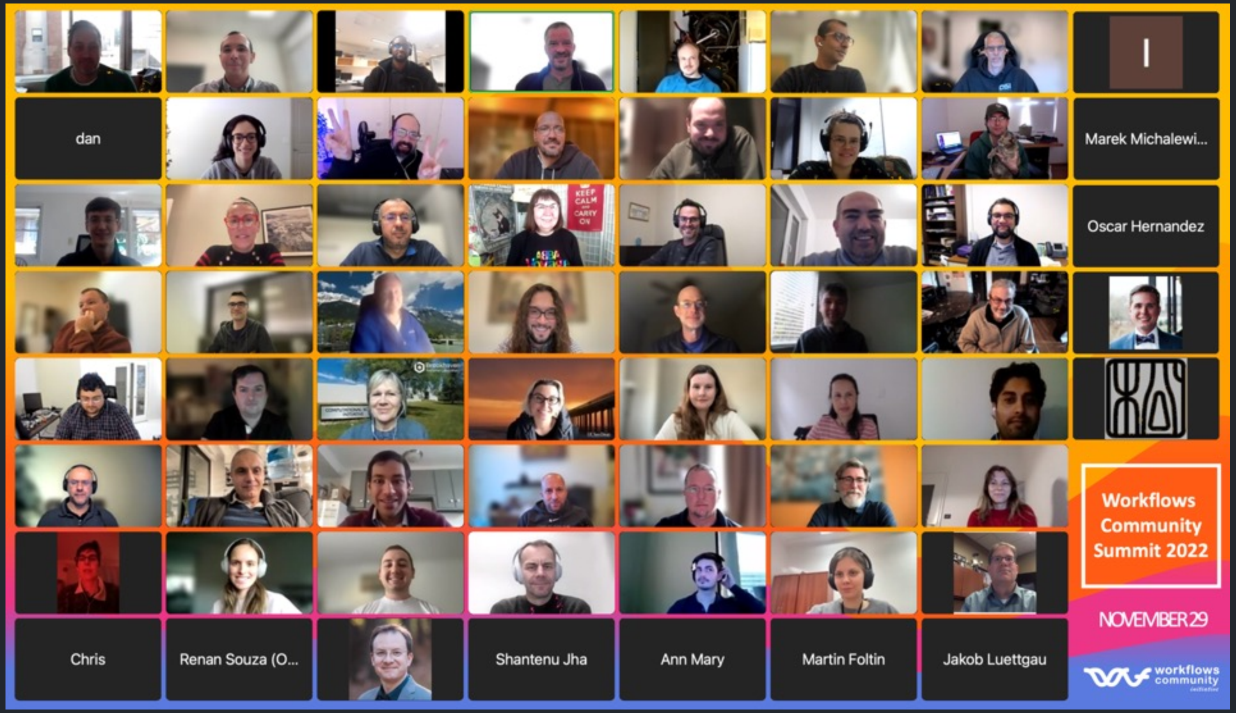
4 research frameworks

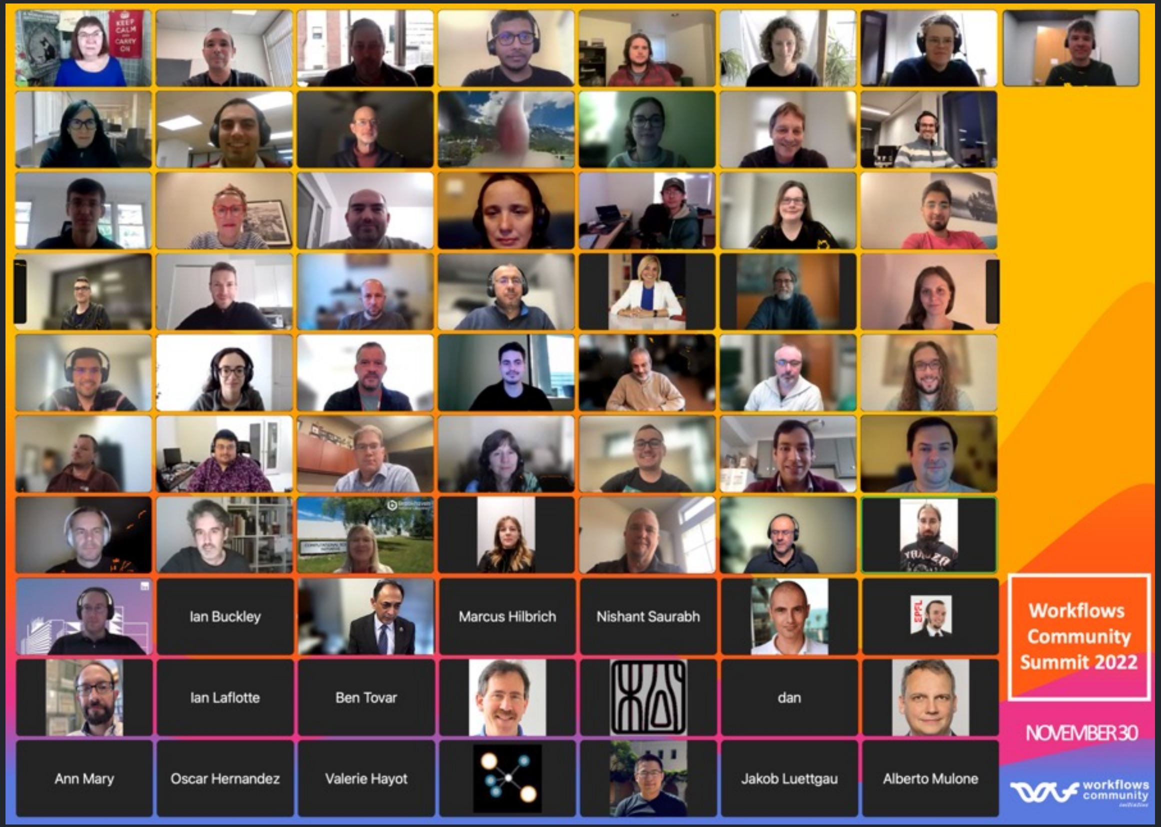
we are always seeking for volunteers…

Workflows Community Summit 2022

workflows community initiative
https://workflows.community

Nov 29 and 30, 2022





100+ international participants

Revised Recommendations

Specifications, standards, and APIs

Standards are often constrained or **hard to implement**
Define **common terms**, building blocks, and concepts

High performance data management and in situ workflows

Edge to cloud continuum and data exchange through data objects
Data usage is more **fine-grain** than a typical HPC code
Adaptive **compression** to reduce data necessary to represent the problem domain

HPC and Quantum workflows

Community do not know how to transfer information to the QC system
Limited resources (hard to access), long queues & expensive
Heterogeneity in quantum devices (vendor specific APIs)

DOI: 10.5281/zenodo.7750669

Revised Recommendations

FAIR workflows

Standards for expressing the **inputs** of the workflow

Limited availability of **metadata**

FAIR data and FAIR workflows are **intertwingled**

Workflows for continuum and cross-facility computing

Describe **aggregate I/O** needs of a workflow

Coordinate **communication** between sites (different **security** domains)

People who design **experiment facilities** are not necessarily computing experts

DOI: 10.5281/zenodo.7750669

AI Workflows Characteristics

workflows community *initiative*

https://workflows.community

DOI: 10.5281/zenodo.7750669

Fine-grained data management

may significantly impair the performance of the shared filesystem
(e.g., image training sets)

Large volumes of data and metadata

NVMes have significantly improved I/O throughput but it is still a bottleneck for AI workloads

Pseudo-random access to datasets

Need for optimizing access to microservices and performing data caching operations
(human-in-the-loop may aggravate the problem)

Recommendations:

- Better understanding of the **requirements** of AI workflows
- Development of AI **workflow benchmarks**

AI Workflows Characteristics

workflows community initiative

https://workflows.community

DOI: 10.5281/zenodo.7750669

AI-enabled-workflows

Workflows for AI
Develop the AI

AI-integrated workflows
Problem-solving with AI

AI-enhanced-workflow
Workflow engine is enhanced with AI

Sub-categories

Inner Loops
ML replacement of subroutines, adapting, and parallel training

Outer Loops
Reinforcement and active learning

Coupled Loops
Digital twins – component of the workflow is controlled by an AI method

Recommendations:

- Refine the **motifs** that would capture the unique requirements of each class
- Common **terminology** and definitions of categories of AI workflows
- Smooth the **integration** (along with the deployment) paths across these categories and their associated components

Sustaining Workflows and Application Services

sustaining **workflows**
& application **services**
https://swas.center

SWAS brings together **academia**, **national labs**, and **industry** to create a sustainable software ecosystem supporting the myriad software and services used in **workflows** as well as the workflow orchestration software itself

Target Software

Workflow Systems
Data Management Frameworks
Visualization Frameworks
AI/ML Tools *(used in modern workflows)*
Application services *(including Cloud services)*

Stakeholder Communities

Workflows and application services
Focus on general and specific domains, non-expert and expert users, and offer configuration-based interfaces, graphical interfaces, domain-specific languages, or programming language libraries or APIs

Science and engineering communities
Understand their current, imminent, and future workflow needs and challenges, and provide guidance for application, infrastructure, and software development

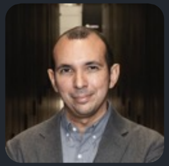
Computing centers and facilities operators
Support the use and deployment of workflow and application services, and provide training to foster proper adoption of workflow tools and therefore offering pathways for sustainability

An Overview of AI Workflows

for HPC Systems

Thank you!

Questions?



Rafael Ferreira da Silva, Ph.D.
Senior Research Scientist – Oak Ridge National Laboratory
Executive Director – Workflows Community Initiative
https://rafaelsilva.com

This research used resources of the Oak Ridge Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC05-00OR22725.

OAK RIDGE National Laboratory

U.S. DEPARTMENT OF ENERGY

The Networking and Information Technology Research and Development (NITRD) Program

Mailing Address: NCO/NITRD, 2415 Eisenhower Avenue, Alexandria, VA 22314

Physical Address: 490 L'Enfant Plaza SW, Suite 8001, Washington, DC 20024, USA Tel: 202-459-9674, Fax: 202-459-9673, Email: nco@nitrd.gov, Website: https://www.nitrd.gov